Exploring and Addressing General Limitations of Compound Potency Predictions Using Machine Learning

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Tiago Borges Janela
aus Barreiro, Portugal

Bonn 2024

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftliche
Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter/Betreuer:	Prof. Dr. rer. nat. Jürgen Bajorath
Gutachter:	Prof. Dr. rer. nat. Holger Fröhlich
Tag der Promotion:	4. Februar 2025
Erscheinungsjahr:	2025

The research towards this thesis was carried out at the Department of Life Science Informatics and Data Science at the b-it Institute of the University of Bonn under the supervision of Prof. Dr. Jürgen Bajorath.

Abstract

Compound potency prediction is a major task in computational drug discovery. Regression models based on machine learning (ML) approaches have become popular for small molecule potency predictions. Recently, deep learning (DL) methods have introduced novel architectures and data representations that have been applied to molecular potency predictions. Upon introducing a new computational approach, initial performance assessment is carried out using benchmark studies. Conventional benchmark calculations use compound potency data against a specific target divided into training sets for model generation and test sets for performance assessment over several rounds of cross-validation. Under these conditions, performance differences between prediction models are often negligible and do not translate into a successful application in prospective tasks. The mechanisms underlying these small performance differences are yet to be determined. This dissertation investigates the intrinsic limitations of current benchmark settings for compound potency predictions using ML models. The first study compares traditional ML, DL and control models' performance under different test conditions for several compound activity classes. Next, potency predictions are extended to a wide range of activity classes using ML and control models. The impact of data composition and potency ranges on prediction accuracy is determined based on different data set generation strategies. At this stage, limitations associated with potency prediction benchmarks, such as limited differences between predictive ML/DL and control models are uncovered. Furthermore, ML/DL and control models are derived with original and modified training sets of increasing compound sizes. Prediction performance is determined over several potency sub-ranges to rationalize the unveiled benchmark limitations. Moreover, the impact of structural analogs on prediction models is determined using a newly designed compound pair-based evaluation scheme to monitor performance over increasing compound potency differences. Additionally, a novel DL method for compound potency predictions is introduced and compared to state-of-the-art ML models for the prediction of potent compounds. Finally, alternative evaluation schemes are explored and possible future steps toward better benchmark systems for ML potency predictions are discussed. Taken together, this thesis uncovers current limitations of benchmark systems for comparing ML models and offers

alternative approaches to better determine compound potency prediction performance.

In memory of my beloved grandmother Maria Dionísia Borges

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor Prof. Dr. Jürgen Bajorath for the opportunity to embark on this scientific endeavor and the advice and guidance during my doctoral studies.

I thank Prof. Dr. Holger Fröhlich for reviewing my dissertation as a co-referent. I also thank Prof. Dr. Diana Imhof and Prof. Dr. Finn Hansen for accepting to be members in my PhD committee.

To all my current colleagues, including Dr. Elena Xerxa, Dr. Andrea Mastropietro, Hengwei Chen, Sanjana Srinivasan, Lisa Piazza and Selina Voßen in the Life-Science Informatics group my sincerest appreciation. A special thank you to Alec Lamens for your personal and scientific support and all our many laughs. To Jannik P. Roth, thank you for our brainstorming discussions and for making me a marathon runner. Moreover, I would like to thank Dr. Martin Vogt for his scientific advice and suggestions. And to all my former colleagues, including Dr. Javed Iqbal, Dr. Kosuke Takeuchi, Dr. Huabin Hu, Dr. Salvatore Galati, Dr. Nicola Gambacorta, Christian Feldmann, Friederike Schwarz, Sabrina Mendonça and Oliver Laufkötter an enormous thank you for all the indispensable support given to me during this time.

Lastly, I owe my deepest gratitude to my family and friends for their continuous encouragement and personal support throughout this journey.

Contents

List of abbreviations

1D, 2D, 3D	One-, two-, three-dimensional
AC	Activity cliff
ADMET	Absorption, distribution, metabolism, excretion, toxicity
AI	Artificial intelligence
AS	Analog series
CADD	Computer-aided drug design
CCR	Compound core-relationship
CVAE	Conditional variational autoencoder
DL	Deep learning
DNN	Deep neural network
DT	Decision tree
ECFP	Extended-connectivity fingerprint
FP	Fingerprint
GCN	Graph convolutional network
GNN	Graph neural network
k-NN	k-Nearest neighbor
-Log	Negative decadic logarithmic
LogP	Logarithmic octanol-water partition coefficient
MACCS	Molecular access system
MAE	Mean absolute error
ML	Machine learning
MMP	Matched molecular pair
MMS	Matched molecular series
MR	Median regression
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship
R^2	Coefficient of determination
RECAP	Retrosynthetic combinatorial analysis procedure
RF	Random forest
RFR	Random forest regression

RMSE	Root mean squared error
SAR	Structure-activity relationship
SHAP	Shapley addictive exPlanations
SMILES	Simplified molecular input line entry system
SPFP	Structure-potency fingerprint
SV	Shapley values
SVM	Support vector machines
SVR	Support vector regression
Tc	Tanimoto coefficient
VAE	Variational autoencoder

Chapter 1

Introduction

1.1 Drug discovery

Small molecule drug discovery is an expensive, time-consuming and complex process with the main objective of identifying molecules that could potentially treat a pathological condition. A drug discovery program involves investigating proteins or genes and intracellular processes associated with a disease condition for the identification of potential therapeutic targets.[1] In the next step, further validation ensures the association between the target protein and the corresponding molecular mechanisms underlying the disease.[2] This stage is followed by hit identification, where compounds are tested for activity against the validated target. During the hit identification stage, high-throughput screening techniques, powered by automated robotics, are employed to screen for potential active compounds in large chemical libraries.[1,3] Next, in the hit-to-lead stage, hit compounds are subject to re-evaluation (*in vitro*) and exploratory analysis of associated compound series. Lead compounds are subsequently selected based on desired properties. Exemplary properties include compound activity against the protein target of interest, the absorption, distribution, metabolism, excretion, toxicity (ADMET) profile and other physicochemical properties. In the lead optimization phase, lead compounds are subject to property optimization by chemical modification. Additionally, pharmacokinetic/pharmacodynamic and dose-response studies are carried out using in vitro and in vivo assays.[4] After these research stages, a preclinical candidate is selected to obtain regulatory authorization to advance to clinical trials. Following a successful trial outcome, regulatory agencies can approve the clinical compound allowing the

drug to enter the market. Pharmacovigilance studies will continue following market introduction to monitor, detect and prevent potential safety risks.[5] The described pipeline of drug discovery is an uncertain, costly and long process. It has been estimated that bringing a novel drug to the market might cost up to 4.5 billion dollars, depending on the therapeutic area and type of drug.[6] A large proportion of the costs is associated with compounds that fail along the drug discovery pipeline.

1.1.1 Computational-aided compound design in drug discovery

As the costs to bring a novel drug to market have steadily risen over the years, pharmaceutical companies have looked into the design and implementation of new strategies that improve and speed up the entire development pipeline.[7,8] One of the main areas that has further advanced research and development in small-molecule drug discovery is computational-aided drug design (CADD). CADD has been effectively applied to speed up processes while reducing experimental costs.[9,10] Here, computational approaches have been used to complement the various steps in the drug discovery pipeline.[9] For instance, during the “hit" stage, in silico screening of compound libraries prior to high-throughput screening can narrow down the number of compounds that require screening. Additionally, CADD tools can aid in lead compound identification efforts throughout the hit-to-lead stage by, for example, predicting the physiochemical characteristics and ADMET profiles.[11]

1.1.2 Molecular potency prediction

In computational medicinal chemistry, molecular property prediction is a major task during several drug development stages (hit-to-lead and lead optimization).[12,13] Commonly, quantitative structure-property relationship (QSPR) approaches are used to predict molecular physicochemical (e.g., aqueous solubility) and physiological (e.g., ADME) properties based on numerical descriptions of molecular structure. In addition to these properties, compound biological activity against a given target (often a protein) can be modeled using

quantitative structure-activity relationship (QSAR) methods.[14–16] For example, standard QSAR approaches employ simple linear regression models derived for series of analogs. These linear models are applied to predict the potency of newly generated compounds for the corresponding compound series. Hence, linear QSAR calculations constitute a methodology in ligand-based drug design that is limited to the prediction of compounds with similar structures. Conversely, in structure-based drug design, prediction methods aim to estimate compound affinities based on modeled or experimental three-dimensional structures of ligand-protein complexes. For instance, molecular docking calculations aim to correctly predict ligand conformations and poses within modeled binding sites. Here, a multitude of scoring functions are used to approximate ligand binding affinities that are commonly either force-field-based, knowledge-based, or empirical principles.[17–19] These approaches are widely applied for hit identification, lead optimization and structure-based virtual screening.[17,20] For these applications, docking scores are mainly used to rank ligand binding conformations, however, they only provide rough approximations of compound binding energies.[17]

On the other hand, at a more advanced level, free energy methods try to estimate binding free energies for ligand-protein complexes using thermodynamic cycles.[21] Therefore, free energy perturbation calculations can be performed to determine relative binding affinities between similar compounds. Relative free energy calculations are more computationally demanding than scoring function-based estimations. Furthermore, approaches that combine quantum mechanics with molecular mechanics are used to estimate relative compound binding energies and potency values.[22] Here, the target binding site and corresponding ligand are modeled using quantum mechanical calculations. At the same time, the remainder of the complex is kept stable to reduce computational time.[23,24]

Non-linear machine learning (ML) models are mainstream in computational drug discovery for ligand-based potency predictions. The ability of ML methods to model structurally diverse compound data, where non-linear structure-activity relationships (SARs) are present, sets them apart from structure-based approaches and standard QSAR.[14] In recent years, deep learning (DL) approaches have gained popularity for property/potency predictions due to increasing computational resources, data availability and model versatility.[14,25]

Together, these approaches pave a new way to tackle challenges faced during the drug discovery process.

1.2 Molecular representations

The chemical representation of molecules in a human and computer-readable format is crucial for drug discovery and chemoinformatics. Several molecular representations have been introduced for ML applications to improve computational efficiency, storage and performance.[26] For instance, the Simplified Molecular Input Line Entry System (SMILES)[27] is one of the most popular molecular linear representations, due to easy interpretability and computational efficiency.[26] SMILES can encode atoms/bond types, charge and stereochemistry, among others, in a string-format notation. Commonly, the direct application of SMILES strings for ML models requires a conversion into a machine-readable form (tokenization).[28] On the other hand, molecules can be represented using molecular graphs. In graph representations, a graph is defined as $G = (V, E)$, where V represents the nodes and E the edges.[29] For small molecules, atoms correspond to the nodes and the bonds to the edges.

Based on these representations, one-dimensional (1D), two-dimensional (2D) and three-dimensional (3D) numerical descriptors can be derived. For 1D numerical descriptors, simple molecular properties such as atom and bond counts or molecular weight are calculated from the molecular formula. In addition, 2D numerical descriptors are derived from molecular graphs that encode topological features and physicochemical properties, such as the octanol-water partition coefficient (logP), are approximated from the graphs using computational models.[30] Compounds can also be represented using 2D fingerprints (FPs), which encode molecular structural information as a binary vector indicating the absence (0) or presence (1) of specific structural features, as described in **Figure 1**.[31] Among different FP types, substructure key-based and topological-based FPs are widely used for chemoinformatics tasks. For substructure key-based FPs, molecules are encoded using a predefined substructure dictionary, where each bit position corresponds to a single structural key. The Molecular ACCess System (MACCS)[32] is one of the most popular keyed FPs. The publicly available version of MACCS is based on 166 different struc-

tural patterns. Alternatively, topological-based FPs are characterized by encoding structural features using a hashing function. For instance, extended-connectivity FPs (ECFPs),[33] based on the Morgan algorithm,[34] encode circular atom environments within a specified diameter. Commonly, hashed encodings of atom environments up to a radius of 2 are used to generate ECFPs of a fixed size of 1024 or 2048 bits.

Additionally, 3D representations of molecules extend 2D structural information by incorporating 3D properties, such as molecular conformation and topology.[35,36] However, the need for conformation estimation, whether through experimentation or computation approaches, together with the molecular conformation variability, limits their use in large-scale ML calculations.[37]

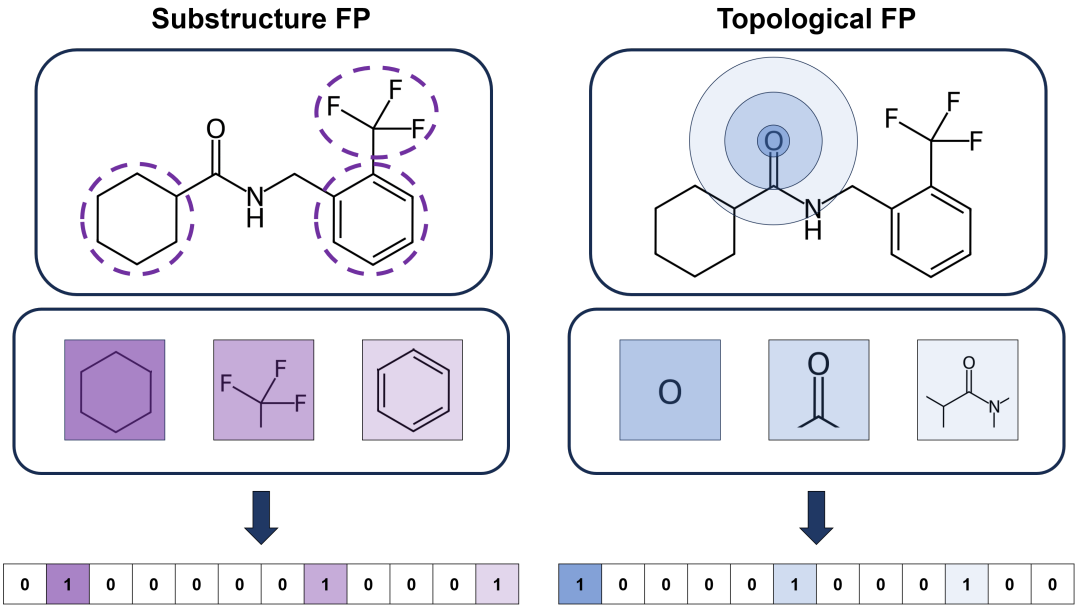


Figure 1: 2D molecular fingerprints. An exemplary keyed substructure FP (left) and a topological FP (right) are illustrated. For the keyed FP, if a substructure is present the bit is set to 1 (shades of purple) or if absent the bit is set to 0 (white). For the topological FP, a local atom environment of radius 2 is illustrated. Here, the environment for each diameter corresponds to an FP present feature in shades of blue (1), while absent features are depicted in white (0).

1.3 Molecular similarity

Molecular similarity is an essential concept in chemoinformatics and also of high importance in drug design.[38,39] In medicinal chemistry, the principle of similarity-property states that structurally similar compounds may display similar properties.[40] Many computational approaches, such as QSAR modeling or ligand-based virtual screening rely on this assumption.[41,42] Various similarity coefficients and compound representations are employed to quantify molecular similarity.

1.3.1 Fingerprint similarity

Two-dimensional molecular FPs combined with the Tanimoto coefficient (Tc),[43] also referred to as the Jaccard index, are very popular for compound similarity calculations.[44,45] The Tc measures the percentage of common substructures between two molecules and is defined as:

$$\mathrm{Tc}\,(A, B) = \frac{c}{a + b - c}$$

where a and b represent the number of features present in compounds A and B, respectively, and c represents the number of features common to compounds A and B.

Therefore, Tc values vary from zero to one indicating either no feature overlap or a perfect match between molecular FPs, respectively. Moreover, different FPs often display different similarity value distributions for the same set of compounds. For example, ECFP value distributions tend to be more narrow and shifted towards smaller Tc values (0.0 - 0.2), compared to MACCS distribution values which display a wider spread and are centered around higher Tc values (0.4 - 0.5) for random non-related small molecules.[45,46] However, this type of similarity assessment also carries its limitations. For instance, the inability to efficiently distinguish between active and inactive compounds based on similarity to active reference molecules alone, for large-scale ligand-based virtual screening.[47]

1.3.2 Matched molecular pairs

Matched molecular pairs (MMPs) are defined as pairs of compounds that share a common structure termed the MMP core (scaffold) and have a chemical change at a single site.[48] For MMPs, single-site substitutions result in a chemical transformation, as illustrated in **Figure 2a**.[49] The MMP concept is important for assessing molecular similarity and guiding compound design during the lead optimization stage.[50]

For a given library of molecules, the computational enumeration of these compound pairs can be achieved using fragmentation algorithms. According to Hussain and Rea, a suitable method for MMP extraction requires computationally efficiency in order to be applied to large compound data sets.[48] Such a method should have the ability to enumerate all MMPs present in a given database.[48] Traditionally, efficient MMP generation is based on the systematic fragmentation of exocyclic single bonds.[48] Another popular approach for MMP extraction relies on searching for the maximum common subgraph (MCS).[51] Here, the objective is to identify the largest common substructure between compound pairs, corresponding to the compound core structure.[52] However, as a result of their combinatorial nature, these approaches can have the disadvantage of being computationally expensive when applied to ultra-large compound sets. Furthermore, MMP extraction can be performed using a retrosynthetic combinatorial analysis procedure (RECAP).[53] The RECAP fragmentation algorithm employs a set of retrosynthetic rules corresponding to 11 original bond cleavage options that generate more chemically accessible RECAP-MMPs.[53,54] The resulting MMP compounds sharing a common core structure can be combined into a matching molecular series (MMS). An MMS is defined as a set of (two or more) compounds sharing a common core distinguished by different substituents at a single substitution site.[55]

MMP compounds derived from these approaches can be used for the identification of activity cliffs (ACs). ACs are defined as pairs of structurally similar compounds (analogs) with large potency differences.[56,57] A commonly applied threshold of a 100-fold change in potency is used to identify ACs.[58,59] Consequently, MMP compounds characterized by these activity differences are designated as MMP-cliffs,[60] as illustrated by **Figure 2b**. For QSAR modeling, the prediction of ACs/MMP-cliffs continues to be a challenging task, as pre-

diction models display difficulties in correctly capturing the structure-activity landscape for these compounds.[57,61,62]

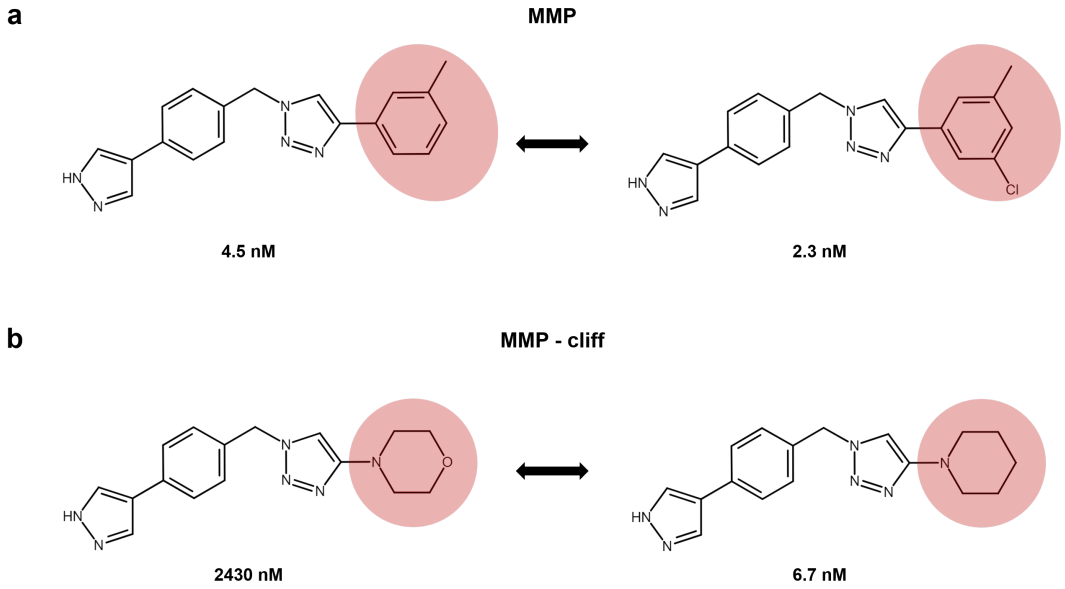


Figure 2: Exemplary MMP and MMP-cliff. A representation of an MMP (a) and MMP-cliff (b) is shown. For each MMP compound an exemplary activity value is provided in nanomolar. Substitution sites are highlighted (red). Marvin was used for drawing and displaying chemical structures.[63]

1.3.3 Analog series

In contrast to an MMS, an analog series (AS) consists of compounds that are characterized by the same core structure containing one or multiple substitution sites with different R-groups (substituents), as shown in **Figure 3**. The exploration of ASs and the design of new analogs is a central aspect during hit-to-lead and lead optimization stages to find and prioritize potential candidate compounds.[64] Traditionally, R-group tables are used to monitor the evolution of ASs.[65] Moreover, computational approaches such as linear and non-linear QSAR analysis have been applied to support the design analogs by predicting the potency of new analogs.[14] Systematic identification and analysis of multiple AS can be a challenging task, especially for large compound libraries.[66] Computational approaches that identify and extract AS from such databases are available. Analogs can be obtained using the MMP concept to generate AS

comprising of single substitutions (MMS), as discussed above. Moreover, AS with multiple substitution sites can be identified using computational methods such as the compound-core relationship (CCR) algorithm. The CCR method defines AS by performing a systematic compound fragmentation at one or multiple substitution sites generating unique core structures and corresponding R-groups substituents.[67] Ideally, generated AS core structures and substituents can be displayed in R-group tables. Additionally, AS can be used to assess QSAR prediction ability on novel compound series.

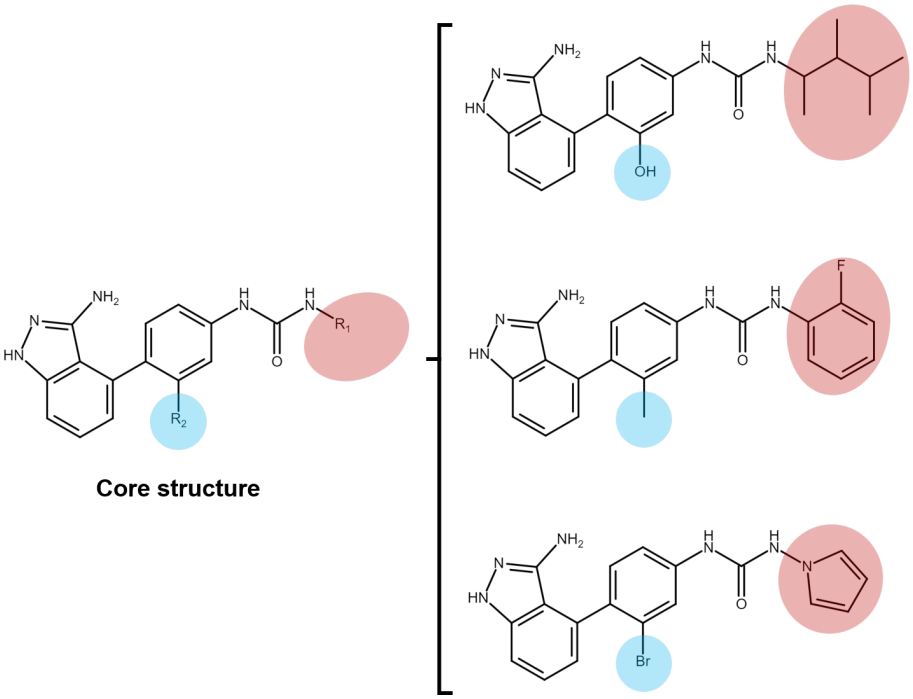


Figure 3: Analog series. An AS comprising three structurally analogous compounds and corresponding core structure is shown. In this case, analogs correspond to ASs with two substitution sites (R_1 and R_2) outlined in red and blue, respectively. Marvin was used for drawing and displaying chemical structures.[63]

1.4 Benchmark calculations

For compound potency predictions, the performance of novel ML models is assessed using benchmark calculations. In benchmark studies, ML models of different complexity depending on sets of (hyper-)parameters are generated using curated compound data partitions for training and test sets, over several

rounds of cross-validation, as illustrated in **Figure 4**. For each cross-validation iteration, parameter optimization is performed using the training sets. To this end, training sets are sub-divided into internal training and test sets multiple times using different parameter settings; models are trained on the internal training sets and evaluated on the remaining data (internal test sets). Optimal parameters are determined by averaging the results of internal test sets for each parameter setting and selecting the best. Final models are derived using original training sets and the optimal parameters. Thereafter, the resulting models are evaluated using various performance metrics analyzing prediction performance over the different test sets. Benchmark calculations include a large number of variables such as the partitioning strategy, selected algorithms and data curation process. These variables are discussed below in more detail.

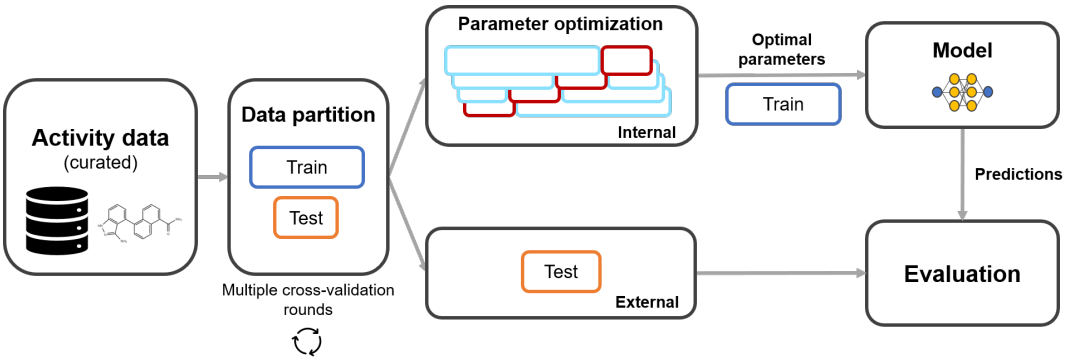


Figure 4: Conventional benchmark scheme. An exemplary benchmark system is shown for compound potency predictions. Activity data is partitioned into external training (blue) and test (orange). Parameter optimization is performed by dividing the training set into several internal training (cyan) and test (red) sets in order to determine optimal parameters. The final model is derived using the best parameters and original training (blue) set. Model evaluation is carried out by calculating the appropriate regression metrics for the external test (orange) set predictions.

1.4.1 Activity data

For benchmark studies, the availability of highly curated compound activity data for a target is required. Such data can be obtained either from proprietary sources of large pharmaceutical companies or, as is common in the academic world, be extracted from public data sources. As described above, activity data originate from biological assays measured as potency, which is characterized by

the compound concentration required to produce an effect of a given magnitude.[68] Several potency measures such as IC_{50}, K_i and K_d are often used for QSAR modeling.[69] For example, K_i and K_d describe the inhibition and dissociation constants, respectively. They represent assay-independent measurements. On the other hand, IC_{50} the half maximal inhibitory concentration, is an assay-dependent measure that corresponds to the compound concentration required to reduce the activity of a protein by half. Typically, potency values in molar concentration are recorded as negative decadic logarithmic (-log) potency values (pIC_{50}, pK_i, pK_d).[69] Activity data can be extracted from publicly available repositories like, for example, ChEMBL.[70] The ChEMBL database comprises manually curated bioactivity information for drug-like compounds based on more than 1.6 million assays. Its major activity data source is the medicinal chemistry literature, covering assays at different stages of compound development reported in publications.[70]

The design of a high-confidence activity data set from the ChEMBL database requires some crucial curation steps. Therefore, compounds with a molecular mass of less than 1000 daltons (small molecules) are selected. Only single human protein assays are retrieved with the highest ChEMBL assay confidence score, ensuring an activity annotation against a single direct protein (excluding homologs) in a binding assay, and a numerical activity value (e.g., K_i, K_d, IC_{50}). Compounds with potency values higher than 10 M or below 10 pM are typically removed, as relevant activity ranges from a -logIC_{50} value of 5 (micromolar) to 11 (nanomolar). Additionally, if multiple potency annotations are available for a compound and a specific protein, the annotations are only considered if they fall within one order of magnitude (10-fold), in which case the values are averaged, otherwise, the annotations are discarded. Subsequently, the removal of pharmaceutical anti-targets is performed. In drug discovery, an anti-target is described as a protein (receptor, enzyme) vital for the organism's function inhibition of which causes unwanted and adverse effects.[71] Therefore, protein targets such as cytochrome P450, UDP-glucuronosyltransferase, hERG, P-glycoprotein and albumin are disregarded. This step is followed by the removal of compounds known for potentially interfering with biological assays. This task is performed using publicly available filters such as Eli Lilly Medic-

inal Chemistry Rules,[72] Pan Assay Interference Compounds (PAINS),[73] and colloidal aggregators.[74]

The resulting high-confidence data set can be used for ML purposes. Here, the curated data ensures only relevant molecular data is applied to ML methods and the noise in the data is reduced to a minimum.

1.4.2 Compound data partitions

Compound data partitions for ML models can greatly impact prediction performance during benchmark calculations. Several cross-validation techniques are available, such as random-, analog/scaffold- and temporal splits. Random data splits aim to create training and test sets sampled from the same data distribution. However, this approach tends to overestimate ML performance, due to the presence of similar compounds, often analogs, in training and test sets. Therefore, it is not providing a realistic evaluation, given that most prospective applications of ML models focus on exploring novel chemical space. In contrast, analog/scaffold data partitions generate training and test sets comprising unique AS for each set, hence, removing the structural bias present in random splitting.[42] Additionally, temporal-based data partitioning provides a useful evaluation for prospective applications. Here, data splits are based on compound activity measurement dates. Thus, training sets are generated using data collected up to a certain point in time with the remaining data used for model validation. However, this methodology is generally not applicable to publicly available data sets.[75]

1.4.3 Performance metrics

In a regression setting, model prediction performance is usually assessed using the mean absolute error (MAE), root mean squared error (RMSE) and the coefficient of determination (R^2). For MAE and RMSE the lower the value the higher is prediction performance. Compared to MAE, the RMSE strongly penalizes larger prediction errors and, therefore, is more sensitive to outliers. For both metrics, interpretation should be accompanied by the specific domain knowledge, due to the metrics dependency on the modeled data distri-

bution. Furthermore, R^2 provides a relative error measure compared to MAE and RMSE. An R^2 value of 0 or negative shows that model accuracy is similar to or worse than predicting the mean of the dataset for every test instance. Meanwhile, an R^2 value of 1 (maximum) indicates a perfect prediction model. MAE, RMSE and R^2 are defined as:

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

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

where y_i and $\hat{y}_i$ are the observed and predicted values, for instance, i, respectively, and n is the total number of instances. For R^2, $\bar{y}$ is the mean observed value for the modeled data set.

1.4.4 Control calculations

In benchmark ML studies, simple control calculations are essential to set minimum performance requirements for predictive models. Two common baseline techniques are y-randomization[76] and central tendency regression. The method of y-randomization consists of shuffling the instance labels (y) for the data set, followed by model retraining and cross-evaluation. In central tendency regression, baseline models are based on assigning, for example, the median (MR) or mean training set value to each test instance. In the case of potency prediction, y-randomization generates random structurepotency relationships within dataset compounds, whereas MR models assign the median potency value of training compounds to each test compound.

1.4.5 Statistical testing

For benchmark calculations, statistical testing should be carried out to assess claims of a model's superior performance compared to other established approaches. Statistical tests are employed in hypothesis testing to determine differences between two or more populations.[77] The test generates a probability value (p-value) used to assess the statistical significance of possible performance differences. Several statistical tests are available in the literature, ranging from parametric (e.g., Students t-test[78]) to non-parametric (e.g., Wilcoxon signed-rank[79]), which can be selected based on whether certain data assumptions are known or unknown. Moreover, when performing multiple comparisons, p-values need to be adjusted to reduce the chance of one model being randomly picked as superior compared to another. Hence, p-value adjustments can be performed using corrections, such as the Bonferroni or Holms methods.[80] Here, it is important to distinguish between statistical significance and functional relevance of the difference between two populations. For example, statistical significance does not always translate to better methodological performance in prospective applications, especially in drug discovery.

1.5 Machine learning

Over the years, a variety of ML methods have been applied to molecular property predictions, in drug discovery.[81,82] For example, traditional ML models such as support vector machines (SVMs)[83] or random forests (RFs),[84] employ fixed molecular representations including MACCS[32] or ECFP[33] as input features that are mapped to the corresponding property values. Most recently, a surge in computational resources enabled DL architectures to be derived for quantitative molecular property predictions. Among these approaches, graph neural networks (GNNs) and transformer models have gained popularity to address these tasks.[85,86] The following section will further describe some of these state-of-the-art methods in more detail.

1.5.1 *k*-nearest neighbor

The objective of k-nearest neighbor (k-NN) algorithm is to find the closest (most similar) n samples to the query instance, as illustrated in **Figure 5**. This method can be used for classification and regression tasks.[87,88] For regression problems, the predicted test value is the average of the values from the closest n-training samples. As the choice of the number of nearest neighbors influences the approach's performance, the best number k of nearest neighbors should be always explored using parameter optimization techniques. For compound potency predictions, k-NN is considered a control model, due to its simplicity. In chemoinformatics, k-NN models have been applied in a variety of tasks, from virtual screening[89] to molecular property predictions.[90]

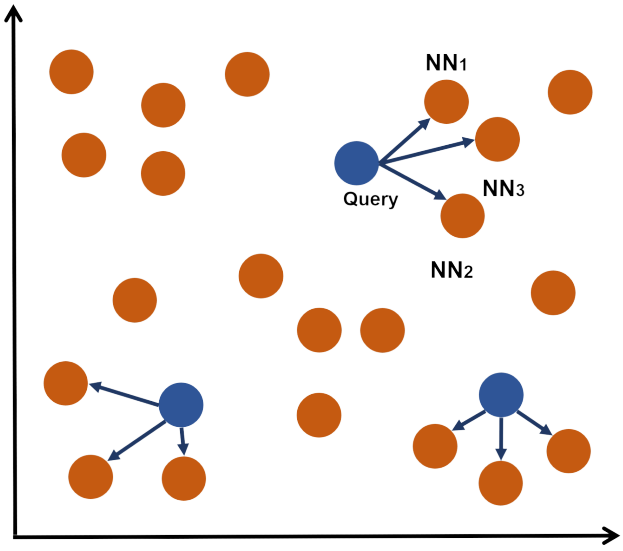


Figure 5: *k*-nearest neighbor algorithm. A k-NN algorithm employing three nearest neighbors (orange) for the prediction of a numerical property of a query instance (blue) label. The final prediction is given by the average value of the corresponding nearest neighbors.

1.5.2 Support vector machines

For chemical data, SVMs have been the model of choice for several applications including compound classification, ranking and multi-target activity prediction.[91] SVMs are supervised ML models that map training data to a defined feature space using kernel functions.[92] The "kernel trick" describes the use

of non-linear kernels if linear separation is not feasible in the original feature space. The model objective is to find the best hyperplane that separates the majority of training instances. Support vector regression (SVR) is an adaptation of the SVM methodology for numerical value predictions, as shown in **Figure 6**.[83] For SVR, the regression hyperplane is defined by (4), which aims to approximate the training instances to the observed labels during the model optimization.

$$y = < w, x > + b \tag{4}$$

where, w corresponds to a weight vector, x to a data instance, $< ., . >$ to the scalar product and b to the bias.

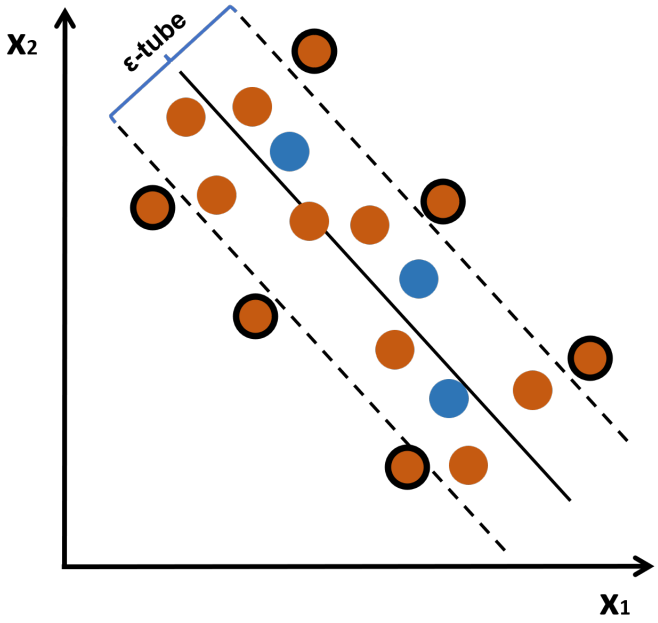


Figure 6: Support vector regression algorithm. For SVR, a decision function is generated based on training instances (orange). Support vectors (training instances with black outlines) are illustrated outside the ϵ-tube. The test instances are represented with blue circles.

SVR employs an ϵ-insensitive tube[93] in which the tube width determines the maximum tolerated error between the input and output values. Moreover, training instances present outside of the ϵ-tube represent the support vectors. For chemical data, diverse kernel functions can be employed, such as linear,

polynomial, radial basis function and the most popular for molecular FPs, the Tanimoto kernel.[94]

1.5.3 Random forest

An RF is a supervised ML method based on an ensemble of decision trees (DT).[84] During the training process, each tree is generated using a bagging procedure, where samples are randomly selected with replacement from the training set. Moreover, at each node split in the tree, the best split is determined by randomly sampling a subset of the available features. These procedures introduce variability among trees while reducing inter-tree correlation.[95] For regression tasks, the final prediction corresponds to the average output of all the trees in the forest, as illustrated in **Figure 7**. RFs have been widely applied in many chemoinformatic-related prediction tasks.[96]

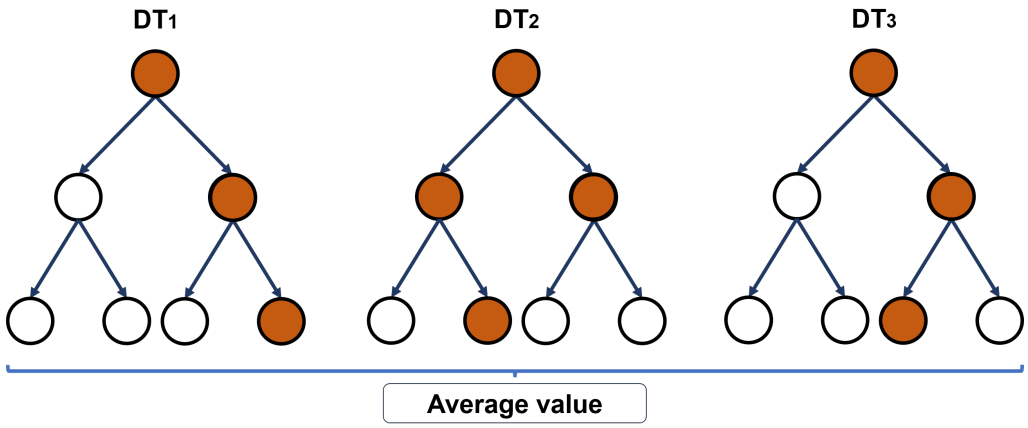


Figure 7: Random forest regression. Shown is a schematic representation of an RF algorithm generated using an ensemble of DTs for a regression task. The predicted value for each DT is derived from the root to the tree leaf node, highlighted using orange circles. The final predicted value is determined as the average value of all trees in the forest.

1.5.4 Deep neural networks

A feedforward deep neural network (DNN) is a DL method that derives a non-linear relationship between input values and corresponding output values (y) by repeatedly applying parameterized mathematical functions to intermediate outcomes that represent the layers of the DNN.[97,98] A typical DNN ar-

chitecture comprises several layers of computational neurons, consisting of an input layer, multiple connected hidden layers and an output layer, as illustrated in **Figure 8**.[97] Each network neuron is defined by the mathematical function in (5). During training, the networks input weights (w_i) and biases (b) for each neuron y according to equation (5) are continuously adjusted to minimize errors between the predicted and observed labels. Here, the x_i are the values of the previous layer (or the input layer in the case of the first hidden layer) and f represents the fixed activation function. This optimization is performed via backpropagation,[99] which involves propagating the gradients of the cost function backward through the network to update the weights and biases.

$$y = f\left(\sum_i w_i x_i + b\right) \tag{5}$$

In computational medicinal chemistry, DNN architectures have been successfully applied in many prediction tasks, such as QSAR modeling,[100] multi-task ADMET[101] and chemical image recognition.[102]

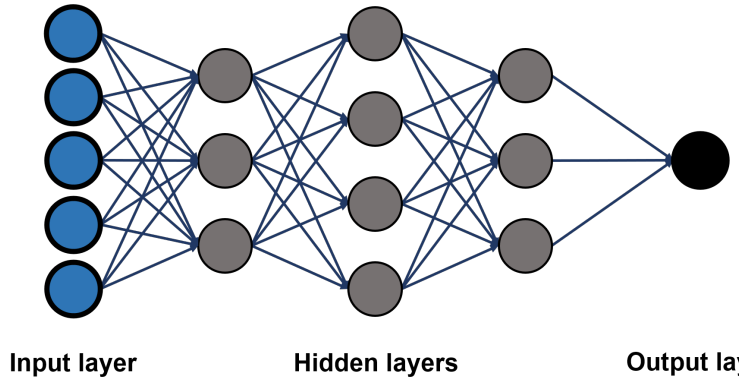


Figure 8: Deep neural network. A feedforward DNN architecture consisting of one input layer (blue nodes), three hidden layers (gray nodes) and one output layer (black node) is illustrated. Each network node is denoted as a circle and represents a neuron. In binary classification and regression tasks, the output layer consists of a single neuron.

1.5.5 Graph neural networks

Graph neural networks (GNNs) are DL methods used to learn from molecular graph representations.[103] For example, graph convolutional networks

(GCNs) employ a convolutional process to derive generalized atom environment representations similar to the Morgan algorithm used in circular FPs or ECFPs.[33,34,104] The GCN's main architecture consists of graph convolutional, pooling, gathering and fully connected dense layers, depicted in **Figure 9**. Through the input layer, molecular graphs are submitted to a graph convolutional layer where convolutional operations extract node representations by considering the weighted average values for each atom and neighborhood atom features. This process is repeated using several convolutional layers, while pooling operations (e.g., max or sum pooling) are performed to aggregate the resulting node and edge representations into a unified graph-level representation. Next, a gathering layer generates the neural FP by summing the generated atom feature vectors. Finally, the molecular FP vectors are mapped to the corresponding labels using a DNN model.[105] Hence, molecular graph convolutions present an excellent alternative to traditional FP descriptors for molecular prediction tasks. In chemistry, representation learning using GNNs has been evolving and has been used in multiple applications including molecular property predictions,[106] drug-target interactions,[107] ADMET prediction[108] and *de novo* drug design.[109]

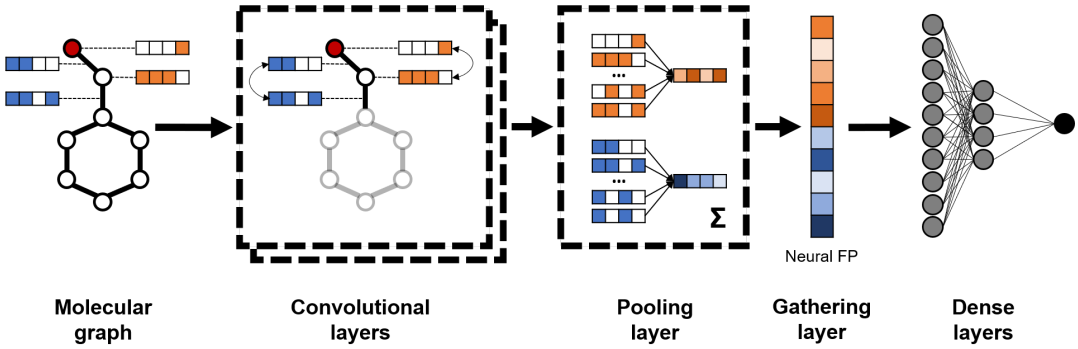


Figure 9: Graph convolutional network. Shown is a schematic illustration of a GCN architecture including convolutional, pooling, gathering and dense layers. Node (atom) and edge (bond) information are represented by orange and blue numerical vectors, respectively. The node and edge representations are concatenated to generate a neural FP. This fingerprint is used as an input sample for a feedforward DNN to generate the predicted output.

1.5.6 Variational autoencoders

A variational autoencoder (VAE) is a probabilistic DL model that can generate new samples based on a learned data distribution.[110] The VAE architecture comprises a recognition network (probabilistic encoder), mapping the high-dimensional input values into a low-dimensional continuous latent space (z). The decoder network reconstructs latent space samples into the original data dimension. During training, the encoder and decoder are optimized by maximizing the evidence lower bound,[111,112] by minimizing Kullback-Leibler divergence[113] between the input and latent distributions and the reconstruction loss. A variety of VAE architectures have been explored in the literature.[114] For instance, conditional VAEs (CVAEs) apply input labels that condition (c) the latent space generation and the corresponding output. Thus, CVAEs use non-random sampling compared to VAEs to generate new instances, as illustrated in **Figure 10**.[115] In recent years, VAEs have been applied in many chemical tasks including generative compound design,[116] compound property predictions[117] and chemical reaction design.[118]

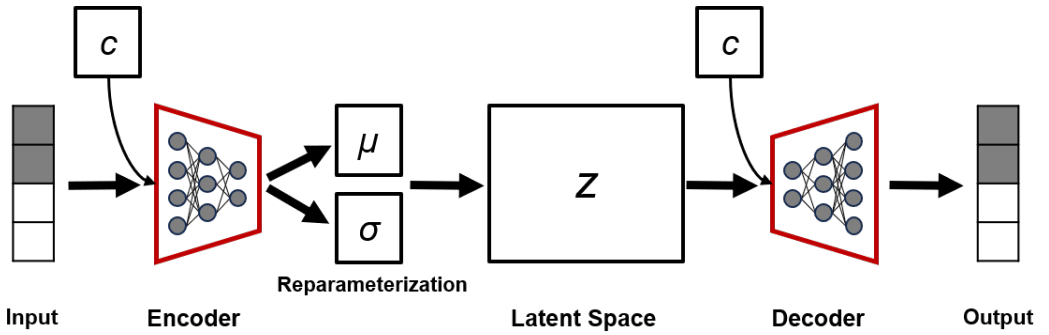


Figure 10: Conditional variational autoencoder. A scheme of the CVAE algorithm consisting of an input layer, followed by the encoder module, a latent space layer (z), a decoder module and an output layer. Here, and are the mean and standard deviation of a Gaussian distribution ($N(0,1)$).

1.5.7 Model explanations

The interpretation of learning characteristics for ML predictions is crucial in pharmaceutical research. Over the years, several methods have been developed to explain model predictions ranging from global to local explanation ap-

proaches.[119] For example, the Shapley value (SV) formalism, derived from game theory,[120] has been widely explored to explain ML models. In the context of ML, SVs quantify the individual contributions of present and absent features for a given test instance. While this approach is feasible for smaller feature sets, SV calculations can become computationally infeasible for larger feature sets due to their combinatorial nature. Therefore, approximation approaches such as Shapley Addictive exPlanations (SHAP)[121] are typically employed. SHAP values are based on a local model for a given instances feature space. In chemoinformatics, SV/SHAP values have been used to explain compound activity predictions,[122] multi-target activity and compound potency predictions.[123]

1.6 Thesis outline

This dissertation aims to provide a better understanding of the general limitations of benchmarking compound potency predictions using state-of-the-art ML models. The dissertation is divided into eight chapters. *Chapters 2* to *Chapter 7* consist of six original publications representing the core of this thesis.

- *Chapter 2* reports benchmarking of ML and control models for compound potency predictions. Therefore, activity classes comprising active compounds against different pharmaceutical targets are generated. Subsequently, predictions under different interpolative and extrapolative test conditions are compared.

- In *Chapter 3*, compound potency prediction benchmark calculations are extended to a large number of activity classes. Here, specific data set modifications are designed to evaluate the influence of potency ranges and data composition in prediction performance for ML and control models.

- *Chapter 4* rationalizes the limitations of compound potency prediction benchmark calculations, uncovered in previous chapters. The prediction performance of ML and control models is monitored across different potency sub-ranges for several activity classes. Moreover, it is investigated if the data distributions across compound potency sub-ranges influence

model predictions by generating training sets of increasing compound sizes.

- In *Chapter 5*, the effect of structural analogs in compound potency predictions is investigated. A compound pair-based test system is designed to evaluate compound predictions over increasing potency differences. In addition, ML predictions are rationalized using an explainable artificial intelligence (AI) method.

- *Chapter 6* introduces a novel DL-based methodology for compound potency prediction. Therefore, this new approach is compared against state-of-the-art ML models across different activity classes. Additionally, the ability of models to correctly predict the most potent test compounds is assessed.

- In *Chapter 7*, a review of the current limitations of compound potency predictions is presented. Multiple fundamental limitations unveiled in the previous chapters are described. Furthermore, a potential alternative benchmark system to compare potency prediction models is introduced. Finally, future directions on how to better assess prediction performance in practical applications are discussed.

- Finally, *Chapter 8* summarizes the main findings of this dissertation and addresses their impact on small-molecule drug discovery.

Chapter 2

Simple Nearest-Neighbour Analysis Meets the Accuracy of Compound Potency Predictions Using Complex Machine Learning Models

The following chapter summarizes the research published as
Janela, T.; Bajorath, J. Simple Nearest-Neighbour Analysis Meets the Accuracy of Compound Potency Predictions Using Complex Machine Learning Models. *Nat. Mach. Intell.* **2022**, *4*, 1246-1255. DOI: 10.1038/s42256-022-00581-6

Author contributions: Tiago Janela: Methodology, Data, Code, Investigation, Analysis, Writing - review and editing. **Jürgen Bajorath**: Conceptualization, Methodology, Analysis, Writing - original draft, Writing - review and editing.

2.1 Summary

In CADD, compound potency prediction is of the highest interest. For this task, several approaches can be employed, from structure- to ligand-based methodologies,[15,124] often based on state-of-the-art ML algorithms.[81] As the development of new AI applications increases, the use of complex model architectures for molecular property predictions has also surged, attempting to improve current state-of-the-art performance.[82,125] In compound potency prediction, ligand-based approaches frequently aim to extract nonlinear relationships between compound structure descriptors and corresponding activity. For this purpose, ML models are built using a set of active compounds against a specific protein with the intent to subsequently predict the potency of novel chemical entities. To evaluate new predictive methods, benchmark calculations are carried out using active compounds, partitioned into training and test sets, for the derivation and evaluation of ML models respectively, over several rounds of cross-validation. In this chapter, benchmark calculations for ML models, including RF regression (RFR), SVR, DNN and GCN, concerning potency prediction are reported. A variety of data sets are evaluated. Additionally, baseline calculations are performed using k-NN, MR and random regression. Furthermore, the ability of these models to extrapolate beyond the data upon which they were trained is assessed through the use of specifically designed data sets.

For compound potency prediction, benchmark calculations were performed for 10 randomly selected activity classes comprising highly curated compound activity data extracted from the ChEMBL database. Therefore, original (complete) sets were created corresponding to the entire data set for each activity class. The data sets were used to derive and evaluate ML models over multiple independent prediction trials. Model performance was assessed using MAE and RMSE for all prediction experiments. For the complete sets, accurate ML models showed very similar performance ($\sim$0.5 MAE), yet SVR slightly outperformed RFR, DNN and GCN models based on statistical analysis. These trends were shown to be consistent across the different activity classes. Furthermore, simple k-NN models reached or even surpassed the accuracy of complex ML

models. Moreover, MR control models achieved an accuracy of ∼0.8-1.0 MAE, across tested activity classes. Thereafter, to assess the effect of data sparseness on model learning, size-diverse training sets consisting of compounds with higher chemical diversity and random equally sized-reduced sets were generated for all activity classes. For sized-reduced training sets, prediction errors slightly increased with increasing performance differences between k-NN/SVR and DNN/GCN. Furthermore, for diverse sets, similar trends were observed for k-NN and SVR. DNN and GCN models approached MR performance for some activity classes. Furthermore, compound data sets were divided into AS to assess the performance of complex ML and k-NN models in a prospective scenario. For this purpose, hold-out sets comprising the largest AS from the original set were generated for each class. Subsequently, models were derived using the remaining AS, and single prediction trials were performed on the hold-out sets. For the majority of the tested classes, similar performance was again observed across all prediction models. Additionally, the ability of each model to extrapolate was tested. For each class, the 10% most potent compounds were selected and used as a test set. The remaining compounds were used for model derivation and predictions were performed on the potent test sets. Again, k-NN prediction accuracy rivaled complex ML models under these conditions. Furthermore, randomized predictions were explored. Accordingly, k-NN and SVR models were generated by randomly assigning potency values to training and test compounds. For MR and random predictions (k-NN/SVR), prediction accuracy differences of ∼0.5 MAE compared to original ML models were observed. This analysis demonstrated that k-NN models showed similar performance compared to complex ML models. This prediction performance could be attributed to similar compounds often having similar potency values. Additionally, prediction performance for k-NN and ML models (∼0.5) compared to MR and randomized models (∼0.9) was separated only by small error margins. Based on these findings, current benchmarking calculations for compound potency predictions require further evaluation. The use of simple k-NN models as a reference method is recommended, together with model evaluation on highly potent and structurally unique compounds set not used in model building. The following chapter extends the compound potency prediction benchmark to a larger set of activity classes.

Chapter 3

Large-Scale Predictions of Compound Potency with Original and Modified Activity Classes Reveal General Prediction Characteristics and Intrinsic Limitations of Conventional Benchmarking Calculations

The following chapter summarizes the research published as
Janela, T.; Bajorath, J. Large-Scale Predictions of Compound Potency with Original and Modified Activity Classes Reveal General Prediction Characteristics and Intrinsic Limitations of Conventional Benchmarking Calculations. *Pharmaceuticals* **2023**, *16*, 530. DOI: 10.3390/ph16040530

Author contributions: Tiago Janela: Methodology, Software, Formal analysis, Investigation, Writing - original draft preparation, Writing - review and editing. **Jürgen Bajorath**: Conceptualization, Methodology, Formal analysis, Writing - original draft preparation, Writing - review and editing.

3.1 Summary

As shown in the previous chapter, complex ML and simple k-NN models achieved similar accuracies in compound potency predictions. Additionally, the analysis showed no advantage in using DL models (DNN/GCN) compared to traditional ML (SVR/RFR) and k-NN models. Moreover, MR and random regression models displayed prediction accuracies of ∼0.9 MAE compared to the performance of accurate ML models of ∼0.5. In this chapter, the previous potency prediction analysis was extended to evaluate the distribution of potency prediction accuracies across a large range of activity classes. Moreover, it was investigated if activity classes' potency ranges and respective compositions may be affecting and thus limiting the relevance of current benchmark calculations.

To further expand the previous analysis (*Chapter 2*), high-confidence data for 376 activity classes comprising at least 50 active compounds was extracted from ChEMBL database. Based on this data, ML (SVR) and control models (1-NN, k-NN, MR) were derived over multiple independent trials and MAE was calculated to assess the corresponding model performance. For the selected activity classes, similar prediction accuracies were observed between SVR, 1-NN and k-NN models with overall errors within one order of magnitude (MAE < 1), in line with the observations from *Chapter 2*. In addition, for the 45 largest activity classes, the effect of different data partition sizes on prediction performance was investigated. Therefore, SVR, 1-NN, k-NN and MR models were implemented using training and test sets with two different size splits (80/20% and 50/50%). For both partition sizes, meaningful and stable compound predictions were obtained with, again, relatively similar performance. Thus, predictions were not significantly affected by varying training set sizes. Considering these results, the data sets for all 45 activity classes were selectively modified to assess possible differences in prediction performance between SVR and controls. The first data set modification was focused on activity class potency distributions. In medicinal chemistry, most potency distributions have a higher prevalence of lowly potent compounds (micromolar) compared to highly potent (nanomolar), resulting in skewed distributions. The predominance of

micromolar potency values might explain the performance of k-NN and MR in comparison to SVR. For this purpose, data sets with balanced potency distributions were generated by evenly populating potency sub-ranges with available compounds. Hence, the tendency of potency datasets to show a skewed distribution towards low micromolar potency values was eliminated. Considering that the balancing procedure reduced the size of the original sets, as a control, identical-size data sets were created with original potency distributions. All models were built with these newly created balanced sets and prediction performance was evaluated. As expected, the median potency values for training sets increased, inevitably slightly increasing MR prediction errors. Nevertheless, SVR performance continued to be comparable to 1-NN and k-NN models. The second modification was designed to examine the effect of removing nearest neighbor compounds on the ability of models to accurately predict compound potency by generating data sets with a reduced number of close neighbors. To account for data set size, a control set was derived by randomly sampling 50% of compounds from the original set. For all models, prediction error increased for decreasing numbers of close neighbors. Notably, the performance of 1-NN and k-NN was still comparable to SVR models across activity classes. Considering the previous findings, AS splitting was explored as another type of data set modification. Training and test sets composed of compounds with unique core structures were derived, resulting in no scaffold overlap among sets. For this setup, errors increased compared to the control sets of the same size. Nevertheless, differences in performance between 1-NN and k-NN and SVR models remained similar. Taken together, this chapter reflects previous observations on a larger scale that prediction performance for different methodologies cannot be realistically assessed by using conventional benchmark settings. Models tested based on these modified data sets showed notably increased prediction errors associated with more difficult test conditions. However, compound predictions were stable and relatively insensitive to the modifications, given that SVR and k-NN demonstrated similar prediction differences. Therefore, the development of meaningful benchmark schemes required further consideration. In the following chapter, the current limitations of compound potency predictions under benchmark scenarios are uncovered and rationalized.

Chapter 4

Rationalizing General Limitations in Assessing and Comparing Methods for Compound Potency Prediction.

The following chapter summarizes the research published as
Janela, T.; Bajorath, J. Rationalizing General Limitations in Assessing and Comparing Methods for Compound Potency Prediction. *Sci. Rep.* **2023**, *13*, 17816. DOI: 10.1038/s41598-023-45086-3

The publication reprint is available in Appendix C. Reprinted with permission from "Janela, T.; Bajorath, J. *Sci. Rep.* **2023**, *13*, 17816". Copyright 2023 Springer Nature.

Author contributions: Tiago Janela: Study design and conduction, Formal analysis, Manuscript preparation. **Jürgen Bajorath**: Study design and conduction, Formal analysis, Manuscript preparation.

4.1 Summary

In *Chapter 3*, benchmark calculations of compound potency prediction were further extended to a large number of activity classes using various ML methods and controls. Thereby, a broader yet detailed overview of model performance was obtained. The results were in line with the findings described in previous chapters. The impact of activity class composition and respective potency value distribution was investigated as a possible cause for the observed limitations. Thus, several activity class modifications were performed comprising balancing data sets according to potency ranges, removal of compound nearest neighbors and analog partitioning for training and test sets. Accordingly, activity classes were benchmarked with modified data sets. Even though prediction accuracy decreased, error margins between methodologies were still very similar. Based on these findings, the benchmark limitations of compound potency predictions needed to be further explored. In this chapter, different ML methods were used to explore the limiting factors behind benchmarking for potency predictions by focusing on the impact of potency sub-ranges and respective value distributions.

To investigate the influence of potency value distributions and potency sub-ranges on compound potency predictions, 8 activity classes were used to derive traditional ML (SVR, RFR) and control (1-NN, 3-NN and MR) models for independent trials. Model prediction performance was evaluated based on MAE, RMSE and squared Pearson correlation coefficient. In line with previous findings (*Chapters 2* and *3*), ML, 1-NN and 3-NN had comparable accuracy across the entire potency range (-logIC_{50}: 5 - 11) in contrast to MR models for all activity classes. Overall, SVR slightly outperformed RFR, 1-NN and 3-NN models based on statistical testing. Subsequently, predictions were assessed using a more detailed view by dividing test compounds into the respective experimentally defined potency sub-ranges (5 - 7, 7 - 9 and 9 - 11). For weakly (5 - 7) and highly potent (9 - 11) sub-ranges, slightly larger prediction errors were observed for ML, 1-NN and 3-NN models, possibly due to increased data sparseness. Meanwhile, MR displayed a significant increase in prediction errors, associated with the increase in distance to median potency values for all activity classes. On the other hand, compounds falling in the potency sub-range (7

- 9) showed a decrease in prediction errors relative to the global accuracy (5 - 11) and were similarly close to MR controls. In order to investigate underlying learning characteristics for each potency sub-range, training sets of increasing size were generated by uniform sampling of compounds for each sub-range. For three activity classes with sufficient numbers of highly potent compounds in the range 9 - 11, nine balanced training sets of increasing size were created. The remaining compounds were used to assemble test sets with balanced potency sub-ranges. For low and high potency sub-ranges (5 - 7 and 9 - 11), small training sets comprising 6-18 compounds were shown to be insufficient for ML models to predict compound potency accurately. Here, significantly reduced prediction performance was observed compared to models trained on larger training sets. Notably, prediction performance remained stable for the median potency range (7 - 9) across training sets of increasing sizes. For this sub-range, the prediction error was relatively small for SVR and RFR models approaching the median potency values of the training sets. Fundamentally, ML models were not required to learn in order to correctly predict compound potency. Furthermore, for the sub-range 7 - 9, ML models and MR achieved the best accuracy compared to 1-NN and 3-NN predictions. Calculations were repeated for imbalanced training sets producing comparable results. These observations showed that global prediction accuracy (5 - 11) was mainly determined by compounds falling into the intermediate potency sub-range (7 - 9) for different methodologies regardless of the differences in potency distributions within activity classes. In contrast, different characteristics were observed for highly and weakly potency sub-ranges where larger prediction errors were consistently detected. Therefore, the low prediction errors systematically recorded for the intermediate potency sub-range (7 - 9) originated the negligible performance differences observed for the different prediction methods (ML and controls) on the global potency range (5 - 11). Taken together, these findings provided a clear rationale why traditional benchmark calculations were unable to assess compound potency prediction methods in a meaningful way. In the next chapter, compound potency predictions are evaluated in the presence of structural analogs.

Chapter 5

Anatomy of Potency Predictions Focusing on Structural Analogues with Increasing Potency Differences Including Activity Cliffs

The following chapter summarizes the research published as
Janela, T.; Bajorath, J. Anatomy of Potency Predictions Focusing on Structural Analogues with Increasing Potency Differences Including Activity Cliffs. *J. Chem. Inf. Model.* **2023**, *63*, 7032-7044. DOI: 10.1021/acs.jcim.3c01530

Author contributions: Tiago Janela: Methodology, Software, Formal analysis, Investigation, Writing - original draft preparation, Writing - review and editing. **Jürgen Bajorath**: Conceptualization, Methodology, Formal analysis, Writing - original draft preparation, Writing - review and editing.

5.1 Summary

In the previous chapter, the limitations and challenges of potency benchmarks were discussed. In this chapter, the goal was to find an alternative evaluation methodology to compare prediction models. The new test system was designed to systematically evaluate the prediction accuracy of ML and control models for potency difference intervals assigned to individual MMPs, including ACs. Furthermore, an explainable AI methodology was used to gain a better understanding of the ML model predictions.

To investigate potency predictions over increasing potency difference intervals between close analogs, 10 activity classes selected from high-confidence data were used for MMP extraction and modeling. To this end, MMS with diverse potency value ranges were identified using the CCR algorithm.[67] Consequently, MMP training and test sets were built for all activity classes using stratified and random data splits. For stratified partitioning, each MMP compound was assigned to either the training or test set. Alternatively, for random partitioning MMP compounds were randomly selected and assigned to either training or test sets. Based on the MMP data sets, compound potency predictions were performed using ML models (SVR, RFR) and controls (1-NN, k-NN, MR) over independent trials. Prediction performance was subsequently evaluated using MAE and subject to statistical significance testing. The analysis for global compound predictions showed that ML, 1-NN and k-NN displayed very similar accuracies resulting in meaningful predictions for different activity classes. Moreover, MR controls yielded higher prediction errors across compound classes consistent with results from previous chapters (*2*, *3* and *4*). Overall, stratified MMP-splitting showed slightly higher accuracy compared to random partitioning, possibly due to the guaranteed presence of close analogs (MMPs) between training and test sets. This initial assessment of the global prediction performance provided a baseline for the evaluation of prediction accuracy based on MMP potency differences. In order to design a novel MMP-based evaluation system, test compound distributions were generated over increasing MMP potency difference intervals (e.g., [0, 0.5], (0.5, 1.0]). For the selected activity classes, test sets showed comparable compound distributions. For most test

compounds, very small potency differences were observed compared to the corresponding MMP partner. With increasing potency differences the number of test compounds decreased rapidly, resulting in very small numbers of AC compounds. For all activity classes, similar results were observed for stratified and random MMP splits based on this evaluation scheme. For test compounds, prediction accuracy decreased with increased MMP potency differences. For test ACs large prediction errors were detected, demonstrating the failure of ML models and controls to correctly predict AC compound potency. To study which features might be determining the MMP predictions, cumulative SV (SVR)[126] and SHAP (RFR)[127] values across all MMP test compounds were determined. Therefore, four groups of features were considered including features present and absent in both MMP compounds, present in one compound or absent in only one MMP compound. Based on this feature analysis, MMP compound predictions were mainly determined by the positive contribution of shared MMP features and the negative contribution of absent features in the corresponding compounds. These findings provided additional evidence for the similarity of potency predictions for MMP compounds. Taken together, the newly introduced MMP-based evaluation system allowed the monitoring of prediction performance over increasing MMP potency difference intervals. Initial performance assessment across multiple activity classes and MMP data sets showed that compound predictions made by ML models and controls produced similar results, consistent with previous findings. However, when combining MMP-based analysis with stratified partitioning it was revealed that prediction accuracy decreased with increasing potency differences. This indicated that the ML models had a notable tendency to predict potency from close analog compounds in training sets. From this analysis, it is recommended that model accuracy for compound potency predictions should be monitored based on the similarity between training and test compounds. In the following chapter, a new compound potency prediction methodology using a structure-activity FP coupled with a DL approach is introduced.

Chapter 6

Predicting Potent Compounds Using a Conditional Variational Autoencoder Based upon a New Structure-Potency Fingerprint

The following chapter summarizes the research published as
Janela, T.; Takeuchi, K.; Bajorath, J. Predicting Potent Compounds Using a Conditional Variational Autoencoder Based upon a New Structure-Potency Fingerprint. *Biomolecules* **2023**, *13*, 393. DOI: 10.3390/biom13020393

Author contributions: **Tiago Janela**: Methodology, Resources, Investigation, Formal analysis, Writing - review and editing. **Kosuke Takeuchi**: Methodology, Resources, Investigation, Formal analysis, Writing - review and editing. **Jürgen Bajorath**: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review and editing.

6.1 Summary

In the previous chapters, the general limitations associated with benchmark calculations for compound potency predictions were uncovered using ML and control models. Under current benchmark conditions, ML and DL models such as SVR, RFR, DNN and GCN were very competitive, however, the introduction of novel approaches for compound potency prediction continued to be of interest. In this chapter, a new methodology for compound potency predictions was developed by using a novel FP, named the structure-potency FP (SPFP), which was combined with a CVAE. The SPFP was designed to combine potency information with compound structure as a unique representation. The design of SPFP consists of a combination of an extended connectivity FP (structure module) and a potency module comprising a cumulative range encoding of the potency values. A CVAE was trained with SPFP to predict the potency module of test compounds based on the structural module. The ability of SPFP-CVAE for potency prediction was investigated by comparing prediction performance against state-of-the-art models (SVR, RFR, DNN) and respective k-NN, mean regression control models. In addition, traditional ML and DL models were evaluated in the ability to predict the most potent test compounds.

Therefore, high-confidence activity data was extracted from ChEMBL database, followed by the random selection of 10 activity classes used for model generation, optimization and evaluation. To design the SPFP potency module, various bit schemes (e.g., single value, value range, or cumulative) were initially explored to encode potency values into a relevant potency interval corresponding to the -logIC_{50} range from 5 to 11. Among all tested approaches, cumulative potency encoding displayed the best prediction stability across different activity classes. Additionally, different bit sizes (100, 500, 1000) were tested to assess potential resolution limits for the potency module. For the different bit sizes, prediction performance was very similar among the evaluated potency modules. Based on these results, the final potency module size was set to a minimum of 100 bits combined with a cumulative encoding scheme. For this potency module size, each bit position represented 0.06 log units which inherently restricted potency predictions to 6% of a log unit. This resolution was considered appro-

priate for the methodology since it matched the general interval of experimental accuracy limits. The final SPFP size comprised 2148 bits positions (2048 + 100). Subsequently, for each activity class, SPFP-CVAE, SVR, RFR, DNN and control (k-NN, mean regression) models were generated and prediction performance was evaluated using MAE and RMSE over multiple independent prediction trials. As an additional control, SPFP-CVAE randomized models were built using randomly shuffled compound potency values for each activity class. For traditional ML, DL and simple k-NN models, similar and meaningful prediction accuracies were observed for different activity classes. SVR models displayed slightly higher accuracy compared to SPFP-CVAE and the remaining ML models, confirmed by significance statistical analysis. Consistent with the findings from previous chapters, no performance advantage between DL and ML models was detected. Furthermore, mean regression and randomized SPFP-CVAE models showed higher prediction errors compared to the accurate prediction models for these activity classes. These two control models provided an upper prediction performance limit for model comparisons. Since predictions of highly potent compounds were most challenging ML, DL and k-NN models were trained on original sets and evaluated for their ability to predict the 10% most potent compounds present in the test sets. Here, the prediction error for the most potent compounds was higher than for the corresponding global predictions. However, similar prediction performance was observed for SVR, RFR, SPFP-CVAE and k-NN models. Taken together, SPFP-CVAE was introduced as a novel framework for compound potency predictions, rivaling the performance of state-of-the-art ML (SVR, RFR) and DL (DNN) models under the current benchmark conditions. This prediction framework provides a possible alternative to the current supervised ML methods. The proposed approach can be further explored and extended to other molecular property prediction tasks. The next chapter summarizes the current limitations of compound potency predictions under benchmark settings and discusses alternative evaluation schemes.

Chapter 7

Uncovering and Tackling Fundamental Limitations of Compound Potency Predictions Using Machine Learning Models

The following chapter summarizes the research published as
Janela, T.; Bajorath, J. Uncovering and Tackling Fundamental Limitations of Compound Potency Predictions Using Machine Learning Models. *Cell Reports Physical Science* **2024**, *5*, 101988. DOI: 10.1016/j.xcrp.2024.101988

Author contributions: **Tiago Janela**: Illustration preparation, Writing - original draft, Writing - review and editing. **Jürgen Bajorath**: Writing - original draft, Writing - review and editing.

7.1 Summary

In the previous chapters, limitations of compound potency benchmarks using ML, DL and control models were investigated across many activity classes. These studies revealed general limitations preventing a reliable comparison of potency prediction models. As a consequence, alternative benchmark calculations should be explored. In this chapter, the general limitations of compound potency prediction benchmarks are described and summarized. In addition, an alternative benchmark scheme yielding a more reasonable evaluation of state-of-the-art prediction models is introduced. Finally, this chapter includes future directions for improving compound potency predictions.

Based on the results of the previous chapters, several critical observations were made in the analysis of compound potency prediction benchmark calculations. The first observation was that k-NN models approached or reached the prediction performance of complex ML and DL models (*Chapter 2*). For several of the studied activity classes, accurate prediction models achieved, on average, MAE values of ∼0.5 and approaching experimental limits (∼0.3).[69,128] Moreover, observed differences in prediction accuracies between ML, DL and k-NN models were only ∼0.1-0.2 MAE. SVR displayed slightly higher accuracy compared to k-NN, RFR, DNN and GCN models. These findings were further explored in *Chapter 3* and similar trends were observed for a much larger set of activity classes. Second, random regression models produced errors of ∼0.9 MAE for several activity classes, in contrast to ∼0.5 from the original prediction models. Hence, only small error differences separated randomized models and ML/DL models, as described in *Chapter 2*. Next, ML models predictions were shown to be biased toward median potency values associated with compound potency value distributions. Prediction performance was assessed across different potency intervals, as detailed in *Chapter 4*. Larger ML prediction errors were observed for potency sub-ranges 5 - 7 and 9 - 11. In contrast, the intermediate potency sub-range 7 - 9 yielded the lowest prediction errors similar to the MR control. Prediction accuracy for the potency sub-range 7 - 9 was similar to the global prediction accuracy (5 - 11), given that the majority of the compounds were present in this potency interval. Finally, ML predictions

were found to be biased by the presence of structural analogs. As described in *Chapter 5*, a test system based on MMP compounds was created to determine the effects of the presence of analogs in training and test sets. This analysis exposed the tendency of ML models to predict test compound potency based on the potency values from closest training analogs. Therefore, the presence of close analogs should be limited and can be achieved using AS-based data partitioning. Based on these previous findings, alternative benchmark systems must be explored to further improve method comparisons. As a proof-of-concept, a new test system was introduced by generating ML models in the presence of inactive compounds (with -logIC_{50} set to 0). These ML models were derived for training sets with an increasing number of inactive compounds using two different inactive selection strategies. Inactive compounds were either selected randomly selecting compounds from different activity classes (random selection) or from a single activity class (homogeneous selection). For the studied activity classes, prediction performance for active compounds decreased with increasing numbers of inactive compounds. In addition, increasing differences between ML and control models were observed. DNN models displayed larger prediction errors compared to SVR and RFR models. Moreover, prediction models trained with randomly selected inactive compounds achieved lower prediction accuracy compared to homogeneous models because this first selection method increased the training compound diversity. In contrast, simple k-NN calculations were less affected by the addiction of inactive training compounds, given that this addition did not substantially replace nearest neighbors of active training compounds. Therefore, benchmark calculations with the addition of inactive training compounds resulted in larger differences between prediction models compared to conventional benchmark systems. Finally, with the introduction of novel benchmark concepts for model comparison, computational approaches should be more closely combined with experimentation. Prospective prediction of experimentally confirmed potent compounds provides an ultimate measure for the relevance of ML models and enables more realistic methodological comparisons.

Chapter 8

Conclusion

Compound potency prediction is of major importance in computational medicinal chemistry. Over the years, an increasing number of novel prediction methodologies have been introduced following the introduction of DL architectures in pharmaceutical research. Together with the increase in available high-quality data, these approaches aim to reduce experimental cost and time requirements. Typically, new prediction methods are evaluated and compared to state-of-the-art models using conventional benchmark calculations. These comparison schemes usually rely on highly curated data partitioned into training and test sets for model building and evaluation over multiple rounds of cross-validation. Model performance is evaluated with commonly used regression metrics and statistical significance analysis. ML models of increasing complexity often display only very small performance differences compared to simpler control models under these conditions. Reasons for these observations were unknown. Thus, to accurately assess the quality of existing and novel predictive approaches, further exploration of benchmark calculations was required. Without properly defined benchmark conditions and clearly stated limitations, obtaining a reliable assessment of state-of-the-art prediction models is impossible. The quality of compound data sets used for benchmark calculations must be carefully determined. Furthermore, it must be analyzed how the presence or absence of closely related compounds in training and test sets might affect predictions. Therefore, limitations of conventional compound potency benchmark predictions are investigated in detail and from different perspectives. In the first study (*Chapter 2*), compound potency predictions were performed

using ML, DL and simple control models and evaluated under different benchmark conditions. For 10 activity classes, ML (SVR, RFR) and DL (DNN, GCN) approaches achieved accurate predictions. However, DL models did not significantly increase the prediction accuracy of other ML models. Overall, SVR achieved slightly better performance compared to RFR, DNN and GCN models. Notably, simple k-NN models reached or surpassed the performance of increasingly complex ML and DL models, under these evaluation conditions. Furthermore, MR and random regression control calculations showed prediction accuracy approaching ML and DL performance, within less than 0.5 orders of magnitude. Furthermore, for test sets of structurally unique and highly potent compounds, ML, DL and simple k-NN displayed similar prediction accuracy, followed by MR models for the majority of activity classes. Based on these findings, further evaluation of benchmark calculations for compound potency predictions was required. In *Chapter 3*, compound potency prediction analysis was further extended to 376 activity classes using SVR and control models. SVR, 1-NN and k-NN models displayed similar prediction performance across all activity classes. In addition, the influence of training and test set sizes on model performance was investigated. SVR, 1-NN and k-NN models were derived for the largest 45 activity classes using differently sized training and test partitions (80/20% and 50/50%). Under these conditions, differences between prediction methods were negligible, indicating little influence of varying training set sizes in model performance. Furthermore, the effects of activity class composition and potency range distribution as potentially limiting factors in benchmark calculations were investigated. SVR, 1-NN and k-NN models were developed for data sets with a reduced number of close neighbors and balanced potency distributions. For these two data set modifications, prediction errors increased, however, differences in performance between SVR, 1-NN and k-NN remained minimal. This large-scale analysis showed that conventional benchmark calculations for comparing predictive models had general limitations, reinforcing the need for further investigations. In the third study (*Chapter 4*), the impact of potency value distributions and corresponding sub-ranges on compound potency prediction was investigated. ML and control models were generated for 8 suitable activity classes and predictive performance was assessed for the entire potency range (-logIC_{50}: 5 - 11). Overall, SVR, RFR, 1-NN and 3-NN mod-

els achieved similar prediction accuracy, consistent with previous observations. Moreover, test compounds were grouped in potency sub-ranges (5 - 7, 7 - 9, 9 - 11) and prediction performance was re-evaluated for each potency sub-range. For weakly (5 - 7) and highly (9 - 11) potent compounds, larger prediction errors were observed compared to the intermediate potency sub-range (7 - 9), which displayed consistently high accuracy similar to the entire potency range (5 - 11). Moreover, performance differences largely increased between MR and remaining ML models, for the outer potency sub-ranges (5 - 7 and 9 - 11) compared to the 7 - 9 sub-range where ML and MR control models closely matched prediction performance. Furthermore, training sets of increasing size were generated to analyze the learning characteristics of each potency sub-range. ML and control models were developed using these training sets and prediction performance was evaluated for each potency sub-range individually. SVR, RFR, 1-NN and 3-NN models showed an incremental increase in prediction accuracy for the potency sub-ranges 5 - 7 and 9 - 11, together with increasing performance differences compared to MR models when training set sizes were augmented. Surprisingly, for potency sub-range 7 - 9, the performance of SVR and RFR models remained stable and close to the median training value for increasing training set sizes. Thus, no learning was required for ML models to correctly predict compound potency in the intermediate sub-range. These findings provided a clear explanation as to why compound potency predictions could not be adequately evaluated using standard benchmark calculations. Therefore, in the next study (*Chapter 5*), a new test scheme for monitoring model performance over potency difference intervals of MMP compounds was explored. Accordingly, ML and control models were derived using MMP-based data sets using different MMP data partitions. Similar to previous observations, SVR, RFR, 1-NN and k-NN models showed comparable prediction accuracy for random and stratified data partitions on all activity classes. Moreover, for MMP compounds, prediction errors increased with increasing potency for SVR, RFR, 1-NN, k-NN and MR control models. Additionally, the models' inability to correctly predict the potency of AC compounds was highlighted. Thereafter, SV/SHAP formalism was employed to identify features driving MMP compound predictions of SVR and RFR models. ML feature analysis showed that MMP potency predictions were determined by positive contributions of fea-

tures shared between MMP compounds and negative contributions of absent features in MMP compounds independent of their potency differences. These observations provided an additional rationale for prediction errors over increasing potency difference intervals. In *Chapter 6*, a novel DL-based methodology was introduced for compound potency prediction based on an FP design encoding both compound structure and potency (SPFP) information in combination with a CVAE model (termed SPFP-CVAE). In contrast to learning relationships between compound structures and potency values, SPFP-CVAE provided a framework for the prediction of specific bit settings in the potency module from settings in a corresponding structural module. Thus, potency module predictions were derived by sampling from the CVAE decoder architecture using the FP structural module. Therefore, SPFP-CVAE's ability to predict compound potency was compared to state-of-the-art ML, DL and control models. The SPFP-CVAE methodology performed similarly to SVR, RFR, k-NN and DNN models across all studied activity classes. In light of benchmark limitations, the prediction of highly potent compounds was of particularly high interest. Therefore, SPFP-CVAE was further evaluated and compared to other ML models. Prediction accuracy for the most potent compounds was comparable for SPFP-CVAE, SVR, k-NN and DNN models. Finally, *Chapter 7* summarized the general limitations of conventional benchmark calculations for compound potency predictions uncovered in the previous chapters. Importantly, k-NN models reached or surpassed the performance of more complex ML models and only confined accuracy differences were observed for ML and control models compared to randomized models (*Chapters 2* and *3*). Furthermore, ML predictions were biased by median potency values (*Chapter 4*) and available structural analogs (*Chapter 5*). Moreover, a novel proof-of-concept regression benchmark system was introduced (*Chapter 6*). ML and control models were derived in the presence of inactive training compounds and the performance of active test compounds was evaluated. Increasing the number of inactive training compounds increased prediction errors for all approaches and larger performance differences were consistently observed for SVR, RFR, DNN and k-NN control models. Finally, future directions for compound potency predictions were discussed. These included the need for novel approaches for model performance assessment and experimental evaluation of predicted potent compounds. In

conclusion, this dissertation explored the current state of conventional benchmark calculations for compound potency predictions using ML models. As chemical libraries increase in size, the use of ML methods will further increase. Consequently, rigorous benchmarking of proposed state-of-the-art methods is essential. Conventional benchmark calculations are currently insufficient to assess the performance of ML methods in potency prediction in a meaningful way. The limitations unveiled herein emphasize the need for novel benchmark systems capable of reliably comparing different models, especially for their ability to predict potent compounds in prospective applications. Assessing the ability of ML models to correctly detect potent compounds should outweigh statistical analysis of benchmark performance. In addition, the availability of highly curated and standardized data sets for ML will be essential for the field to further progress. This especially applies to academia, which mainly relies on publicly available data.

Bibliography

[1] Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of Early Drug Discovery. *British Journal of Pharmacology* **2011**, *162*, 1239–1249.

[2] Tabana, Y.; Babu, D.; Fahlman, R.; Siraki, A. G.; Barakat, K. Target Identification of Small Molecules: An Overview of the Current Applications in Drug Discovery. *BMC Biotechnology* **2023**, *23*, 44.

[3] Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. *Nat. Rev. Drug Discov.* **2003**, *2*, 369–378.

[4] Barcelos, M. P.; Gomes, S. Q.; Federico, L. B.; Francischini, I. A. G.; Hage-Melim, L. I. d. S.; Silva, G. M.; de Paula da Silva, C. H. T. In *Research Topics in Bioactivity, Environment and Energy: Experimental and Theoretical Tools*; Taft, C. A., de Lazaro, S. R., Eds.; Springer International Publishing: Cham, 2022; pp 481–500.

[5] Jeetu, G.; Anusha, G. Pharmacovigilance: A Worldwide Master Key for Drug Safety Monitoring. *J. Young Pharm.* **2010**, *2*, 315–320.

[6] Schlander, M.; Hernandez-Villafuerte, K.; Cheng, C.-Y.; Mestre-Ferrandiz, J.; Baumann, M. How Much Does It Cost to Research and Develop a New Drug? A Systematic Review and Assessment. *PharmacoEconomics* **2021**, *39*, 1243–1269.

[7] Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discov.* **2010**, *9*, 203–214.

[8] Kiriiri, G. K.; Njogu, P. M.; Mwangi, A. N. Exploring Different Approaches to Improve the Success of Drug Discovery and Development Projects: A Review. *Future Journal of Pharmaceutical Sciences* **2020**, *6*, 27.

[9] Bender, A.; Cortés-Ciriano, I. Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 1: Ways to Make an Impact, and Why We Are Not There Yet. *Drug Discovery Today* **2021**, *26*, 511–524.

[10] Wallach, I. et al. Ai Is a Viable Alternative to High Throughput Screening: A 318-Target Study. *Sci. Rep.* **2024**, *14*, 7526.

[11] Sabe, V. T.; Ntombela, T.; Jhamba, L. A.; Maguire, G. E. M.; Govender, T.; Naicker, T.; Kruger, H. G. Current Trends in Computer Aided Drug Design and a Highlight of Drugs Discovered Via Computational Techniques: A Review. *European Journal of Medicinal Chemistry* **2021**, *224*, 113705.

[12] Sadybekov, A. V.; Katritch, V. Computational Approaches Streamlining Drug Discovery. *Nature* **2023**, *616*, 673–685.

[13] Bajorath, J. Computer-Aided Drug Discovery. *F1000Research* **2015**, *4*.

[14] Cherkasov, A. et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010.

[15] Lewis, R. A.; Wood, D. Modern 2D QSAR for Drug Discovery. *WIREs Computational Molecular Science* **2014**, *4*, 505–522.

[16] Lill, M. A. Multi-Dimensional QSAR in Drug Discovery. *Drug Discovery Today* **2007**, *12*, 1013–1017.

[17] Guedes, I. A.; Pereira, F. S. S.; Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.* **2018**, *9*, 1089.

[18] Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.

[19] Liu, J.; Wang, R. Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **2015**, *55*, 475–482.

[20] Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys. Rev.* **2017**, *9*, 91–102.

[21] Williams-Noonan, B. J.; Yuriev, E.; Chalmers, D. K. Free Energy Methods in Drug Design: Prospects of Alchemical Perturbation in Medicinal Chemistry. *J. Med. Chem.* **2018**, *61*, 638–649.

[22] Gleeson, M. P.; Gleeson, D. QM/MM Calculations in Drug Discovery: A Useful Method for Studying Binding Phenomena? *J. Chem. Inf. Model.* **2009**, *49*, 670–677.

[23] Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angewandte Chemie International Edition* **2009**, *48*, 1198–1229.

[24] Zhou, T.; Huang, D.; Caflisch, A. Quantum Mechanical Methods for Drug Design. *Curr. Top. Med. Chem.* **2010**, *10*, 33–45.

[25] Tropsha, A.; Isayev, O.; Varnek, A.; Schneider, G.; Cherkasov, A. Integrating QSAR Modelling and Deep Learning in Drug Discovery: The Emergence of Deep QSAR. *Nat. Rev. Drug Discov.* **2024**, *23*, 141–155.

[26] David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular Representations in AI-Driven Drug Discovery: A Review and Practical Guide. *J. Cheminf.* **2020**, *12*, 56.

[27] Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

[28] Krenn, M. et al. Selfies and the Future of Molecular String Representations. *Patterns* **2022**, *3*, 100588.

[29] Kay, E. Graph Theory with Applications. *Journal of the Operational Research Society* **1977**, *28*, 237–238.

[30] Leo, A.; Jow, P. Y.; Silipo, C.; Hansch, C. Calculation of Hydrophobic Constant (log P) from pi and f Constants. *J. Med. Chem.* **1975**, *18*, 865–868.

[31] Bajusz, D.; Rácz, A.; Héberger, K. In *Comprehensive Medicinal Chemistry III*; Chackalamannil, S., Rotella, D., Ward, S. E., Eds.; Elsevier: Oxford, 2017; pp 329–378.

[32] Accelrys, MACCS keys. MDL Information Systems, Inc. 2011.

[33] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

[34] Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

[35] Lim, J.; Ryu, S.; Park, K.; Choe, Y. J.; Ham, J.; Kim, W. Y. Predicting DrugTarget Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. *J. Chem. Inf. Model.* **2019**, *59*, 3981–3988.

[36] Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. ProteinLigand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957.

[37] Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in Drug Design - A Review. *Current Topics in Medicinal Chemistry* **2010**, *10*, 95–115.

[38] Medina-Franco, J. L.; Maggiora, G. M. *Chemoinformatics for Drug Discovery*; John Wiley & Sons, Ltd, 2013; pp 343–399.

[39] Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

[40] Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of Molecular Diversity Descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.

[41] Bajorath, J. In *Bioinformatics: Volume II: Structure, Function, and Applications*; Keith, J. M., Ed.; Springer: New York, NY, 2017; pp 231–245.

[42] Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

[43] Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants: The Computer Is Programmed to Simulate the Taxonomic Process of Comparing Each Case with Every Other Case. *Science* **1960**, *132*, 1115–1118.

[44] Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

[45] Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry: Miniperspective. *J. Med. Chem.* **2014**, *57*, 3186–3204.

[46] Jasial, S.; Hu, Y.; Vogt, M.; Bajorath, J. Activity-Relevant Similarity Values for Fingerprints and Implications for Similarity Searching. *F1000Research* **2016**, *5*, 591.

[47] Venkatraman, V.; Gaiser, J.; Demekas, D.; Roy, A.; Xiong, R.; Wheeler, T. J. Do Molecular Fingerprints Identify Diverse Active Drugs in Large-Scale Virtual Screening? (No). *Pharmaceuticals* **2024**, *17*, 992.

[48] Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339–348.

[49] Yang, Z.; Shi, S.; Fu, L.; Lu, A.; Hou, T.; Cao, D. Matched Molecular Pair Analysis in Drug Discovery: Methods and Recent Applications. *J. Med. Chem.* **2023**, *66*, 4361–4377.

[50] Dossetter, A. G.; Griffen, E. J.; Leach, A. G. Matched Molecular Pair Analysis in Drug Discovery. *Drug Discovery Today* **2013**, *18*, 724–731.

[51] Raymond, J. W.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching of Chemical Structures. *J. Comput. Aided Mol. Des.* **2002**, *16*, 521–533.

[52] Sheridan, R. P.; Miller, M. D. A Method for Visualizing Recurrent Topological Substructures in Sets of Active Molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 915–924.

[53] Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP–Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

[54] León, A. d. l. V. d.; Bajorath, J. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. *Med. Chem. Commun.* **2013**, *5*, 64–67.

[55] Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure-Activity Relationship Trailing. *J. Med. Chem.* **2011**, *54*, 2944–2951.

[56] Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18–28.

[57] Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.

[58] Stumpfe, D.; Hu, H.; Bajorath, J. Evolving Concept of Activity Cliffs. *ACS Omega* **2019**, *4*, 14360–14368.

[59] Stumpfe, D.; Bajorath, J. Monitoring Global Growth of Activity Cliff Information over Time and Assessing Activity Cliff Frequencies and Distributions. *Future Medicinal Chemistry* **2015**, *7*, 1565–1579.

[60] Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138–1145.

[61] Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of Activity Cliffs Using Support Vector Machines. *J. Chem. Inf. Model.* **2012**, *52*, 2354–2365.

[62] van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **2022**, *62*, 5938–5951.

[63] Marvin; ChemAxon, Ltd: Budapest, 2024. https://chemaxon.com/marvin (accessed 2024-08-23).

[64] Yoshimori, A.; Bajorath, J. Computational Analysis, Alignment and Extension of Analogue Series from Medicinal Chemistry. *Future Sci. OA* **2022**, *8*, FSO804.

[65] Wassermann, A. M.; Bajorath, J. Directed R-Group Combination Graph: A Methodology To Uncover StructureActivity Relationship Patterns in a Series of Analogues. *J. Med. Chem.* **2012**, *55*, 1215–1226.

[66] Stumpfe, D.; Dimova, D.; Bajorath, J. Computational Method for the Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. *J. Med. Chem.* **2016**, *59*, 7667–7676.

[67] Naveja, J. J.; Vogt, M.; Stumpfe, D.; Medina-Franco, J. L.; Bajorath, J. Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound-Core Relationship Method. *ACS Omega* **2019**, *4*, 1027–1032.

[68] Waldman, S. A. Does Potency Predict Clinical Efficacy? Illustration Through an Antihistamine Model. *Ann. Allergy Asthma Immunol.* **2002**, *89*, 7–11.

[69] Landrum, G. A.; Riniker, S. Combining IC_{50} or K_i Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model.* **2024**, *64*, 1560–1567.

[70] Zdrazil, B. et al. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Research* **2024**, *52*, D1180–D1192.

[71] Kabir, A.; Muth, A. Polypharmacology: The Science of Multi-Targeting Molecules. *Pharmacol. Res.* **2022**, *176*, 106055.

[72] Bruns, R. F.; Watson, I. A. Rules for Identifying Potentially Reactive or Promiscuous Compounds. *J. Med. Chem.* **2012**, *55*, 9763–9772.

[73] Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740.

[74] Tingle, B. I.; Tang, K. G.; Castanon, M.; Gutierrez, J. J.; Khurelbaatar, M.; Dandarchuluun, C.; Moroz, Y. S.; Irwin, J. J. ZINC-22A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *J. Chem. Inf. Model.* **2023**, *63*, 1166–1176.

[75] Landrum, G. A.; Beckers, M.; Lanini, J.; Schneider, N.; Stiefl, N.; Riniker, S. SIMPD: An Algorithm for Generating Simulated Time Splits for Validating Machine Learning Approaches. *J. Cheminf.* **2023**, *15*, 119.

[76] Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.

[77] Rainio, O.; Teuho, J.; Klén, R. Evaluation Metrics and Statistical Tests for Machine Learning. *Sci. Rep.* **2024**, *14*, 6086.

[78] Student. The Probable Error of a Mean. *Biometrika* **1908**, *6*, 1–25.

[79] Conover, W. J. On Methods of Handling Ties in the Wilcoxon Signed-Rank Test. *Journal of the American Statistical Association* **1973**, *68*, 985–988.

[80] Aickin, M.; Gensler, H. Adjusting for Multiple Testing When Reporting Research Results: The Bonferroni Vs Holm Methods. *Am. J. Public Health* **1996**, *86*, 726–728.

[81] Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

[82] Walters, W. P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc. Chem. Res.* **2021**, *54*, 263–270.

[83] Drucker, H.; Surges, C. J.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. Advances in Neural Information Processing Systems. 1997; pp 155–161.

[84] Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

[85] Li, H.; Zhang, R.; Min, Y.; Ma, D.; Zhao, D.; Zeng, J. A Knowledge-Guided Pre-Training Framework for Improving Molecular Representation Learning. *Nat Commun* **2023**, *14*, 7568.

[86] Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. 2020, arXiv:2010.09885, arXiv.org e-Print archive. http://arxiv.org/abs/2010.09885.

[87] Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory* **1967**, *13*, 21–27.

[88] Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* **1992**, *46*, 175.

[89] Mostafa, A. A.; Alhossary, A. A.; Salem, S. A.; Mohamed, A. E. GBO-KNN a New Framework for Enhancing the Performance of Ligand-Based Virtual Screening for Drug Discovery. *Expert Systems with Applications* **2022**, *197*, 116723.

[90] Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. *J. Chem. Inf. Model.* **2008**, *48*, 220–232.

[91] Rodríguez-Pérez, R.; Bajorath, J. Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *J. Comput. Aided Mol. Des.* **2022**, *36*, 355–362.

[92] Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273–297.

[93] Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Statistics and Computing* **2004**, *14*, 199–222.

[94] Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Networks* **2005**, *18*, 1093–1110.

[95] Breiman, L. Bagging Predictors. *Machine Learning* **1996**, *24*, 123–140.

[96] Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 468–481.

[97] Nielsen, M. A. *Neural Networks and Deep Learning*; Determination Press: San Francisco, CA, USA, 2015; Vol. 2018.

[98] Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, 2016; Vol. 1.

[99] Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533–536.

[100] Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263–274.

[101] Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, *59*, 1253–1268.

[102] Rajan, K.; Zielesny, A.; Steinbeck, C. Decimer: Towards Deep Learning for Chemical Image Recognition. *J. Cheminf.* **2020**, *12*, 65.

[103] Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput. Aided Mol. Des.* **2016**, *30*, 595–608.

[104] Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. Advances in Neural Information Processing Systems. 2015.

[105] Rittig, J. G.; Gao, Q.; Dahmen, M.; Mitsos, A.; Schweidtmann, A. M. *Graph Neural Networks for the Prediction of Molecular Structure-Property Relationships*; 2023; pp 159–181.

[106] Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A Compact Review of Molecular Property Prediction with Graph Neural Networks. *Drug Discovery Today: Technologies* **2020**, *37*, 1–12.

[107] Torng, W.; Altman, R. B. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *J. Chem. Inf. Model.* **2019**, *59*, 4131–4149.

[108] Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *J. Med. Chem.* **2020**, *63*, 8835–8848.

[109] Xiong, J.; Xiong, Z.; Chen, K.; Jiang, H.; Zheng, M. Graph Neural Networks for Automated *De Novo* Drug Design. *Drug Discovery Today* **2021**, *26*, 1382–1393.

[110] Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. 2022, arXiv:1312.6114. arXiv.org e-Print archive. http://arxiv.org/abs/1312.6114.

[111] Rezende, D. J.; Mohamed, S.; Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. 2014, arXiv:1401.4082, arXiv.org e-Print archive. http://arxiv.org/abs/1401.4082.

[112] Doersch, C. Tutorial on Variational Autoencoders. 2021, arXiv:1606.05908, arXiv.org e-Print archive. http://arxiv.org/abs/1606.05908.

[113] Kullback, S.; Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* **1951**, *22*, 79–86.

[114] Wei, R.; Garcia, C.; El-Sayed, A.; Peterson, V.; Mahmood, A. Variations in Variational Autoencoders - A Comparative Evaluation. *IEEE Access* **2020**, *8*, 153651–153670.

[115] Sohn, K.; Lee, H.; Yan, X. Learning Structured Output Representation using Deep Conditional Generative Models. Advances in Neural Information Processing Systems. 2015.

[116] Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

[117] Tevosyan, A.; Khondkaryan, L.; Khachatrian, H.; Tadevosyan, G.; Apresyan, L.; Babayan, N.; Stopper, H.; Navoyan, Z. Improving VAE Based Molecular Representations for Compound Property Prediction. *J. Cheminf.* **2022**, *14*, 69.

[118] Tempke, R.; Musho, T. Autonomous Design of New Chemical Reactions Using a Variational Autoencoder. *Commun. Chem.* **2022**, *5*, 1–10.

[119] Rodríguez-Pérez, R.; Bajorath, J. Explainable Machine Learning for Property Predictions in Compound Optimization. *J. Med. Chem.* **2021**, *64*, 17744–17752.

[120] Shapley, L. S. In *Contributions to the Theory of Games, Volume II*; Kuhn, H. W., Tucker, A. W., Eds.; Princeton University Press, 2016; pp 307–318.

[121] Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 2017.

[122] Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. *J. Med. Chem.* **2020**, *63*, 8761–8777.

[123] Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Machine Learning Models Using Shapley Values: Application to Compound Potency and Multi-Target Activity Predictions. *J. Comput. Aided Mol. Des.* **2020**, *34*, 1013–1026.

[124] Mobley, D. L.; Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annual Review of Biophysics* **2017**, *46*, 531–558.

[125] Li, Y.; Hsieh, C.-Y.; Lu, R.; Gong, X.; Wang, X.; Li, P.; Liu, S.; Tian, Y.; Jiang, D.; Yan, J.; Bai, Q.; Liu, H.; Zhang, S.; Yao, X. An Adaptive Graph Learning Method for Automated Molecular Interactions and Properties Predictions. *Nat. Mach. Intell.* **2022**, *4*, 645–651.

[126] Feldmann, C.; Bajorath, J. Calculation of Exact Shapley Values for Support Vector Machines with Tanimoto Kernel Enables Model Interpretation. *iScience* **2022**, *25*, 105023.

[127] Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67.

[128] Brown, S. P.; Muchmore, S. W.; Hajduk, P. J. Healthy Skepticism: Assessing Realistic Model Performance. *Drug Discovery Today* **2009**, *14*, 420–427.

Appendix A

Simple Nearest-Neighbour Analysis Meets the Accuracy of Compound Potency Predictions Using Complex Machine Learning Models

nature machine intelligence

Article

https://doi.org/10.1038/s42256-022-00581-6

Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models

Received: 13 July 2022

Accepted: 3 November 2022

Published online: 14 December 2022

Check for updates

Tiago Janela & Jürgen Bajorath

Compound potency prediction is a popular application of machine learning in drug discovery, for which increasingly complex models are employed. The general aim is the identification of new chemical entities that are highly potent against a given target. The relative performance of potency prediction models and their accuracy limitations continue to be debated in the field, and it remains unclear whether deep learning can further advance potency prediction. We have analysed and compared approaches of varying computational complexity for potency prediction and shown that simple nearest-neighbour analysis consistently meets or exceeds the accuracy of machine learning methods regarded as the state of the art in the field. Moreover, completely random predictions using different models were shown to reproduce experimental values within an order of magnitude, resulting from the potency value distributions in commonly used compound data sets. Taken together, these findings have important implications for typical benchmark calculations to evaluate machine learning performance. Simple controls such as nearest-neighbour analysis should generally be included in model evaluation. Furthermore, the narrow margin separating the best and completely random potency predictions is unrealistic and requires the consideration of alternative benchmark criteria, as discussed herein.

In cheminformatics and medicinal chemistry, the prediction of compound potency or other molecular properties plays a central role. For potency prediction, ligand- and structure-based approaches are applied[1–5], many of which employ machine learning (ML)[5]. As in other areas where artificial intelligence has become a focal point, complex deep learning architectures are increasingly used for potency/property prediction, in both structure- and ligand-based modelling[5–12]. However, despite apparent advances, some of these predictions are also controversially viewed[13–15].

Ligand-based potency prediction accounting for nonlinear structure–activity relationships is a mainstay in cheminformatics. To this end, supervised ML models are derived on the basis of sets of known active compounds to predict the potency of new molecules. While prospective applications to identify novel active compounds represent the ultimate goal, model performance is initially assessed via benchmarking, which is a prerequisite for reporting new computational approaches. In typical benchmark settings, compound data sets with activity against a particular target (often termed activity classes) are divided into training and test sets and predictions are evaluated using cross-validation protocols. This route is conventionally followed when reporting new methods and prediction models.

In our in-house efforts to develop computational approaches for the identification of novel active compounds and prediction of their potency, we have investigated hundreds of activity classes with

Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany. e-mail: bajorath@bit.uni-bonn.de

Table 1 | Activity classes and performance comparison

ChEMBL target ID	Target name	Present study			Sakai et al.[12]	
		Number of compounds	kNN (MAE)	GCN (MAE)	Number of compounds	GCN (MAE)
220	Acetylcholinesterase	1,699	0.49±0.019	0.53±0.045	9,737	0.57±0.019
230	Cyclooxygenase-2	1,166	0.43±0.033	0.49±0.036	5,085	0.68±0.031
260	MAP kinase p38 alpha	1,351	0.45±0.018	0.52±0.03	4,518	0.54±0.017
262	Glycogen synthase kinase-3 beta	1,000	0.52±0.023	0.58±0.031	2,702	0.67±0.022
279	Vascular endothelial growth factor receptor 2	2,273	0.56±0.018	0.63±0.05	8,936	0.55±0.012
284	Dipeptidyl peptidase IV	1,316	0.48±0.026	0.55±0.035	4,517	0.58±0.010
1865	Histone deacetylase 6	1,034	0.49±0.020	0.50±0.027	2,725	0.47±0.023
2409	Epoxide hydratase	1,227	0.63±0.028	0.59±0.024	—	—
4005	PI3-kinase p110-alpha subunit	1,262	0.47±0.028	0.53±0.063	5,699	0.48±0.012
4822	Beta-secretase 1	1,116	0.44±0.023	0.47±0.033	7,554	0.57±0.028
Mean value		—	0.48±0.023	0.53±0.039	—	0.57±0.019

Activity classes investigated in our study are reported. For nine of these ten classes, potency value predictions using GCN were also carried out in an independent study[12]. For these classes, the GCN predictions are compared. In addition, the results of kNN calculations are reported. For all predictions, s.d. values are provided. Mean performance values are calculated for the nine common classes.

methods of greatly varying complexity. In these studies, we have frequently observed that methodological complexity does not scale with prediction accuracy. Herein, we report representative potency predictions demonstrating that a simplistic *k*-nearest-neighbour (kNN) approach consistently meets or exceeds the accuracy of advanced ML methods, including support vector regression (SVR), a widely applied standard in the field, random forest regression (RFR), deep neural network (DNN) and graph convolutional neural network (GCN) with representation learning. DNN and GCN represent increasingly popular deep learning methods for molecular potency/property prediction. Furthermore, we also report different control calculations to determine the intrinsic accuracy limitations of compound potency predictions based on available activity data, yielding some surprising results. Taken together, these findings also suggest that conventional benchmark criteria to assess the predictive performance of ML models require careful reconsideration.

Results

Compound data sets and potency value distributions

From ChEMBL[16], we randomly selected ten activity classes comprising at least 1,000 compounds meeting predefined data confidence criteria (Methods). Large activity classes were selected to ensure the availability of reasonably sized training sets for deep learning. Table 1 (left-hand side) summarizes the composition of these compound classes. As expected for large activity classes originating from medicinal chemistry, the compounds were active against popular pharmaceutical targets.

Figure 1a compares the potency (pIC_{50}) value distributions of the ten randomly selected classes, for which a lower potency threshold of 10 μM was applied, and Fig. 1b shows the potency value distribution of each class. As typically observed for compounds from medicinal chemistry sources, there was substantial overlap between potency values in the micromolar to high-nanomolar range, but there were also large class-dependent differences (Fig. 1a). Median potency values fell into the pIC_{50} range 6–8 (Fig. 1b).

Extended Data Fig. 1 shows structural similarity versus potency difference plots for the activity classes, revealing the presence of many structurally diverse compounds with varying potency differences and decreasing numbers of structurally similar compounds per class. As a general trend, increasing structural similarity corresponded to decreasing potency differences (with exceptions), as expected.

ML models

For each activity class, we generated SVR, RFR, DNN and GCN potency prediction models using standard protocols for the complete data set and size-reduced sets of random composition or increased diversity (Methods). As a control, kNN predictions were carried out in which the potency value of the one, two or three most similar training compounds was assigned to each test compound (averaged if necessary; Methods). As an additional control, a 'median regressor' (MR) was applied that assigned the median potency value of a given training set to each test compound.

Both kNN and MR calculations provided reference points for predictions with ML models of increasing complexity. Independent potency prediction trials using different models were evaluated on the basis of the mean absolute error (MAE) and root mean square error (RMSE), as shown in Fig. 2 and Extended Data Fig. 2, respectively. For both measures, the same trends were observed, with a further increase of RMSE compared with MAE values by ~0.1–0.2 log units.

For the randomly selected activity classes, very similar results were obtained. The MAE/RMSE distributions were generally narrow, except for DNN/GCN and some of the size-reduced training sets, indicating stable predictions over independent trials. As expected, differences between the value distributions in boxplots were at least moderately statistically significant in the majority of cases (Supplementary Table 1), However, for original training sets (comprising 80% of the compounds per class, Methods), prediction accuracy of ~0.5 MAE was observed for all activity classes. The prediction accuracy of SVR was overall slightly superior to that of RFR, DNN and GCN. Importantly, the accuracy of the simple kNN predictions was very similar to or better than the accuracy of the ML models. For the baseline MR predictions assigning a constant median potency value to all test compounds, a class-dependent accuracy of ~0.8–1.0 MAE was obtained.

For nine of the ten activity classes, independent GCN predictions were reported in a large-scale study by Sakai et al.[12]. The GCN results are compared in Table 1 (right-hand side). Sakai et al. used many more active compounds for modelling because these investigators did not apply specific data confidence criteria as in our analysis (Methods). Nonetheless, the accuracy of the independent GCN predictions was encouragingly similar. However, kNN calculations were overall more accurate than GCN predictions (Table 1).

We also derived models for training sets of reduced size to investigate the potential influence of data sparseness on learning.

https://doi.org/10.1038/s42256-022-00581-6

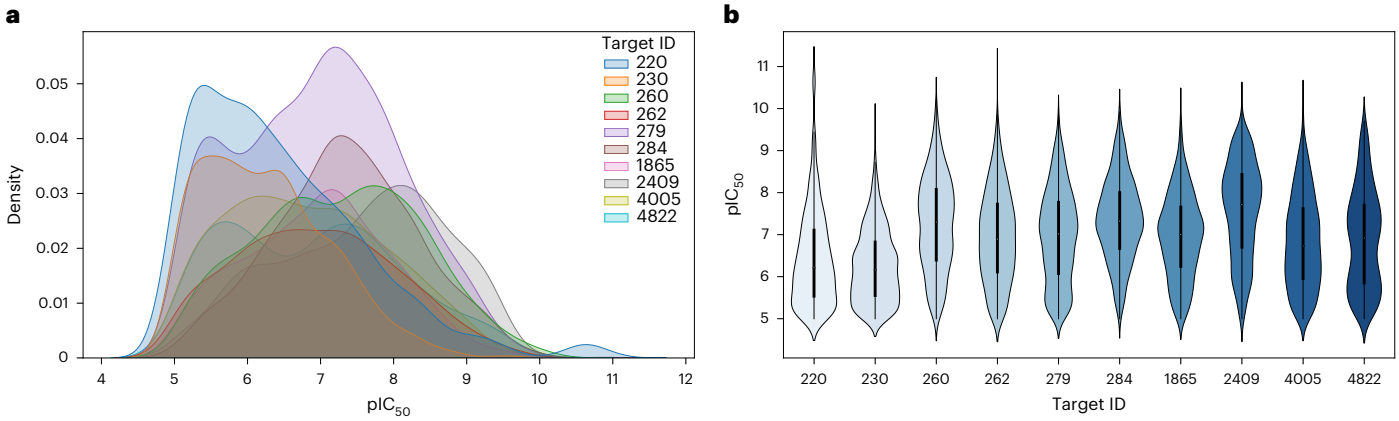


Fig. 1 | Potency value distributions of activity classes. a, The distributions of ten randomly selected activity classes are compared in a density plot, in which the data distribution is determined by a kernel density estimation. **b**, Violin plots report the potency value distribution of each class. In a violin plot, a value distribution is represented by its maximum value (upper thin line), upper quartile (upper thick line), median value (white dot), lower quartile (lower thick line) and minimum value (lower thin line). On each side, a density plot is shown. The number of samples used to generate the violin plots is reported in Table 1 (*n*, number of compounds).

For randomly selected size-reduced training sets (20% of the compounds, Methods), a gradual increase in MAE/RMSE values was observed, as expected (Fig. 2). Here, the gap between the prediction accuracy of kNN/SVR on the one hand and DNN/GCN on the other widened. Moreover, we generated structurally diverse training sets of the same size (20% of the compounds) through dissimilarity selection (Methods). For diverse training sets, the same trends were observed. In this case, DNN approached and GCN partly exceeded the MAE level of MR. However, even for diverse sets, which were designed to principally disfavour kNN, the accuracy of kNN calculations remained closely similar to SVR, with 0.6–0.8 MAE (Fig. 2).

Predictions for unique and highly potent compounds

As another control for kNN predictions, each activity class was partitioned into analogue series (Methods), representing a form of clustering sensitive to medicinal chemistry applications (compounds from analogue series share the same core structure), and the largest analogue series was removed as a hold-out set from each class. Then, models were derived on the basis of all remaining compounds and used to predict the potency of the hold-out set in an individual trial. The predictions were evaluated on the basis of the MAE and RMSE, as shown in Fig. 3 and Extended Data Fig. 3, respectively. With the exception of two activity classes (target IDs 2409 and 4822) where kNN prediction accuracy decreased to ~1.5 MAE, similar prediction accuracy was again observed for all models.

Furthermore, extrapolative predictions were attempted after removing the most potent 10% of compounds from each activity class as a hold-out set, training models on the remaining (less potent) compounds, and predicting the potency of the hold-out sets. Figure 4 and Extended Data Fig. 4 show the results evaluated on the basis of MAE and RMSE, respectively. Naturally, under extrapolative conditions, the simple MR would be expected to yield the largest errors, as observed. In addition, for two of ten activity classes (target IDs 230 and 4822), DNN achieved best performance by a margin of ~0.5 MAE. In these two cases, the most potent training compounds were overpredicted, resulting in more accurate predictions of structurally analogous highly potent test compounds. In the remaining cases, the performance of methods including kNN was again very similar, with MAE ranging from ~1 to 2, depending on the activity class.

Randomized models

Given that the baseline MR approach reached an accuracy of ~1.0 MAE, that is, within one order of magnitude (tenfold) of experimental potency values, we also investigated completely random prediction models for kNN and SVR that were obtained by random shuffling of potency values across training and test sets. The potency value of each compound was randomly assigned to another, thus generating random structure–potency relationships for training. These randomizing models were applied to predict a randomized test set. Figure 5 shows the results obtained for fully randomized kNN and SVR predictions compared with MR (which remained constant). Very similar results were obtained when models derived from randomized training sets were applied to original test sets. Strikingly, random models yielded ~0.8–1.0 MAE across all data sets. MAE values for kNN slightly increased relative to SVR and MR, which were very similar across all activity classes. Thus, best predictions obtained for complete data sets and random predictions were only separated by ~0.5 MAE.

Discussion

Prediction of compound potency and other molecular properties is one of the major applications of ML in cheminformatics, medicinal chemistry and drug design. In the artificial intelligence era, complex computational methods are often employed for this purpose. Thus, while SVR is a widely recognized standard for potency prediction in the field, predictions using various DNN/GCN architectures are increasingly reported. The initial evaluation of new computational approaches typically relies on compound activity classes and conventional benchmark settings.

The results reported herein point to two critical issues in compound potency predictions that are currently little considered. First, our analysis shows that there is little, if any benefit in using complex ML models for potency predictions compared with simple kNN calculations. Second, best-performing ML models and completely random predictions are only distinguished by a small MAE margin corresponding to a less than tenfold difference in potency relative to experimental observations. Both of these issues require further consideration.

kNN analysis has been successfully used previously in chemical similarity searching and compound classification. The underlying principle is commonplace in medicinal chemistry: many similar compounds (such as structural analogues) have similar potencies. Notable exceptions are activity cliffs (that is, structural analogues with large potency differences)[17]. However, since only ~5% of bioactive compounds participate in the formation of activity cliffs across different activity classes[17], their influence on potency prediction accuracy on the basis of statistical modelling is for the most part negligible. This is especially the case for large data sets that are dominated by compounds falling into the

https://doi.org/10.1038/s42256-022-00581-6

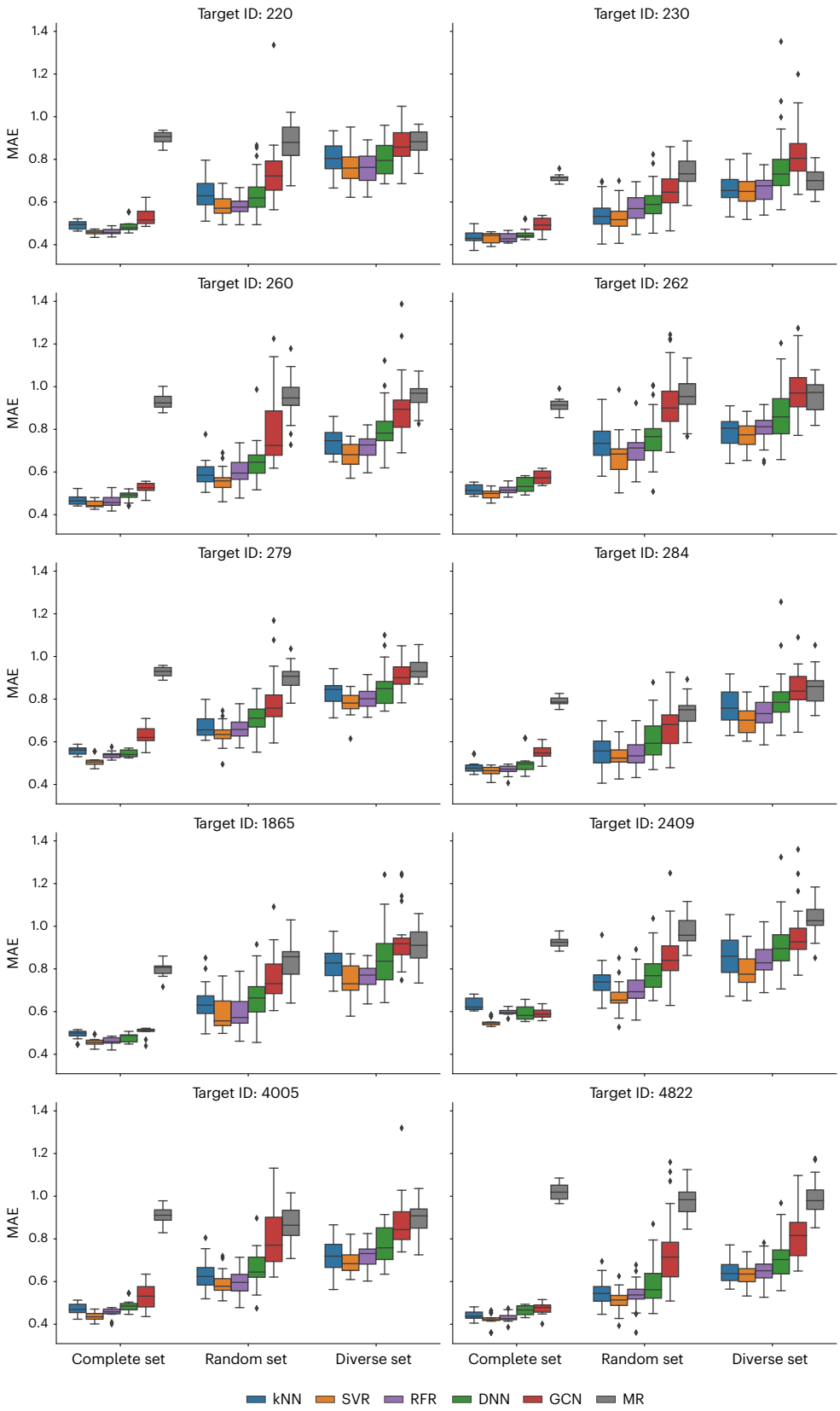


Fig. 2 | Prediction accuracy. Boxplots report the distribution of MAE values for ten independent potency prediction trials on different activity classes (identified by ChEMBL target IDs according to Table 1) using different models (kNN, SVR, RFR, DNN, GCN and MR). Results of predictions are reported for complete training sets (complete set) and size-reduced training sets (random and diverse sets, respectively). In a boxplot, a value distribution is represented by its maximum value (upper whisker, corresponding to 25%), upper quartile (upper boundary of the box), median value (horizontal line), lower quartile (lower boundary of the box), corresponding to the interquartile range (50%), and minimum value (lower whisker, 25%). Individual values classified as statistical outliers are shown as diamonds.

https://doi.org/10.1038/s42256-022-00581-6

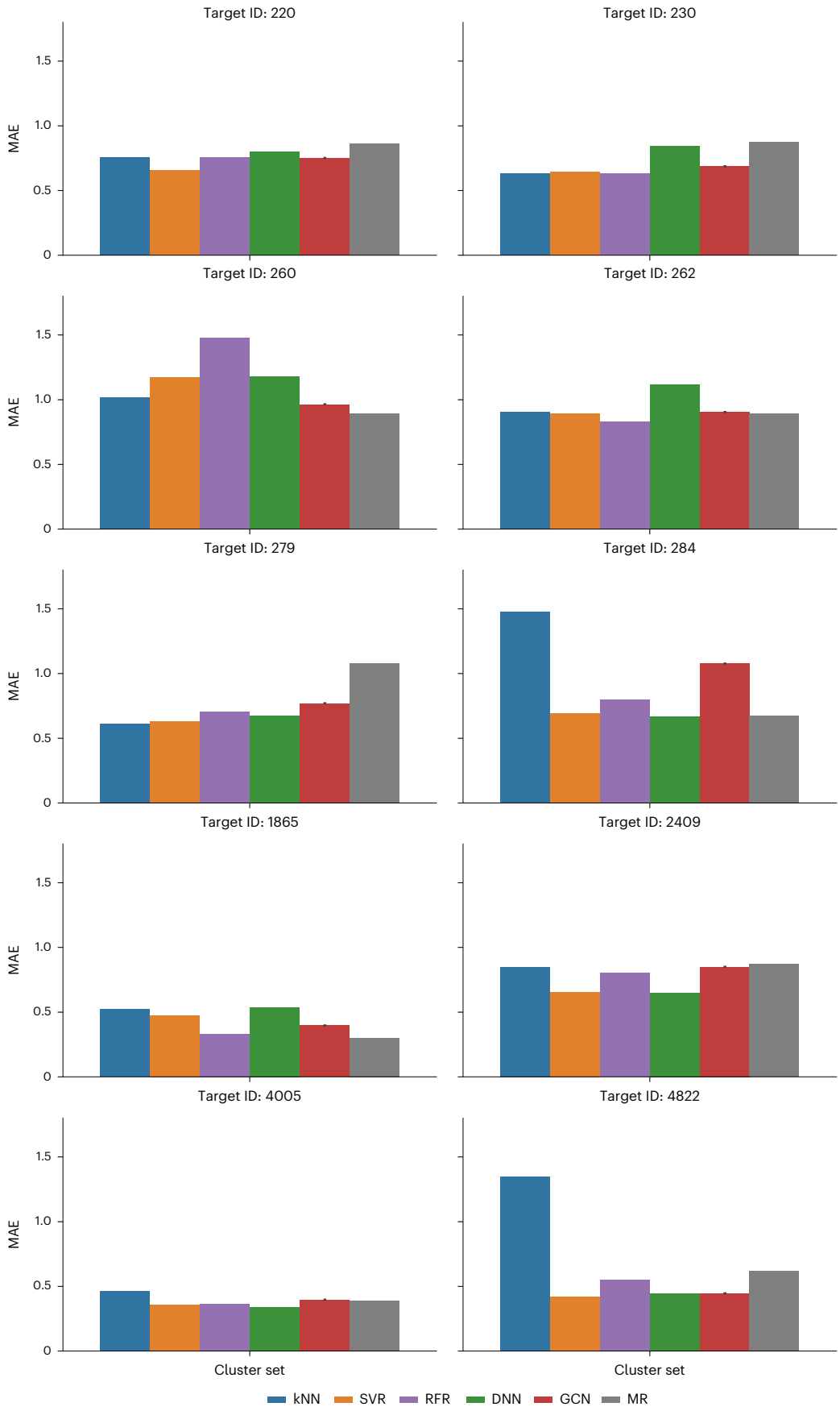


Fig. 3 | Prediction accuracy for unique hold-out sets. MAE values are reported for a prediction trial on a structurally unique hold-out set (cluster set) from each activity class using different models (kNN, SVR, RFR, DNN, GCN and MR) derived from complete training sets.

https://doi.org/10.1038/s42256-022-00581-6

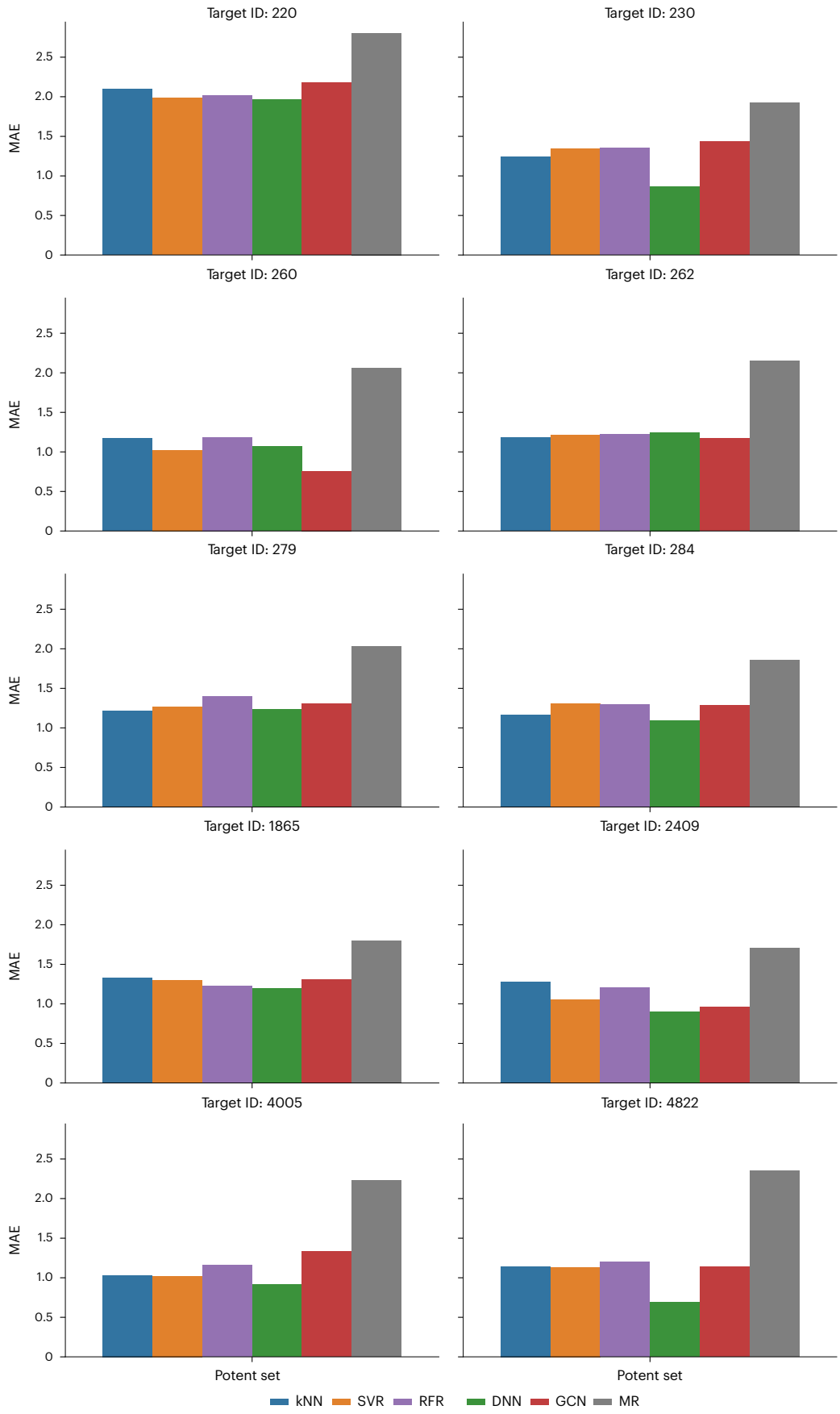


Fig. 4 | Prediction accuracy for most potent compounds. MAE values are reported for a prediction trial on the most potent compounds held out from each activity class (potent set) using different models (kNN, SVR, RFR, DNN, GCN and MR) derived from complete training sets.

https://doi.org/10.1038/s42256-022-00581-6

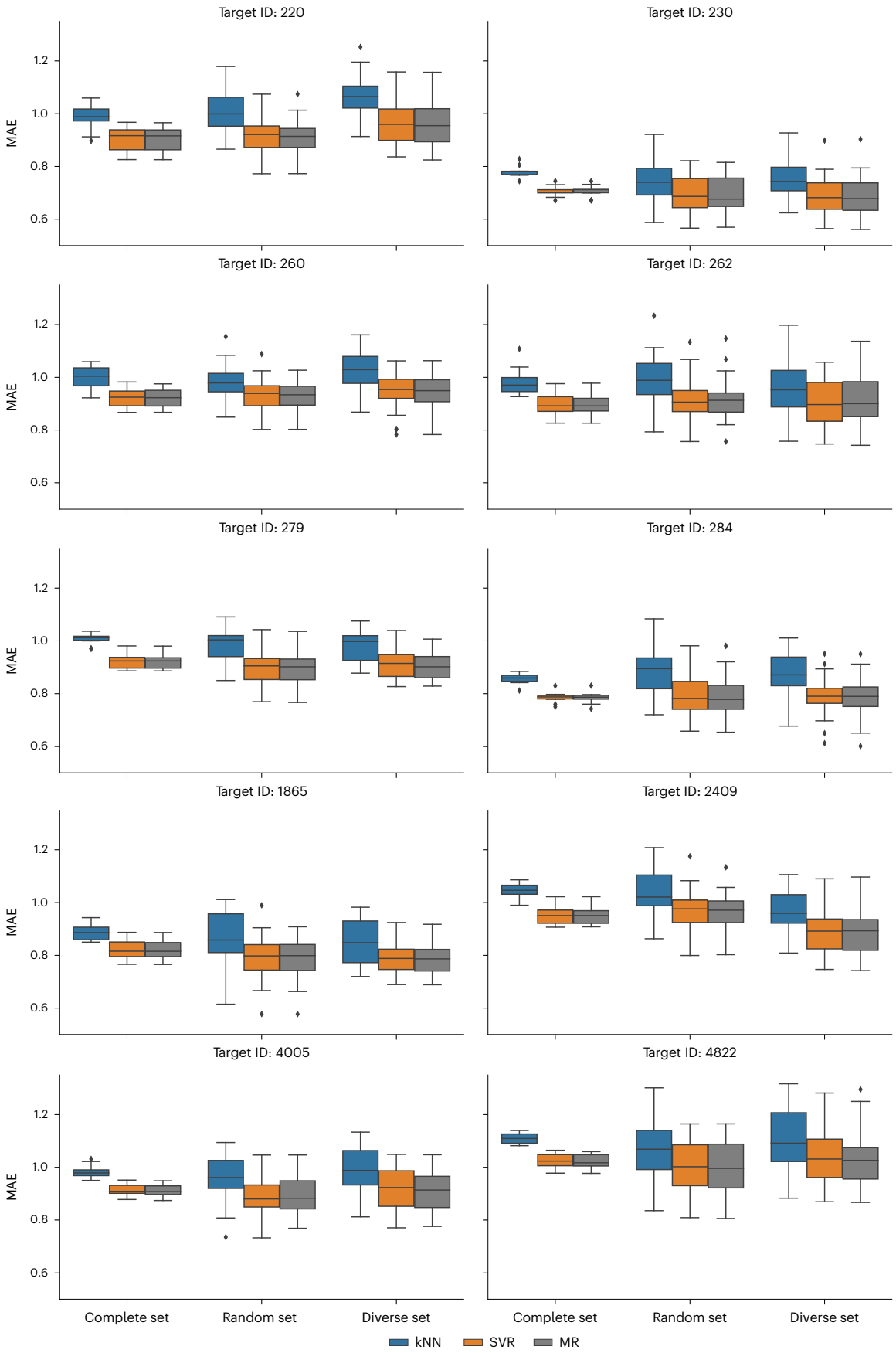


Fig. 5 | Performance of random prediction models. Boxplots report the distribution of MAE values for ten independent potency prediction trials on different activity classes using randomized kNN and SVR models in comparison with MR predictions according to Fig. 2. The boxplot elements are defined according to Fig. 2.

micro- to high-nanomolar range, which is typically the case for activity classes available for benchmarking. Clearly, simple approaches such as kNN calculations should generally be used as a reference for the evaluation of new computational approaches for compound potency/property predictions. The results presented herein indicate that it might be difficult to firmly establish advantages of ML methods over these simple predictions. For size-reduced and structurally diverse training sets, kNN rivalled SVR and performed better than DNN/GCN, despite the use of large numbers of compounds for learning[12]. Even for hold-out sets of structurally unique or most potent compounds overall similar performance was observed for potency prediction models of different complexities.

Furthermore, the small margin separating best and random predictions revealed a major shortcoming of conventional benchmarking. Simply assigning the median potency value of a training set to any test compound (MR) produced an MAE of 0.8–1.0. Notably, this error range was closely matched by completely random predictions. These findings are a direct consequence of the potency value distributions of activity classes from medicinal chemistry, as also shown herein. In practical applications, consistently predicting the potency of new compounds within one order of magnitude (tenfold) would be a considerable success. However, in benchmark settings, random predictions artificially yield this level of 'pseudoaccuracy'. Hence, under these conditions, it is very difficult, if not impossible, to assess the 'true' performance of computational methods.

Consequently, we might preferentially concentrate on prospective applications to predict and experimentally verify the potency of novel compounds. However, in the virtual screening literature, a frequent misconception is that a computational approach is 'validated' if one or more new active compounds are identified. However, this is not the case unless it is conclusively shown that simpler methods do not identify the same or similar compounds. In prospective potency prediction, this would also require the use of reference methods such as kNN.

There are other challenging prediction tasks that require special consideration. For example, late-stage lead optimization data from medicinal chemistry typically contain many very similar compounds with comparable (often relatively high) potencies and only a few 'outliers' representing activity cliffs. From a statistical point of view, the prevalence of such structure–potency relationships also favours simple kNN predictions, but they are not applicable to search for the most interesting analogues forming activity cliffs. This leaves much room for methods that are capable of quantitatively accounting for statistically underrepresented instances of high discontinuity in structure–potency relationships.

In conclusion, in light of the findings reported herein, it is evident that benchmark settings for computational potency predictions and the apparent performance of complex ML models require re-evaluation. An incremental step forward might be focusing methodological assessment on the prediction of hold-out sets containing structurally unique and highly potent compounds not considered during model derivation. This would at least alleviate some of the limitations caused by global potency value distributions and compound similarity relationships in benchmark data sets and address the most important practical goal of computational potency prediction. Furthermore, benchmark data sets might be designed to equally populate binned potency intervals with compounds from different series. Together with the use of kNN as a general reference method, more meaningful benchmark settings would help to avoid overestimation of potency prediction models and provide a more realistic assessment of their potential for practical applications.

Methods

Compounds and activity data

Activity classes were extracted from ChEMBL (version 30)[16]. Bioactive compounds for which direct interactions with a human target protein were reported at the highest level of confidence (target confidence score 9) and with a numerically specified potency (IC_{50}) value (standard relation '=') were selected (IC_{50} values were recorded as the negative decadic logarithm, pIC_{50}). Only compounds with a molecular weight of less than 1,000 Da and pIC_{50} values falling into the range of 5–11 were retained. Furthermore, compounds designated as 'potential transcription error', 'inconclusive' or 'not active' were omitted. Finally, potential assay interference compounds were removed using public filters[18–20]. On the basis of these criteria, ten activity classes containing a minimum of 1,000 qualifying compounds were randomly selected, yielding a total of 13,444 compounds.

Data set design

For each activity class, training sets of different compositions and size were randomly selected, including 'original' training sets comprising 80% of the compounds per class (see below), size-reduced sets with 20% of the compounds and equally sized sets (20%) with increased chemical diversity. These diverse sets were generated using the MaxMin dissimilarity algorithm[21]. Initially, pairwise compound similarity was systematically calculated using the Tanimoto coefficient[22] and a seed compound was randomly selected. Then, another compound was selected on the basis of the maximum Tanimoto distance to the seed, followed by the next compound with largest distance to a compound present in the evolving set. Distance-based compound selection was repeated until the set contained a predefined number of compounds corresponding to 20% of the activity class. For independent models and prediction trials, multiple sets were generated on the basis of different seed compounds.

Machine learning

For compound potency prediction, different ML models were generated.

Support vector regression. SVR is a variant of support vector machines that minimizes the error between predicted and observed values by deriving an epsilon-insensitive tube (ε-tube) based on the training instances[23,24]. During training, the samples are projected into a higher-dimensional feature space using kernel functions. The width of the ε-tube determines the error margin and penalizes samples falling outside the tube[24]. During model optimization, the ε-tube margin is adjusted by varying the cost parameter (C) that regulates the trade-off between the training errors and margin size.

The regularization hyperparameter C was optimized with the values of 0.001, 0.01, 0.1, 1, 10, 100 and 1,000. SVR models with the Tanimoto kernel[25] were generated using scikit-learn[26].

Random forest regression. RFR is a supervised ML method based on an ensemble of decision trees. During training, each tree is generated by node splitting using randomly selected training samples with bootstrapping[27]. The final predictions are derived as the mean value across all trees in the forest. The number of decision trees (25, 100, 200), the minimal number of samples for a leaf node (1, 2, 5) and the minimal number of samples for a split (2, 3, 5) were used as search parameters. RFR models were optimized and generated using scikit-learn[26].

Deep neural network. A feedforward DNN is a deep learning method that maps an input value to its output value by employing a nonlinear activation function $f(x)$. The basic DNN architecture consists of an input layer, multiple fully connected hidden layers with a variable number of neurons, and an output layer. Computational neurons are defined using the following equation: $y = f(\sum_i x_i w_i + b)$ (refs. [28,29]). During training, the neuron-associated weights (w_i) and biases (b) are iteratively updated to minimize the deviation between the predicted and observed output values. This optimization process involves calculation of the gradient of the cost function and backpropagation through the network until a minimal error is obtained.

For the DNN models, several architectures were evaluated by varying the number of hidden layers (2 or 3) and the number of neurons per layer (100–500) and testing different learning rates (0.1, 0.01, 0.001). The Adam[30] optimizer was employed for network training, the hyperbolic tangent (tanh) and rectified linear unit (ReLU) were tested as activation functions and the batch size was set to 32. Hyperparameters were optimized on the basis of internal validation with 80% and 20% training data split. Models were derived over 200 epochs and the early-stopping criterion was applied to minimize potential overfitting. All DNN models were implemented using TensorFlow[31] and Keras (https://keras.io/).

Graph convolutional neural network. GCN is a DNN variant that learns object representations directly from graphs[32,33], defined as $G = (V, E)$, with V and E being a set of vertices (nodes) and edges, respectively. In molecular graphs, nodes correspond to atoms and edges to bonds. Learned features comprise local neighbourhoods of atoms annotated with properties (for example, atom type, valence and aromaticity) that are obtained by graph message passing through convolutional layers and assembled by a graph-pooling layer. The representation format is a neural fingerprint that is generated by combining all node-level feature vectors. The terminal GCN output layer generates the predicted property value associated with a learned molecular graph representation. Similarly to the DNN optimization procedure, the training error is minimized through a cost function.

The channel width of graph convolutional layers was determined using value settings of (64, 64), (256, 256), (512, 512) and (1,024, 1,024). For the atom-level dense layer, the number of channels was optimized (64, 256, 512 or 1,024). Furthermore, different learning rates (0.01, 0.001) and dropout values (0, 0.25, 0.5) were investigated. As for DNNs, Adam was used as the optimization method. ReLU and tanh were used as activation functions for the convolutional and graph-pooling layers, respectively. The GNN architecture for potency prediction included two convolutional layers. GCN parameters were optimized using an 80:20 training data split. Training was carried out with batch normalization for a maximum of 200 epochs with early-stopping option. GCN models were built using DeepChem (https://deepchem.io/).

***k*-nearest neighbour.** kNN is a non-parametric regression method that predicts test instances based on the shortest distance (highest similarity) to training samples[34]. For example, for 1-NN, the potency value of the most similar training compounds is assigned to a test compound; for 3-NN, the average potency of the top-three most similar compounds is predicted for the test compound.

For kNN predictions, the one, three and five most similar compounds were evaluated to identify the best-performing k value. kNN calculations were carried out using scikit-learn.

Median regression. As a control, random predictions were carried out by assigning the median potency value of a given data set to each test compound from this set. Hence, all compounds were predicted to have the same (median) potency.

Random predictions. As another control, randomized predictions were carried out by random shuffling of potency values across compounds of each training set, which is often referred to as y randomization[35].

Calculation protocol. For all supervised ML models, a uniform calculation protocol was used. For original randomly selected training sets, the remaining 20% of the compounds per class were used as test sets. For size-reduced training sets, 5% of the compounds per class were randomly selected as test sets (ensuring constant training-to-test set ratios). Hyperparameters were optimized using tenfold internal cross-validation via grid search to minimize the model error. For each data set and method, ten individual prediction trials with independently derived models were carried out.

Molecular representations, similarity calculations and analogue series

For kNN, SVR and DNN, compounds were represented using the standard extended connectivity fingerprint with diameter 4 (ECFP4)[36]. The folded 2,048-bit version was generated using RDKit (http://www.rdkit.org/). For kNN, Tanimoto similarity[22] was calculated on the basis of ECFP4. For GCN models, compounds were transformed into a binary vector comprising 75 atom features using DeepChem. For each activity class, analogue series were systematically identified using the compound–core relationship algorithm[37].

Performance measures

To assess model performance, MAE and RMSE were calculated to compare predicted and observed potency values:

$$\mathrm{MAE}\,(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

$$\mathrm{RMSE}\,(y,\hat{y}) = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n}}. \quad (2)$$

Here y is the experimental and $\hat{y}$ the predicted potency value.

For the complete compound sets, statistical significance assessment of value distributions from predictions was based on MAE values using the non-parametric Wilcoxon test[38]. The alpha threshold was set to 0.05. The P values were compared with alpha ($P \leq 0.05$) and the null hypothesis was rejected/accepted.

Data availability

Publicly available compounds and activity data including compound activity classes and sets of analogue series extracted from these classes were obtained from ChEMBL using the data selection and calculation protocols provided in Compounds and activity data and Molecular representations, similarity calculations and analogue series. In addition, all data sets used for the calculations reported herein are freely via the following link: https://github.com/TiagoJanela/ML-for-compound-potency-prediction. Source data are provided with this paper.

Code availability

All calculations were carried out using public domain programs and computational tools.

Additional code used for our calculations is freely available via the following link: https://github.com/TiagoJanela/ML-for-compound-potency-prediction. The code is also available at https://doi.org/10.5281/zenodo.7238586 (ref. [39]).

References

1. Gleeson, M. P. & Gleeson, D. QM/MM calculations in drug discovery: a useful method for studying binding phenomena? *J. Chem. Inf. Model.* **49**, 670–677 (2009).
2. Mobley, D. L. & Gilson, M. K. Predicting binding free energies: frontiers and benchmarks. *Annu. Rev. Biophys.* **46**, 531–558 (2017).
3. Li, H., Sze, K. H., Lu, G. & Ballester, P. J. Machine-learning scoring functions for structure-based virtual screening. *WIREs Comput. Mol. Sci.* **11**, e1478 (2021).
4. Lewis, R. A. & Wood, D. Modern 2D QSAR for drug discovery. *WIREs Comput. Mol. Sci.* **4**, 505–522 (2014).
5. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
6. Lavecchia, A. Deep learning in drug discovery: opportunities, challenges and future prospects. *Drug Discov. Today* **24**, 2017–2032 (2019).

7. Walters, W. P. & Barzilay, R. Applications of deep learning in molecule generation and molecular property prediction. *Acc. Chem. Res.* **54**, 263–270 (2020).
8. Torng, W. & Altman, R. B. Graph convolutional neural networks for predicting drug–target interactions. *J. Chem. Inf. Model.* **59**, 4131–4149 (2019).
9. Son, J. & Kim, D. Development of a graph convolutional neural network model for efficient prediction of protein–ligand binding affinities. *PLoS ONE* **16**, e0249404 (2021).
10. Li, Y. et al. An adaptive graph learning method for automated molecular interactions and properties predictions. *Nat. Mach. Intell.* **4**, 645–651 (2022).
11. Fang, X. et al. Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.* **4**, 127–134 (2022).
12. Sakai, M. et al. Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. *Sci. Rep.* **11**, 525 (2021).
13. Chen, L. et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS ONE* **14**, e0220113 (2019).
14. Yang, J., Shen, C. & Huang, N. Predicting or pretending: artificial intelligence for protein–ligand interactions lack of sufficiently large and unbiased datasets. *Front. Pharmacol.* **11**, e69 (2020).
15. Volkov, M. et al. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *J. Med. Chem.* **65**, 7946–7958 (2022).
16. Bento, A. P. et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42**, D1083–D1090 (2002).
17. Stumpfe, D., Hu, Y., Dimova, D. & Bajorath, J. Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *J. Med. Chem.* **57**, 18–28 (2014).
18. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**, 2719–2740 (2010).
19. Bruns, R. F. & Watson, I. A. Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.* **55**, 9763–9772 (2012).
20. Irwin, J. J. et al. An aggregation advisor for ligand discovery. *J. Med. Chem.* **58**, 7076–7087 (2015).
21. Ashton, M. et al. Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quant. Struct. Relatsh.* **21**, 598–604 (2002).
22. Willett, P., Barnard, J. M. & Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998).
23. Drucker, H., Surges, C. J. C., Kaufman, L., Smola, A. & Vapnik, V. Support vector regression machines. In *Proc. Ninth International Conference on Neural Information Processing Systems* (eds Jordan, M. I. & Petsche, T.) 155–161 (MIT Press, 1997).
24. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
25. Ralaivola, L., Swamidass, S. J., Saigo, H. & Baldi, P. Graph kernels for chemical informatics. *Neural Netw.* **18**, 1093–1110 (2005).
26. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
27. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
28. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
29. Nielsen, M. A. *Neural Networks and Deep Learning* (Determination, 2015).
30. Kingma, D. P. & Ba, J. L. Adam: a method for stochastic optimization. In *Third International Conference on Learning Representations (ICLR) 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (eds Bengio, Y. & LeCun, Y.) (2015).
31. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. In *OSDI'16: Proc. 12th USENIX Conf. Operating Systems Design and Implementation* (chairs Keeton, K. & Roscoe, T.) 265–283 (USENIX Association, 2016).
32. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw. Learn. Syst.* **20**, 61–80 (2009).
33. Duvenaud, D. K. et al. Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process. Syst.* **28**, 2224–2232.
34. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**, 175–185 (1992).
35. Rücker, C., Rücker, G. & Meringer, M. *y*-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.* **47**, 2345–2357 (2007).
36. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
37. Naveja, J. J. et al. Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. *ACS Omega* **4**, 1027–1032 (2019).
38. Conover, W. J. On methods of handling ties in the Wilcoxon signed-rank test. *J. Am. Stat. Assoc.* **68**, 985–988 (1973).
39. Janela, T. ML-for-compound-potency-prediction. *Zenodo* https://doi.org/10.5281/zenodo.7238586 (2022).

Acknowledgements

We thank C. Feldmann, A. Lamens, F. Siemers and M. Vogt for helpful discussions.

Author contributions

Conceptualization, J.B.; methodology, T.J. and J.B.; data and code, T.J.; investigation, T.J.; analysis, T.J. and J.B.; writing—original draft, J.B.; writing—review and editing, T.J. and J.B.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-022-00581-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00581-6.

Correspondence and requests for materials should be addressed to Jürgen Bajorath.

Peer review information *Nature Machine Intelligence* thanks Alexander Tropsha and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Appendix B

Large-Scale Predictions of Compound Potency with Original and Modified Activity Classes Reveal General Prediction Characteristics and Intrinsic Limitations of Conventional Benchmarking Calculations.

pharmaceuticals

Article

Large-Scale Predictions of Compound Potency with Original and Modified Activity Classes Reveal General Prediction Characteristics and Intrinsic Limitations of Conventional Benchmarking Calculations

Tiago Janela and Jürgen Bajorath *

Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 6, D-53115 Bonn, Germany; janela@bit.uni-bonn.de
* Correspondence: bajorath@bit.uni-bonn.de; Tel.: +49-228-7369-100

Abstract: Predicting compound potency is a major task in computational medicinal chemistry, for which machine learning is often applied. This study systematically predicted compound potency values for 367 target-based compound activity classes from medicinal chemistry using a preferred machine learning approach and simple control methods. The predictions produced unexpectedly similar results for different classes and comparably high accuracy for machine learning and simple control models. Based on these findings, the influence of different data set modifications on relative prediction accuracies was explored, including potency range balancing, removal of nearest neighbors, and analog series-based compound partitioning. The predictions were surprisingly resistant to these modifications, leading to only small error margin increases. These findings also show that conventional benchmark settings are unsuitable for directly comparing potency prediction methods.

Keywords: compound potency predictions; activity classes; machine learning; nearest neighbor controls; benchmark calculations

Citation: Janela, T.; Bajorath, J. Large-Scale Predictions of Compound Potency with Original and Modified Activity Classes Reveal General Prediction Characteristics and Intrinsic Limitations of Conventional Benchmarking Calculations. *Pharmaceuticals* **2023**, *16*, 530. https://doi.org/10.3390/ph16040530

Academic Editor: Krzysztof Marciniec

Received: 17 March 2023
Revised: 27 March 2023
Accepted: 31 March 2023
Published: 2 April 2023

1. Introduction

Compound potency prediction is of major interest in medicinal chemistry and drug design. Many different computational methods have been introduced for potency predictions based on structures of ligand-target complexes or small molecules [1–11]. These approaches have different computational complexity and sophistication. Traditionally, quantitative structure–activity relationship (QSAR) methods have played a major role in medicinal chemistry [1]. Classical QSAR models are based on two-dimensional representations of small molecules, typically employ numerical descriptors of molecular structure and chemical properties, and represent linear regression models to predict the potency of newly designed compounds to extend analog series. Thus, QSAR only applies to congeneric compounds if linear structure–activity relationships (SARs) exist [1]. In addition, for the estimation of interaction energies from experimental or modeled protein–ligand complexes, a variety of scoring functions were developed that are, for the most part, based on force fields from molecular mechanics [2]. Estimating interaction energies using scoring functions of different designs and complexity is used as a rough approximation of binding (free) energies and relative potencies of other ligands (without calculating exact potency values). Scoring functions apply to diverse compounds and are critically important to prioritize putative ligands from structure-based virtual screening, despite their approximate nature [2]. At a higher level of sophistication, free energy methods attempt to calculate exact binding free energy values from protein–ligand complexes based on thermodynamic principles [3]. Particularly popular in medicinal chemistry and drug design are free energy perturbation methods to calculate relative binding free energies of congeneric compounds

based on molecular dynamics simulations by "alchemically" transforming one analog into another. Compared to force field calculations, relative free energy calculations are computationally very expensive. Although free energy methods have been available for three or four decades, they have been increasingly applied in recent years in drug discovery due to advances in computational power and conformational sampling procedures [3].

Furthermore, in structure-based design, binding energy, and compound potency values can also be calculated using methods that combine molecular mechanics (MM) treatment of protein–ligand complexes with quantum mechanical (QM) representations of narrowly defined ligand binding sites (termed QM/MM approaches) [4]. The underlying idea is achieving accurate energy calculations in binding sites through quantum mechanics while reducing computational costs for the remainder of complexes to render the calculations feasible. For ligand-based potency prediction, machine learning (ML) methods play a major role [5–7]. Therefore, suitable ML methods must be applicable for regression. Compared to QSAR, the major attraction of computationally more complex ML approaches is their ability to account for non-linear SARs and predict potency values of structurally diverse compounds. Non-linear SARs are typically observed in medicinal chemistry when optimizing compound series, which intrinsically limits the applicability domain of classical QSAR. Accordingly, ML regression models have become very popular for compound potency prediction. The majority of approaches include ML mainstay methods such as random forest regression [6] or support vector regression (SVR) [7,8]. Over the years, SVR has become the probably most frequently used ML approach for numerical potency prediction and a standard in the field. Recently, deep neural networks (DNNs) have also been increasingly applied for this task [9–11]. Many different DNN architectures can be adapted for numerical property predictions, including compound potency. This methodological versatility is a major attraction of DNNs [9–11]. Moreover, DNNs enable the evaluation of new concepts for potency prediction. For example, convolutional neural networks can predict numerical properties from voxel representations of ligand binding sites. For chemical applications, graph neural networks have become increasingly popular, and they have also been adapted for ligand affinity predictions. Therefore, graph representations of molecular interactions are extracted from structures of protein–ligand complexes and used as input for deep graph neural networks to predict the affinity of small molecular ligands. Exploring novel concepts for potency predictions is still in its early stages (and some findings are controversial). Hence, it will take time until these approaches mature. While DNN calculations are computationally much more expensive compared to other ML approaches, they are not necessarily superior for potency prediction [12], as further discussed below.

The prediction of compound potency (and other biological or physico-chemical molecular properties) is carried out to benchmark or calibrate computational approaches and, in addition, prospectively predict novel active compounds. While prospective applications are naturally most interesting in medicinal chemistry and drug discovery, benchmarking is essential for the initial evaluation of predictive models but insufficient to ensure successful applications. Typical benchmark conditions for numerical potency prediction involve using sets of specific active compounds (often termed activity classes) with varying potency divided into training sets for model derivation and test sets for evaluation, usually with cross-validation on the basis of multiple independent prediction trials. Analogous benchmark settings are applied to assess compound classification models (derived, for example, to distinguish between active and inactive compounds).

Recently, we have shown that potency prediction methods of varying computational complexity display similar predictive performance [12]. Specifically, k-nearest neighbor (kNN) analysis was found to reproduce experimental potency values of test compounds within an order of magnitude comparable to increasingly complex ML methods, including DNNs. In 1-NN analysis, test compounds are compared to training set compounds via similarity calculations, and the potency value of the most similar training compound is assigned to a given test compound. For 10 different activity classes, there was no advantage

of DNNs over SVR and kNN calculations, with SVR achieving the overall best performance, albeit by only small margins [12]. Hence, simple predictions were often as accurate as increasingly complex ML methods. Furthermore, assigning the median potency value of a training set to any test compound, corresponding to median regression (MR), often approached the accuracy of ML models. Moreover, randomized prediction models often reproduced experimental potency values within an order of magnitude, and there was only a confined prediction error interval into which random and ML predictions fell.

Questions raised by these observations included whether these findings might generalize across large numbers of activity classes and whether their composition and/or potency ranges might limit benchmarking evaluations. Therefore, in this study, we have systematically investigated compound potency predictions on an unprecedentedly large scale and designed specific data set modifications to investigate their influence on the prediction accuracy of different reference methods. Potency predictions were surprisingly stable across hundreds of compound classes, and relative method performance was largely resistant to specific data set modifications. Furthermore, predictions using ML and simple control models were only distinguished by small error margins, revealing intrinsic limitations of conventional benchmark calculations.

2. Results

2.1. Study Concept

First, we aimed to obtain a global view of potency prediction characteristics and relative accuracies of selected methods. Therefore, we carried out systematic potency value predictions on 376 qualifying activity classes from medicinal chemistry sources [13] using SVR and controls, including 1-NN, additional kNN, and MR calculations (see Methods). The activity classes were curated, ensuring high-confidence potency data were available for all compounds. SVR was selected as the overall preferred ML approach in our previous comparison [12]. Second, based on the obtained results, we then investigated the influence of specific data set modifications on relative prediction accuracies.

2.2. Large-Scale Predictions

Predictions were assessed by calculating the mean absolute error (MAE) for predicted and experimental logarithmic potency values (see Section 4). Given the very large number of activity classes and calculations, all results are made available in a data deposition via the following link: https://uni-bonn.sciebo.de/s/vU5vnG5wjQPTpd1 (accessed on 28 March 2023)). In addition, for representative subsets of activity classes, results are reported in the following and as Supplementary Materials.

For the 376 activity classes, the results of the predictions were surprisingly similar. While MAE values varied moderately across different classes, it was generally observed that 1-NN/kNN predictions approached or met SVR performance, consistent with our earlier observations for 10 activity classes [12]. In addition, most predictions produced meaningful results, with median MAE values over multiple independent trials falling within one order of magnitude, corresponding to less than 10-fold prediction error. Notably, for best predictions, MAE values of 1 or larger were not observed for any activity class. Supplementary Figure S1 shows the results for the 45 largest activity classes that were representative of all 376 activity classes. Hence, only limited class-dependent differences were detected.

Supplementary Figure S1a,b compare predictions for the 45 activity classes based upon 80/20% and 50/50% training/test compound splits, respectively. Again, these predictions yielded very similar results. Hence, different training set sizes had little influence on the predictions. Thus, the predictions were stable, as indicated by narrow MAE value distributions across different trials.

Figure 1 shows exemplary compounds from eight of the 45 activity classes (and reports their targets) used in the following to illustrate results obtained for the 45 largest classes. In addition, Figure 2 shows the results of the original predictions for the eight activity

classes and 80/20% splits, illustrating trends commonly observed for all classes. Although SVR mostly achieved the highest accuracy (lowest MAE values), followed by kNN/1-NN, the differences between median values were typically only very small, ~0.1 MAE or even less. Statistically significant differences were only observed for about half of the classes (Wilcoxon test, p-value < 0.005; see Section 4). Even the simplistic MR prediction, assigning the constant median potency value of the training set to all test compounds, typically yielded prediction accuracies close to 1.0 MAE. Thus, these findings revealed that (i) even simple control predictions generally produced fairly accurate results and that (ii) there was no sufficient separation between SVR and kNN or MR controls to enable a realistic assessment of ML potency prediction methods. Across as many as 376 different activity classes, essentially no cases were detected where prediction accuracy was low and simple controls failed compared to SVR.

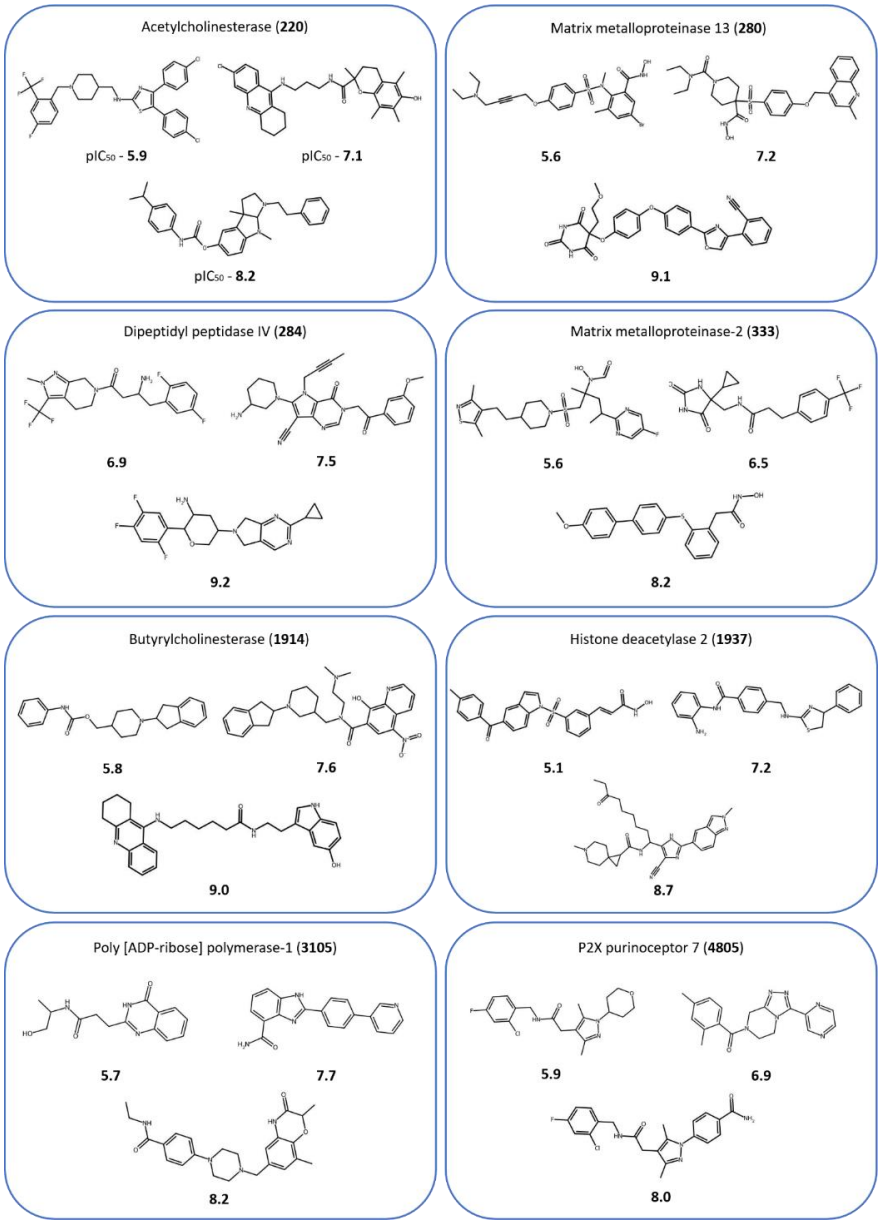


Figure 1. Compounds from selected activity classes. For eight activity classes, exemplary compounds are shown with their logarithmic potency values (pIC_{50}). For each class, the target name and ChEMBL ID (in parentheses) are provided.

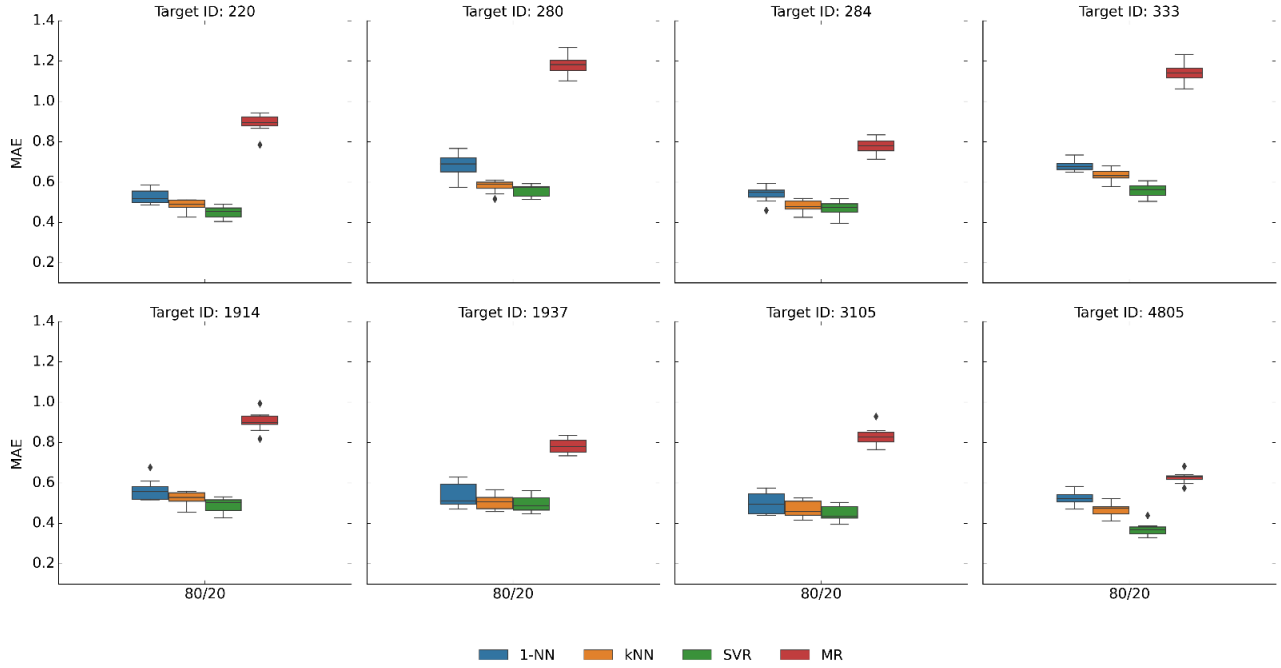


Figure 2. Prediction accuracy. Boxplots report the distribution of MAE values for potency predictions over 10 independent trials on the eight activity classes in Figure 1 using 1NN, kNN, SVR, and MR models (applying a training/test set compound split of 80:20%). In boxplots, the upper and lower whiskers indicate maximum and minimum values, the boundaries of the box represent the upper and lower quartiles, values classified as statistical outliers are shown as diamonds, and the median value is indicated by a horizontal line.

These findings raised the question of whether the activity classes could be modified in specific ways to increase the prediction accuracy separation of SVR and the kNN controls and hence obtain an improved basis for methodological comparisons. These modifications altered the original composition of activity classes by design, thus producing model data sets. The predictions were then repeated on the resulting variants of the 45 largest activity classes. The following shows representative results for the subset of eight activity classes.

2.3. Potency Range Balancing

We first determined the potency value distributions across the largest activity classes. As shown in Supplementary Figure S2a, potency distributions in activity classes from medicinal chemistry are not uniform but skewed because most compounds are generally active in the low micromolar range. Therefore, we reasoned that the dominance of compounds with micromolar potency values might explain the strong performance of kNN and MR relative to SVM. Consequently, we generated activity class variants with balanced potency distributions (see Section 4), as shown in Supplementary Figure S2b. In the modified data sets, most potency sub-ranges were evenly populated (except sub-ranges containing limited numbers of most potent compounds). Thus, balancing eliminated the bias of potency value distributions towards the low micromolar range. We then repeated the predictions on the balanced activity class variants. Since balancing inevitably led to a reduction in data set size, we also generated equally sized data sets with original potency distribution as a control (50/50% training/test compound splits). Figure 3 reports the results for the predictions on balanced data sets that were similar to those of the original predictions. As a consequence of potency balancing, the median potency values of the training set increased, which also increased the MAE of MR in several cases. However, the performance of kNN/1-NN compared to SVR essentially remained constant.

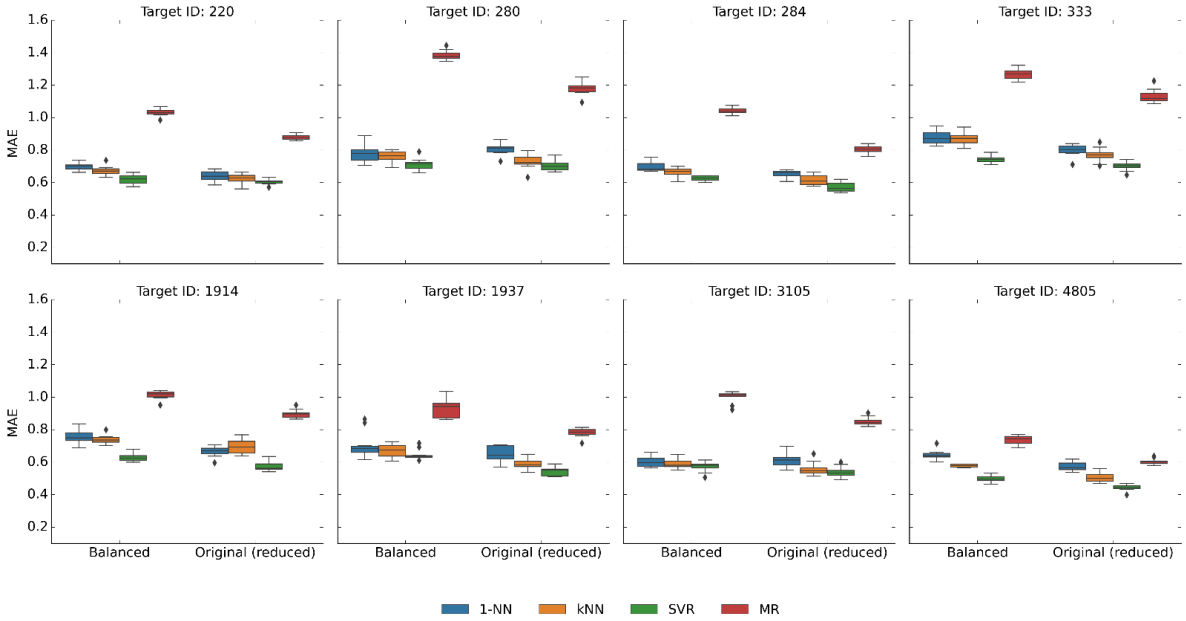


Figure 3. Prediction accuracy for activity classes with balanced potency value distributions. Boxplots report the distribution of MAE values over 10 independent trials for the eight activity classes after balancing their potency value distributions. As a control, results are reported for the original data sets that were reduced by random compound removal to the same size as the balanced sets. In boxplots, the upper and lower whiskers indicate maximum and minimum values, the boundaries of the box represent the upper and lower quartiles, values classified as statistical outliers are shown as diamonds, and the median value is indicated by a horizontal line.

2.4. Removal of Nearest Neighbors

In light of these findings, we systematically removed nearest neighbors from the original activity classes. Therefore, exhaustive pairwise compound similarity calculations were carried out for each class; compounds were ranked according to highest similarity to nearest neighbors, and the top 50% of compounds from the ranking were removed from the data sets. As a size control, data sets containing half of the original compounds were randomly selected. Figure 4 shows the results of predictions after nearest neighbor removal and equally sized control data sets (all 45 activity classes produced equivalent results). Nearest neighbor removal generally increased median MAE values for all methods by ~0.1–0.2 and slightly broadened value distributions (such that the predictions became again more similar to MR). However, even the removal of 50% of most similar compounds was insufficient to significantly reduce the performance of kNN/1-NN relative to SVR, an unexpected finding.

2.5. Analog Series-Based Data Partitioning

Another structural data set modification was carried out by extracting all analog series from each activity class, then partitioning the complete series into training and test sets (to obtain ~80/20% compound splits). Accordingly, there was no analog overlap between the sets. Accordingly, training and test compounds had distinct core structures. Because most compounds from medicinal chemistry belong to analog series (resulting from chemical optimization efforts), analog series-based partitioning was generally applicable to activity classes. Figure 5 shows the results of predictions for these activity class variants and equally sized subsets of the original data sets used as a control (equivalent results were again obtained for all 45 activity classes). Under these conditions, median MAE values also increased by ~0.1–0.2 relative to the controls. The value distributions generally broadened (as one might expect for independent trials using training and test sets of unique analog series composition). Broader distributions are indicative of more variable (less stable)

predictions, which complicates the comparison of different methods. However, despite analog series partitioning, the predictive performance of SVR and kNN/1-NN remained very similar.

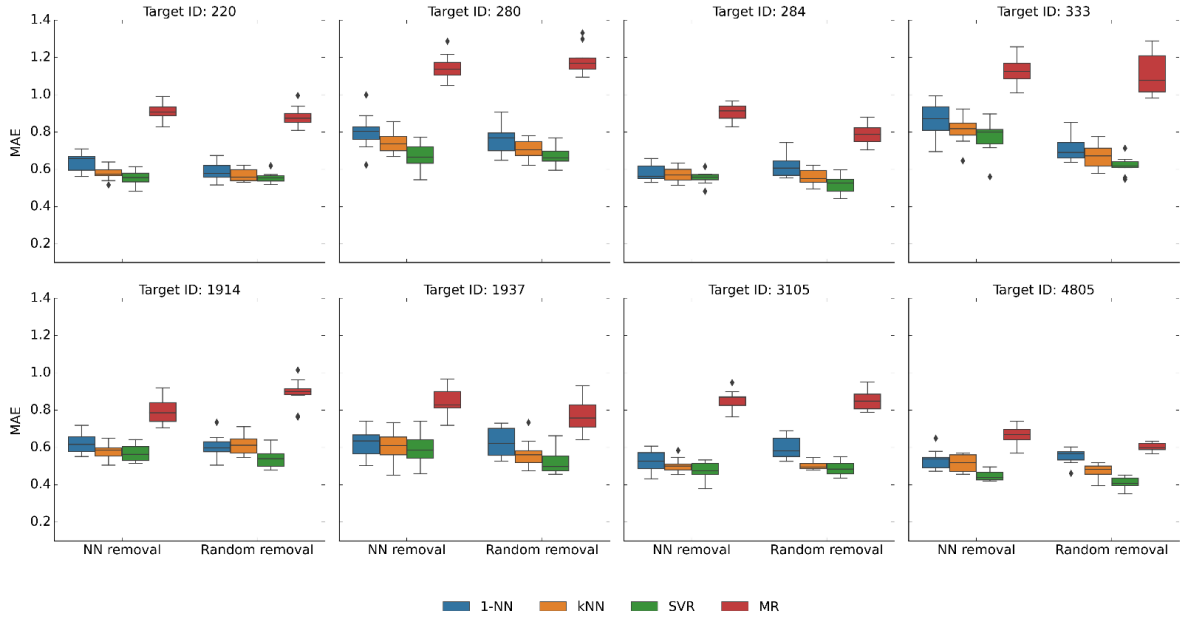


Figure 4. Prediction accuracy after removal of nearest neighbor relationships. Boxplots report the distribution of MAE values over 10 independent trials for the eight activity classes after removal of 50% of nearest neighbors and control data sets after random removal of 50% of the compounds. In boxplots, the upper and lower whiskers indicate maximum and minimum values, the boundaries of the box represent the upper and lower quartiles, values classified as statistical outliers are shown as diamonds, and the median value is indicated by a horizontal line.

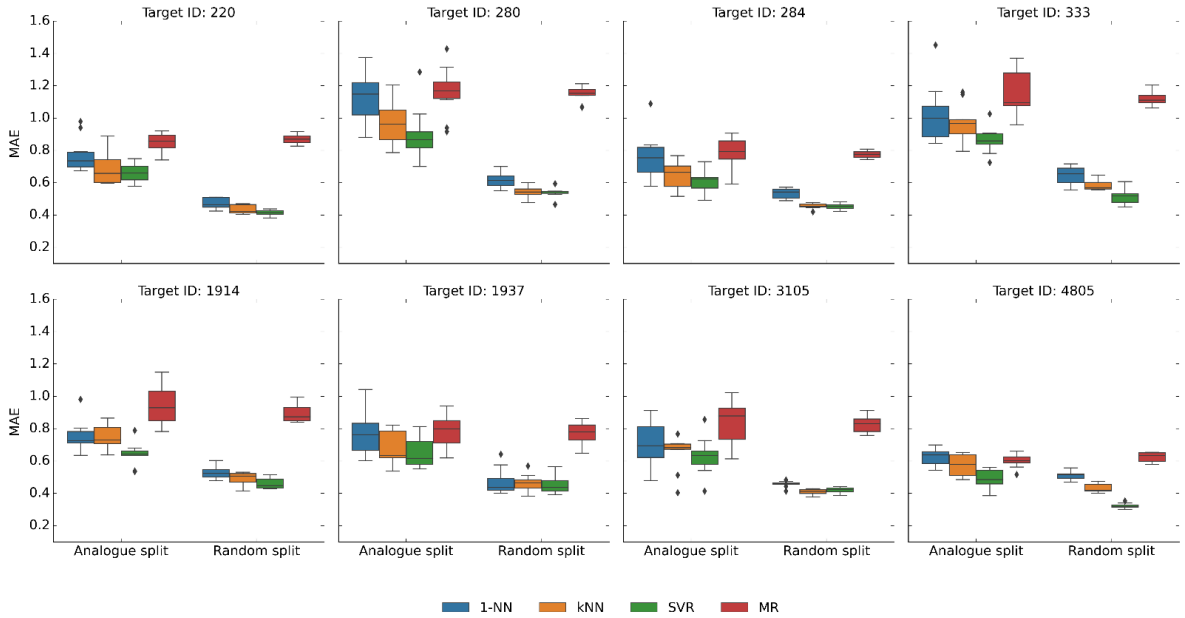


Figure 5. Prediction accuracy after analog series partitioning. Boxplots report the distribution of MAE values over 10 independent trials for the eight activity classes using training and test sets (~80:20% compound split) consisting of distinct analog series. As a control, results are reported for original training and test sets of exactly the same size. In boxplots, the upper and lower whiskers indicate maximum and minimum values, the boundaries of the box represent the upper and lower quartiles, values classified as statistical outliers are shown as diamonds, and the median value is indicated by a horizontal line.

3. Discussion

Our current study was designed in light of previous observations that simple 1-NN calculations often approached or met the accuracy of increasingly complex ML methods in compound potency predictions. To better understand these prediction characteristics and explore consequences for benchmark comparisons of different methods, we have carried out systematic potency value predictions on 376 activity classes with sufficient numbers of compounds using a preferred ML approach and simple controls, including kNN and MR calculations. Activity classes were curated to ensure that high-confidence activity measurements were available for all compounds, thus avoiding potential bias of predictions due to limited data quality. Our calculations most likely represent one of the largest (if not the largest) compound potency prediction campaigns reported to date. The results of the global predictions were surprisingly similar across a large number of activity classes from three points of view. First, there were only little activity class-specific differences in prediction patterns and accuracy; second, most predictions had limited error margins falling well within an order of magnitude; third, in accordance with our earlier observations, kNN calculations consistently rivaled SVR performance, and there was only a small error range separating prediction accuracy including MR, the most control. Thus, global potency predictions were surprisingly stable and accurate for methods of different complexity. These findings implied that calculations on activity classes from medicinal chemistry might generally produce predictions that are too similar for a realistic assessment and comparison of different potency prediction methods. Accordingly, the results also call the relevance of conventional benchmark settings into question. Benchmark calculations are essential for assessing basic method performance but must also reliably quantify relative differences in the accuracy of alternative approaches. Therefore, we then explored (i) possible reasons for the success of simple potency prediction approaches and (ii) ways in which activity classes and calculation conditions might be modified to increase the difficulty and sensitivity of benchmarking using model data sets. Specifically, we balanced potency distributions in activity classes, removed large numbers of nearest neighbors from them, and trained and tested models on structurally distinct compound sets obtained by analog series partitioning. Predictions on model data sets were again unexpectedly robust. Notably, while minor increases in prediction errors were observed for modifications rendering the predictions more challenging, none of these operations led to a significant difference in relative performance between SVR and kNN. The observed stability and robustness of the predictions on original and modified activity classes can be positively viewed because promising predictions are obviously possible with rather different approaches and using data set variants of varying composition. However, for conventional benchmarking, the implications are profound. Based on the findings reported herein, benchmark calculations on activity classes from medicinal chemistry, even if specifically modified to increase prediction challenges, do not enable sound comparisons of different methods because alternative predictions, including simple controls, are only differentiated by small error margins. A potential reason for this might include the prevalence of structurally related compounds with similar potency in activity classes (originating from chemical optimization efforts) or the under-representation of highly potent compounds in data sets (representing the most attractive prediction targets). As shown herein, however, predictions were resistant to substantial structural modifications of activity classes. Thus, from this point of view, our study should raise awareness of these issues and trigger attempts to develop fundamentally different concepts for evaluating and comparing potency prediction methods, providing opportunities for future investigations.

4. Materials and Methods

4.1. Compound Activity Data

From ChEMBL release 30 [13], bioactive compounds of less than 1000 Da with standard potency measurements (IC_{50}) and a numerical specified potency value (standard relation '=') were retrieved. Potency values were recorded as the negative decadic logarithm. Only

compounds with direct interactions (target relationship type: "D") against human proteins at the highest confidence level (target confidence score: 9) and pIC_{50} values ranging from 5 to 11 were considered. Additionally, measurements labeled "potential transcription error" and "potential author error" were removed. In addition, potential assay interference compounds were removed using public filters and tools [14–16].

Based on these selection criteria, 91,733 compounds belonging to 376 activity classes containing at least 50 compounds were obtained for the analysis. The largest 45 activity classes consisted of at least 500 compounds each (yielding 40,440).

4.2. Compound Sets with Balanced Potency Distribution

The 45 largest activity classes were balanced to obtain an even potency value distribution across the entire potency range, yielding reduced data sets of 50% of the original size. These balanced data sets were generated by dividing the potency range of each class into a maximum of six equally sized bins (for logarithmic potency values of 5–6, 6–7, 7–8, 8–9, 9–10, and 10–11). The average number of compounds per bin was calculated by dividing the number of available compounds by the number of bins. The bins were subsequently populated with compounds until the number was equal to the calculated average. For bins representing highest potency values, the number of available compounds was often insufficient to satisfy this criterion. In this case, other potency bins for which compounds were still available were uniformly populated until the final size of the balanced set was equal to 50% of the original compound set.

4.3. Model Building and Implementation

For model building and evaluation, training and test sets were generated using random and analog series-based compound partitioning. For each activity class, compounds were randomly partitioned to obtain 80/20% and 50/50% training/test compound splits. In addition, analog series comprising at least two compounds were systematically extracted from activity classes using the compound–core relationship algorithm [17]. Remaining singletons were discarded. The analog series were then partitioned into training and test sets corresponding to ~80/20% training/test compound splits such that both sets consisted of unique analog series with no analog overlap between sets.

4.3.1. Support Vector Regression

SVR is an extension of the support vector machine algorithm for supervised learning that derives a regression function through the generation of an ε-insensitive tube using training data. If a linear data separation is not feasible in the original feature space, a kernel function is employed to project the data to a high-dimensional space where linear separation might become possible [7,8]. For SVR, the regularization hyper-parameter C was determined by testing (0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 10, 100, and 10,000) values. SVR models were derived using the Tanimoto kernel [18].

4.3.2. k-Nearest Neighbor Regression

kNN is a non-parametric supervised learning method that ranks training compounds based on increasing molecular similarity (decreasing distance). For a test compound, the potency is then determined based on the potency values of the k top-ranked compounds from the training set [19]. For kNN, the best-performing k values were determined for one, three, and five top-ranked compounds by averaging potency values for three and five compounds. In addition to applying optimized kNN values, 1-NN predictions were consistently reported for all activity classes. For compound comparison, Tanimoto similarity was calculated using the folded 2048-bit version of the extended connectivity fingerprint with bond diameter 4 (ECFP4) [20] generated with RDKit [21].

4.3.3. Median Regression

The MR control calculation uniformly assigns the median potency value of the training set to each test set compound. This approach was employed as a control calculation.

4.3.4. Hyperparameter Optimization

For parameter optimization, kNN and SVR were submitted to a grid search with 5-fold internal cross-validation implemented using scikit-learn [22].

4.4. Molecular Representation

For modeling, compounds were represented using the folded 2048-bit version of ECFP4 generated using RDKit.

4.5. Performance Metric

To evaluate model performance, the mean absolute error (MAE) was calculated by comparing predicted and experimental test compound potency values. The calculations were carried out using scikit-learn. MAE is defined as

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (1)$$

where n is the number of compounds, and y and $\hat{y}$ are the experimental and predicted potency values, respectively.

Increasing MAE values indicate decreasing prediction accuracy and vice versa.

4.6. Statistical Significance Testing

Statistical significance evaluation of differences between MAE value distributions was carried out using the Wilcoxon test [23]. The alpha value with Bonferroni correction (n = 10) was set to 0.005 and compared to the respective p-value ($p < 0.005$).

5. Conclusions

In this work, we have systematically investigated compound potency predictions on nearly 400 different activity classes using ML and simple control models. In accord with earlier observations, methods of different complexity produced overall similar prediction accuracy differentiated by only small error margins, as demonstrated now on a very large scale. Moreover, relative method performance remained stable despite specific potency range and structural data set modifications designed to increase the difficulty of the calculations. Taken together, our findings clearly indicate that conventional benchmark calculations are not a realistic indicator of differences in the predictive performance of alternative computational methods. Therefore, future research in this area should focus on exploring and devising new concepts for benchmarking potency prediction methods.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ph16040530/s1, Figure S1: Prediction accuracy, Figure S2: Potency value distributions.

Author Contributions: Conceptualization, J.B.; methodology, T.J. and J.B.; software, T.J.; formal analysis, T.J. and J.B.; investigation, T.J.; writing—original draft preparation, T.J. and J.B.; writing—review and editing, T.J. and J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article and supplementary material.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lewis, R.A.; Wood, D. Modern 2D QSAR for Drug Discovery. *WIREs Comput. Mol. Sci.* **2014**, *4*, 505–522. [CrossRef]
2. Guedes, I.A.; Pereira, F.S.S.; Dardenne, L.E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.* **2018**, *9*, e1089. [CrossRef] [PubMed]
3. Williams-Noonan, B.J.; Yuriev, E.; Chalmers, D.K. Free Energy Methods in Drug Design: Prospects of "Alchemical Perturbation" In Medicinal Chemistry. *J. Med. Chem.* **2018**, *61*, 61638–61649. [CrossRef]
4. Gleeson, M.P.; Gleeson, D. QM/MM Calculations in Drug Discovery: A Useful Method for Studying Binding Phenomena? *J. Chem. Inf. Model.* **2009**, *49*, 670–677. [CrossRef]
5. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug. Discov.* **2019**, *18*, 463–477. [CrossRef] [PubMed]
6. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958. [CrossRef]
7. Drucker, H.; Burges, C. Support Vector Regression Machines. *Adv. Neural Inform. Proc. Syst.* **1997**, *9*, 155–161.
8. Smola, A.J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
9. Hou, F.; Wu, Z.; Hu, Z.; Xiao, Z.; Wang, L.; Zhang, X.; Li, G. Comparison Study on the Prediction of Multiple Molecular Properties by Various Neural Networks. *J. Phys. Chem. A* **2018**, *122*, 9128–9134. [CrossRef]
10. Feinberg, E.N.; Sur, D.; Wu, Z.; Husic, B.E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V.S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530. [CrossRef]
11. Walters, W.P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc. Chem. Res.* **2020**, *54*, 263–270. [CrossRef]
12. Janela, T.; Bajorath, J. Simple Nearest Neighbor Analysis Meets the Accuracy of Compound Potency Predictions Using Complex Machine Learning Models. *Nat. Mach. Intell.* **2022**, *4*, 1246–1255. [CrossRef]
13. Bento, A.P.; Gaulton, A.; Hersey, A.; Bellis, L.J.; Chambers, J.; Davies, M.; Krüger, F.A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090. [CrossRef]
14. Baell, J.B.; Holloway, G.A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740. [CrossRef]
15. Bruns, R.F.; Watson, I.A. Rules for Identifying Potentially Reactive or Promiscuous Compounds. *J. Med. Chem.* **2012**, *55*, 9763–9772. [CrossRef]
16. Irwin, J.J.; Duan, D.; Torosyan, H.; Doak, A.K.; Ziebart, K.T.; Sterling, T.; Tumanian, G.; Shoichet, B.K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58*, 7076–7087. [CrossRef]
17. Naveja, J.J.; Vogt, M.; Stumpfe, D.; Medina-Franco, J.L.; Bajorath, J. Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound-Core Relationship Method. *ACS Omega* **2019**, *4*, 1027–1032. [CrossRef] [PubMed]
18. Ralaivola, L.; Swamidass, S.J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Netw.* **2005**, *18*, 1093–1110. [CrossRef] [PubMed]
19. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175–185.
20. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef]
21. RDKit: Cheminformatics and Machine Learning Software. 2013. Available online: http://www.rdkit.org (accessed on 1 July 2022).
22. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
23. Conover, W.J. On Methods of Handling Ties in the Wilcoxon Signed-Rank Test. *J. Am. Stat. Assoc.* **1973**, *68*, 985–988. [CrossRef]

Appendix C

Rationalizing General Limitations in Assessing and Comparing Methods for Compound Potency Prediction.

scientific reports

Check for updates

OPEN

Rationalizing general limitations in assessing and comparing methods for compound potency prediction

Tiago Janela & Jürgen Bajorath✉

Compound potency predictions play a major role in computational drug discovery. Predictive methods are typically evaluated and compared in benchmark calculations that are widely applied. Previous studies have revealed intrinsic limitations of potency prediction benchmarks including very similar performance of increasingly complex machine learning methods and simple controls and narrow error margins separating machine learning from randomized predictions. However, origins of these limitations are currently unknown. We have carried out an in-depth analysis of potential reasons leading to artificial outcomes of potency predictions using different methods. Potency predictions on activity classes typically used in benchmark settings were found to be determined by compounds with intermediate potency close to median values of the compound data sets. The potency of these compounds was consistently predicted with high accuracy, without the need for learning, which dominated the results of benchmark calculations, regardless of the activity classes used. Taken together, our findings provide a clear rationale for general limitations of compound potency benchmark predictions and a basis for the design of alternative test systems for methodological comparisons.

In computer-aided drug discovery, the prediction of compounds that are active against given targets and the prediction of compound potency are central tasks[1,2]. For the quantitative prediction of compound potency, methods of greatly varying complexity have been introduced, ranging from linear regression techniques to deep machine learning[3–8]. For modeling of non-linear structure-activity relationships and potency prediction, machine learning has generally become the prevalent approach, for which a variety of algorithms are available[1,5]. Despite the increasing popularity of deep neural networks[7,8], mainstay approaches such as random forest regression (RFR)[9] or support vector regression (SVR)[10] continue to be widely used.

Computational methods for qualitative compound activity or quantitative potency predictions must generally be evaluated in benchmark settings using known active compounds. For activity prediction, classification models are often trained to separate sets of compounds that are active against different targets, termed activity classes, from randomly assembled compounds. Hence, activity classes represent target-based compound data sets (target sets). For potency prediction, regression models are derived for individual activity classes to predict potency values of test sets extracted from these classes. Care should be taken to limit model derivation and evaluation to compounds for which well-defined potency measurements of the same type are available that can be directly compared. Hence, data curation plays an important role.

Although benchmarking is not a reliable indicator for the success or failure of alternative approaches in practical applications, it represents an essential first step in performance evaluation and comparison of different methods. However, for compound potency prediction, principal limitations of benchmark calculations were recently uncovered[11]. Specifically, it was shown that (i) different machine learning methods including deep neural networks produced very similar potency predictions on different activity classes; (ii) simple k-nearest neighbor (kNN) assignments, carried out as a control, correctly predicted potency values within an order of magnitude, comparable to increasingly complex machine learning methods; (iii) random predictions often also reproduced experimental potency values within an order of magnitude; (iv) prediction errors of all methods fell into a small interval[11]. Overall, SVR predictions were slightly more accurate than those obtained with other

B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Department of Life Science Informatics and Data Science, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 5/6, 53115 Bonn, Germany. ✉email: bajorath@bit.uni-bonn.de

methods including deep neural networks, but observed differences were only marginal[11]. Hence, typical benchmark calculations yielded predictions of comparable accuracy using distinct methods of varying computational complexity as well as random predictions. It follows that standard benchmark calculations are not suitable for assessing the predictive performance of machine learning methods in a meaningful way. The generality of these unexpected findings was further investigated by systematic potency predictions using machine learning methods and controls on 367 activity classes covering all pharmaceutical target classes including, among others, diverse enzymes, different types of receptors, and ion channels, for which qualifying potency measurements were available, which yielded very similar results[12]. Taken together, these findings suggested that the intrinsic limitations of benchmark potency predictions might be a consequence of the composition of activity classes originating from medicinal chemistry sources and their potency value distributions. Therefore, activity classes were modified in different ways including removal of nearest neighbors, partitioning of compounds into training and test sets based on analogue series (thereby avoiding "data leakage", that is, the use of analogous compounds for training and testing), and balancing of compound numbers across different potency levels[12]. Then benchmark calculations were repeated with modified activity classes. However, the predictions were surprisingly stable and largely insensitive to these data set modifications, leading to only small increases in error margins that were very similar for different methods[12]. Thus, reasons for the very similar performance of different methods and simple controls in compound potency predictions remained elusive.

Therefore, we have further investigated potential reasons for the limitations of compound potency predictions. Since the predictions were essentially insensitive to structural modifications of activity classes, we have conducted an in-depth analysis of the influence of potency value distributions and potency sub-ranges in activity classes on compound potency predictions using different approaches, as reported herein.

Methods

Compounds and activity data

From ChEMBL (version 30)[13], compounds with reported direct interactions (target relationship type: "D") with human targets at the highest confidence level (target confidence score: 9), a molecular mass of at most 1000 Da, and available numeric IC_{50} values (recorded as negative decadic logarithmic pIC_{50} values) in the range of 5–11 were extracted. Compounds with measurements flagged as "potential transcription error" or "potential author error" were discarded as well as compounds with assay interference potential detected using available filters[14–16]. We searched for activity classes for which at least 75 compounds falling into each of the three potency pIC_{50} sub-ranges 5–6.9, 7–8.9, and 9–11 were available, leading to the identification of eight classes comprising a total of 9301 compounds. In the following, for simplicity, these sub-ranges are referred to as 5–7, 7–9, and 9–11. Figure 1 shows exemplary compounds for each class and specifies the target names.

Training and test sets

For each activity class, training and test sets for 10 independent prediction trials were obtained by random compound partitioning into 50% training and 50% test data. Hence these training and test sets were not balanced across the three potency sub-ranges. Supplementary Table S1 reports the proportions of compounds falling into each potency sub-range for all activity classes. For the three activity classes with the largest number of compounds in the potency sub-range 9–11 (highly potent compounds), nine training sets of increasing size were generated (for 10 independent trials) by uniformly sampling compounds for each potency sub-range. Smallest training sets consisted of only six compounds (two from each potency sub-range), followed by training sets with 12 compounds (four per potency sub-range), 18, 30, 48, 78, 126, 204, and 330 compounds. After building the largest training set (330 compounds), the remaining compounds, were used to build the test set with balanced potency sub-ranges (with respect to sub-range 9–11, containing the smallest number of compounds per sub-range for the three activity classes).

For comparison, corresponding predictions were also carried out for imbalanced training sets of increasing size and imbalanced test sets.

Molecular representation

For machine learning, compounds were represented using the folded 2048-bit version of the extended connectivity fingerprint with bond diameter 4 (ECFP4)[17] generated with RDKit[18].

Machine learning models

Given that potency prediction results were very similar using methods of different complexity and simple controls[11], machine learning models were built using SVR, the overall preferred approach, and RFR for comparison.

Hyperparameter optimization

For hyperparameter optimization, a grid search with 3-split cross-validation was performed using *scikit-learn*[19] based on training data. Therefore, training sets were divided into 50% training and 50% validation data. For balanced training sets, the splits were stratified by potency range.

Support vector regression

SVR is a variant of the support vector machine algorithm for supervised learning that derives a hyperplane based on training instances to reduce the error between observed and predicted values. A kernel function is used to project samples from the original dimension into a higher-dimensional feature space[10,20]. For SVR, the

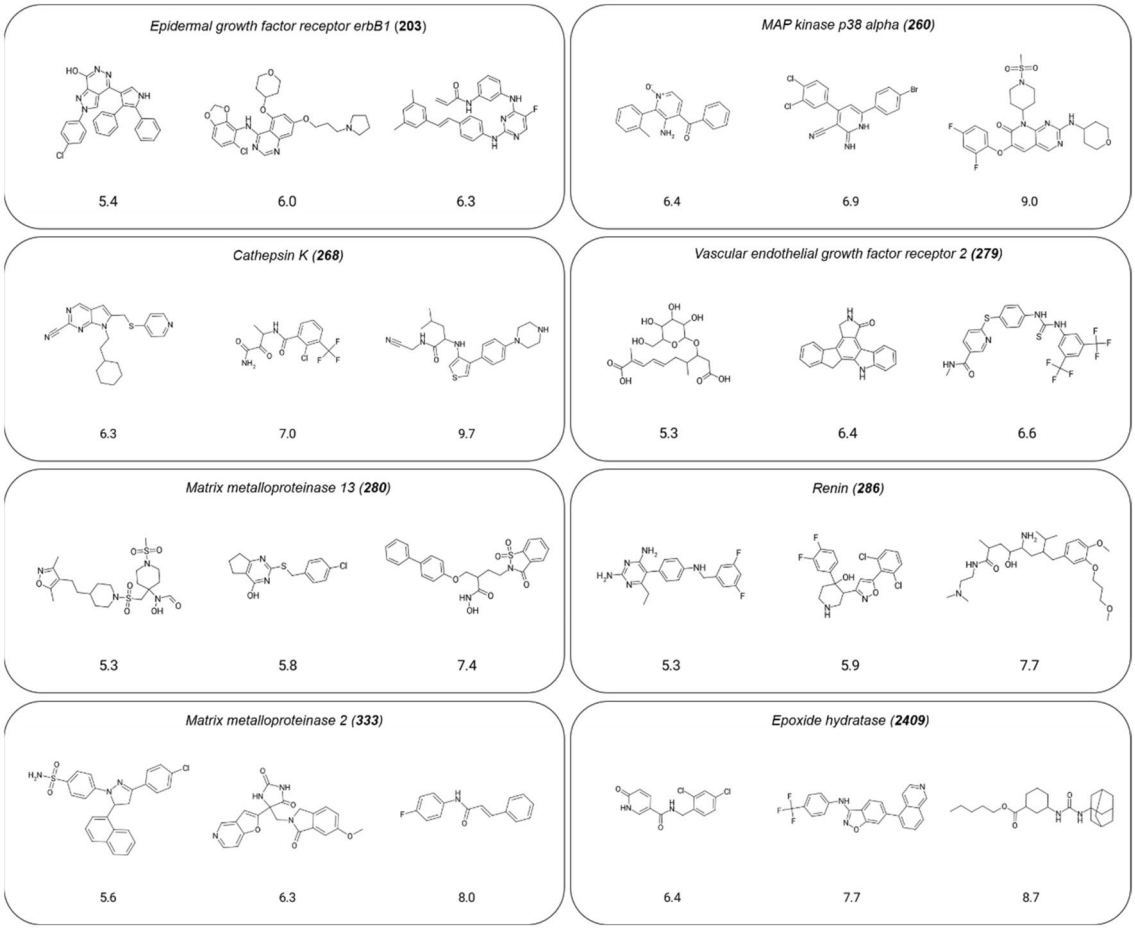


Figure 1. Activity classes. For each of the eight activity classes (target sets), the target name and ChEMBL target ID (in parentheses) are provided and exemplary structurally diverse compounds are shown. For each compound, the pIC_{50} value is reported.

cost parameter C was optimized with the values of 1, 10, 100, and 1000. Models with the Tanimoto kernel[21] were built using *scikit-learn*.

Random forest regression

RFR is a machine learning method employing an ensemble of decision trees. Each tree model was built by randomly sampling a subset of training compound using bootstrapping[9,22]. Numerical values were predicted as the average value of all individual trees. For RFR, the number of trees (50, 100, 200), minimum number of samples per split (2, 3, 5, 10), minimum sample per leaf (1, 2, 5, 10), and maximal number of features for achieving the best split (sqrt, log2) were optimized.

Controls

Nearest neighbor calculations

k-NN is a regression technique that selects for each test instance the k nearest neighbors from the training set and assigns the potency value of the most similar training compound to the test instance (1-NN) or averages the potency values for the k (> 1) most similar training compounds[23]. For comparing test and training set compounds, Tanimoto similarity[24] was calculated based on ECFP4. 1-NN and 3-NN calculations were carried out with *scikit-learn*.

Median regression

Median regression (MR), the simplest possible control, assigns the median potency value of the training set to each test compound as the predicted value.

Performance metrics

Prediction accuracy was evaluated using the mean absolute error (MAE), root mean squared error (RMSE), and squared Pearson correlation coefficient (r^2). Training of machine learning models was guided by MAE values or, as a control, R^2 (coefficient of determination).

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \quad (1)$$

$$RMSE(y,\hat{y}) = \sqrt{\sum_{i=1}^{n}\frac{\left(y_i - \hat{y}_i\right)^2}{n}} \quad (2)$$

$$r^2 = \left(\frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}\right)^2 \quad (3)$$

For MAE and RMSE, n is the number of compounds, and y and $\hat{y}$ are the experimental and predicted potency values, respectively. For r^2, m_x is the mean of vector x and m_y the mean of vector y.

Statistical significance testing

The Wilcoxon signed-rank test[25] was used to assess the statistical significance of observed differences between MAE, RMSE and r^2 value distributions. The p-value ($p < \alpha$) was compared to an alpha value of 0.005 with Bonferroni correction (n = 10).

Results

Compound potency value distributions

We first determined the potency value distributions of the activity classes, as shown in Fig. 2a and Fig. 2b for the three classes with largest numbers of compounds in potency (pIC_{50}) sub-range 9–11 and Supplementary Fig. S1a and Fig. S1b for all classes. The center of the entire potency range 5–11 corresponding to compounds with intermediate potency contained the majority of compounds in all classes. In each case, the median potency of all classes fell into the pIC_{50} interval 7–8. However, there were clear activity class-dependent differences in potency value distributions, with different peaks in the distributions.

Pairwise similarity was then separately calculated for all compounds falling into each of the potency sub-ranges 5–7, 7–9, and 9–11. Figure 2c and Supplementary Fig. S1c show that the resulting similarity value distributions were comparable for all activity classes and also comparable for each class across the different potency sub-ranges. As expected, some activity classes were structurally more homogeneous than others in individual sub-ranges (such as class 203 in Fig. 2), but large differences in compound similarity value distributions across different potency sub-ranges were not observed. Hence, there was no apparent relationship between intra-class compound similarity and differences in potency value distributions between the activity classes.

Compound potency predictions

For the eight activity classes, potency predictions were carried out using the SVR, RFR, 1-NN, 3-NN, and MR approaches. Prediction accuracy was assessed on the basis of MAE, RMSE, and r^2 calculations. Figure 3 shows the results for the three activity classes with the largest numbers of highly potent compounds and Supplementary Fig. S2 compares the results for all activity classes based on MAE (Fig. S2a), RMSE (Fig. S2b), and r^2 values (Fig. S2c). Consistent with earlier observations[11,12], the performance of all methods across the entire potency range was comparable for all activity classes and varying training and test set ratios. The predictions were stable, as indicated by very narrow error distributions over independent trials, and reached reasonable accuracy, with MAE and RMSE values generally smaller than 0.8 and 1.0, respectively (except for MR, as further discussed below). Hence, the different methods generally predicted potency values well within an order of magnitude (tenfold). Lowest prediction errors detected were ~ 0.4 and ~ 0.5 for MAE and RMSE, respectively. SVR predictions were overall slightly more accurate than RFR and 1-/3-NN calculations. As a control, the machine learning models were also retrained using R^2 as a cost function and the predictions using these models were assessed based on MAE values. As shown in Supplementary Fig. S2a and Fig. S2d, the results obtained for alternatively trained models using MAE or R^2 for alternatively trained models were nearly identical.

Importantly, while the majority of differences between MAE and RMSE value distributions for all pairwise comparisons of methods were statistically significant, as shown in Supplementary Fig. S3a and S3b, respectively, differences in mean prediction errors of all methods were confined to ~ 0.1 units and thus essentially negligible. These results were stable for varying training and test set ratios. In addition, Supplementary Fig. S3c shows that most differences between r^2 values for potency ranges 5–7 and 9–11 were not statistically significant, hence indicating the presence of strong correlation.

The predictions were then separately compared for all test compounds falling into each of the three potency sub-ranges, as also shown in Fig. 3 and Supplementary Fig. S2, which provided a more differentiated view of the results. For weakly potent (sub-range 5–7) and highly potent (9–11) compounds, prediction errors increased by up to ~ 0.2 units for SVR, RFR, and 1-/3-NN. For MR, MAE/RMSE values up to 2.0 were observed because the median potency value of all activity classes fell into the pIC_{50} range of 7–8 (see above). By contrast, for test compounds in potency sub-range 7–9, prediction errors further decreased for all methods by ~ 0.1 units compared to the global accuracy (potency range 5–11) and was closely matched by MR. The comparison in Fig. 3 indicated that the global prediction accuracy of all methods was essentially determined by the similarly low prediction error observed for all methods in intermediate potency sub-range 7–9 where all compound potency values tended to be close to the median.

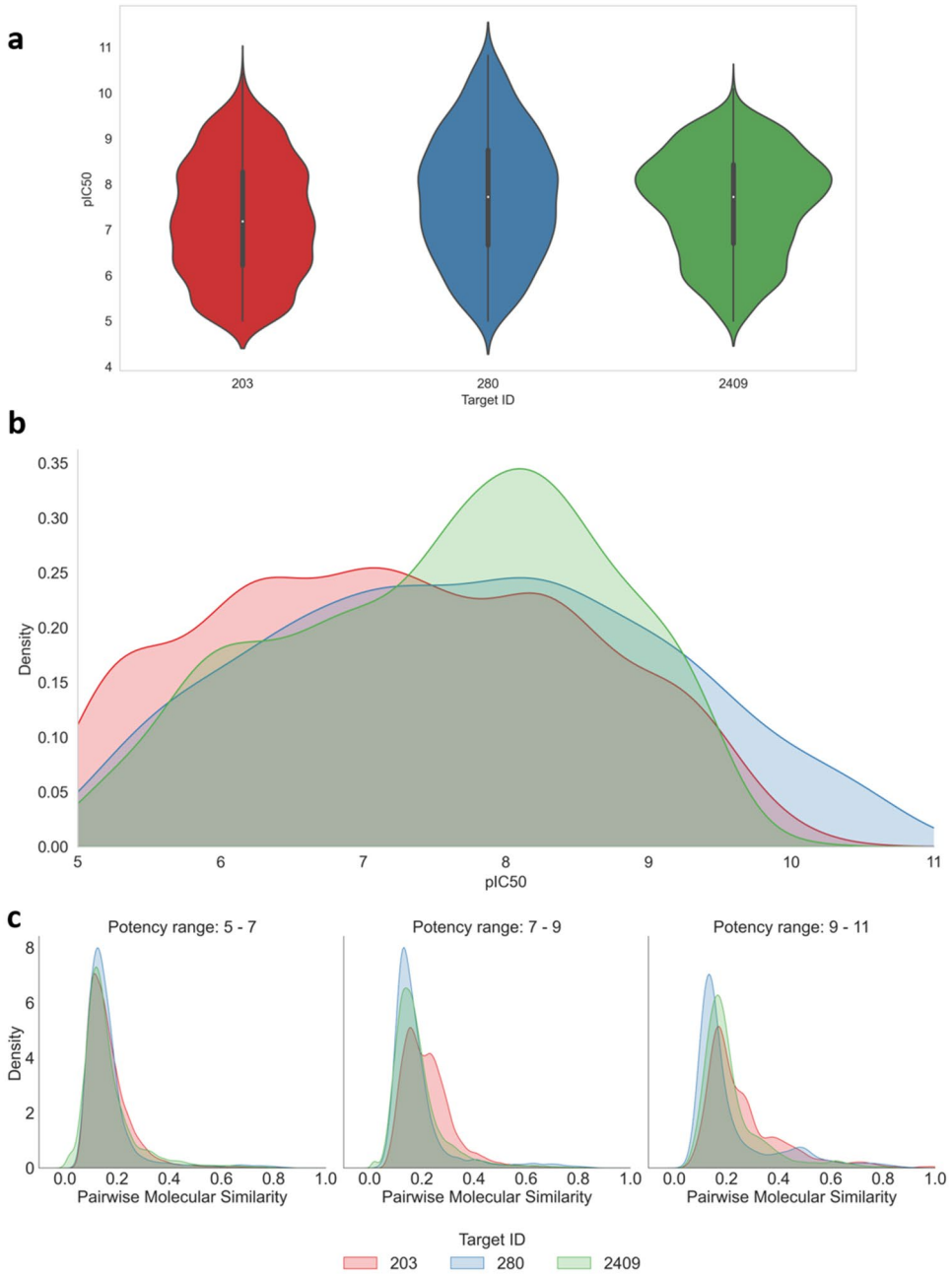


Figure 2. Potency value and pairwise molecular similarity distributions. For the three activity classes with the largest numbers of compounds in potency sub-range 9–11, (**a**) violin plots report the potency value distributions across the three potency sub-ranges (5–7, 7–9, 9–11). In a violin plot, a value distribution is represented by its maximum value (upper thin line), upper quartile (upper thick line), median value (white dot), lower quartile (lower thick line) and minimum value (lower thin line). On each side of the vertical line, a density plot is shown. In (**b**), density plots obtained by kernel density estimation compare the potency distributions across the entire potency range. In (**c**), density plots report the distributions of pairwise Tanimoto similarity values for compounds populating the three potency sub-ranges.

Furthermore, as shown in Fig. 3, calculation of r^2 for predicted and experimental potency values revealed positive correlation across the entire potency range for SVR, RFR and 1-/3-NN predictions (as anticipated, given the low prediction errors and large sample sizes). For the three potency sub-ranges, correlation was significantly lower, which was at least in part attributable to the small sample sizes for the low and high potency sub-ranges. Largest correlation was observed for the mid sub-range (7–9), consistent with the low prediction errors in this range. Importantly, r^2 calculations did not lead to a larger separation between the performance of different models. Thus, correlation analysis mirrored the observed prediction characteristics discussed above.

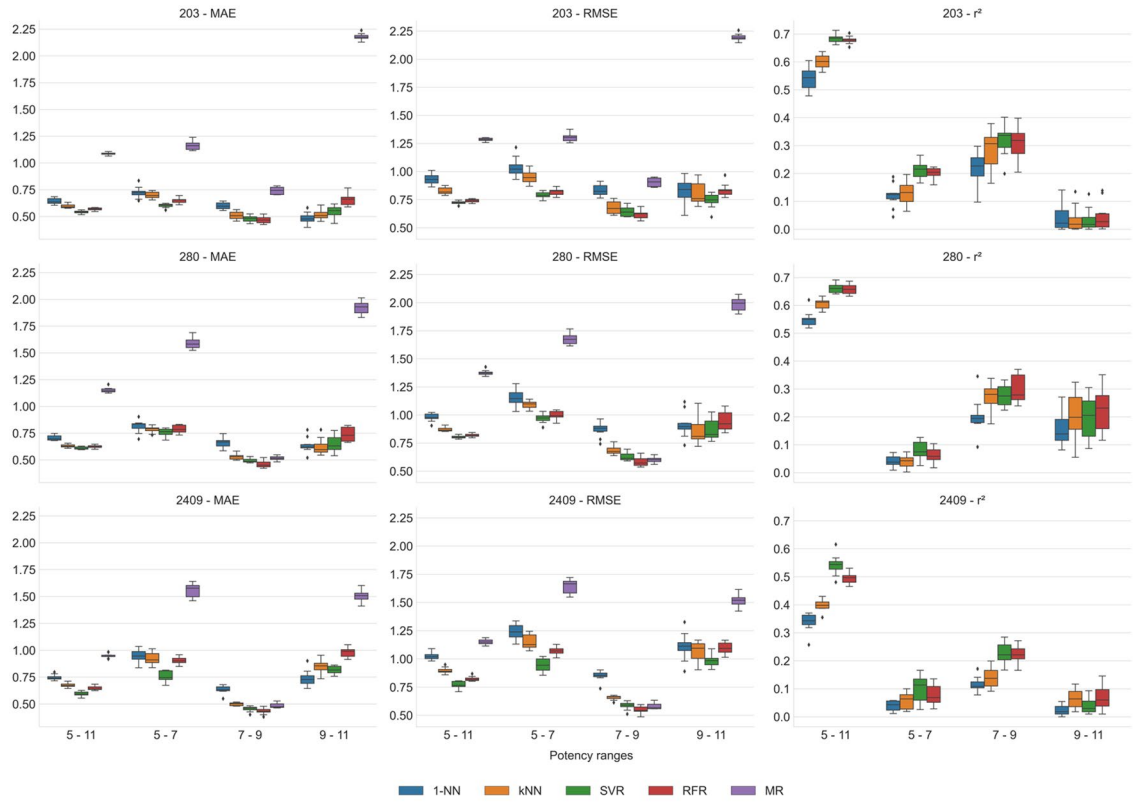


Figure 3. Prediction accuracy. Boxplots report the distribution of MAE (left), RMSE (middle), and r^2 values (right) for potency predictions over 10 independent trials with constantly sized (imbalanced) training sets using 1-NN, 3-NN, SVR, RFR, and MR for three activity classes. In each case, predictions are reported for the entire potency range (5–11) and test compounds with experimental potency falling into the three sub-ranges. In boxplots, the upper and lower whiskers indicate maximum and minimum values, the boundaries of the box represent the upper and lower quartiles, values classified as statistical outliers are shown as diamonds, and the median value is indicated by a horizontal line.

Potency value sub-range dependence of predictions

To further investigate the apparent dependence of the predictions on the compound potency sub-ranges of the activity classes, we generated training sets with balanced sub-range populations of increasing size for the three activity classes for which sufficient numbers of highly potent compounds (see Methods) were available and repeated the predictions for each sub-range. Size variation of training sets was introduced to examine data requirements for the predictions and learning characteristics of the methods. Figure 4 shows the results of sub-range based potency predictions.

For weakly potent (sub-range 5–7) and highly potent (9–11) compounds, smallest training sets of 6–18 compounds produced median MAE values of ~ 2.0 (corresponding to ~ 100-fold potency prediction errors) for all methods and yielded broad MAE value distributions, indicating unstable predictions that were often comparable MR. As expected, very small training sets were insufficient for machine learning and a median MAE of ~ 2.0 essentially represented the upper limit of prediction errors observed under these conditions. When the size of training sets further increased, the predictions for weakly and highly potent compounds became more stable and accurate for SVR, RFR, and 1-/3-NN, as indicated by increasing separation from the MR values, and approached the accuracy level observed in the global predictions (Fig. 3). Hence, for weakly and highly potent test compounds, prediction accuracy clearly increased with the size of training sets with balanced potency sub-ranges, as expected. Notably, the relative performance of the different methods remained comparable as training set sizes and prediction accuracy increased.

By contrast, distinct prediction characteristics were observed for test compounds in potency sub-range 7–9. Here, the prediction errors were constantly small, independent of training set size, and the accuracy achieved by SVR and RFR was very close to the median potency values of the training sets. Thus, in this case, essentially no learning was required and prediction accuracy was constantly high for MR across all training sets. Whereas 1-/3-NN closely matched SVR/RFR predictions for highly and weakly potent compounds, NN calculations mostly yielded larger errors in the intermediate potency sub-range, especially 1-NN. However, most of these NN calculation errors were comparable to the best predictions achieved with all methods for highly and weakly potent test compounds based on largest training sets. Moreover, SVR, RFR, and MR predictions in the potency sub-range 7–9 were consistently the by far most accurate predictions that were obtained. As an additional control,

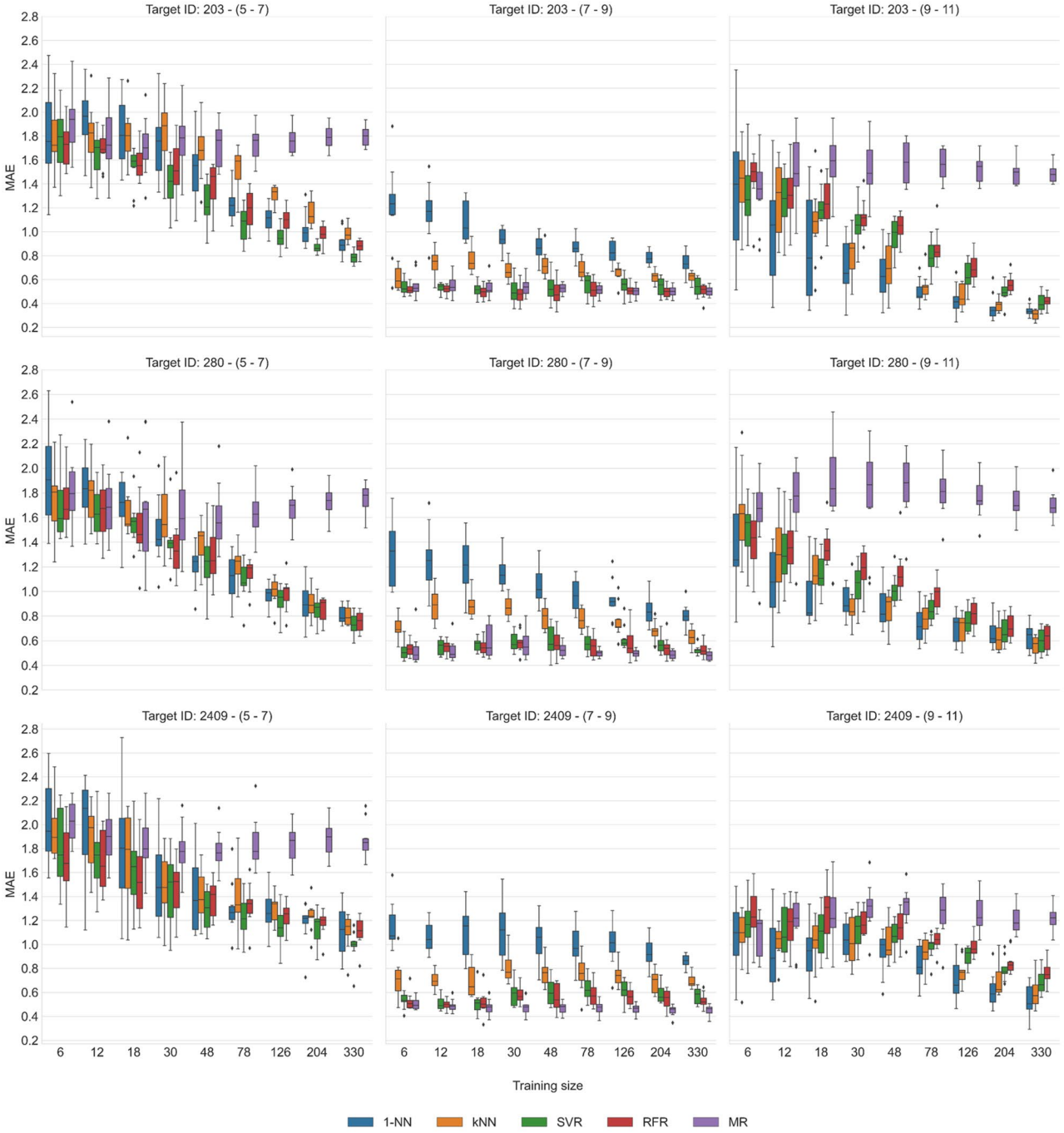


Figure 4. Prediction accuracy for training sets of increasing size. Boxplots report the distribution of MAE values for potency predictions over 10 independent trials with potency sub-ranged balanced training sets of increasing size using 1-NN, 3-NN, SVR, RFR, and MR for three activity classes. The predictions were separately carried out for each potency sub-range.

we also repeated the potency sub-range predictions with imbalanced training sets of increasing size, as shown in Supplementary Fig. S4, yielding the same trends.

The analysis of the potency sub-range dependence of the predictions clearly demonstrated that they were largely determined by compounds falling into the intermediate potency range. Here, predictions for machine learning models were consistently most accurate. However, there was essentially no learning required because the predictions were independent of training set sizes and closely matched the median potency values of the training sets. Hence, in the intermediate potency sub-range, predictions yielded artificially low errors, due to narrow local potency value distributions around the median. In original activity classes, the majority of compounds fell into the intermediate potency range, which strongly dominated global potency predictions.

Conclusion

Compound potency predictions play an important role in computer-aided drug discovery. Benchmark calculations are essential and widely applied for an initial assessment and comparison of predictive methods, prior to practical applications. However, previous studies have revealed general limitations of benchmark evaluation of potency prediction methods. Increasingly complex machine learning methods and simple control calculations displayed similar performance in test calculations on many different activity classes, and even random predictions were only separated from machine learning results by small error margins. As a consequence, benchmarking can currently not reliably assess the predictive performance and relative differences between alternative methods; a conundrum for method development and evaluation. Since the performance of distinct potency prediction approaches was comparable for many different activity classes (as also shown herein), these observations must in principle be attributable to intrinsic features of activity classes such as structural composition or potency value distributions, as we have reasoned. However, origins of apparent artifacts in benchmarking potency prediction methods have remained unknown so far, presenting a substantial problem for the field. Therefore, we have designed test calculations to directly investigate the influence of potency value distributions and sub-range effects on compound potency predictions. Although potency value distributions of activity classes differed, predictions were largely determined by very low errors consistently detected in the intermediate potency (pIC_{50}) sub-range 7–9 into which median potency values of different activity classes fell. These prediction characteristics fundamentally differed from those observed for weakly and highly potent compounds. Machine learning predictions in the intermediate potency sub-range consistently and closely matched median potency values of training sets even under learning conditions where predictions of weakly and highly potent compounds essentially failed. The dominance of very low errors in the intermediate potency sub-range led to closely comparable results of different approaches in global potency predictions and provided a clear rationale for the artificial outcome of benchmark calculations including the low error margins hindering methodological comparisons. Taken together, the results of our analysis explain in detail why conventional benchmark settings do not provide a realistic assessment of compound potency prediction methods and provide a basis for future work investigating alternative approaches for more reliable methodological comparisons.

Data availability

Calculations were carried out using publicly available software and compound data. Code used for this analysis and the curated activity classes are freely available via the following links: https://github.com/TiagoJanela/Limitations-compound-potency-predictions and https://zenodo.org/badge/latestdoi/663107456.

Received: 7 July 2023; Accepted: 16 October 2023
Published online: 19 October 2023

References

1. Bajorath, J. Computer-aided drug discovery. *F1000Research* https://doi.org/10.12688/f1000research.6653.1 (2015).
2. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
3. Lewis, R. A. & Wood, D. Modern 2D QSAR for drug discovery: QSAR for drug discovery. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 505–522 (2014).
4. Williams-Noonan, B. J., Yuriev, E. & Chalmers, D. K. Free energy methods in drug design: Prospects of "alchemical perturbation" in medicinal chemistry. *J. Med. Chem.* **61**, 638–649 (2018).
5. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
6. Feinberg, E. N. *et al.* PotentialNet for molecular property prediction. *ACS Cent. Sci.* **4**, 1520–1530 (2018).
7. Hou, F. *et al.* Comparison study on the prediction of multiple molecular properties by various neural networks. *J. Phys. Chem. A* **122**, 9128–9134 (2018).
8. Walters, W. P. & Barzilay, R. Applications of deep learning in molecule generation and molecular property prediction. *Acc. Chem. Res.* **54**, 263–270 (2021).
9. Svetnik, V. *et al.* Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958 (2003).
10. Drucker, H., Surges, C. J. C., Kaufman, L., Smola, A. & Vapnik, V. Support vector regression machines. *Adv. Neural. Inform. Proc. Syst.* **9**, 155–161 (1997).
11. Janela, T. & Bajorath, J. Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models. *Nat. Mach. Intell.* **4**, 1246–1255 (2022).
12. Janela, T. & Bajorath, J. Large-scale predictions of compound potency with original and modified activity classes reveal general prediction characteristics and intrinsic limitations of conventional benchmarking calculations. *Pharmaceuticals* **16**, 530 (2023).
13. Bento, A. P. *et al.* The ChEMBL bioactivity database: An update. *Nucleic Acids Res.* **42**, 1083–1090 (2014).
14. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**, 2719–2740 (2010).
15. Bruns, R. F. & Watson, I. A. Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.* **55**, 9763–9772 (2012).
16. Irwin, J. J. *et al.* An aggregation advisor for ligand discovery. *J. Med. Chem.* **58**, 7076–7087 (2015).
17. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
18. RDKit: Cheminformatics and Machine Learning Software. http://www.rdkit.org . Accessed 1 June 2022.
19. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
20. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
21. Ralaivola, L., Swamidass, S. J., Saigo, H. & Baldi, P. Graph kernels for chemical informatics. *Neural Netw.* **18**, 1093–1110 (2005).
22. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
23. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**, 175–185 (1992).
24. Willett, P., Barnard, J. M. & Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998).
25. Conover, W. J. On methods of handling ties in the Wilcoxon signed-rank test. *J. Am. Stat. Assoc.* **68**, 985–988 (1973).

Acknowledgements
The authors thank Alec Lamens and Jannik P. Roth for helpful discussions.

Author contributions
Both authors contributed to designing and conducting the study, analyzing the results, and preparing the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-45086-3.

Correspondence and requests for materials should be addressed to J.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix D

Anatomy of Potency Predictions Focusing on Structural Analogues with Increasing Potency Differences Including Activity Cliffs.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

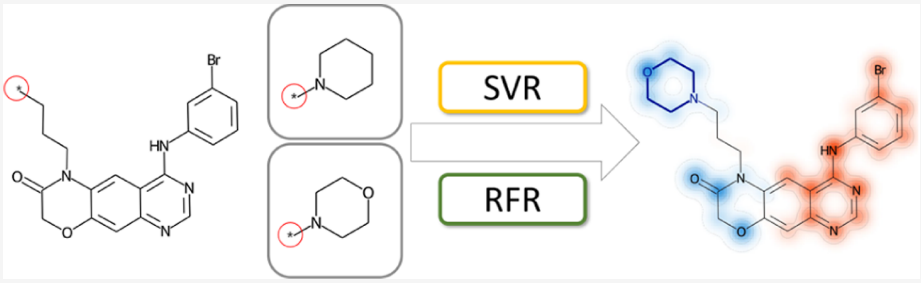
pubs.acs.org/jcim Article

Anatomy of Potency Predictions Focusing on Structural Analogues with Increasing Potency Differences Including Activity Cliffs

Tiago Janela and Jürgen Bajorath*

Cite This: J. Chem. Inf. Model. 2023, 63, 7032−7044 Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: Potency predictions are popular in compound design and optimization but are complicated by intrinsic limitations. Moreover, even for nonlinear methods, activity cliffs (ACs, formed by structural analogues with large potency differences) represent challenging test cases for compound potency predictions. We have devised a new test system for potency predictions, including AC compounds, that is based on partitioned matched molecular pairs (MMP) and makes it possible to monitor prediction accuracy at the level of analogue pairs with increasing potency differences. The results of systematic predictions using different machine learning and control methods on MMP-based data sets revealed increasing prediction errors when potency differences between corresponding training and test compounds increased, including large prediction errors for AC compounds. At the global level, these prediction errors were not apparent due to the statistical dominance of analogue pairs with small potency differences. Test compounds from such pairs were accurately predicted and determined the observed global prediction accuracy. Shapley value analysis, an explainable artificial intelligence approach, was applied to identify structural features determining potency predictions using different methods. The analysis revealed that numerical predictions of different regression models were determined by features that were shared by MMP partner compounds or absent in these compounds, with opposing effects. These findings provided another rationale for accurate predictions of similar potency values for structural analogues and failures in predicting the potency of AC compounds.

INTRODUCTION

Compound potency predictions play a central role in drug design. Widely used potency prediction approaches range from linear quantitative structure−activity relationship (QSAR) methods and scoring functions for quantifying ligand−target interactions to complex free-energy (perturbation) and different machine learning (ML) methods for nonlinear regression modeling.[1−8] While standard QSAR and free-energy perturbation approaches are typically limited to predictions of congeneric compounds, appropriately trained ML models can also be applied to predict the potency of structurally diverse compounds. Popular ML methods for regression modeling and prediction of numerical potency values include support vector regression (SVR),[9] random forest regression (RFR),[10] and various deep neural network (DNN) architectures that have recently gained in popularity.[7,8,11−14]

The initial assessment and comparison of computational methods for predicting compound potency (and molecular properties in general) typically require benchmark calculations, in which models are derived and evaluated based on labeled data such as compounds with known activity against given targets and available experimentally determined potency values. While such benchmark calculations are generally important for assessing model performance prior to practical (prospective) applications, benchmarking of potency prediction methods has intrinsic limitations.[15,16] For example, the assignment of potency values of nearest neighbors in training sets to test compounds (k-nearest neighbor (*k*-NN, *k*NN) analysis or the even simpler assignment of training set median potency values to test compounds (median regression) often

Received: September 22, 2023
Revised: September 30, 2023
Accepted: October 26, 2023
Published: November 9, 2023

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

approaches the predictive performance of increasingly complex ML models.[15] Furthermore, ML-based and randomized potency value predictions are often only separated by narrow error margins of 1 to 2 orders of magnitude,[15] which leads to artificially favorable predictions in benchmark settings. At least in part, these limitations result from compound potency and similarity distributions in target-based compound sets (often termed activity classes) that are commonly used for benchmarking.[15,16] As a possible alternative, potency predictions might be focused on identifying highly potent compounds,[17] taking into account that it might be difficult to precisely predict their potency values, given that their magnitude is statistically underrepresented in activity classes.

Moreover, there are other principal limitations associated with compound potency predictions that are a direct consequence of varying compound structure−activity relationship (SAR) characteristics. For example, standard QSAR methods rely on the presence of continuous SARs in compound sets, where small chemical modifications of congeneric compounds lead to gradual changes in potency.[18] By contrast, SAR discontinuity, where a small chemical modification leads to changes in potency of different magnitudes,[18] falls outside the applicability domain of QSAR methods. The presence of SAR discontinuity typically requires the application of nonlinear ML models for potency prediction. However, such ML models often also strike their limits when encountering activity cliffs (ACs), representing the pinnacle of SAR discontinuity. ACs are defined as pairs or groups of structurally similar compounds (structural analogues) with large differences in potency.[19,20] Because ACs capture extreme SAR discontinuity, they present particularly challenging test cases for QSAR and ML predictions. Accordingly, QSAR models typically produce significant prediction errors for AC compounds.[19−22] However, in most activity classes, only ∼5% of pairs of structural analogues represent ACs with an at least 100-fold difference in compound potency.[23,24] Thus, ACs are generally rare in compound data sets, and their prediction errors might therefore often not significantly affect the overall prediction accuracy observed in potency benchmarks. ACs can be identified on the basis of pairwise molecular similarity calculations or by enumerating pairs of structural analogues and comparing their potency.[20,23] For the computational detection of structural analogues, the matched molecular pair (MMP) concept[25] is readily applicable. MMPs are defined as pairs of compounds that are only distinguished by a chemical modification at a single site,[25] which provides a sound basis for defining ACs because potency changes can consistently be attributed to the replacement of an individual substituent. Accordingly, MMPs capturing an at least 100-fold difference in analogue potency, so-called MMP-cliffs,[26] have become a widely used AC definition in the field.[23]

Beginning in 2012, various attempts have been made to predict ACs.[27−38] Most of these studies have attempted to predict compound pairs forming ACs (often applying the MMP-cliff definition) and distinguish them from pairs of compounds with small potency differences.[27,29−35] For these purposes, ML classification models were used, often producing high accuracy in distinguishing ACs from other pairs of similar compounds. Recently, DNN variants have been used to predict ACs from molecular images[32,33] or graphs using representation learning.[34,35] By contrast, only few attempts have thus far been made to predict the actual potency value of AC compounds using ML regression models and/or DNNs of varying complexity.[22,36−38] The results of the currently most comprehensive study have confirmed the challenges in accurately predicting the potency of AC compounds and have shown that standard ML regression models yielded overall better performance than DNNs.[38]

Given the limitations of potency benchmark calculations and challenges for potency predictions as a consequence of SAR discontinuity, it is meaningful to consider alternative evaluation criteria and system set-ups. We have conceived a new test system to systematically determine the accuracy of potency predictions for structural analogues with increasing potency differences, including ACs. The approach enabled the assessment of prediction accuracy for increasingly challenging test compounds. In addition, we have adapted an explainable artificial intelligence (XAI) approach to better understand how predictions obtained using different ML regression models were determined.

METHODS

Compounds and Activity Data. From ChEMBL (release 33),[39] activity classes comprising compounds with a molecular mass of less than 1000 Da were extracted. Undesired targets such as drug-metabolizing cytochrome P450 isoforms, hERG, and serum-albumin were not considered. Compounds flagged as "not active", "inactive", "inconclusive", "potential author error", or "potential transcription error" were disregarded. In addition, only compounds with direct target interactions (target relationship type: "D") tested in a single-protein assay with the highest ChEMBL assay confidence score of 9 were retained. Furthermore, for each compound, the availability of an IC_{50} potency measurement with a specific value ("=") of at least 10 μM and at most 10 pM was required (recorded as a pIC_{50} value). Hence, active compounds fell into the pIC_{50} range of 5 to 11. Of note, for our current analysis, preference was given to IC_{50} values over other measurements (including equilibrium constants) because large numbers of qualifying compounds were required and IC_{50} values were by far the most frequently available measurement. In addition, in potency prediction studies, the use of different types of potency measurements that cannot be directly compared should be avoided. Therefore, IC_{50} values were exclusively used herein (although equilibrium constants are, in principle, assay-independent). If multiple IC_{50} value measurements were available for a given compound, they were averaged to yield the final potency annotation if all values fell into the same order of magnitude (or if all remaining values fell into the same order of magnitude after the largest or smallest value was removed as a likely outlier). Finally, activity classes were screened for potential assay interference compounds using Lilly medicinal chemistry rules[40] and filters for pan-assay interference compounds[41] and aggregators.[42] After these compound selection and data curation criteria were applied, the 10 largest activity classes were retained for the systematic extraction of MMP and regression modeling. Table 1 summarizes the composition of these activity classes after data curation prior to MMP analysis.

Reported are the activity classes used to derive MMP-based data sets for compound potency prediction.

Matched Molecular Pairs. From each activity class, MMPs were extracted using the compound−core relationship (CCR) algorithm.[43] The CCR method systematically identifies analogue series with single or multiple substitution sites in compound data sets and was applied here to identify all

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

Table 1. Activity Classes

target name	target ID	# compounds
epidermal growth factor receptor erbB1	203	1586
acetylcholinesterase	220	1898
MAP kinase p38 alpha	260	1495
vascular endothelial growth factor receptor 2	279	2475
dipeptidyl peptidase IV	284	1359
histone deacetylase 1	325	1990
histone deacetylase 6	1865	1494
epoxide hydratase	2409	1410
hepatocyte growth factor receptor	3717	1288
PI3-kinase p110-alpha subunit	4005	1534

matching molecular series (MMSs),[44] that is, analogue series with a single substitution site. MMS were obtained using the CCR approach by systematically fragmenting all combinations of exocyclic bonds in compounds according to retrosynthetic rules,[43] yielding core structures and substituents. The core structure was required to be at least twice as large as a substituent fragment that was permitted to consist of a maximum of 13 non-hydrogen atoms. To prevent potential compound overlap between different MMSs, the compound with the smallest number of non-hydrogen atoms was omitted from each series. For each MMS, compound pairs representing MMPs were enumerated such that each compound occurred in only a single MMP (hence avoiding compound overlap between MMPs). For each activity class, this MMS-based sampling procedure was conducted 10 times, thus generating 10 different MMP data sets per class in which a given compound appeared in only a single MMP. Each of these data sets was used to derive 10 independent training and test sets for regression modeling, as described in the following, hence ensuring that there was no compound overlap between training and test sets in independent trials.

Following the MMP-cliff definition,[26] MMPs analyzed herein were considered ACs if the two MMP compounds had an at least 100-fold difference in potency against their target protein. Figure 1 shows exemplary MMP, including an MMP-cliff.

Training and Test Sets. For MMP data sets from each activity class, training and test sets were assembled by using stratified and random data partitioning for the generation of MMP-based regression models and control calculations. For stratified splitting, the compounds forming each MMP were assigned to the training and test sets, respectively. In other words, each MMP was divided into training and test compounds. Thus, stratified sampling ensured that for each training compound, at least one close structural analogue was present in the test set and vice versa. As a control, a random split was carried out by pooling all MMP compounds, followed by random sampling of compounds for training and test sets. In both cases, a training/test data ratio of 50/50% was consistently applied to generate data sets for 10 independent prediction trials.

Random Forest Regression. RFR is a supervised ML algorithm that derives an ensemble of decision trees by randomly sampling training instances using bootstrapping.[10] Test compound predictions are obtained as the average value over all decision trees. For RFR, the number of trees (50, 100, 200), maximal number of features for achieving the best split (sqrt, log2), minimum number of samples per split (2, 3, 5, 10), and minimum sample per leaf (1, 2, 5, 10) were used for optimization. Models were implemented using *scikit-learn*.[45]

Support Vector Regression. SVR is a nonlinear learning method that maps training instances into a higher-dimensional feature space using kernel functions.[9] Herein, the Tanimoto kernel[46] was used, and SVR models were built with *scikit*-learn. The C parameter, which determines the trade-off between the regularization term and the loss function, was optimized with the values of 0.001, 0.1, 1, 10, 100, 1000, and 10,000.

Molecular Representation. For ML models, the folded 2048-bit version of the extended connectivity fingerprint with bond diameter 4 (ECFP4),[47] calculated with RDKit,[48] was used to represent MMP compounds.

Hyperparameter Optimization. For hyperparameter optimization, grid search and 5-fold internal cross-validation were performed using *scikit*-learn. Preferred parameters were selected based on the average error across all optimization trials.

Control Calculations. *Nearest Neighbor Calculations.* *k*-NN predicts the potency values for test compounds by searching for the respective potency values of the *k* nearest neighbors. For 1-NN, the prediction corresponds to the potency value of the closest (most similar) training compound. For *k*-NN ($k > 1$), the final prediction is obtained by averaging the potency values of the *k* most similar training compounds.[49] For NN assessment, Tanimoto similarity[50] was calculated based on ECFP4. In the case of *k*-NN (*k*NN), for *k*, the better performing value of 3 or 5 was used, as determined with *scikit*-learn. Notably, 1-NN and *k*NN assignments are not influenced by learning from MMPs or ACs and hence represent a meaningful control for ML regression models.

Median Regression. MR was used as a control calculation. In MR, the value predicted for each test compound corresponds to the median potency value of the training set.

Model Explanation. To explain predictions of ML models, the Shapley value (SV) formalism originating from game

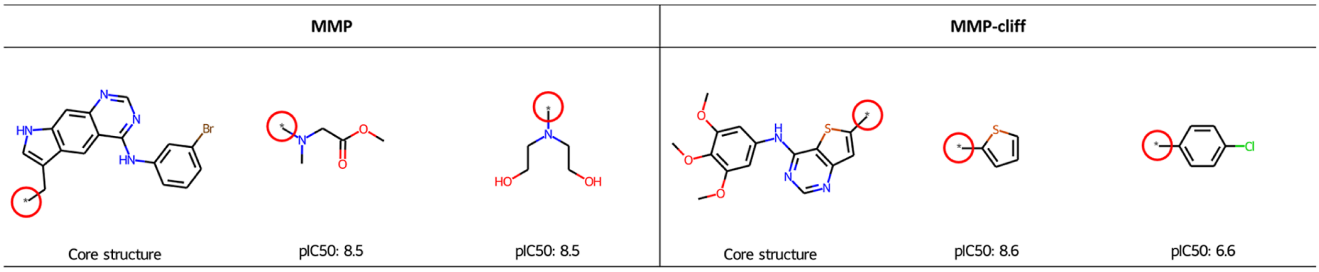


Figure 1. Exemplary MMPs and AC. Exemplary MMPs are shown using the core structure and respective substituents of the paired compounds. The MMP on the right represents an MMP-cliff. Substitution sites are encircled (red). For each compound, the pIC_{50} value is reported.

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

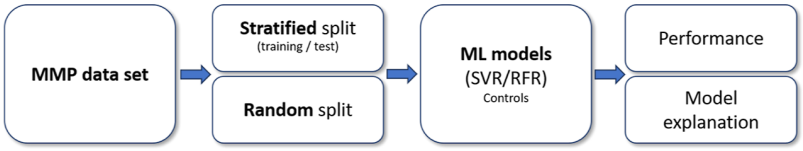


Figure 2. Analysis workflow. This flowchart summarizes the different stages of the analysis, from the generation of MMP data sets (see also Figure 1) and compound pair-based partitioning for assembling training and test sets to potency predictions monitored across subsequent potency difference intervals and their explanation.

theory[51] was applied. In ML, SVs quantify the contribution of features that are present or absent in compounds to predictions of individual test instances. In potency prediction, the expected value, which would be predicted if no features were available, corresponds to the mean of the potency value distribution of the training set. For each test compound, the expected value and the sum of all individual feature contributions quantified using SVs yield the predicted potency value. Given the exponential computational time requirements for increasing numbers of features, exhaustive SV calculations become infeasible for most ML predictions. Therefore, the Shapley Addictive exPlanations (SHAP) approach[52] is typically applied that approximates SVs as SHAP values by deriving a local model in feature space in the vicinity of a given test instance. Currently, the calculation of exact SV values is feasible only for a few ML methods. For SVR using the Tanimoto kernel and binary (present/absent) features, exact SVs can be calculated using the SVETA algorithm.[53] Furthermore, for decision tree methods, TreeExplainer[54] applies the SHAP formalism based on decision tree paths that do not depend on missing features to calculate feature importance values that correspond to exact SVs. For our SVR and RFR predictions, exact SVs and SHAP values were calculated using SVETA or TreeExplainer with "interventional" feature perturbation,[54] respectively. For each compound in a randomly selected test set of an activity class, instance-based cumulative SVs and SHAP values were calculated as the sum of all individual contributions of the present or absent features for the visualization of feature contributions.

Performance Metric. Model performance was evaluated using the conventional mean absolute error (MAE) defined as

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (1)$$

In this equation, n represents the number of compounds, and y and $\hat{y}$ are the experimental and predicted potency values, respectively.

Statistical Significance Testing. The Wilcoxon signed-rank[55] test was used to evaluate the statistical significance of observed differences between MAE value distributions. The alpha value was set to 0.05 and compared to the p-value ($p < \alpha$).

RESULTS AND DISCUSSION

Study Concept. We aimed to analyze compound potency predictions over increasing potency difference intervals between structural analogues, including ACs. Therefore, from different activity classes, compound data sets were extracted that exclusively consisted of pairs of structural analogues (MMPs) with varying potency values. Stratified partitioning of MMP compounds principally ensured that training and test sets corresponded to structural analogues. These MMP-based data sets were then used for deriving and evaluating different ML models and controls. Initially, global potency predictions were carried out using these data sets to provide a reference point for the subsequent assessment of prediction accuracy for MMP subsets of corresponding training and test compounds with increasing potency differences. Therefore, potency difference intervals were consistently defined for MMP subsets, ranging from compound pairs having the same or very similar potency to ACs with the largest potency differences. To rationalize prediction characteristics of different regression models, XAI analysis was carried out, including systematic feature importance assessment for test sets and feature mapping for individual test set compounds. Figure 2 summarizes the analysis scheme.

Global Prediction Accuracy. For the MMP-based data sets derived from 10 large activity classes, systematic compound potency predictions were carried out using SVR and RFR models and 1-NN, k-NN, and MR controls. There was no need to include more complex (DNN) models in the comparison since the performance of such models was at best comparable to but mostly worse than SVR or RFR in previous potency prediction studies.[15,38] The results of our calculations are reported in Figure 3. The ML regression models and controls displayed very similar prediction accuracy across all activity classes. Overall, SVR achieved the best performance, closely followed by RFR, 1-NN, and k-NN, typically with low MAE values ranging from 0.4 to 0.6. Even MR yielded potency values for test compounds with MAE values of mostly 0.8 to 0.9. The results obtained for our MMP-based data sets were fully consistent with previous potency predictions using unmodified activity classes.[15] Figure 3 also shows that the predictions were generally stable, given the narrow MAE value distributions over independent trials. While differences between distributions obtained using different methods were in most cases statistically significant ($p < 0.05$), absolute differences between predicted values and their medians were only minute, typically within a 0.1 MAE. Thus, prediction accuracy was generally well within 1 order of magnitude (10-fold), reflecting consistently meaningful predictions, with only small activity-class-dependent variations. Moreover, the results were also very similar for stratified and random sampling of training and test compounds from MMPs, with stratified sampling yielding overall slightly lower MAE values, as further illustrated in Figure 4. Thus, while stratified sampling ensured that each individual MMP was split between the training and test sets, very similar results were obtained when MMP training and test compounds were randomly sampled, also reflecting the general stability of the predictions.

Compound Distribution. We next determined the distribution of test compounds over the MMP-based potency difference ranges. Therefore, the potency difference between compounds forming each MMP was determined; the observed potency differences were divided into equally sized bins, and all

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

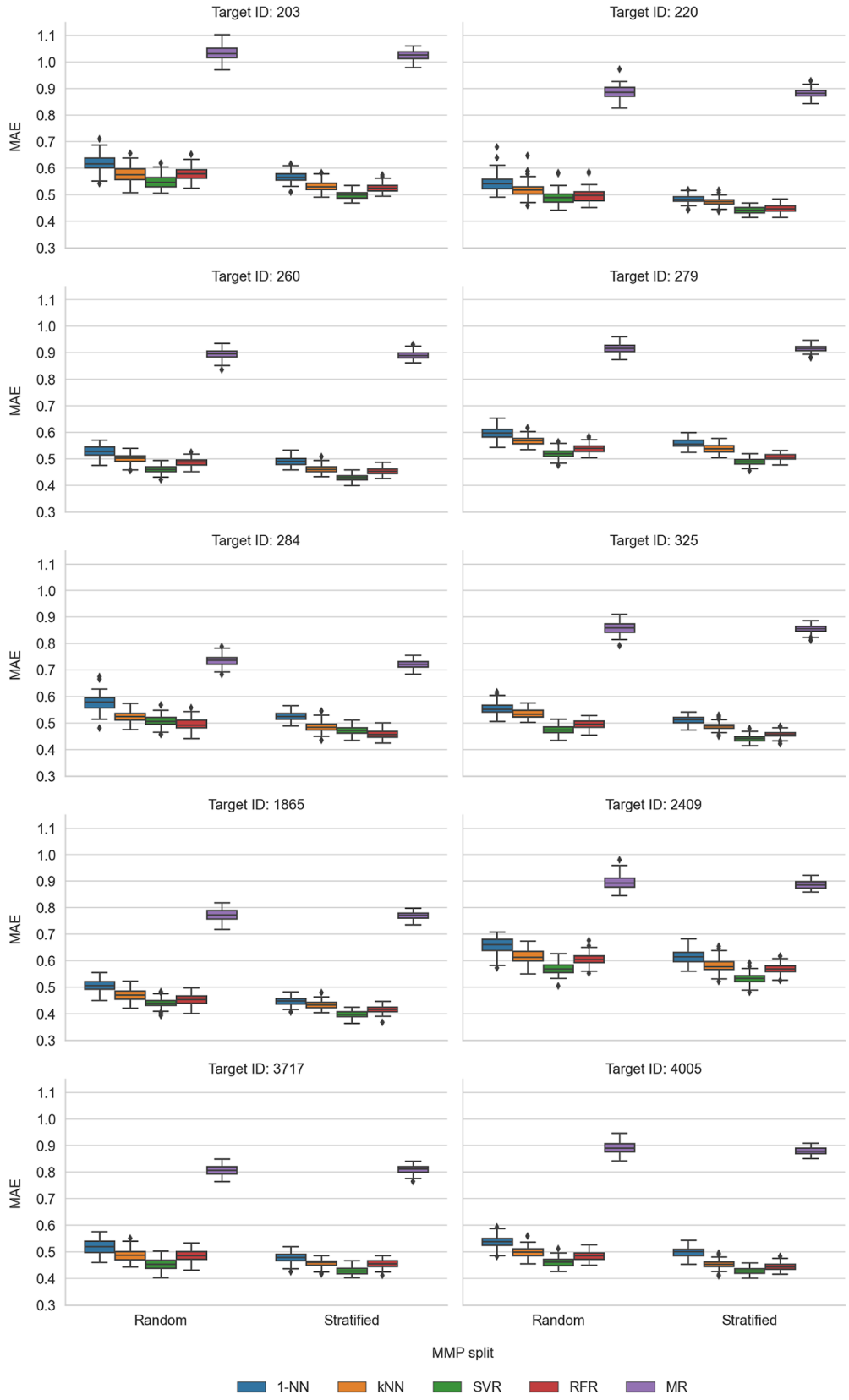


Figure 3. Global prediction accuracy. For each activity class (identified using its Target ID), boxplots report the distribution of MAE values over 10 independent trials for potency predictions based on random (left) or stratified (right) partitioning of MMP compounds. Five different methods are applied (color-coded according to the legend). In boxplots, the lower and upper whiskers indicate minimum and maximum values; the boundaries of the box represent the lower and upper quartiles; values classified as statistical outliers are shown as diamonds, and the median value is indicated by a horizontal line.

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

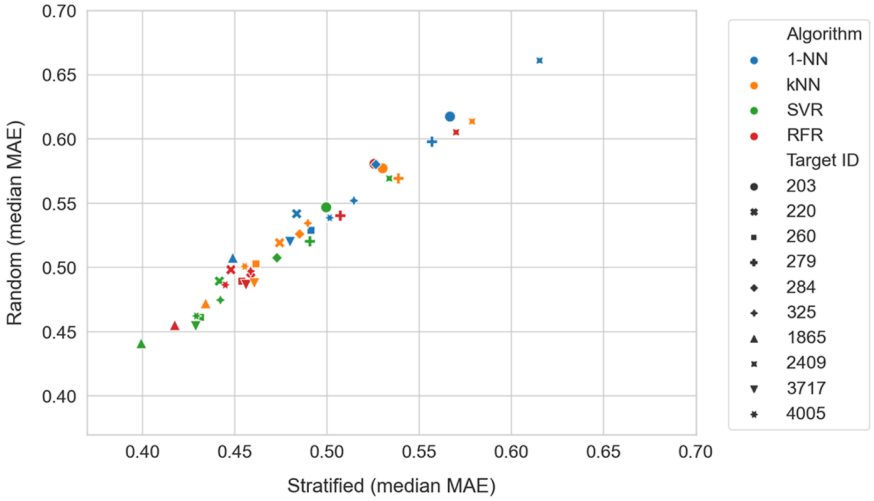


Figure 4. Stratified versus random sampling. The scatter plot compares the median MAE values after stratified or random partitioning of training and test compounds for the calculations reported in Figure 3. Activity classes are represented by different symbols that are color-coded by prediction methods according to the inset on the right.

MMPs were assigned to the corresponding bin. Then, each test compound was mapped to the potency difference bin containing the MMP from which it originated. The results are shown in Figure 5. For all activity classes, similar distributions of test compounds over potency difference intervals were observed. The majority of test compounds had MMP partners with closely similar potencies, falling within 0.5 orders of magnitude. The number of test compounds then rapidly declined over potency difference intervals at increasing potency. Following the MMP-cliff definition applied herein, the AC range began at a potency difference of 2 orders of magnitude (100-fold). Figure 5 shows that all activity classes contained only small numbers of AC test compounds, mostly on the order of 10−20. In a few instances, MMP-cliffs captured compound pairs with more than 1000- and up to 10,000-fold differences in potency. Overall, test compounds having MMP partners with small potency differences within 1 order of magnitude clearly dominated the distribution.

Prediction Accuracy Across Increasing Potency Difference Intervals. The potency-difference-based organization of MMPs enabled us to monitor prediction accuracy across increasing potency difference ranges, thereby generating a detailed view of these predictions as an alternative to the assessment of global potency prediction accuracy. For all activity classes, the prediction accuracy of test compounds falling into different MMP-based potency difference ranges is shown in Figure 6a,b. Closely corresponding trends were observed for all activity classes and stratified vs random compound partitioning. For test compounds having MMP partners with very similar potency, the predictions were accurate, regardless of the absolute potency values, with median MAEs consistently close to 0.5. For test compounds from MMPs capturing larger potency differences, the prediction accuracy gradually decreased. Due to the rapidly decreasing sample size over increasing potency difference intervals, the distributions capturing independent trials notably widened. For the confined numbers of AC compounds, large prediction errors of 2 to 3 or even 4 orders of magnitude were frequently observed, reflecting a general failure in reliably predicting the potency of AC compounds.

Taken together, the findings in Figure 6 clearly indicate that the predictions were largely determined by the presence of structural analogues of test compounds in training sets. Regardless of the methodology, values close to the potency of these training compounds were assigned to corresponding test compounds. For most of the test compounds, structural analogues had very similar potency (Figure 5), leading to accurate predictions that dominated global prediction accuracy (Figure 3), yielding very similar results for the different activity classes. This prediction phenotype also explained the success of *k*-NN potency value assignments compared to ML regression models.

Given the strong tendency of essentially all approaches to extrapolate from close structural analogue(s) in training sets and assign similar potency values to corresponding test compounds, prediction errors consistently increased across increasing potency difference intervals. In Figure 6, most differences between the MAE value distributions for the small samples of AC compounds were not statistically significant ($p < 0.05$), owing to the presence of generally large errors. Because AC compounds represented only a very small proportion of the test compounds in the different activity classes, their prediction errors did not notably affect global prediction accuracy (Figure 3), which was clearly dominated by structural analogues having similar potency.

Features Determining Predictions. We further analyzed the potency predictions by using SV/SHAP calculations for the SVR and RFR regression models. Figure 7 shows the results of test-compound-based cumulative SV (SVR) and SHAP value (RFR) analysis across all activity classes. For MMPs from which the test compounds originated, four different categories of representation (fingerprint) features were distinguished including features that were present in both MMP training and test compounds (red distributions in Figure 7), absent in both compounds (blue), present in only one of the MMP compounds (green), or absent in one compound (magenta). Hence, the two latter categories represent MMP compound-specific features. In Figure 7, red cumulative feature distributions capture highest positive SV/SHAP values compared to other distributions meaning that features present in both training and test compounds increased the expected

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

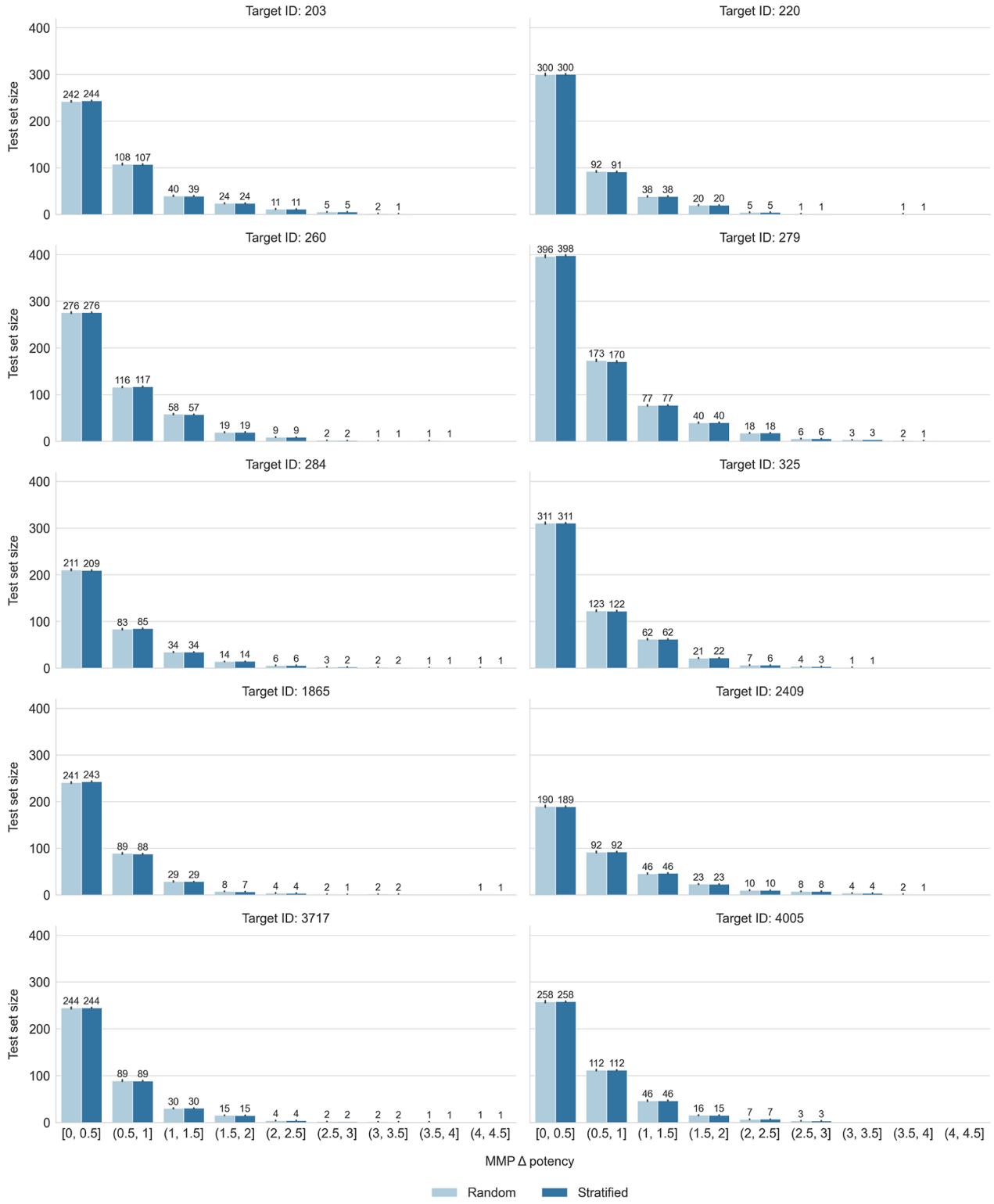


Figure 5. Distribution of test compounds over increasing MMP potency difference intervals. For each activity class, the barplot represents the mean distribution of test set compounds after random or stratified partitioning over increasing MMP potency difference intervals (x-axis: MMP (Δ) potency). Each MMP was divided into a training and test compound, and the test compound was assigned to the potency difference interval into which its MMP fell. For example, [0, 0.5] means that the potency difference between the MMP training and test compound was at most 0.5 orders of magnitude, and (4, 4.5) means that the potency difference was between 4 and 4.5 orders of magnitude. The AC range begins with the (2, 2.5] interval. The average number of test compounds per MMP-based potency interval is reported at the top of each bar.

value of the predictions. By contrast, blue cumulative feature distributions result from features that were absent in both training and test compounds. These distributions captured mostly negative SV/SHAP values and thus decreased the

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

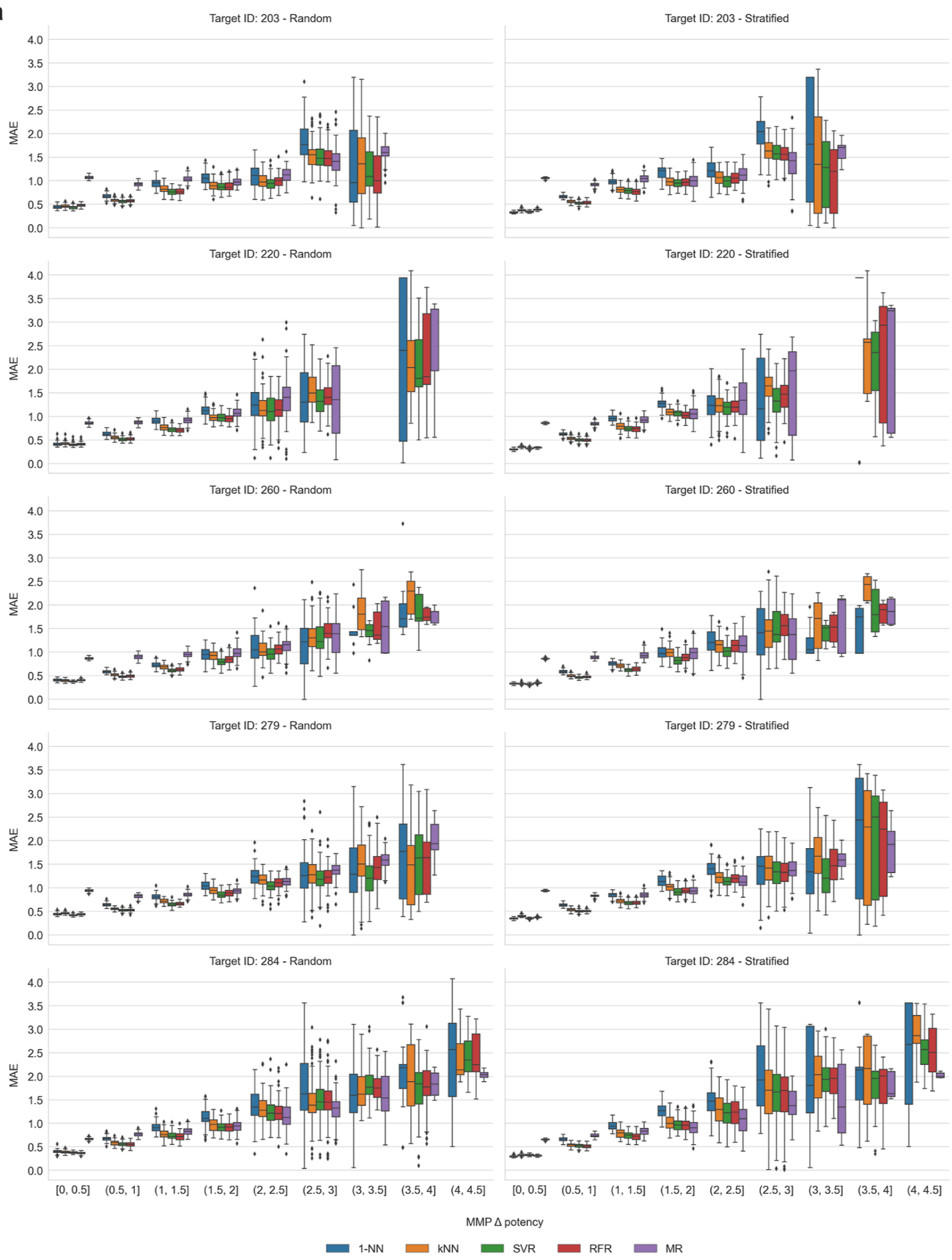


Figure 6. Prediction accuracy over increasing MMP potency difference intervals. Boxplots report the distribution of MAE values for test compounds falling into increasing MMP potency difference intervals according to Figure 5 after random (left) or stratified (right) partitioning. The representation corresponds to Figure 3. In panels (a) and (b), results for five activity classes are shown.

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

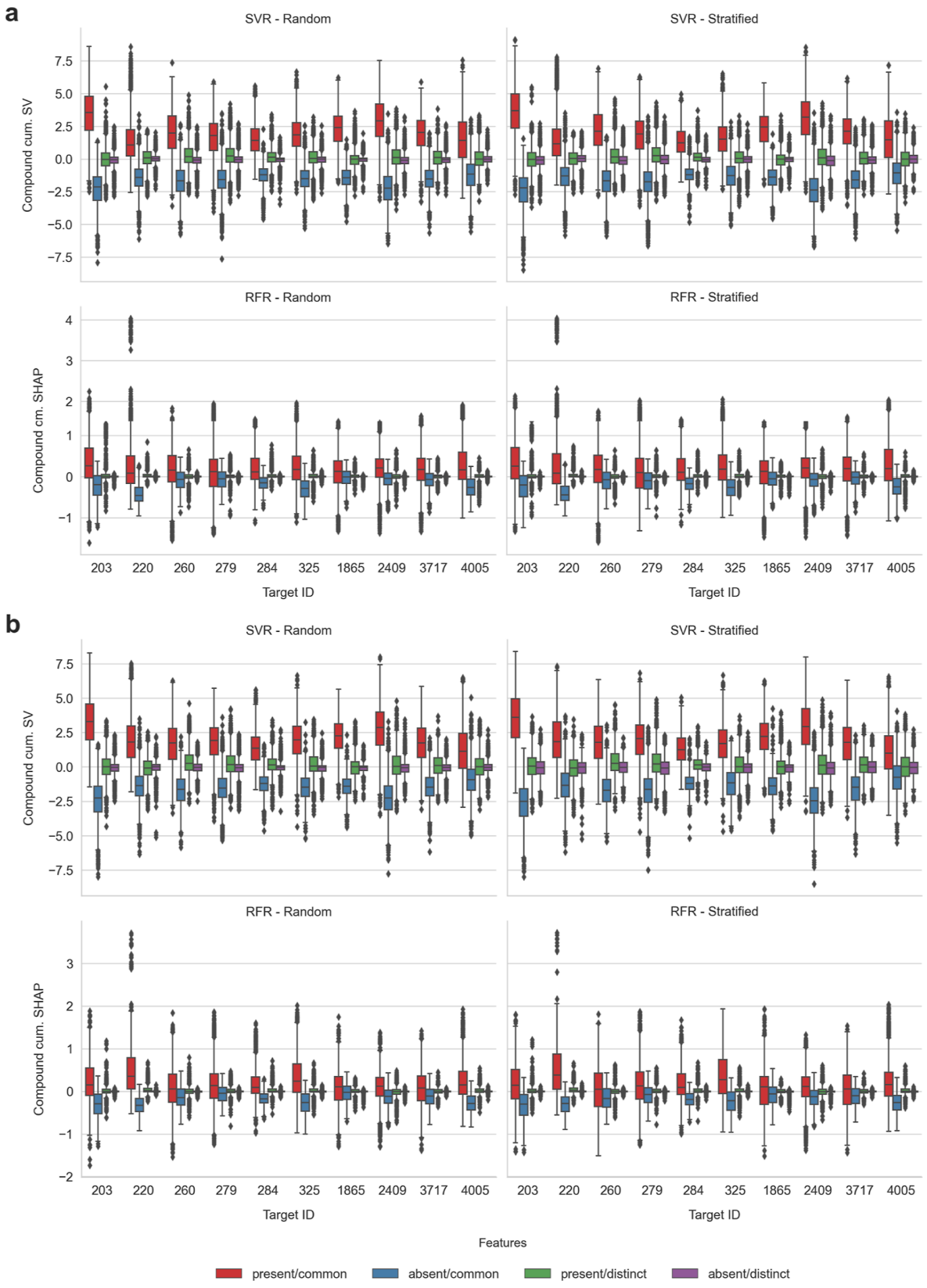


Figure 7. Distribution of Shapley feature contributions for test compounds. For each activity class, boxplots report the cumulative SV (for SVR) and SHAP value (for RFR) contributions for different feature subsets and (a) correctly (residual ≤0.5) and (b) incorrectly (residual >0.5) predicted test compounds over 10 independent trials. Boxplots are color-coded by feature subsets accounting for features that were present or absent in both compounds forming an MMP (present/common or absent/common) or only present or absent in one of the MMP partners (present/distinct or absent/distinct).

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

expected values of the predictions. Cumulative contributions of zero had no influence on the predictions. Rather unexpectedly, for MMP compounds, essentially the same feature contribution trends were consistently detected (with some variation in value magnitudes), regardless of the regression model, sampling strategy, or activity class (Figure 7). Hence, fingerprint features shared by MMP compounds made positive contributions to the potency value predictions of test compounds, whereas features absent in both compounds made negative contributions. By contrast, present or absent features that were unique to an MMP compound had only very little or no influence on the predictions (with cumulative contributions close to zero, except for outliers of the distributions).

Accordingly, predictions of test compounds were consistently driven by features shared with training compounds in the same way, that is, by balancing the positive contributions of features present in both compounds with the negative contributions of features absent in both compounds. Since the sum of the expected value and the positive and negative cumulative feature contributions corresponded to the predicted numerical value of a compound in the SV/SHAP explanation of regression models, these feature contributions resulted in similar potency values for the corresponding training and test compounds (as rigorously determined by stratified partitioning of MMP), regardless of their potency differences. For the majority of test compounds, this tendency led to accurate predictions. However, for AC compounds, this inevitably led to prediction errors of increasing magnitude. Thus, the SV/SHAP value analysis provided another intuitive explanation for the observed prediction characteristics over increasing potency difference intervals, as discussed above.

Feature Mapping. Feature contributions to the prediction of individual test instances can be visualized by mapping of structural features that are present in test compounds (but not absent) and color-coding their effects. Figure 8 compares feature contributions for test compounds that were correctly predicted using the SVR and RFR models (Figure 8a) or incorrectly predicted (Figure 8b). Features that increased or decreased the expected value of the predictions had positive or negative SV/SHAP values, respectively (and are colored red and blue, respectively). Their sum resulted in the cumulative contribution. Notably, feature mapping takes all present features into account, including those that are shared by MMP compounds and unique to test compounds. The comparison of feature contributions of individual test compounds for SVR and RFR models revealed some model-dependent differences, as one would expect, but also closely corresponding feature contributions, both for correctly and incorrectly predicted compounds. The comparison also showed that different model-dependent feature contributions could yield the same predicted potency values. It is also evident that many contributing features mapped to the core structures of test compounds that are shared in MMPs, rather than the substituents unique to test compounds. For accurate predictions of AC compounds, the distinguishing substituents would principally be required to make strong contributions, hence pointing at another possible origin of prediction errors associated with AC compounds.

CONCLUSIONS

In this study, we carried out an in-depth analysis of potency value predictions using ML models and controls that are affected by limitations originating from potency value

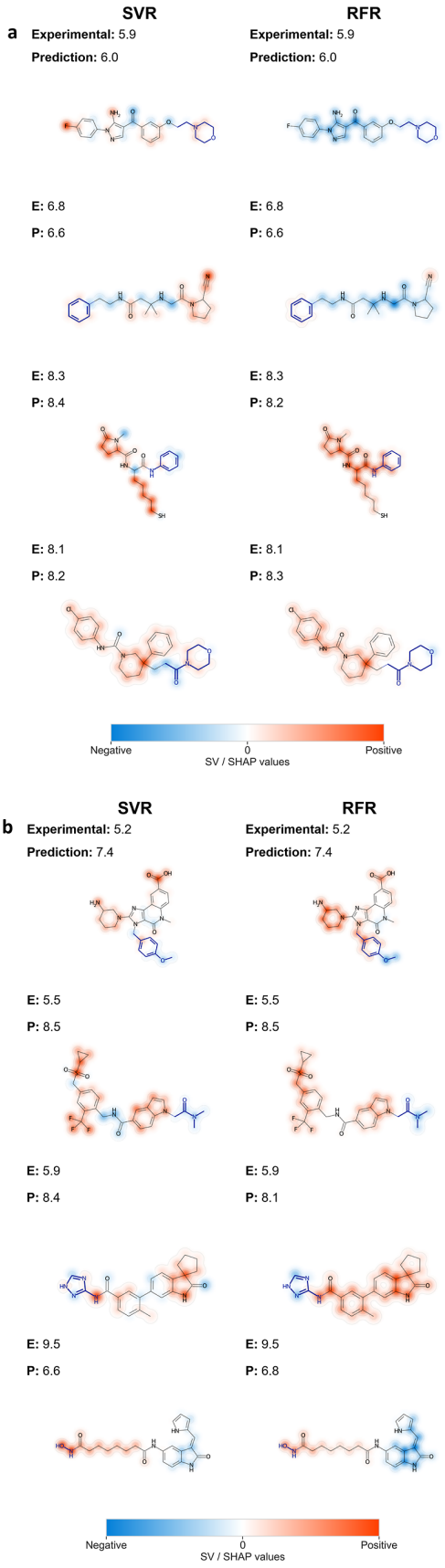


Figure 8. Atom-based feature mapping. SVs (SVR) and SHAP values (RFR) of features present in test compounds (a) correctly and (b) incorrectly predicted using SVR (left) and RFR (right) models were

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

Figure 8. continued

mapped on atoms forming each feature. The resulting cumulative atom-based feature contributions are color-coded using a continuous color spectrum ranging from blue (negative SV/SHAP contribution) over white to red (positive SV/SHAP contribution). Positive and negative contributions account for increases and decreases in potency values, respectively, relative to the expected value of SV/SHAP calculations. Experimental (E) and predicted (P) pIC_{50} potency values are reported, and substructures in test compounds that distinguish them from their MMP partners in training sets are shown in dark blue.

distributions in compound data sets and the presence of ACs. For the analysis, a new MMP-based compound test system was designed that made it possible to monitor the prediction accuracy across increasing MMP-dependent potency difference intervals. The analysis was complemented with SV/SHAP-based quantification of cumulative instance-based feature contributions to further rationalize the predictions. Potency predictions using the MMP-based data sets, regression models, and controls produced very similar results for different activity classes, consistent with previous observations made for unmodified compound data sets. A key finding of the MMP-based analysis in combination with stratified compound partitioning was that all methods displayed a strong tendency to predict similar potency values for corresponding training and test compounds representing close structural analogues. Given the statistical dominance of analogue pairs with the same or similar potency, this tendency consistently led to promising global prediction accuracy, camouflaging smaller numbers of problematic predictions. However, potency-interval-dependent evaluation of the MMP-based predictions clearly revealed that prediction errors steadily increased with increasing potency differences between corresponding training and test compounds, culminating in 100- to 1000-fold errors frequently observed for AC compounds. Moreover, quantitative assessment of cumulative feature contributions using the SV/SHAP formalism revealed that predictions of MMP compounds were overall consistently determined by positive contributions of features shared by MMP compounds and negative contributions of features absent in these compounds, regardless of their potency differences. These observations provided another rationale for the prediction of similar potency values for the corresponding compounds.

Taken together, the picture emerging from these findings is very clear. If test compounds have close structural analogues in training sets, then ML models tend to predict the training set value for the test instance. Increasing potency differences between structural analogues then leads to prediction errors of increasing magnitude, ultimately resulting in the inability of ML models to predict the potency of (statistically under-represented) AC compounds in any meaningful way. If compound data sets used for potency prediction exercises are rich in structural analogues having comparable potency, which is usually the case for activity classes originating from compound optimization efforts in medicinal chemistry, the assessment of global potency prediction accuracy using different models tends to produce similar results and overestimates the accuracy of the calculations. Hence, prediction accuracy should best be separately monitored for structurally similar training and test compounds with increasing potency differences, as well as for test compounds with no structural counterparts in training sets. Focusing predictions on test instances that are distinct from training compounds provides meaningful opportunities for follow-up investigations.

■ ASSOCIATED CONTENT

Data Availability Statement

All data and code used for the analysis are freely available via the following links: https://github.com/TiagoJanela/MMP-potency-prediction, https://zenodo.org/badge/latestdoi/695169027.

■ AUTHOR INFORMATION

Corresponding Author

Jürgen Bajorath − *Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53115 Bonn, Germany; Lamarr Institute for Machine Learning and Artificial Intelligence, Rheinische Friedrich-Wilhelms-Universität Bonn, D-53115 Bonn, Germany;* orcid.org/0000-0002-0557-5714; Email: bajorath@bit.uni-bonn.de

Author

Tiago Janela − *Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53115 Bonn, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c01530

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Lewis, R. A.; Wood, D. Modern 2D QSAR for Drug Discovery. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 505−522.
(2) Liu, J.; Wang, R. Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **2015**, *55*, 475−482.
(3) Guedes, I. A.; Pereira, F. S. S.; Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol* **2018**, *9*, 1089.
(4) Mobley, D. L.; Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* **2017**, *46*, 531−558.
(5) Williams-Noonan, B. J.; Yuriev, E.; Chalmers, D. K. Free Energy Methods in Drug Design: Prospects of "Alchemical Perturbation" in Medicinal Chemistry: Miniperspective. *J. Med. Chem.* **2018**, *61*, 638−649.
(6) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discovery* **2019**, *18*, 463−477.
(7) Sadybekov, A. V.; Katritch, V. Computational Approaches Streamlining Drug Discovery. *Nature* **2023**, *616*, 673−685.
(8) Walters, W. P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc. Chem. Res.* **2021**, *54*, 263−270.
(9) Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199−222.
(10) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(11) Hou, F.; Wu, Z.; Hu, Z.; Xiao, Z.; Wang, L.; Zhang, X.; Li, G. Comparison Study on the Prediction of Multiple Molecular Properties by Various Neural Networks. *J. Phys. Chem. A* **2018**, *122*, 9128−9134.

(12) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520−1530.
(13) Jiménez-Luna, J.; Pérez-Benito, L.; Martínez-Rosell, G.; Sciabola, S.; Torella, R.; Tresadern, G.; De Fabritiis, G. DeltaDelta Neural Networks for Lead Optimization of Small Molecule Potency. *Chem. Sci.* **2019**, *10*, 10911−10918.
(14) Rodríguez-Pérez, R.; Bajorath, J. Evaluation of Multi-Target Deep Neural Network Models for Compound Potency Prediction under Increasingly Challenging Test Conditions. *J. Comput. Aided Mol. Des.* **2021**, *35*, 285−295.
(15) Janela, T.; Bajorath, J. Simple Nearest-Neighbour Analysis Meets the Accuracy of Compound Potency Predictions Using Complex Machine Learning Models. *Nat. Mach. Intell.* **2022**, *4*, 1246−1255.
(16) Janela, T.; Bajorath, J. Large-Scale Predictions of Compound Potency with Original and Modified Activity Classes Reveal General Prediction Characteristics and Intrinsic Limitations of Conventional Benchmarking Calculations. *Pharmaceuticals* **2023**, *16*, 530.
(17) Chen, H.; Bajorath, J. Designing Highly Potent Compounds Using a Chemical Language Model. *Sci. Rep.* **2023**, *13*, 7412.
(18) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure-Activity Landscapes. *Drug Discovery Today* **2009**, *14*, 698−705.
(19) Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
(20) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18−28.
(21) Sheridan, R. P.; Karnachi, P.; Tudor, M.; Xu, Y.; Liaw, A.; Shah, F.; Cheng, A. C.; Joshi, E.; Glick, M.; Alvarez, J. Experimental Error, Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure-Activity Relationship Models? *J. Chem. Inf. Model.* **2020**, *60*, 1969−1982.
(22) Dablander, M.; Hanser, T.; Lambiotte, R.; Morris, G. M. Exploring QSAR Models for Activity-Cliff Prediction. *J. Cheminf.* **2023**, *15*, 47.
(23) Stumpfe, D.; Hu, H.; Bajorath, J. Evolving Concept of Activity Cliffs. *ACS Omega* **2019**, *4*, 14360−14368.
(24) Stumpfe, D.; Bajorath, J. Monitoring Global Growth of Activity Cliff Information over Time and Assessing Activity Cliff Frequencies and Distributions. *Future Med. Chem.* **2015**, *7*, 1565−1579.
(25) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.
(26) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138−1145.
(27) Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of Activity Cliffs Using Support Vector Machines. *J. Chem. Inf. Model.* **2012**, *52*, 2354−2365.
(28) Husby, J.; Bottegoni, G.; Kufareva, I.; Abagyan, R.; Cavalli, A. Structure-Based Predictions of Activity Cliffs. *J. Chem. Inf. Model.* **2015**, *55*, 1062−1076.
(29) Horvath, D.; Marcou, G.; Varnek, A.; Kayastha, S.; de la Vega de León, A.; Bajorath, J. Prediction of Activity Cliffs Using Condensed Graphs of Reaction Representations, Descriptor Recombination, Support Vector Machine Classification, and Support Vector Regression. *J. Chem. Inf. Model.* **2016**, *56*, 1631−1640.
(30) Tamura, S.; Miyao, T.; Funatsu, K. Ligand-Based Activity Cliff Prediction Models with Applicability Domain. *Mol. Inf.* **2020**, *39*, 2000103.
(31) Tamura, S.; Jasial, S.; Miyao, T.; Funatsu, K. Interpretation of Ligand-Based Activity Cliff Prediction Models Using the Matched Molecular Pair Kernel. *Molecules* **2021**, *26*, 4916.
(32) Iqbal, J.; Vogt, M.; Bajorath, J. Learning Functional Group Chemistry from Molecular Images Leads to Accurate Prediction of Activity Cliffs. *Artif. Intell. Life Sci.* **2021**, *1*, 100022.
(33) Iqbal, J.; Vogt, M.; Bajorath, J. Prediction of Activity Cliffs on the Basis of Images Using Convolutional Neural Networks. *J. Comput. Aided Mol. Des.* **2021**, *35*, 1157−1164.
(34) Park, J.; Sung, G.; Lee, S.; Kang, S.; Park, C. ACGCN: Graph Convolutional Networks for Activity Cliff Prediction between Matched Molecular Pairs. *J. Chem. Inf. Model.* **2022**, *62*, 2341−2351.
(35) Tamura, S.; Miyao, T.; Bajorath, J. Large-Scale Prediction of Activity Cliffs Using Machine and Deep Learning Methods of Increasing Complexity. *J. Cheminf.* **2023**, *15*, 4.
(36) Jiménez-Luna, J.; Skalic, M.; Weskamp, N. Benchmarking Molecular Feature Attribution Methods with Activity Cliffs. *J. Chem. Inf. Model.* **2022**, *62*, 274−283.
(37) Chen, H.; Vogt, M.; Bajorath, J. DeepAC - Conditional Transformer-Based Chemical Language Model for the Prediction of Activity Cliffs Formed by Bioactive Compounds. *Digital Discovery* **2022**, *1*, 898−909.
(38) van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **2022**, *62*, 5938−5951.
(39) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.
(40) Bruns, R. F.; Watson, I. A. Rules for Identifying Potentially Reactive or Promiscuous Compounds. *J. Med. Chem.* **2012**, *55* (22), 9763−9772.
(41) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53* (7), 2719−2740.
(42) Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58* (17), 7076−7087.
(43) Naveja, J. J.; Vogt, M.; Stumpfe, D.; Medina-Franco, J. L.; Bajorath, J. Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound-Core Relationship Method. *ACS Omega* **2019**, *4*, 1027−1032.
(44) Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure-Activity Relationship Trailing. *J. Med. Chem.* **2011**, *54*, 2944−2951.
(45) 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, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(46) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Network* **2005**, *18*, 1093−1110.
(47) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(48) *RDKit: Cheminformatics and Machine Learning Software*; GitHub, 2013, http://www.rdkit.org. (accessed 2023-07-01).
(49) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175−185.
(50) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(51) Shapley, L. S. 17. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Vol. II*; Kuhn, H. W., Tucker, A. W., Eds.; Princeton University Press, 1953; pp 307−318.
(52) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*; NIPS, 2017; Vol. *30*.
(53) Feldmann, C.; Bajorath, J. Calculation of Exact Shapley Values for Support Vector Machines with Tanimoto Kernel Enables Model Interpretation. *iScience* **2022**, *25*, 105023.
(54) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56−67.
(55) Conover, W. J. On Methods of Handling Ties in the Wilcoxon Signed-Rank Test. *J. Am. Stat. Assoc.* **1973**, *68* (344), 985−988.

https://doi.org/10.1021/acs.jcim.3c01530
J. Chem. Inf. Model. 2023, 63, 7032−7044

Appendix E

Predicting Potent Compounds Using a Conditional Variational Autoencoder Based upon a New Structure-Potency Fingerprint.

biomolecules

MDPI

Article

Predicting Potent Compounds Using a Conditional Variational Autoencoder Based upon a New Structure–Potency Fingerprint

Tiago Janela, Kosuke Takeuchi and Jürgen Bajorath *

Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 5/6, D-53115 Bonn, Germany

* Correspondence: bajorath@bit.uni-bonn.de; Tel.: +49-228-73-69100

Abstract: Prediction of the potency of bioactive compounds generally relies on linear or nonlinear quantitative structure–activity relationship (QSAR) models. Nonlinear models are generated using machine learning methods. We introduce a novel approach for potency prediction that depends on a newly designed molecular fingerprint (FP) representation. This structure–potency fingerprint (SPFP) combines different modules accounting for the structural features of active compounds and their potency values in a single bit string, hence unifying structure and potency representation. This encoding enables the derivation of a conditional variational autoencoder (CVAE) using SPFPs of training compounds and apply the model to predict the SPFP potency module of test compounds using only their structure module as input. The SPFP–CVAE approach correctly predicts the potency values of compounds belonging to different activity classes with an accuracy comparable to support vector regression (SVR), representing the state-of-the-art in the field. In addition, highly potent compounds are predicted with very similar accuracy as SVR and deep neural networks.

Keywords: bioactive compounds; potency prediction; fingerprints; machine learning; conditional variational autoencoder

Citation: Janela, T.; Takeuchi, K.; Bajorath, J. Predicting Potent Compounds Using a Conditional Variational Autoencoder Based upon a New Structure–Potency Fingerprint. *Biomolecules* **2023**, *13*, 393. https://doi.org/10.3390/biom13020393

Academic Editors: Umesh Desai, Daniel Afosah and Mire Zloh

Received: 14 December 2022
Revised: 7 February 2023
Accepted: 16 February 2023
Published: 18 February 2023

1. Introduction

Compound potency prediction is a major task in chemoinformatics and computational medicinal chemistry. For potency prediction, both structure- and ligand-based approaches are available. Structure-based methods attempt to predict small molecule (ligand) potency on the basis of experimental (or modeled) three-dimensional (3D) structures of ligand–target complexes. Ideally, such predictions aim to calculate the free energy of binding [1,2], for example, by applying alchemical free energy perturbation methods [2]. These calculations are challenging due to their high computational costs and the need to achieve consistent accuracy across different targets and compound classes [1]. Alternatively, scoring functions of different levels of sophistication are used to approximate ligand binding energies [3–6].

At the other end of the methodological spectrum reside classical ligand-based approaches for 2D and 3D quantitative structure–activity relationship (QSAR) modeling, which derive linear descriptor-based models for predicting potency values of congeneric compounds (structural analogues) [7,8]. Furthermore, for ligand-based modeling of non-linear SARs and potency prediction, random forest (RF) regression [9] and, in particular, support vector regression (SVR) have become preferred machine learning approaches [10,11]. While SVR typically produces statistically sound prediction models, it also displays a tendency to under-predict the individual most potent compounds because they are often algorithmically classified as outliers [12].

The increasing popularity of deep machine learning in pharmaceutical research [13–17] is also influencing structure- and ligand-based potency prediction. One of the attractions of deep learning is the ability to derive new object representations from input data such

as molecular graphs, thereby alleviating the need to use pre-conceived molecular descriptors for prediction tasks. Suitable deep neural network (DNN) architectures have been adapted for developing scoring functions [6,7] or deriving ligand–target binding energy models [18–21]. Despite their apparent success, such models are in part controversially viewed due to the observed strong dependence of their performance on varying training set composition [22,23], resulting from the memorization of training data leading to apparently accurate predictions that do not depend on correctly accounting for ligand–target interactions [22–24]. Similar observations have also been made for deep compound classification models with limited generalization ability [25]. In addition to studying ligand–target interactions, DNNs are also intensely investigated for ligand-based molecular property predictions including potency [26–29]. To these ends, various DNN architectures and learning strategies have been adapted. However, on data sets from medicinal chemistry, which are often limited in size, DNN-based property prediction models often do often exceed—or even meet—the performance of simpler models [29,30]. Hence, for both compound property and potency prediction, no firm conclusion can currently be drawn concerning the potential superiority of DNNs over standard approaches. We have recently shown that k-nearest neighbor (kNN) analysis meets the accuracy of other ML methods in potency prediction [31]. Moreover, randomized predictions typically reproduce experimental potency values within an order of magnitude, which is a direct consequence of potency value distributions in compound activity classes commonly used for benchmarking [31]. Hence, the best ML models and random predictions are only distinguished by a small margin of maximally one order of magnitude, representing a general limitation associated with the benchmarking of potency prediction methods. This needs to be taken into consideration when evaluating these methods, calling for the inclusion of simple controls such as kNN.

In this work, we introduce a novel concept for compound potency prediction that combines a special fingerprint (FP), termed structure–potency FP (SPFP), with a deep learning approach. FPs accounting for chemical structure and topology are a mainstay for chemical similarity searching [32,33]. SPFP is the first FP representation specifically designed to combine compound structure and potency information in a modular format. Using SPFP, a conditional variational autoencoder (CVAE) [34,35] is trained to predict potency from chemical structure using the structural module of test compounds as input. Given the uniform structure–potency bit string encoding, SPFP–CVAE models do not depend on class labels or associated variables for learning.

2. Materials and Methods

2.1. Compound Activity Classes

Bioactive compounds were extracted from ChEMBL (version 28) [36]. The compounds with a reported direct target interaction (target confidence score: 9) and a numerically specified potency (pIC50) value (standard relation: "=") were initially retrieved. Then, the compounds with a molecular weight less than 1000 Da and potency values falling into the pIC50 range from 5 to 11 were selected. All the compounds with interactions labeled as "inactive", "not active", "inconclusive", "potential transcription error", or "pan assay interference compounds" (PAINS) [37] were discarded. Furthermore, the PAINS filter from RDKit, a filter based on liability rules from medicinal chemistry [38], and the aggregation advisor [39] were applied to remove compounds with potential assay interference characteristics. On the basis of these criteria, 132,175 unique compounds were obtained with activity against 1315 human targets. The qualifying compounds were organized into target-based activity classes (pharmaceutical anti-targets were omitted). A set of 10 activity classes was randomly selected from the large pool, comprising 18,231 unique compounds (Table 1) and used for activity class-based model building, hyperparameter optimization, and model evaluation.

Table 1. Activity classes. Ten activity classes used for deriving and evaluating activity class-based prediction models are reported.

Target Name	Target ID	# Compounds
Beta-secretase 1	4822	2270
11-beta-hydroxysteroid dehydrogenase 1	4235	2232
Phosphodiesterase 10A	4409	2109
Acetyl-CoA carboxylase 2	4829	1811
Dipeptidyl peptidase IV	284	1709
Sodium channel protein type IX alpha subunit	4296	1703
Tyrosine-protein kinase SYK	2599	1616
Vascular endothelial growth factor receptor 2	279	1614
Epidermal growth factor receptor erbB1	203	1606
Vanilloid receptor	4794	1562

2.2. Model Building and Evaluation

For each activity class, training and test sets were randomly assembled to yield a constant 90:10 compound partition. Across all models, the predictive performance was evaluated over 10 independent trials using different performance measures. For 80:20 compound data partitions used as a control, nearly identical results were obtained.

2.2.1. Conditional Variational Autoencoder

CVAE [40] is an adaptation of the variational autoencoder (VAE) [41], a supervised deep learning algorithm for generative modeling that constructs a conditioned data representation into a continuous latent variable (z). The probabilistic encoder q(z | X, c) (recognition network) uses a condition vector (c) to map the input data to a Gaussian distribution, p(z | c)∼N(0, I) (prior network) into the latent space. The decoder p(X | z, c) then reconstructs data samples from the conditioned latent space to obtain the original input representation (dimensionality). The encoder and decoder are trained with the objective of optimizing the evidence lower bound (ELBO) of the input data [42,43]. During training, the conditioned encoder learns to approximate a latent variable distribution by minimizing the Kullback–Leibler (KL) divergence [44] between data distributions in the original and latent space. The decoder is trained to minimize the reconstruction error of the data representation.

The CVAE encoder and decoder networks consisted of three hidden layers, with 512, 256, and 128 neurons, respectively. For hyper-parameter optimization, a grid search protocol was applied to determine the number of neurons for the latent layer (16, 32, and 64). Different learning rates (0.1, 0.01, and 0.001), dropout rates (0 and 0.5), and batch sizes (16, 32, and 64) were evaluated. Network training was performed with Adam [45] optimizer and the hyperbolic tangent (tanh) was used as the activation function. The parameters β (1 and 2) and σ (0.01, 0.1, and 1) were tested. The learning rate was steadily reduced, during training, to improve learning, and the models were run for a maximum of 150 epochs or until convergence was reached with the early stopping option to avoid network overfitting. The CVAE cost function was computed as the mean of the reconstruction (binary cross-entropy) loss and KL divergence loss.

2.2.2. Support Vector Regression

The support vector regression (SVR) is a variant of the supervised support vector machine algorithm that derives an ε-insensitive tube based on the training data for the prediction of numerical values, with the maximum permitted error provided by the width of the ε tube [10,11].

For SVR, the cost hyper-parameter C was optimized by testing (0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 10, 100, and 10000) values. The SVR models were built using the Tanimoto kernel [46].

2.2.3. Random Forest Regression

Random forest regression (RFR) is a machine-learning algorithm based on an ensemble of decision trees. During model training, each tree is created by splitting the respective node and bootstrapping aggregation is used to randomly select the training instances. The mean value across all decision trees is used to determine the final prediction [47].

In RFR parameter optimization, the number of decision trees (50, 100, and 200), the minimal number of samples for a split (2, 3, 5, and 10), and the minimum number of leaf-node samples (1, 2, 5, and 10) were used as the search parameter space.

2.2.4. Deep Neural Network

DNN is a deep learning method capable of mathematically modeling data using a non-linear activation function through the neurons of the network's fully connected layers. The network learning process consists of interactively determining the difference between the observed and predicted values, using a stochastic gradient descent algorithm to minimize the loss function until it converges to a specific minimum value [48,49].

The DNN models were trained using several network architectures by varying the different numbers of hidden layers (2 and 3) with hyperbolic tangent (tanh) activation, and the network neurons (100–500). Grid searches were performed for different batch sizes (16 and 64), dropout (0 and 0.5), and learning rates (0.1, 0.01, and 0.001). The networks were trained using an Adam optimizer for a maximum of 200 epochs with early termination.

2.2.5. k-Nearest Neighbor Ranking

kNN is a supervised learning method that ranks the training compounds based on increasing the fingerprint similarity (shortest distance). For the final prediction, the k-top training compounds potency value is accessed (e.g., 1-NN—potency value, and 3-NN—average potency) and assigned to the test compound [50]. For kNN optimization, the optimal k values were evaluated with 1, 3, and 5 top-rated compounds.

2.2.6. Mean Regression

The mean regressor (MR) approach is based on assigning the mean potency value of the training set to each compound present in the test set. This method was used as a control calculation to generate the random predictions.

2.2.7. Random Predictions

A y-randomization control was performed by the random reassignment of potency values across the compounds from each activity class (random shuffling) [51].

2.2.8. Hyperparamters and Implementation

The kNN, SVR, RF, and SPFP–CVAE model hyperparameters were optimized using an internal five-fold cross-validation, whereas the DNN parameter optimization was performed with an internal 90:10 training–validation split. The SVR, RFR, kNN, and MR models were generated using scikit-learn [52]. The CVAE and DNN models were implemented with Keras [53] and Tensorflow [54].

2.2.9. Evaluation Metrics

The performances of all the models were evaluated by calculating the mean absolute error (MAE) and root mean squared error (RMSE) for predicted and experimental test compound potency values using scikit-learn.

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

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

where n is the number of compounds, and y and ŷ are the experimental and predicted potency values, respectively.

An assessment of the statistical significance was performed for the value distributions from predictions based on MAE and RMSE values using the nonparametric Wilcoxon test [55]. The null hypothesis was either rejected or accepted, by setting alpha to 0.05 and comparing it to the respective *p*-value ($p \leq 0.05$).

2.3. Molecular Representation

The compounds were represented using a folded version of the extended connectivity fingerprint with bond diameter 4 (ECFP4) [56], which is a generally preferred topological descriptor for many chemoinformatics applications, consisting of layered atom environments, consisting of 2048 bits. The ECFP4 fingerprint was generated using RDKit [57].

The scripts for the reported calculations and the curated activity classes are available from the authors upon request.

3. Results and Discussion

3.1. Concept of Potency Prediction Based on Fingerprint-Based Potency Encoding

The introduction of a structure–potency fingerprint (SPFP) provided the basis for a new approach in potency prediction. The underlying idea was to unify structural and potency encodings in a modular fingerprint representation of a constant format such that the potency module representing a numerical value could be predicted from the structural module of test compounds using deep learning. This unified and intuitive modular encoding of compound structure and potency enabled the derivation of a chemical language model such as a CVAE using SPFPs of training compounds to predict the potency module of test compounds using only their structure module as input.

An extended connectivity fingerprint with a constant size of 2048 bits represented the structure module of SPFP that was combined with a newly designed potency module for representing compound potency values. We defined two principal requirements for the potency module. Hence, it was required to, first, represent the biologically relevant large (negative decadic logarithmic) potency range from 5 to 11 and, second, encode the potency values at a meaningful resolution such that accurate predictions could in principle be obtained. Therefore, alternative single value, value range, and cumulative coding schemes suitable for bit string representations were initially investigated and cumulative value range encoding was found to be the most robust approach (that is, yielding the most stable predictions across independent trials). Accordingly, contiguous segments of increasing numbers of bits were used to represent increasingly potent compounds populating the entire logarithmic potency range from 5 to 11. For example, Figure 1a,b show how a potency value of 5.2 and 8.0 was encoded by setting on the first four and 51 bits in the potency module, respectively. To meet the second requirement stated above, we set the size of the potency module to a minimum of 100-bit positions such that each individual bit position accounted for 0.06 log units via cumulative potency encoding. Accordingly, the resolution of the potency predictions was intrinsically limited to 6% of a log unit. This level was deemed acceptable for the approach because it fell within the typical range of experimental accuracy limitations. Smaller bit numbers for the potency module would lead to larger resolution limits while larger numbers would further increase the resolution. Therefore, we also tested larger versions of the potency module using the SPFP–CVAE models comprising 500 and 1000 bits, as reported in Figure 2. These control calculations produced very similar results to those obtained for the 100-bit potency module, hence showing that the prediction accuracy could not be further increased by decreasing the resolution limit of the potency encoding and supporting the choice of 100-bit positions for the potency module. Furthermore, for potency predictions using CVAE sampling, a bit

module with a constant format and meaningful size was required to assess the predictions in a meaningful way (see below).

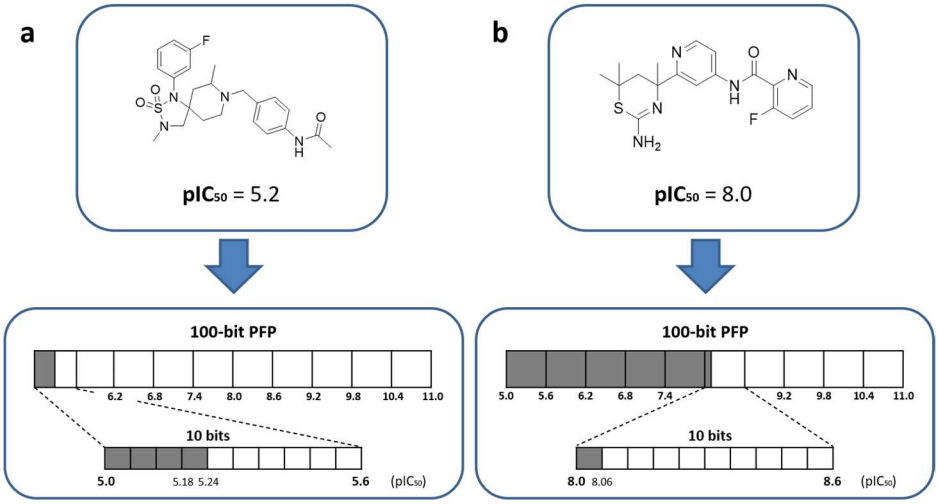


Figure 1. Cumulative potency encoding. (**a**,**b**) illustrate how logarithmic potency values of different magnitude are encoded in the potency module of SPFP.

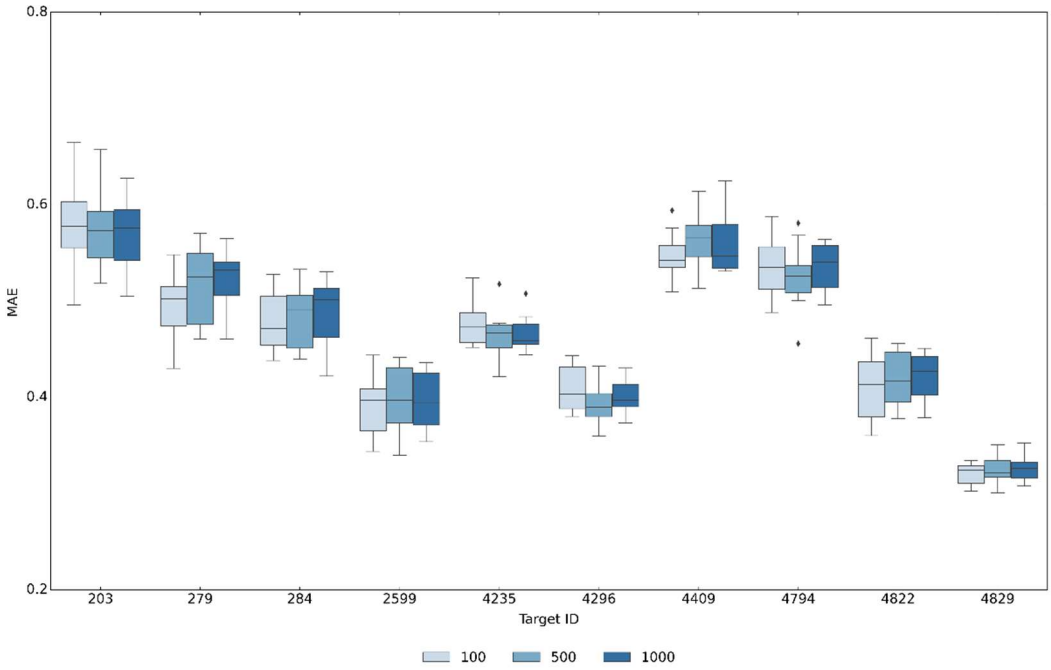


Figure 2. Prediction accuracy for SPSF with differently sized potency modules. Boxplots report mean absolute error (MAE) values of SPFP–CVAE models using alternative SPFP versions with potency modules comprising 100, 500, or 1000 bits evaluated across all activity classes according to Table 1. In boxplots, the upper and lower whiskers indicate maximum and minimum values, the boundaries of the box represent the upper and lower quartiles, values classified as statistical outliers are shown as diamonds, and the median value is indicated by a horizontal line.

3.2. Learning and Prediction Strategy

The CVAE model architecture used here consists of an encoder, latent space layer, and decoder, as illustrated in Figure 3.

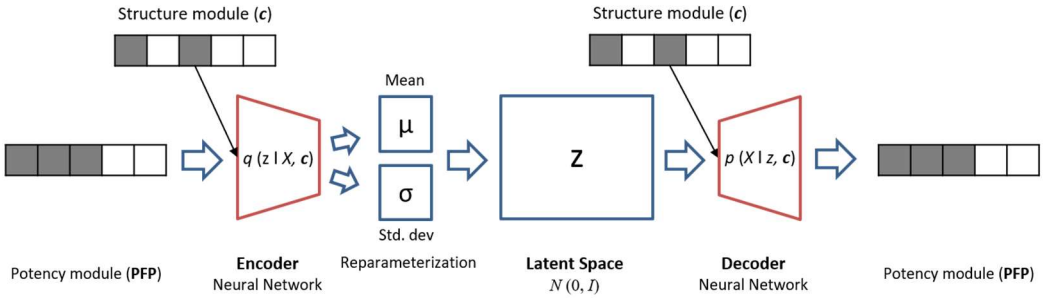


Figure 3. Architecture of the conditional variational autoencoder. The encoder network transforms the potency module (PFP), conditioned by the structure module (c), into a distribution of latent variables (z). The decoder samples a conditioned latent vector from a Gaussian distribution and reconstructs the potency module.

For each compound activity class, a CVAE model was trained to reproduce the complete bit patterns of the potency module, conditioned by the structure module, as illustrated in Figure 4a. Each CVAE model was then used to predict the bit settings of the potency module (PFP). Therefore, the potency values of the test compound were predicted by submitting the structure module (c) to the CVAE decoder to generate the corresponding potency module, as illustrated in Figure 4b. Since the CVAE predictions depended on the sampling of potency modules in latent space, the evaluation criteria for potency module variants were defined. Accordingly, for a given test compound, a sampled potency module was classified as valid if it contained a contiguous bit string in which all bits were set on. If this criterion was met, the predicted potency value was assigned to the center of the respective potency interval (e.g., 5.03 for the [5.0–5.06] interval), resulting in a constant standard deviation of ±0.03 log units for all predictions. By contrast, if the output bits were not contiguous, that is, if they were not consistent with the cumulative encoding of the potency module, the prediction was classified as invalid and the sampling was continued until a valid prediction was obtained, given a maximal number of permitted sampling steps.

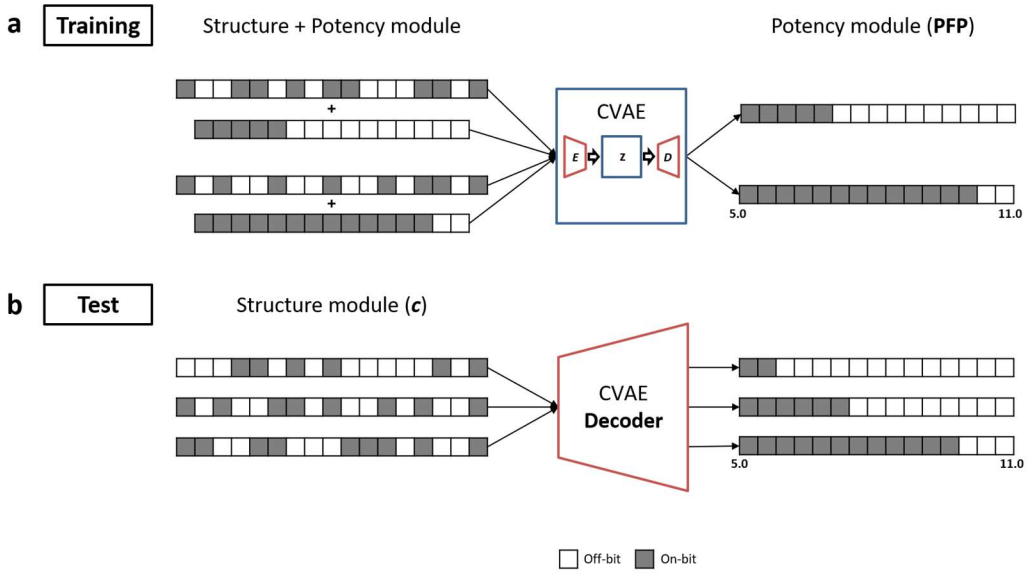


Figure 4. Conditional variational autoencoder modeling. In (**a**,**b**), the CVAE training and prediction strategies are illustrated, as discussed in the text.

3.3. Potency Predictions

For 10 randomly selected compound activity classes, different ML models were generated. Figure 5 shows that the compound potency value distributions of the activity classes were overlapping yet distinct, mostly yielding median potency values in the high

nanomolar range. The comparison also shows that logarithmic potency values below 5 (approaching experimental accuracy limitations) and above 10 (sub-nanomolar potency) were generally sparse.

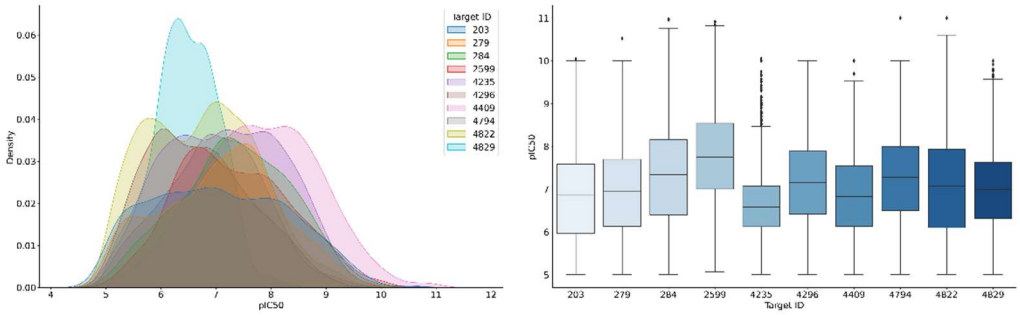


Figure 5. Potency value distribution of activity classes. Kernel density estimation plots (**left**) and boxplots (**right**) compare the potency values distributions of the 10 activity classes. Coloring of boxplots is arbitrary. The horizontal line indicates the median of the value distribution and diamond symbols represent statistical outliers.

Activity class-dependent potency prediction models were then generated for SPFP–CVAE, k-nearest neighbor (kNN) analysis, SVR, RFR, and DNN. These ML approaches currently represent the state of the art in compound potency prediction [31]. In addition, a mean regressor (MR) was applied as a control, which simply assigned the mean potency value of an activity class to all test compounds. The results are reported in Figure 6.

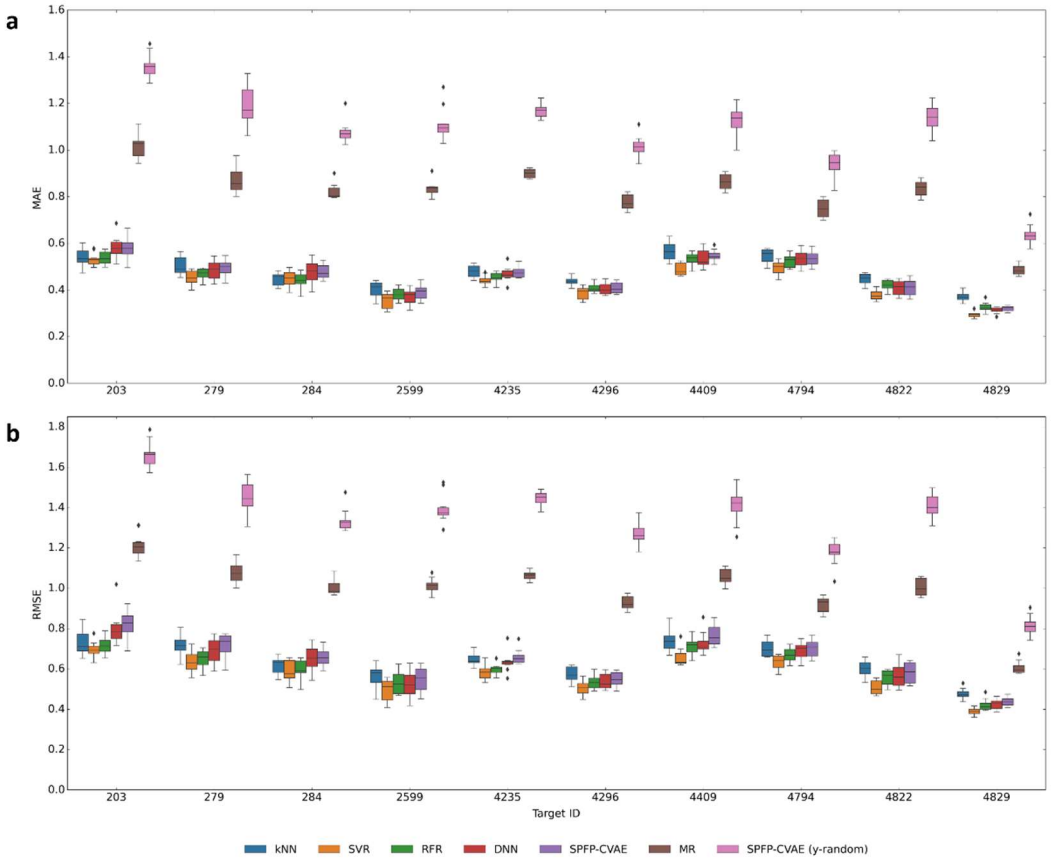


Figure 6. Prediction accuracy. Boxplots report the (**a**) MAE and (**b**) RMSE values for potency predictions using different ML models across all activity classes.

Overall, similar prediction accuracy was observed for the different ML models, regardless of their complexity, mostly with median MAE and RMSE of ~0.4–0.5 and ~0.6–0.7, respectively. As observed before [31], simple kNN-based potency assignments approached or exceeded the prediction accuracy of ML models and there was no advantage of deep learning approaches over other ML methods. Moreover, the MR yielded median MAE and RMSE values of ~0.8–0.9 and ~1.0–1.1, respectively. The performance of randomized SPFP–CVAE models was only slightly worse than MR, mostly with a median MAE value of ~1.0–1.2, owing to the dominance of compounds with potency values between 6 and 8 across all activity classes, as reported in Figure 5. These artificial predictions using MR or randomized models reproduced experimental values within about one order of magnitude, providing a limit for prediction accuracy, while most accurate ML models typically achieved mean MAE value of ~0.4. Hence, there was only a relatively small margin between best and artificial predictions, defining a window of less than one order of magnitude in which model performance must be evaluated [31]. In the previous study, equally curated versions of three activity classes (279, 284, 4822) from a different ChEMBL release were investigated using ML methods with different calculation protocols, yielding prediction accuracies very similar to the values reported herein [31].

Many of the small performance differences observed in Figure 6 were not statistically significant, as reported in Figure 7, while differences between SPFP–CVAE, SVR, and RFR were statistically significant for about half of the activity classes. However, the prediction accuracy of these three approaches was very similar, which was also reflected by the respective *p*-values. Overall, SVR was the preferred approach, but only by a very small margin compared to SPFP–CVAE and other ML methods. For example, the differences in the median MAE between SPFP–CVAE and SVR ranged max. ~0.01–0.02, depending on the activity class, which was marginal at most and would be considered irrelevant for all practical purposes.

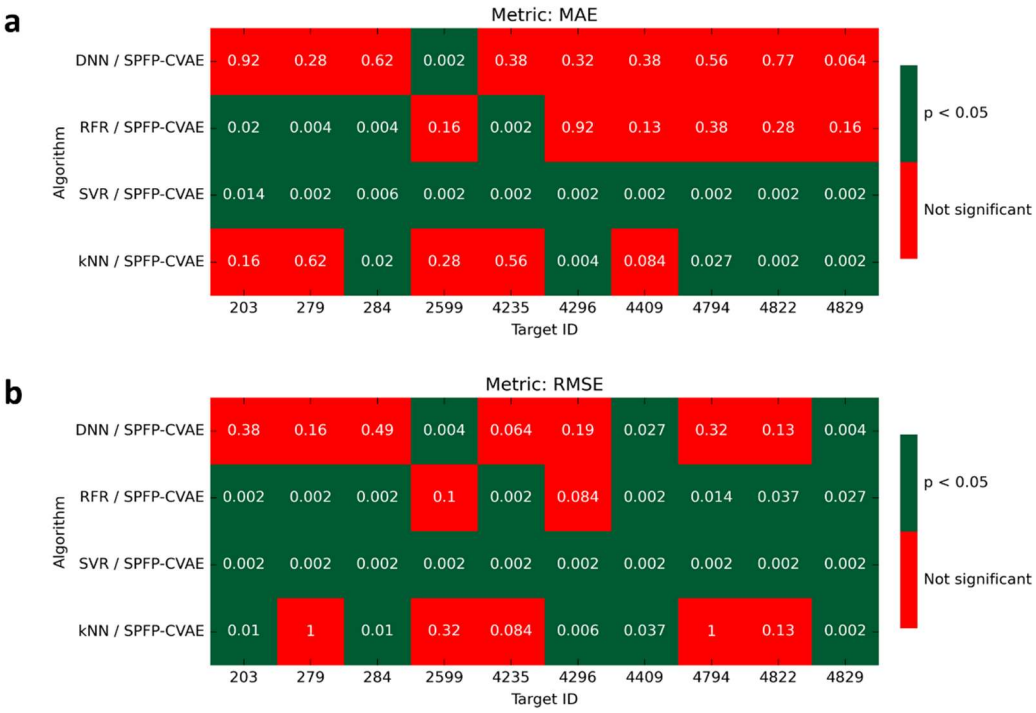


Figure 7. Statistical significance assessment. Statistical significant (Wilcoxon signed-rank) tests based on (**a**) MAE and (**b**) RMSE values were carried out for the performance differences observed between SPFP–CVAE and all other ML models (kNN, SVR, RFR, and DNN). Red cells indicate *p*-values above $\alpha = 0.05$ (no statistical significance) and green cells *p*-values below $\alpha = 0.05$ (statistical significance).

3.4. Predicting Highly Potent Compounds

We then investigated the ability of the different ML methods to predict the 10% most potent compounds in a test set (typically amounting to ~15–20 compounds) using models derived based on the original sets. The results for models derived from original training sets are shown in Figure 8. Due to the small test sample size of these predictions, the MAE and RMSE value distributions were broader than for the global predictions reported in Figure 6.

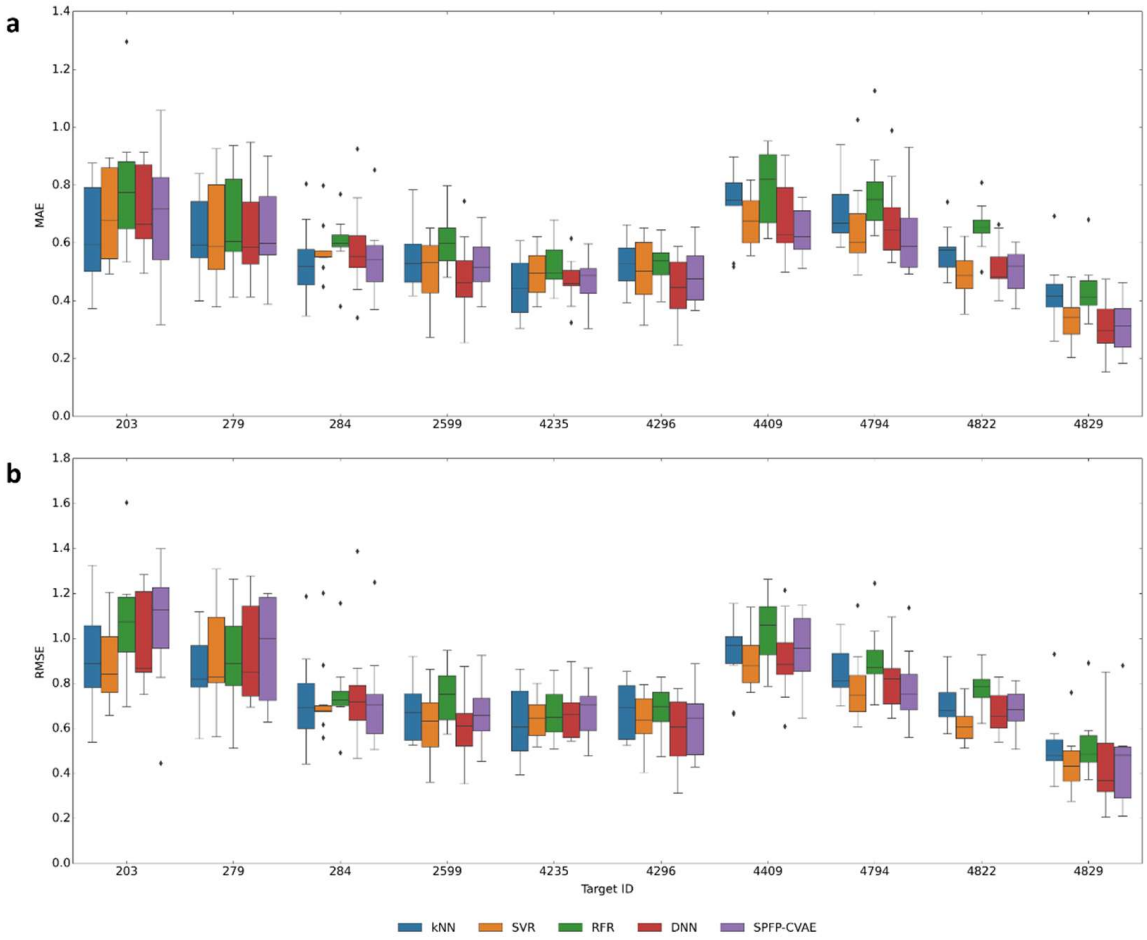


Figure 8. Prediction accuracy for the most potent test compounds. Boxplots report the median MAE (**a**) and RMSE (**b**) values for the 10% most potent test compounds from all classes and different ML models including kNN.

Compared to the global predictions, the median MAE and RMSE value for most potent compounds increased to ~0.6 and ~0.8 or greater, respectively, for about half of the activity classes while the prediction accuracy remained similar to before for the remaining classes. However, the performance of the different ML methods including kNN continued to be comparable (MR was omitted here because of the naturally large deviations for the small number of the most potent compounds). Overall, SVR, SPFP–CVAE, and DNN yielded best predictions with only small (and activity class-dependent) differences between these methods. The predictions for the exemplary compounds are shown in Figure 9.

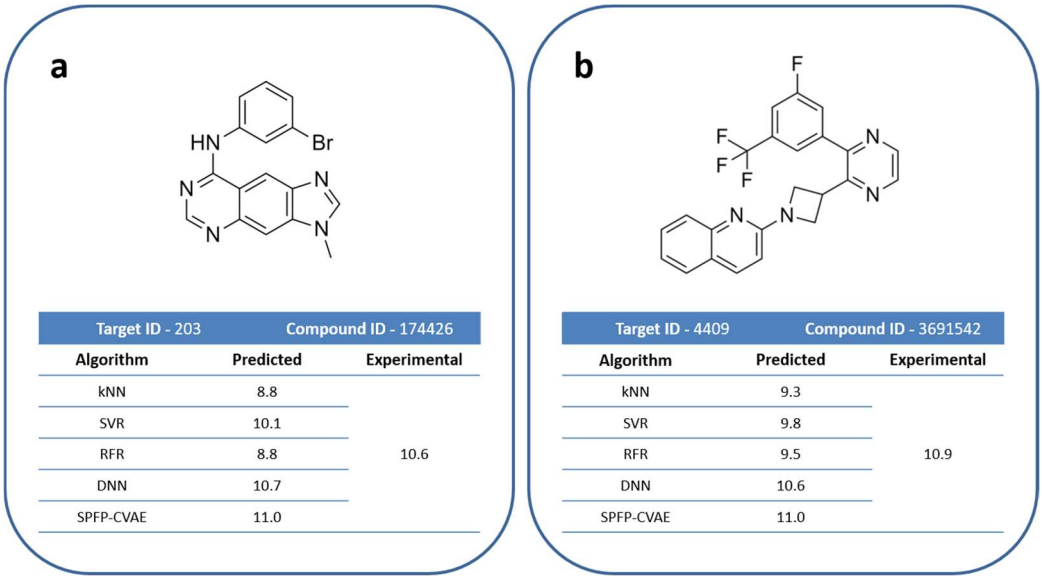


Figure 9. Highly potent compounds. (**a**,**b**) illustrate predictions for two exemplary highly potent test compounds using different methods.

4. Conclusions

Compound potency prediction is an important task in chemoinformatics and medicinal chemistry. For structure- and ligand-based predictions, different methods have been introduced. QSAR techniques including non-linear modeling using machine learning continue to play an important role. Herein, we have introduced a new methodological concept for compound potency prediction that depends on the newly designed SPFP format for structure–potency encoding and CVAE learning. The SPFP–CVAE concept was devised to enable the prediction of bit settings in SPFP potency modules from input structure modules, without learning correlations between structural representations and potency values used as a dependent variable. In activity class-dependent predictions, the SPFP–CVAE approach essentially met SVR performance, representing the current state of the art in the field. Given the general limitations associated with the potency predictions in benchmark settings, we consider the prediction of most potent compounds a particularly meaningful exercise. In this case, SVR, SPFP–CVAE, and DNN achieved comparable accuracy. Taken together, our results indicate that the SPFP–CVAE concept introduced herein provides a new methodological framework for compound potency prediction that can be further explored in various ways. Importantly, FP-based structure–potency encoding, as introduced herein, can be easily modified for different applications, providing a versatile input format for ML.

Author Contributions: T.J. and K.T.: Methodology, Resources, Investigation, and Formal analysis, Writing—review and editing; J.B.: Conceptualization, Methodology, Formal analysis, Writing—original draft, and Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Compound data sets are publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mobley, D.L.; Gilson, M.K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* **2017**, *46*, 531–558. [CrossRef]
2. Williams-Noonan, B.J.; Yuriev, E.; Chalmers, D.K. Free Energy Methods in Drug Design: Prospects of "Alchemical Perturbation" In Medicinal Chemistry. *J. Med. Chem.* **2018**, *61*, 61638–61649. [CrossRef]
3. Liu, J.; Wang, R. Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **2015**, *55*, 475–482. [CrossRef]
4. Gleeson, M.P.; Gleeson, D. QM/MM Calculations in Drug Discovery: A Useful Method for Studying Binding Phenomena? *J. Chem. Inf. Model.* **2009**, *49*, 670–677. [CrossRef]
5. Guedes, I.A.; Pereira, F.S.S.; Dardenne, L.E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.* **2018**, *9*, e1089. [CrossRef]
6. Li, H.; Sze, K.H.; Lu, G.; Ballester, P.J. Machine-Learning Scoring Functions for Structure-Based Virtual Screening. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1478. [CrossRef]
7. Lewis, R.A.; Wood, D. Modern 2D QSAR for Drug Discovery. *WIREs Comput. Mol. Sci.* **2014**, *4*, 505–522. [CrossRef]
8. Akamatsu, M. Current State and Perspectives of 3D-QSAR. *Curr. Top. Med. Chem.* **2002**, *2*, 1381–1394. [CrossRef]
9. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958. [CrossRef]
10. Drucker, H.; Burges, C. Support Vector Regression Machines. *Adv. Neural. Inform. Proc. Syst.* **1997**, *9*, 155–161.
11. Smola, A.J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
12. Balfer, J.; Bajorath, J. Systematic Artifacts in Support Vector Regression-Based Compound Potency Prediction Revealed by Statistical and Activity Landscape Analysis. *PloS ONE* **2015**, *10*, e0119301. [CrossRef]
13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
14. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477. [CrossRef]
15. Lavecchia, A. Deep Learning in Drug Discovery: Opportunities, Challenges and Future Prospects. *Drug Discov. Today* **2019**, *24*, 2017–2032. [CrossRef]
16. Bajorath, J. Deep Machine Learning for Computer-Aided Drug Design. *Front. Drug Discov.* **2022**, *2*, e829043. [CrossRef]
17. Kim, J.; Park, S.; Min, D.; Kim, W.Y. Comprehensive Survey of Recent Drug Discovery Using Deep Learning. *Int. J. Mol. Sci.* **2019**, *22*, e9983. [CrossRef]
18. Jimenez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287–296. [CrossRef]
19. Torng, W.; Altman, R.B. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *J. Chem. Inf. Model.* **2019**, *59*, 4131–4149. [CrossRef]
20. Kwon, Y.; Shin, W.H.; Ko, J.; Lee, J. AK-Score: Accurate Protein-Ligand Binding Affinity Prediction Using an Ensemble of 3D-Convolutional Neural Networks. *Int. J. Mol. Sci.* **2020**, *21*, e8424. [CrossRef]
21. Son, J.; Kim, D. Development of a Graph Convolutional Neural Network Model for Efficient Prediction of Protein-Ligand Binding Affinities. *PLoS ONE* **2021**, *16*, e0249404. [CrossRef]
22. Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C.J.; Duca, J.S.; Hornak, V.; Koes, D.R.; Kurtzman, T. Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. *PLoS ONE* **2019**, *14*, e0220113. [CrossRef]
23. Yang, J.; Shen, C.; Huang, N. Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets. *Front. Pharmacol.* **2020**, *11*, e69. [CrossRef]
24. Volkov, M.; Turk, J.A.; Drizard, N.; Martin, N.; Hoffmann, B.; Gaston-Mathé, Y.; Rognan, D. On the Frustration to Predict Binding Affinities from Protein-Ligand Structures with Deep Neural Networks. *J. Med. Chem.* **2022**, *65*, 7946–7958. [CrossRef]
25. Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization rather than Generalization. *J. Chem. Inf. Model.* **2021**, *58*, 916–932. [CrossRef]
26. Hou, F.; Wu, Z.; Hu, Z.; Xiao, Z.; Wang, L.; Zhang, X.; Li, G. Comparison Study on the Prediction of Multiple Molecular Properties by Various Neural Networks. *J. Phys. Chem. A* **2018**, *122*, 9128–9134. [CrossRef]
27. Feinberg, E.N.; Sur, D.; Wu, Z.; Husic, B.E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V.S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530. [CrossRef]
28. Shen, J.; Nicolaou, C.A. Molecular Property Prediction: Recent Trends in the Era of Artificial Intelligence. *Drug Discov. Today Technol.* **2019**, *32*, 29–36. [CrossRef]
29. Walters, W.P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc. Chem. Res.* **2020**, *54*, 263–270. [CrossRef]
30. Bajorath, J. State-of-the-Art of Artificial Intelligence in Medicinal Chemistry. *Future Sci. OA* **2021**, *7*, FSO702. [CrossRef]
31. Janela, T.; Bajorath, J. Simple Nearest Neighbor Analysis Meets the Accuracy of Compound Potency Predictions Using Complex Machine Learning Models. *Nat. Mach. Intell.* **2022**, *4*, 1246–1255. [CrossRef]
32. Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discov. Today* **2006**, *11*, 1046–1053. [CrossRef]
33. Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold Hopping Using Two-Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010**, *53*, 5707–5715. [CrossRef]

34. Gómez-Bombarelli, R.; Wei, J.N.; Duvenaud, D.; Hernández-Lobato, J.M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T.D.; Adams, R.P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276. [CrossRef]
35. Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular design. *Mol. Inform.* **2018**, *37*, e1700123. [CrossRef]
36. Bento, A.P.; Gaulton, A.; Hersey, A.; Bellis, L.J.; Chambers, J.; Davies, M.; Krüger, F.A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090. [CrossRef]
37. Baell, J.B.; Holloway, G.A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740. [CrossRef]
38. Bruns, R.F.; Watson, I.A. Rules for Identifying Potentially Reactive or Promiscuous Compounds. *J. Med. Chem.* **2012**, *55*, 9763–9772. [CrossRef]
39. Irwin, J.J.; Duan, D.; Torosyan, H.; Doak, A.K.; Ziebart, K.T.; Sterling, T.; Tumanian, G.; Shoichet, B.K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58*, 7076–7087. [CrossRef]
40. Sohn, K.; Lee, H.; Yan, X. Learning Structured Output Representation Using Deep Conditional Generative Models. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), Montreal, Canada, 7–12 December 2015; pp. 3483–3491.
41. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114. Available online: https://arxiv.org/abs/1312.6114 (accessed on 1 June 2022).
42. Doersch, C. Tutorial on Variational Autoencoders. *arXiv* **2016**, arXiv:1606.05908. Available online: https://arxiv.org/abs/1606.05908 (accessed on 1 May 2022).
43. Rezende, D.J.; Mohamed, S.; Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In Proceedings of the 31st International Conference on Machine Learning (ICML), Beijing, China, 21–26 June 2014; pp. 3057–3070.
44. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
45. Kingma, D.P.; Ba, J.L. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
46. Ralaivola, L.; Swamidass, S.J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Netw.* **2005**, *18*, 1093–1110. [CrossRef]
47. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
48. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
49. Nielsen, M.A. *Neural Networks and Deep Learning*; Determination Press: San Francisco, CA, USA, 2015.
50. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175–185.
51. Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357. [CrossRef]
52. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
53. Chollet, F.K. Keras. Available online: https://github.com/fchollet/keras (accessed on 30 July 2022).
54. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A System for Large-Scale Machine Learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
55. Conover, W.J. On Methods of Handling Ties in the Wilcoxon Signed-Rank Test. *J. Am. Stat. Assoc.* **1973**, *68*, 985–988. [CrossRef]
56. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef]
57. RDKit: Cheminformatics and Machine Learning Software. 2013. Available online: http://www.rdkit.org (accessed on 1 July 2022).

Appendix F

Uncovering and Tackling Fundamental Limitations of Compound Potency Predictions Using Machine Learning Models.

Cell Reports
Physical Science

CellPress
OPEN ACCESS

Perspective

Uncovering and tackling fundamental limitations of compound potency predictions using machine learning models

Tiago Janela[1] and Jürgen Bajorath[1,2,3,*]

SUMMARY

Molecular property predictions play a central role in computer-aided drug discovery. Although a variety of physicochemical (e.g., solubility or chemical reactivity) or physiological properties (e.g., metabolic stability or toxicity) can be predicted, biological activity is by far the most frequently investigated compound feature. Activity predictions are carried out in a qualitative (target-based activity, through compound classification) or quantitative (compound potency or ligand-target affinity, through regression modeling) manner. Many studies have evaluated and compared different machine learning methods for activity and potency predictions, recently with a focus on deep learning. Regardless of the methods used, these studies generally rely on conventional benchmark settings. Recent work has shown that potency prediction benchmarks have severe general limitations that have long been unnoticed but prevent a reliable assessment of different methods and their relative performance. In this perspective, we outline general limitations of benchmark settings for compound potency predictions, introduce potential alternatives enabling a more realistic assessment of state-of-the-art predictive models, and discuss future directions for elucidating predictions and further increasing their impact.

INTRODUCTION

Compound potency predictions play a central role in computer-aided drug discovery.[1–3] Over the years, various ligand- and target structure-based potency prediction methods have been introduced, at different levels of computational sophistication. Since the early 1960s,[4] quantitative structure-activity relationship (QSAR) analysis methods have become a foundation of computer-aided drug discovery and continue to be widely applied in the practice of medicinal chemistry, especially during compound optimization.[5,6] The complexity of QSAR-type methods greatly varies, ranging from simplistic multiple linear regression to machine learning (ML) models.[4–6] Standard QSAR techniques relate physicochemical properties of compounds, accounted for through the use of numerical descriptors, to biological activity via linear models. Once calibrated for an evolving compound series, these models are applied to predict the potency of structural analogs in the search of increasingly potent compounds. As such, QSAR predictions are typically confined to congeneric compounds and carried out for one series at a time. For potency predictions on structurally more diverse compounds, ML regression models are derived, as further discussed below. Hence, contemporary QSAR approaches include both linear and non-linear (ML) models.[6]

[1]Department of Life Science Informatics and Data Science, B-IT, University of Bonn, Friedrich-Hirzebruch-Allee 5/6, 53115 Bonn, Germany

[2]Lamarr Institute for Machine Learning and Artificial Intelligence, University of Bonn, Friedrich-Hirzebruch-Allee 5/6, 53115 Bonn, Germany

[3]Limes Institute – Program Unit Chemical Biology and Medicinal Chemistry, University of Bonn, Friedrich-Hirzebruch-Allee 5/6, 53115 Bonn, Germany

*Correspondence: bajorath@bit.uni-bonn.de

https://doi.org/10.1016/j.xcrp.2024.101988

In structure-based compound design, different types of energy calculations are carried out to approximate ligand affinities. For example, ligand docking calculations (that is, placing compounds into target binding sites in a chemically and sterically complementary manner) are carried out for structure-based virtual screening as well as compound optimization.[7] These calculations make use of a variety of scoring functions and energy calculations that are mostly based on molecular mechanics force fields or statistics of ligand-target interactions (such as potentials of mean force).[7–9] Irrespective of how docking scores are designated in the literature (often as energy values), they generally represent rough estimates of ligand-target interaction energies and are primarily used to rank hypothetical binding modes of docked compounds on a relative scale. At a higher level of sophistication, relative binding free energy perturbation calculations are more accurate and best applied to X-ray structures of ligand-target complexes.[10,11] Similar to QSAR predictions, free energy perturbation calculations are essentially restricted to congeneric compounds and used to calculate affinity/potency differences between structural analogs. Free energy calculations following the quantum mechanics/molecular mechanics (QM/MM) approach represent a further advance where a ligand and its immediate protein environment are treated quantum mechanically, while the remainder of the target is treated using molecular mechanics functions.[12,13] In general, free energy calculations have high computational demands and are difficult to generalize across different targets and compound series.

Although structure-based potency predictions and conventional QSAR analysis are conceptually distinct, these approaches have in common that they are generally limited to individual compound series consisting of structural analogs. Therefore, to enable potency predictions based on larger datasets containing structurally diverse compounds, regression models using different ML algorithms have become a mainstay in computer-aided drug discovery.[14] For example, random forest regression (RFR)[15] or support vector regression (SVR)[16] models are, by today's standards, computationally inexpensive and usually straightforward to derive for given target-dependent sets of active compounds (termed activity classes). In contrast to standard QSAR-type modeling, these methods have the principal advantage that they are able to capture non-linear structure-activity relationships (SARs) and thus applicable to structurally diverse compounds. SVR was first applied in drug discovery in the late 1990s and evolved to be a standard for non-linear potency prediction (non-linear QSAR) over time. Furthermore, in recent years, deep neural network (DNN) architectures such as deep feedforward, recurrent, and graph neural networks or transformers have increasingly been used for quantitative molecular property predictions.[14,17–20] Figure 1 shows a schematic representation of the SVR, RFR, and exemplary DNN approaches, and Figure 2 illustrates compound potency prediction using an ML model.

Although promising potency predictions were reported for different DNNs, improvements in accuracy over standard ML methods such as RFR or SVR were often marginal at best for different molecular representations.[17–19,24,25] These findings were at least in part attributable to the situation that datasets of active compounds from early-phase drug discovery are small (on the order of hundreds to thousands of compounds) compared to datasets from other fields such as natural language processing or image analysis that were substantially advanced through DNNs. Compound activity data mostly result from compound (hit-to-lead or lead) optimization projects in medicinal chemistry, which explains data sparseness.

Furthermore, compound predictions do not depend on representation learning (in contrast to, for example, image analysis based on pixel data), given the availability

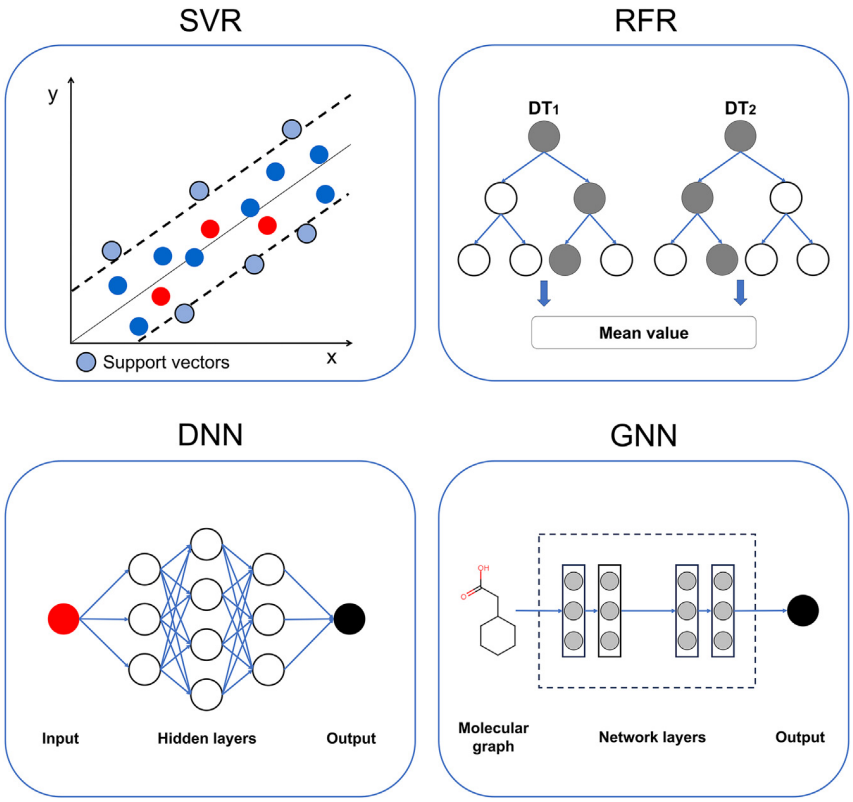


Figure 1. Schematic representation of different machine learning methods

Illustrated are the support vector regression (SVR), random forest regression (RFR), deep feedforward neural network (DNN), and graph neural network (GNN) methods. For SVR, blue circles represent training instances, and light blue support vectors are used to generate the tube around the derived function (hyperplane). For SVR and DNN methods, red circles illustrate test instances. For RFR, DT stands for decision tree, and the tree path from the root to the tree leaf is highlighted using gray circles. Different from the other methods, RFR is an ensemble approach relying on independently derived DT models. RFR and SVR represent adaptations of original classification algorithms for predicting numerical values. DNNs and GNNs employ non-linear activation functions to map an input value to the respective output across computational neuron layers. For GNNs, the input data are a molecular graph. Further methodological details of the RFR, SVR, DNN, and GNN methods are provided in Breiman,[15] Drucker et al.,[16] Khamparia and Singh,[21] and Scarselli et al.,[22] respectively.

of a wealth of pre-defined structural and molecular property descriptors. Clearly, compound data spareness and the use of high-resolution molecular representations do not play into the strengths of deep learning.

Moreover, there are also potential methodological caveats. For instance, as a new approach for structure-based affinity calculations, graph neural networks (GNNs) have been used to predict compound potency values from ligand-target interaction graphs extracted from X-ray structures.[26] Significant correlation between predicted and experimental values was frequently reported.[26] However, these predictions were found to be largely determined artificially by ligand memorization of GNNs and thus did not depend on learning ligand-target interactions.[26,27] Given that similar active compounds often tend to have similar potency, memorizing training compounds that are structurally related to test compounds and assigning the potency of these training samples to test compounds causes these effects.

In general, however, the absence of significant performance increases for potency predictions using computational methods of increasing complexity has fundamental reasons and consequences, as discussed in this perspective.

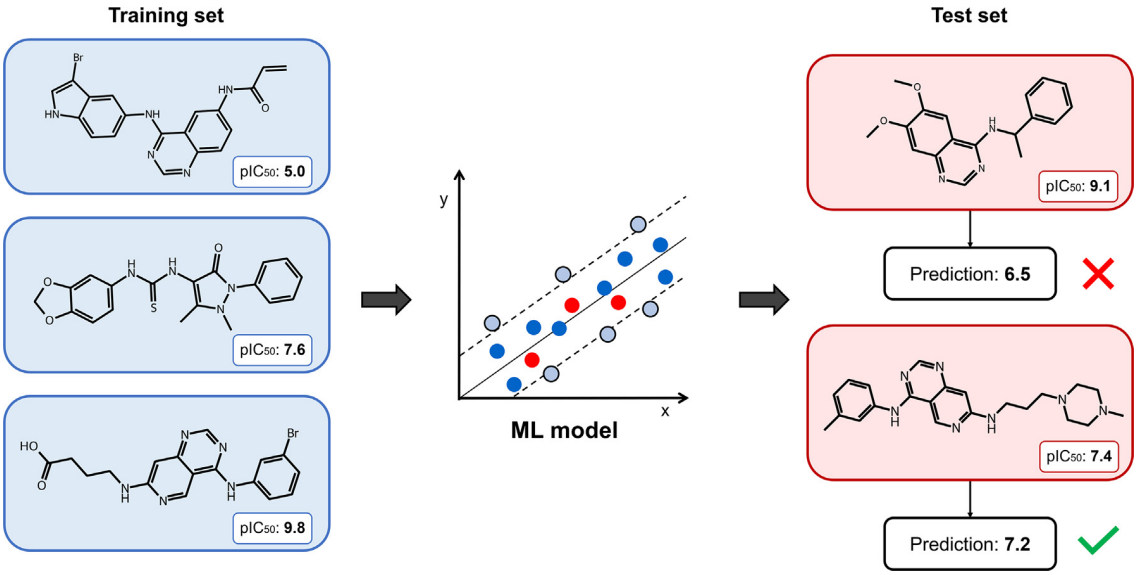


Figure 2. Compound potency predictions with an ML model

The model is derived using training compounds to predict the potency of test instances for a given activity class. These calculations are vulnerable to typically observed potency value distributions in compound datasets from medicinal chemistry and varying structural relationships between training and test instances, as further discussed in the text. IC_{50} represents the half-maximal inhibitory concentration of a compound inhibiting a biological target,[23] providing an assay-dependent measure of potency. pIC_{50} stands for the negative logarithmic IC_{50} value in molar (M) concentration.

In the following sections, we detail intrinsic limitations of potency prediction benchmark calculations, put them into scientific context, and deliberate alternatives for the evaluation of predictive models that might be more informative. Finally, we discuss future directions for the field including the need for novel methodological concepts and approaches integrating computation and experiment.

Uncovering general limitations

For potency prediction using ML models, the performance of computational methods typically relies on benchmark systems and calculations. Benchmarking is of critical importance for the prospective use of novel methodologies. Benchmark studies require reliable compound activity data extracted from, for example, the ChEMBL[28] database, the major public repository of literature and patent data from medicinal chemistry. Like data from any larger compound repository, ChEMBL data are heterogeneous since they originate from different assays carried out in different laboratories and contain inaccuracies due to experimental variance, measurement errors, or ambiguous target assignments.[23] However, ChEMBL is manually curated, actively maintained and expanded, and remains the source of choice for publicly available compounds from medicinal chemistry and associated activity data. Once activity classes have been obtained, the compounds and activity data are partitioned into training and test sets for ML model derivation and evaluation, respectively, conventionally via multiple independent cross-validation iterations using different performance measures. Such benchmark settings are generally applied for the evaluation and comparison of the prediction accuracy of different ML methods, both in compound classification and regression modeling. Classification typically aims at distinguishing between compounds from an activity class for a given target and other randomly selected compounds assumed to be inactive against this target, while regression aims at the prediction of numerical potency value for test compounds from a given activity class. Because activity classes originate from compound optimization efforts, they mostly cover a wide range from high micromolar to

low nanomolar (or sub-nanomolar) potency. Thus, properly derived regression models should be capable of predicting compound potency across the entire range for typically sized activity classes, including weakly and highly potent test compounds. For classification models, a variety of performance metrics are applied to assess prediction accuracy[29]; for regression models, the mean absolute error (MAE) or root-mean-squared error are primarily used.[29] Benchmark publications typically report high prediction accuracy of different models and activity classes, often proposing superiority of new models based on an increase in prediction accuracy of only a few percent.

Given the increasing use of deep learning models in computer-aided drug discovery, potency prediction benchmarks were systematically carried out for ML methods of increasing complexity including RFR, SVR, DNN, and a graph convolutional network (GCN), a GNN variant with representation learning compared to simple controls and randomized regression models.[30–32] As controls for baseline performance, *k*-nearest neighbor (*k*-NN) and median regression (MR) calculations were carried out, which do not require learning. In *k*-NN calculations, the potency assigned to a test compound was averaged over the *k* most similar training compounds. In MR, the median potency value of the training set was assigned to each test compound. Thus, all test compounds were predicted to have the same median potency. Furthermore, in random regression (RR), models were trained after randomly reassigning (shuffling) potency values across compounds. Thus, in RR, SARs, the presence of which provides the basis for learning, were eliminated. Figure 3 shows the results of representative potency predictions.[30–32]

In these potency predictions, several key observations were made.

- k-NN reached or surpassed the accuracy of increasingly complex ML models

In independent trials, stable potency predictions were observed for essentially all models, and differences between alternative models were marginal at best. All observed differences fell well within an order of magnitude. On average, predictions reached MAE values of ~0.5 relative to the experimental data, corresponding to ~3-fold differences in potency, which are often not biologically relevant and approach experimental accuracy limits of ~0.3 log units (corresponding to ~2-fold differences in potency).[23,33] Based on these considerations, the predictions were generally accurate. Furthermore, predictions using different models were only separated by ~0.1–0.2 MAE (except MR). Overall, SVR reached slightly higher performance than DNN/GNN models and *k*-NN calculations (Figure 3A).[30] However, while the differences between model performance distributions had moderate statistical significance,[30] differences between predictions with alternative models essentially fell within experimental accuracy limitations. In large-scale predictions over hundreds of different activity classes covering current pharmaceutical targets, these trends were consistently observed, confirming their generality.[31]

- Only small error margins separated ML models and controls from random predictions

For RR models, prediction yielded mean MAE values of ~0.9 (<10-fold) across different activity classes (Figure 3B), compared to ~0.5 for the original models. Thus, the performance of ML models was separated from randomized predictions by only small margins of maximally ~0.5 MAE (corresponding to ~3-fold differences in potency).[30]

- Predictions of ML models were biased by median potency values

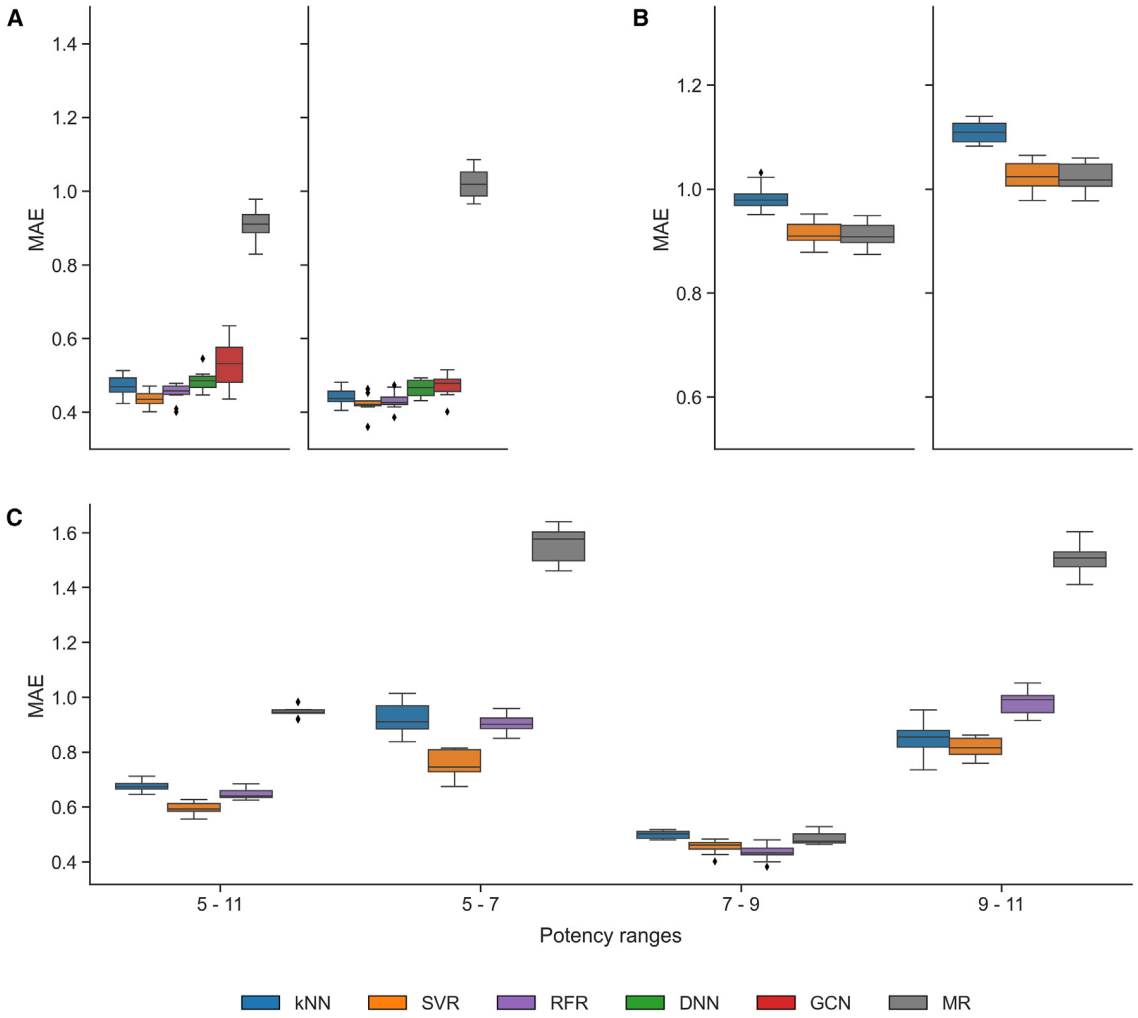


Figure 3. Representative potency predictions for exemplary activity classes

As a measure of prediction accuracy, mean absolute error (MAE) values for logarithmic potencies are reported for (A) different ML models trained on two individual activity classes for targets PI3-kinase p110-alpha subunit (left) and beta-secretase 1 (right),[30,31] (B) corresponding randomized models for these two activity classes,[30,31] and (C) predictions on an exemplary activity class (epoxide hydratase) monitored across potency intervals.[32] ML methods are designated according to Figure 2. In addition, kNN and GCN abbreviate *k*-nearest neighbor and graph convolutional network, respectively. The results of multiple independent predictions are reported in boxplots. In boxplots, the box defines upper and lower quantile and the horizontal line the median value of the distribution. Upper and lower whiskers represent the maximum and minimum value, respectively. Diamond symbols mark statistical outliers. Activity classes in (A) and (B) consisted of 1262 and 1116 compounds, respectively, and 80% vs. 20% training/test data partitions were generated. The activity class in (C) comprised 1227 compounds. In this case, 50% vs. 50% training/test data partitions were generated. The results shown here are representative of all studied activity classes.[30–32]

To further investigate the findings discussed above, predictions using different models were monitored for compounds falling into different potency intervals,[32] as shown in Figure 3C. For low and high potency intervals of 5–7 and 9–11, MAE values were largest for all models and MR, with MAE values up to ~1 and 1.5, respectively (low and high potency values depart from the median). By contrast, for the potency interval 7–9, into which median values (~8) fell, MAEs were lowest for ML models and MR (~0.5) and very similar. Furthermore, prediction accuracy of different models in the potency interval 7–9 mirrored global prediction accuracy, as also shown in Figure 3A. This was the case because the majority of compounds in this and the other activity classes had micromolar to high nanomolar potency and thus

fell into this interval,[32] which is typical for compound activity classes from medicinal chemistry.[32] Thus, data concentration around median values determined global prediction accuracy and biased predictions of ML models.

- Predictions of ML models were biased by structural analogs

We also examined consequences of the presence of structural analogs in training and test sets of ML models on potency predictions.[34] Therefore, a molecular test system was designed comprising pairs of analogs with increasing potency differences that were divided into training and test compounds. Analog pairs were systematically extracted from different activity classes following the matched molecular pair (MMP) formalism.[35] An MMP is defined as a pair of compounds that are only distinguished by a chemical change at a single site (here an R-group replacement).[35] Using this system setup, all ML models showed a clear tendency to predict values close to the potency of training compounds for structural analogs in test sets.[34] Accordingly, the larger the potency differences between analogs were, the less accurate were the predictions. These findings also rationalized notorious difficulties of ML models to predict activity cliffs[36] observed in different studies. Activity cliffs were introduced as pairs of very similar compounds with large potency differences[36] (e.g., >100-fold) and hence represent the pinnacle of SAR non-linearity (discontinuity). However, since activity cliffs occur only infrequently and most structural analogs have comparable potency, ML predictions for the MMP-based test system were accurate for many analogs pairs falling into mid-potency range (logarithmic potency interval 7–9).

However, to avoid potential analog bias ("data leakage") in potency predictions, leading to over-optimistic prediction outcomes, the presence of structural analogs in training and test sets should best be avoided. This can be accomplished by scaffold- or analog-series-based partitioning of training and test data (instead of random splitting), as exemplified by the test system discussed above. Although series-based partitioning also is an approach of choice to derive models for prospective applications, it is not (yet) common practice in molecular ML.

Taken together, these key findings showed that potency predictions using different ML methods and *k*-NN calculations displayed similarly accurate performance largely determined by compounds in the mid-potency range and separated from randomized predictions by only confined margins. It follows that predictions using conventional benchmark settings cannot discriminate between ML methods of different complexity and do not provide a realistic picture of the predictive performance of these approaches.

Accordingly, conclusions drawn from conventional potency prediction benchmarks are questionable at best.

Alternative calculation schemes

The intrinsic limitations of potency benchmark calculations discussed above have long gone unnoticed but present a substantial problem for molecular ML and computer-aided drug discovery. Since the performance of increasingly complex methods was similar for hundreds of activity classes, leading to only little relative differences between prediction accuracy, and only separated from random predictions by small margins, these limitations obviously did not depend on compound classes and their chemical characteristics or specific (target-based) activities. Instead, they largely resulted from potency value distributions commonly observed in activity classes from

medicinal chemistry used for benchmarking, as discussed above. As illustrated in Figure 3C, a characteristic feature of these distributions was the dominance of compound potency values in the micromolar to high nanomolar range (logarithmic potency interval 7–9), which included the median potency value of most activity classes and largely determined global prediction accuracy.

Although the design and evaluation of alternative benchmark schemes is still in its early stages, based on these insights, we can at least point at possible directions for improving the basis of methodological comparisons. For example, the presence of structural analogs in training and test sets of ML models should be avoided. Furthermore, the evaluation of potency predictions should preferentially focus on potent compounds in test sets with significant deviations from the median potency value. For all practical purposes, highly potent compounds represent the most important prediction tasks. As shown above, for highly potent compounds, global prediction accuracy decreased, and performance differences between alternative methods increased, albeit only slightly (Figure 3C). However, such modifications of typical benchmarking will likely not suffice to substantially improve the basis for method evaluation. Instead, new benchmark concepts must be developed, which is not a trivial task. In principle, system setups are required that yield more variable prediction outcomes and that increase the separation of alternative ML methods, simple controls, and random prediction accuracy. Currently, such test systems are unavailable but can be conceived. For example, we have derived potency prediction models for a given activity class in the presence of increasing proportions of presumed *inactive compounds*, that is, compounds from other classes with an assigned potency value of 0 for that activity class. Models were built for training sets with 0%–100% of inactive compounds (added in increments of 10%), with 100% corresponding to the size of the original training set of active compounds. Inactive compounds were either randomly selected from only one other activity class (termed "homogeneous" selection) or different activity classes (random selection). Importantly, inactive training instances expanded the potency range for learning by several orders of magnitude. Figure 4 shows the results of predictions for three exemplary activity classes using models derived in the presence of increasing proportions of inactive training instances.

Compared to the original models, in the presence of increasing proportions of inactive training compounds, the distributions of independent prediction trials widened, prediction accuracy decreased, and the separation between ML models and controls increased. These effects were much stronger for randomly selected inactive training compounds than compounds from homogeneous selection because randomly selected samples further increased the diversity of training sets, which provided additional challenges for ML models. Notably, *k*-NN prediction accuracy was less affected than other models by the addition of diverse inactive compounds to the training set because of the low probability that nearest active neighbors of test compounds were replaced by random training samples. By contrast, large prediction errors were observed for the DNN models, which adapted worse than other ML methods to the presence of increasing numbers of inactive training instances. Exploring reasons for this apparent vulnerability should be of interest.

Hence, the inclusion of inactive compounds in model derivation alters benchmark settings for the evaluation of potency predictions in a meaningful way.

Forward-looking viewpoints

While qualitative compound activity and quantitative potency predictions have already played a central role in computer-aided drug discovery for many years, the relevance of

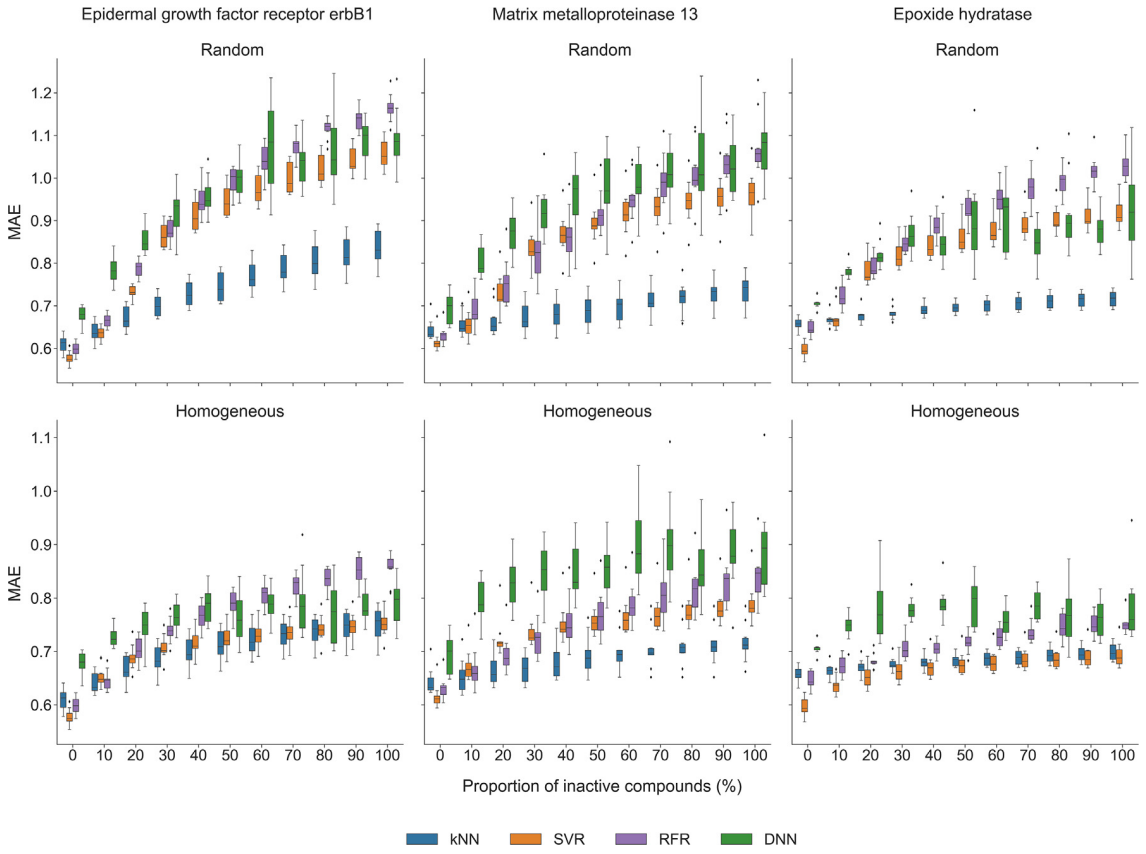


Figure 4. Prediction accuracy of models built in the presence of inactive compounds

Boxplots report the distribution of MAE values for active test compounds using different models derived in the presence of increasing proportions of inactive training compounds (via homogeneous or random selection). Results of pilot calculations are shown for three exemplary previously studied publicly available compound activity classes.[32,34] For the studied activity classes (from left to right), dataset sizes corresponded to 1586, 797, and 1410 active compounds, respectively. In each case, 50% vs. 50% training/test data partitions were generated. For each class, the target name is provided.

such predictions will most likely further increase in the artificial intelligence (AI) era. For example, as chemical source libraries are becoming ultra-large, there will be a need to tightly control the magnitude of biological screening efforts and experimental costs. Although it is known that the accuracy of activity and potency prediction methods observed in benchmark settings is not a reliable indicator of their performance in practical drug discovery applications, reasons for this discrepancy have only been little explored until recently. Going forward, raising further awareness of the limitations of such predictions and conventional benchmark systems will be particularly relevant as demands and expectations will grow. Importantly, intrinsic limitations of potency predictions highlighted herein cannot be compensated for by using prediction models of increasing computational complexity, which is a trend in the AI era. Rather, new concepts for a more meaningful assessment of computational approaches must be devised. For instance, it should be realized and further emphasized that the ultimate goal of such computational efforts is the identification of potent compounds in prospective drug discovery applications. Clearly, this calls for a stronger focus on such applications, complementing and further extending efforts to develop second-generation benchmark schemes enabling more informative methodological comparisons and more reliable performance estimates. However, given that many computational groups driving method development are not able to regularly interface with experiments, especially in academia, prospective applications are not straightforward to implement. Therefore,

interdisciplinary drug discovery environments are challenged to contribute more strongly to advancing the field. In particular, computationally driven case studies with seamless experimental evaluation of predictions in discovery projects are expected to receive wide attention, provide practically relevant results, and balance expectations, especially as the complexity of computational approaches increases. Studies prioritizing small numbers of compounds for experimental evaluation are thought to be more important for advancing the field than the statistical assessment and comparison of global prediction accuracy on different compound classes, provided these studies are rigorously reported to the scientific community. This will require increasing orientation toward open science in traditionally conservative drug discovery environments, but the benefits will be mutual. In any event, it is hoped that increasing awareness of computational limitations of benchmark predictions will lead to more prospective applications accompanying the development of new computational concepts.

ACKNOWLEDGMENTS

The authors thank Jannik Roth and Alec Lamens for helpful discussions.

AUTHOR CONTRIBUTIONS

T.J. and J.B. wrote and edited the manuscript. T.J. prepared the illustrations.

DECLARATION OF INTERESTS

There are no conflicts of interest to declare.

REFERENCES

1. Jorgensen, W.L. (2004). The many roles of computation in drug discovery. Science *303*, 1813–1818. https://doi.org/10.1126/science.1096361.
2. Bajorath, J. (2015). Computer-aided drug discovery. F1000Res. *4*. F1000 Faculty Rev-630. https://doi.org/10.12688/f1000research.6653.1.
3. Sadybekov, A.V., and Katritch, V. (2023). Computational approaches streamlining drug discovery. Nature *616*, 673–685. https://doi.org/10.1038/s41586-023-05905-z.
4. Hansch, C., Maloney, P.P., Fujita, T., and Muir, R.M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature *194*, 178–180. https://doi.org/10.1038/194178b0.
5. Lewis, R.A., and Wood, D. (2014). Modern 2D QSAR for drug discovery. WIREs Comput. Mol. Sci. *4*, 505–522. https://doi.org/10.1002/wcms.1187.
6. Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R., et al. (2014). QSAR Modeling: Where Have You Been? Where Are You Going To? J. Med. Chem. *57*, 4977–5010. https://doi.org/10.1021/jm4004285.
7. Kitchen, D.B., Decornez, H., Furr, J.R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug Discov. *3*, 935–949. https://doi.org/10.1038/nrd1549.
8. Liu, J., and Wang, R. (2015). Classification of current scoring functions. J. Chem. Inf. Model. *55*, 475–482. https://doi.org/10.1021/ci500731a.
9. Marin, E., Kovaleva, M., Kadukova, M., Mustafin, K., Khorn, P., Rogachev, A., Mishin, A., Guskov, A., and Borshchevskiy, V. (2024). Regression-based active learning for accessible acceleration of ultra-large library docking. J. Chem. Inf. Model. *64*, 2612–2623. https://doi.org/10.1021/acs.jcim.3c01661.
10. Abel, R., Wang, L., Harder, E.D., Berne, B.J., and Friesner, R.A. (2017). Advancing drug discovery through enhanced free energy calculations. Acc. Chem. Res. *50*, 1625–1632. https://doi.org/10.1021/acs.accounts.7b00083.
11. Williams-Noonan, B.J., Yuriev, E., and Chalmers, D.K. (2018). Free energy methods in drug design: prospects of "alchemical perturbation" in medicinal chemistry. J. Med. Chem. *61*, 638–649. https://doi.org/10.1021/acs.jmedchem.7b00681.
12. Senn, H.M., and Thiel, W. (2009). QM/MM methods for biomolecular systems. Angew. Chem., Int. Ed. Engl. *48*, 1198–1229. https://doi.org/10.1002/anie.200802019.
13. Zhou, T., Huang, D., and Caflisch, A. (2010). Quantum mechanical methods for drug design. Curr. Top. Med. Chem. *10*, 33–45. https://doi.org/10.2174/156802610790232242.
14. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. *18*, 463–477. https://doi.org/10.1038/s41573-019-0024-5.
15. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
16. Drucker, H., Surges, C.J.C., Kaufman, L., Smola, A., and Vapnik, V. (1996). Support vector regression machines. In Advances in Neural Information Processing Systems (NIPS 1996), pp. 155–161.
17. Hou, F., Wu, Z., Hu, Z., Xiao, Z., Wang, L., Zhang, X., and Li, G. (2018). Comparison study on the prediction of multiple molecular properties by various neural networks. J. Phys. Chem. A *122*, 9128–9134. https://doi.org/10.1021/acs.jpca.8b09376.
18. Walters, W.P., and Barzilay, R. (2021). Applications of deep learning in molecule generation and molecular property prediction. Acc. Chem. Res. *54*, 263–270. https://doi.org/10.1021/acs.accounts.0c00699.
19. Li, H., Zhang, R., Min, Y., Ma, D., Zhao, D., and Zeng, J. (2023). A knowledge-guided pre-training framework for improving molecular representation learning. Nat. Commun. *14*, 7568. https://doi.org/10.1038/s41467-023-43214-1.
20. Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., and Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J. Cheminf. *13*, 12. https://doi.org/10.1186/s13321-020-00479-8.
21. Khamparia, A., and Singh, K.M. (2019). A systematic review on deep learning architectures and applications. Expet Syst. *36*, e12400. https://doi.org/10.1111/exsy.12400.

22. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., and Monfardini, G. (2009). The graph neural network model. IEEE Trans. Neural Network. *20*, 61–80. https://doi.org/10.1109/TNN.2008.2005605.

23. Landrum, G.A., and Riniker, S. (2024). Combining IC_{50} or K_i values from different sources is a source of significant noise. J. Chem. Inf. Model. *64*, 1560–1567. https://doi.org/10.1021/acs.jcim.4c00049.

24. Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D., and Wang, F. (2023). A systematic study of key elements underlying molecular property prediction. Nat. Commun. *14*, 6395. https://doi.org/10.1038/s41467-023-41948-6.

25. van Tilborg, D., Alenicheva, A., and Grisoni, F. (2022). Exposing the limitations of molecular machine learning with activity cliffs. J. Chem. Inf. Model. *62*, 5938–5951. https://doi.org/10.1021/acs.jcim.2c01073.

26. Volkov, M., Turk, J.A., Drizard, N., Martin, N., Hoffmann, B., Gaston-Mathé, Y., and Rognan, D. (2022). On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. J. Med. Chem. *65*, 7946–7958. https://doi.org/10.1021/acs.jmedchem.2c00487.

27. Mastropietro, A., Pasculli, G., and Bajorath, J. (2023). Learning characteristics of graph neural networks predicting protein–ligand affinities. Nat. Mach. Intell. *5*, 1427–1436. https://doi.org/10.1038/s42256-023-00756-9.

28. Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Krüger, F.A., Light, Y., Mak, L., McGlinchey, S., et al. (2014). The ChEMBL bioactivity database: an update. Nucleic Acids Res. *42*, 1083–1090. https://doi.org/10.1093/nar/gkt1031.

29. Bender, A., Schneider, N., Segler, M., Patrick Walters, W., Engkvist, O., and Rodrigues, T. (2022). Evaluation guidelines for machine learning tools in the chemical sciences. Nat. Rev. Chem *6*, 428–442. https://doi.org/10.1038/s41570-022-00391-9.

30. Janela, T., and Bajorath, J. (2022). Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models. Nat. Mach. Intell. *4*, 1246–1255. https://doi.org/10.1038/s42256-022-00581-6.

31. Janela, T., and Bajorath, J. (2023). Large-scale predictions of compound potency with original and modified activity classes reveal general prediction characteristics and intrinsic limitations of conventional benchmarking calculations. Pharmaceuticals *16*, 530. https://doi.org/10.3390/ph16040530.

32. Janela, T., and Bajorath, J. (2023). Rationalizing general limitations in assessing and comparing methods for compound potency prediction. Sci. Rep. *13*, 17816. https://doi.org/10.1038/s41598-023-45086-3.

33. Brown, S.P., Muchmore, S.W., and Hajduk, P.J. (2009). Healthy skepticism: assessing realistic model performance. Drug Discov. Today *14*, 420–427. https://doi.org/10.1016/j.drudis.2009.01.012.

34. Janela, T., and Bajorath, J. (2023). Anatomy of potency predictions focusing on structural analogues with increasing potency differences including activity cliffs. J. Chem. Inf. Model. *63*, 7032–7044. https://doi.org/10.1021/acs.jcim.3c01530.

35. Hussain, J., and Rea, C. (2010). Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J. Chem. Inf. Model. *50*, 339–348. https://doi.org/10.1021/ci900450m.

36. Maggiora, G.M. (2006). On outliers and activity cliffs–why QSAR often disappoints. J. Chem. Inf. Model. *46*, 1535. https://doi.org/10.1021/ci060117s.

Additional Publications

Janela, T.; Takeuchi, K.; Bajorath, J. "Introducing a Chemically Intuitive Core-Substituent Fingerprint Designed to Explore Structural Requirements for Effective Similarity Searching and Machine Learning. *Molecules* **2022**, *27*, 233". DOI: 10.3390/molecules27072331