Artificial Intelligence in the Life Sciences 1 (2021) 100007

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Computational prediction of frequent hitters in target-based and cell-based assays

Conrad Stork[a], Neann Mathai[b], Johannes Kirchmair[a,b,c,*]

[a] Universität Hamburg, Faculty of Mathematics, Informatics and Natural Sciences, Department of Informatics, Center for Bioinformatics, 20146 Hamburg, Germany
[b] Department of Chemistry and Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway
[c] Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

ARTICLE INFO

Keywords:
Machine learning
Frequent hitters
Nuisance compounds
PAINS
Biological assays
High-throughput screening

ABSTRACT

Compounds interfering with high-throughput screening (HTS) assay technologies (also known as “badly behaving compounds”, “bad actors”, “nuisance compounds” or “PAINS”) pose a major challenge to early-stage drug discovery. Many of these problematic compounds are “frequent hitters”, and we have recently published a set of machine learning models (“Hit Dexter 2.0”) for flagging such compounds.

Here we present a new generation of machine learning models which are derived from a large, manually curated and annotated data set. For the first time, these models cover, in addition to target-based assays, also cell-based assays. Our experiments show that cell-based assays behave indeed differently from target-based assays, with respect to hit rates and frequent hitters, and that dedicated models are required to produce meaningful predictions. In addition to these extensions and refinements, we explored a variety of additional setups for modeling, including the combination of four machine learning classifiers (i.e. k-nearest neighbors (KNN), extra trees, random forest and multilayer perceptron) with four sets of descriptors (Morgan2 fingerprints, Morgan3 fingerprints, MACCS keys and 2D physicochemical property descriptors).

Testing on holdout data as well as data sets of “dark chemical matter” (i.e. compounds that have been extensively tested in biological assays but have never shown activity) and known bad actors show that the multilayer perceptron classifiers in combination with Morgan2 fingerprints outperform other setups in most cases. The best multilayer perceptron classifiers obtained Matthews correlation coefficients of up to 0.648 on holdout data. These models are available via a free web service.

Introduction

High-throughput screening (HTS) assay technologies are a cornerstone of modern drug discovery. They allow the biological testing of large numbers of compounds on targets of interest within a short period of time [1]. A major challenge faced in high-throughput screening is false-positive hits resulting from different types of assay interference [2].

Compounds causing assay interference are referred to as “badly behaving compounds”, “bad actors” or “nuisance compounds”. Many of them, but by far not all, are “frequent hitters” (i.e. compounds which show higher-than-expected hit rates in biological assays). This is because not all types of assay interference are frequent events. In fact, many types of assay interference are triggered only by specific conditions.

Importantly, not all frequent hitters are nuisance compounds. Quite on the contrary: frequent hitter behavior can be a result of true promiscuity mediated by “privileged scaffolds” [3]. Privileged scaffolds enable compounds to bind, in a specific manner, to a number of distinct proteins. Such compounds can be particularly useful in the context of polypharmacology and drug repurposing.

An established experimental strategy to discriminate genuine hits from false-positive results is the use of orthogonal and counterscreen assays [4], but even with such an advanced experimental setup some cases of assay interference may not be captured because the underlying mechanisms are manifold.

Given the complexities involved in the conduction and analysis of experimental screens, computational tools to aid the discrimination of genuine hits from false ones are in high demand. Today, a variety of in silico approaches for cherry-picking the most promising hits for follow-up studies are at our disposal [5–10]. We will discuss these briefly in the context of the individual types of assay interference.

The most prominent cause of interference in biological assays (biochemical assays in particular) is related to the formation of aggregates by small molecules that engage in nonspecific interactions with

[*] Corresponding author.
E-mail address: johannes.kirchmair@univie.ac.at (J. Kirchmair).

https://doi.org/10.1016/j.ailsci.2021.100007
Received 14 June 2021; Received in revised form 5 August 2021; Accepted 6 August 2021
Available online 8 August 2021

biomacromolecules [5]. Several computational approaches have been reported for the assessment of small molecules with regard to their risk of forming colloidal aggregates. These tools include Aggregator Advisor [11], ChemAgg [12] and SCAM detective [13]. Aggregator Advisor flags potential aggregators based on their molecular similarity to a set of 12,000 known aggregators, taking logP into account. ChemAgg and SCAM Detective are machine learning models for the classification of small molecules into aggregators and non-aggregators. Whereas ChemAgg is based on a XGBoost model, SCAM detective utilizes a set of random forest models.

A second important cause of assay interference is the chemical reactivity of compounds, in particular that related to electrophilicity [14]. Chemically reactive compounds may bind covalently to biomacromolecules or interact with the assay screening technology in an undesired way. Computational approaches for identifying reactive compounds are mostly based on sets of rules which describe substructures that have been linked to chemical reactivity [15].

Further types of assay interference are covered under the umbrella of the well-known pan-assay interference compounds (PAINS) concept [16]. PAINS are compounds based on molecular scaffolds that have been associated with various types of assay interference. PAINS include redox cycling compounds (e.g. toxoflavins), covalent binders (e.g. isothiazolones or ene-rhodanines), membrane disruptors (e.g. curcumin), metal complex-forming compounds (e.g. hydroxyphenyl hydrazones) and unstable compounds (e.g. phenol-sulfonamides) [17]. The molecular fragments linked to PAINS have been compiled in a collection of several hundred structural patterns, and this collection has been implemented in various in silico platforms and software libraries to offer means for flagging potentially problematic compounds [18]. An alternative approach to flagging potential PAINS was recently presented by Koptelov et al. [19]. They use discriminative subgraph mining to identify characteristic patterns in PAINS and non-PAINS, and utilize these patterns, in combination with numerical descriptors, to derive decision tree models for PAINS prediction.

A number of focused machine learning models have been devised for the identification of compounds that likely cause specific types of assay interference. For example, Luciferase Advisor [20] and ChemFluc [21] are models for the prediction of compounds (luciferase inhibitors) that may interfere with luciferase-based assays. InterPred [22] includes a set of QSAR models for the prediction of luciferase inhibitors and autofluorescence compounds in cell-based and target-based assays.

Several computational tools are in existence that predict frequent hitters independent of the underlying mechanisms (genuine promiscuity; various types of assay interference). For example, researchers at AstraZeneca have derived a statistical model for the prediction of frequent hitters based on their in-house historical bioactivity data [23]. Another statistical model for the prediction of frequent hitters is BADAPPLE [24]. In contrast to the AstraZeneca model, the BADAPPLE model is derived from molecular scaffolds rather than complete molecular structures.

More recently, machine learning has been moved into the focus also in the field of frequent hitter and assay interference prediction. For example, Hit Dexter 2.0 [25], developed by some of us, predicts frequent hitters utilizing a set of extra tree models that are trained on large sets of data extracted from the PubChem Bioassay database [26]. More recently, Feldmann et al. [27] reported a machine learning approach for the prediction of true promiscuous compounds (multi-target compounds) in which they removed likely aggregators and other types of assay interference compounds from the training sets in an effort to work with cleaner sets of promiscuous and non-promiscuous compounds.

Whereas a sizable number of in silico models for the prediction of frequent hitters and badly behaving compounds are at our disposal today, most of them have clear limitations with respect to the coverage of mechanisms of interference and assay technologies. In particular, the existing approaches are focused on, or limited to, biochemical (i.e. target-based) assays and do not adequately represent cell-based assays, which can behave very differently with respect to assay interference.

Table 1
Definitions of values for the manually assigned label “target type”.

Label value	Description
target-based	Assays generating readouts from purified proteins or peptides
cell-based	Assays generating readouts from cells
other	Any other assays such as tissue-based and organism-based assays

In continuation of the further development of Hit Dexter, we present here a refined set of machine learning models for frequent hitter prediction that cover biochemical assays and, for the first time, also cell-based assays. More specifically, we have developed three types of models: (i) models for target-based assays, (ii) models for cell-based assays designed to measure a specific protein-compound interaction, and (iii) models for an extended selection of cell-based assays, covering also cell-based assays designed to measure nonspecific interactions such as toxicity.

Each of the models is derived from a new, large, high-quality data set that we extracted from the PubChem Bioassay database and annotated manually. In addition to the extra tree (ET) classifiers employed previously, we are now exploring also k-nearest neighbors (KNN) classifiers as baseline models, as well as random forest (RF) and multilayer perceptron (MLP) classifiers. The best models presented in this work are available via a free web service at https://nerdd.univie.ac.at/hitdexter3/ and information on the assay data sets is provided as Supporting Information.

Materials and methods

Data set compilation

PubChem Bioassay data selection and annotation

The PubChem Bioassay Database [28–30] was queried for all assays with measured bioactivity data reported for at least 10,000 compounds (i.e., compounds with unique PubChem Compound IDs, CIDs). The data for the selected assays were downloaded and the labels “target type” and “bioactivity type” were assigned manually to each of these assays according to the definitions provided in Tables 1 and 2.

Following manual assay labeling, three different data sets were compiled:

- target-based assay data set: includes all data from assays with “target type” = “target-based” (which implies “bioactivity type” = “specific bioactivity”)
- cell-based assay data set: includes all data from assays with “target type” =“cell-based” AND “bioactivity type” = “specific bioactivity”
- extended assay data set: includes, in addition to the data included in the cell-based assay data set, all data from assays with “target type” = “cell-based”

The individual assays of the target-based assay data set were checked for the availability of Protein Gene Identifier (GI) information, which is utilized to retrieve protein sequence information from the NCBI Protein database [31] (the protein sequence information will be required, in a later step, for protein clustering and to ensure a diverse protein set). Sixty-six assays of the target-based assay data set had no GI or multiple GI annotations and were hence removed. In addition, seven assays of the target-based assay data set, four assays of the cell-based assay data set, and seven assays of the extended cell-based assay data set were removed because of disproportionally high hit rates (i.e. hit rates in excess of the average hit rate plus three standard deviations (σ), calculated over all assays of the respective data set). For the target-based assay data set, the six assays with the highest hit rates are all measuring CYP P450 enzyme activity. In the case of the cell-based assay data set, this concerns four assays, with hit rates of 59%, 55%, 17% and 15% (note that for approximately three quarters of the assays included in the cell-based assay data set their hit rates are below 1%). For the extended cell-based assay data set the seven assays with hit rates above 16% were removed.

Table 2
Definitions of values for the manually assigned label “bioactivity type”.

Label value	Description
specific bioactivity	Assays designed to measure a specific biological property such as the activity of an enzyme. Cytotoxicity assays are not included in this category. Counterscreen assays are included if they measure a specific biological effect. An example of a counterscreen assigned this label value is a luciferase counterscreen that is commonly employed to identify compounds which can cause interference in luciferase-based (bioluminescence) assays
nonspecific bioactivity	Assays that measure cell growth, cell viability, cytotoxicity, cell growth inhibition, or other nonspecific assay readouts
other	Assays that measure physicochemical processes (not bioactivities), DNA or RNA binding, etc.

Table 3
Data set sizes and compounds removed during chemical structure processing.

	Target-based assay data set	Cell-based assay data set	Extended cell-based assay data set
No. compounds in the data set prior to chemical structure processing	1,545,406	1,421,472	1,858,887
No. compounds removed due to invalid SMILES	1	3	9
No. compounds removed due to lack of a single, valid activity outcome[1]	45,184	23,259	53,984
No. compounds removed due to presence of elements uncommon to drug-like compounds	331	381	3151
No. compounds removed by the molecular weight filter	10,847	11,120	22,106
No. compounds in the final data set	1,489,043	1,386,709	1,779,637

[1] Compounds that were removed because of the lack of a valid activity outcome that can be derived from the raw data (i.e. compounds without a single annotated “Active” or “Inactive” assay outcome)

As the last filtering criterion, any assays without at least one compound measured as active and one compound measured as inactive were removed from the data set. For a complete overview of all assays removed during data preparation see Table SI_1.

Chemical structure processing

The SMILES notations of the 1,545,406 compounds covered by the target-based assay data set, the 1,421,472 compounds covered by the cell-based assay data set, and the 1,858,887 compounds covered by the extended cell-based assay data set were retrieved from the PubChem Bioassay database via the PubChem PUG REST interface [32]. The ChEMBL Structure Pipeline [33] (also known as “ChEMBL Compound Curation Pipeline”), was utilized to (i) neutralize charged molecules, (ii) remove salt and solvent components, and (iii) neutralize charged molecules once more (to cover cases where a charged component was removed during step ii). The technical description of this chemical structure preparation procedure is reported in Ref. [33].

Any compounds with molecular weight below 180 or above 900 Da were removed from the data set, as well as any compounds composed of any elements other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br and I. Molecules represented by more than one tautomer were merged to a single representation using the “canonalize” method implemented in the “TautomerEnumerator” class of RDKit [34] (version 2020.09.1). During this procedure the compounds were represented as RDKit molecules and were in a last step converted to canonical SMILES. Further duplicate compounds were removed based on identical SMILES. For an overview of the removed compounds see Table 3. For all additional data sets used within this study, including the ChEMBL 23 database [35], the dark chemical matter (DCM) data set compiled by Wassermann et al. [36], the data set of Dahlin et al. [37] (containing compounds that are known to cause interference in biological assays), and the data set of Borrel et al. [22] (containing compounds that were experimentally confirmed to cause false positive readouts in bioluminescence assays due to luciferase inhibition and/or autofluorescence), the same chemical structure standardization process was performed. Since the data set of Borrel et al. contains only CAS numbers as compound identifiers, the SMILES notations were fetched via the Chemical Identifier Resolver [38].

Extraction of activity data from the selected assays

For each of the selected assays, any compounds consistently (i.e. one or several times) labeled as “Active” were defined as active, and any compounds consistently labeled as “Inactive” were defined as inactive. Any compounds with contradicting assay outcomes (e.g. “Active” and “Inactive”, or “Active” and “Inconclusive”) were removed. A compound is treated as active on a cluster of proteins (see “Protein clustering”) if it is active on at least one protein of that cluster.

In order to ensure the consistency of predictions, compounds with identical Morgan2 fingerprints [39,40] (1024 bits) but differing promiscuity labels (e.g., symmetric molecules) were removed from the respective training set. For any compounds with identical Morgan2 fingerprints only one instance was kept in the respective training set.

Definition of the active-to-tested ratio (ATR)

The hit rate of a compound in biological assays is described as the active-to-tested ratio (ATR; Eq. (1)):

$$ATR = \frac{A}{T}, \tag{1}$$

where A is the number of assays a compound was tested active and T is the total number of assays a compound was tested in. For compounds, the terms hit rate and ATR are used interchangeably in this work.

Protein clustering

Based on the GIs assigned to the individual proteins, the FASTA sequences of the respective proteins were retrieved from the NCBI using the “Entrez” package of Biopython [41] (version 1.78). Protein clustering was performed using cd-hit [42] with the same parameters described in Ref. [25] (sequence identity= 60%; tolerance= 3). This resulted in 273 protein clusters using 296 unique proteins for the target-based assay data set.

Model development and hyperparameter optimization

Prior to model development, a random, stratified split of the data into a training set (90%) and a testing set (10%) was performed with the “train_test_split” method of the “model_selection” module of scikit-learn [43] (version 0.23.2). All models were trained and optimized on the training set. The final models were tested on the test set.

Morgan fingerprints and MACCS keys were calculated with RDKit, whereas 206 2D physicochemical property descriptors (meaning the complete set of available 2D descriptors) were calculated with the Molecular Operating Environment [44] (MOE; version 2020.09).

Default parameters were employed for generating machine learning models for the selection of a suitable set of descriptors, with the following exceptions: For the KNN classifier, the number of nearest neighbors to be taken into account for prediction (n_neighbors) was set to 1; for the RF and the ET classifiers, the class weight (class_weight) was set to “balanced”; for the MLP classifier (implemented in scikit-learn), the number of iterations was set to 1000 as some of the calculations did not converge within the default, 200 iterations.

The generation of the individual models was repeated for ten times, with different random states (i.e. 42 to 51), in order to compute the median and the variance of the performance metrics (details provided in the Results section). The final models were generated with random state=42 and the application of the synthetic minority oversampling technique (SMOTE version 0.7.0) [45].

Performance measurements and variance estimation

The MCC (Eq. (2)) was used as the primary measure of model performance. The MCC is a balanced metric that takes the true positive (TN), false positive (FP), true negative (TN) and false negative (FN) instances into account:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}} \tag{2}$$

The MCC returns values between -1 (total disagreement between prediction and observation) and +1 (perfect agreement).

The area under the (receiver operating characteristic) curve (AUC) was used as an indicator of the ranking performance of the models.

The tests for statistical significance were performed with the “ttest_rel” function of the “scipy.stats” module. The variance in the performance of the models (on the test data) was estimated by testing the models on ten randomly compiled subsets (80%) of the original test set.

Results

Analysis, annotation and refinement of PubChem Bioassay data

In order to develop a better understanding of the relevance of the data available from the PubChem Bioassay database for modeling the frequent hitter behavior of small molecules, we conducted a comprehensive analysis of the chemical and biological data.

With more than 297 Million measured bioactivities, the PubChem Bioassay database is the world’s largest, public collection of bioassay data [30]. It is also one of only a few data resources offering access to a large amount of high-throughput screening data. The number of measured bioactivities recorded per assay varies greatly across the individual assay data sets, from a single compound to 646,275 compounds (Table 4).

We decided to base our work on the 1180 (i.e. 474 + 706) assay data sets containing measurements for at least 10,000 compounds because these data sets offer a good trade-off between data quality and coverage. The vast majority of these data sets have been generated by the most reputable HTS facilities (including the Scripps Research Institute, the Sanford-Burnham Medical Research Institute, The Broad Institute of MIT and Harvard, and the NIH/National Center for Advancing Translational Sciences (NCATS)), for which reason a high standard in HTS can be expected.

Table 4
Size of the PubChem Bioassay data sets.[1]

Number of assays in the PubChem Bioassay database	Number of measured compounds
587,477	1
633,294	2 to 99
5082	100 to 999
1403	1000 to 9999
474	10,000 to 99,999
706	100,000 to 646,275 (maximum)

[1] Numbers referring to the raw, unprocessed PubChem Bioassay database.

In preparation of model development, we manually annotated the 1180 assay data sets according to the “assay type” (i.e. target-based, cell-based, other; see Table 1 for exact definitions) and “bioactivity type” (i.e. specific bioactivity, nonspecific bioactivity, other; see Table 2 for exact definitions). Models for the prediction of frequent hitters in biochemical (i.e. target-based) assays will be built on all (359) assay data sets labeled as “target-based” (which implies the “bioactivity type” value “specific bioactivity”) and annotated with exactly one Protein Gene Identifier (GI; the GI will be utilized later to obtain protein sequence information to quantify the relatedness of proteins; the requirement for assays to be assigned exactly one GI ensures that the assay is designed to measure one particular protein of interest). Similarly, models for cell-based assays designed to measure a specific activity will be built on all (369) assay data sets labeled as “cell-based” AND “specific bioactivity”. Models will also be derived from an extended set of cell-based assays that includes data from an additional 250 cell-based assays labeled “nonspecific bioactivity”. These additional, cell-based assays measure non-specific properties such as cell viability or cytotoxicity. A list of the Assay Identifiers (AIDs) for the three assay data sets is provided in Table SI_2.

We also set steps to address two important biases in the assay data set collection. The first bias results from assays with unusually high hit rates. In target-based assays, high hit rates are often related to the measurement of highly promiscuous proteins such as CYP enzymes. In cell-based assays, high hit rates can be related, for example, to cytotoxicity or high assay sensitivity. Compounds which have been measured, for whatever reason, in several of these assays may, in consequence, be identified as frequent hitters, regardless of whether their activities are focused on a number of closely related proteins or observed across a range of distinct proteins.

The average hit rate of the 359 target-based assays is 0.009. However, a small number of assays has much higher hit rates, up to 0.252 (Fig. 1). Similarly, the average hit rate for the 369 cell-based assays is 0.014, with a small number of assays having much higher hit rates, up to 0.588. For the extended set of 619 cell-based assays, the average hit rate is 0.023, with the maximum at 0.588. For the reasons discussed above we decided to remove any assays with hit rates exceeding the average hit rate plus three σ. This concerned seven, four and seven assays of the target-based, cell-based and extended cell-based assay data set, respectively.

The second bias is introduced by groups of assays measuring related proteins. Related proteins have a high likelihood of binding the same small molecules, meaning that, for example, assay data sets with a strong representation of protein kinase targets will likely show high hit rates for protein kinase inhibitors. Models for frequent hitter prediction that are trained on such data would likely flag any kinase inhibitor as a frequent hitter, which is not the intended behavior of these models.

In order to address the bias introduced by the overrepresentation of groups of related proteins, we clustered the target-based assay data set according to the amino acid sequences of the target proteins (note that the clustering was not performed for the cell-based assays data sets because cell-based assays may report activities for a number of different proteins). More specifically, all data sets related to proteins with an amino acid sequence identity exceeding 60% were merged into a cluster (see Materials and Methods for details). This clustering procedure resulted in 296 protein clusters (starting from 352 proteins covered by the target-based assay data set).

After addressing the two important biases, in the final processing step the molecular structures contained in the data sets were processed and checked for correctness. Any problematic instances were removed, as outlined in Table 3 and described in the Materials and Methods section in full detail. This resulted in a target-based, a cell-based and an extended cell-based assay data set consisting of 1,489,043, 1,386,709 and 1,779,637 unique compounds with at least one confirmed target protein, respectively.

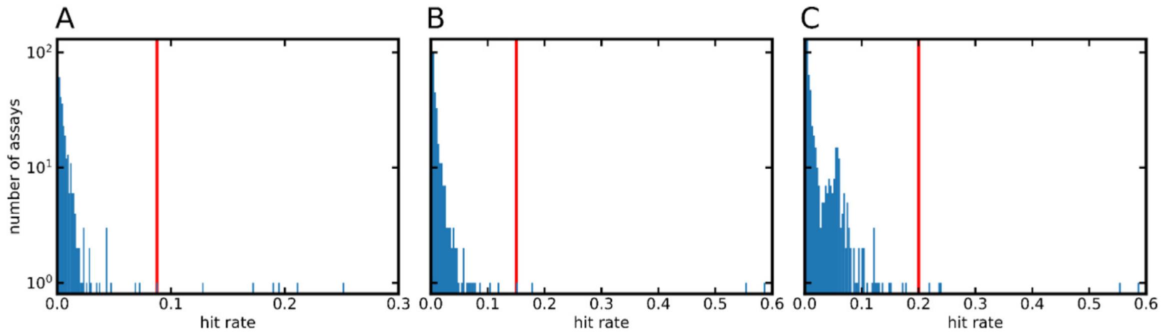

Fig. 1. Histograms (200 bins each) showing the hit rates of the assays included in the (A) target-based, (B) cell-based, and (C) extended cell-based assay data sets. The red line marks the mean hit rate + 3σ. Note that the scales of the x-axes differ for the three diagrams.

Table 5
Composition of the training and test sets.

Data set	Promiscuity class	Class definitions	No. compounds in the training set	No. compounds in the test set
target-based assay data set	HPROM[1]	ATR > 0.053	4614	550
	PROM	ATR > 0.022	20274	2303
	NPROM	ATR < 0.007	219061	24483
cell-based assay data set	HPROM[1]	ATR > 0.058	5578	616
	PROM	ATR > 0.025	24913	2825
	NPROM	ATR < 0.008	226382	25427
extended cell-based assay data set	HPROM[1]	ATR > 0.070	5135	538
	PROM	ATR > 0.030	24673	2776
	NPROM	ATR < 0.010	235241	26398

[1] The compounds labeled as HPROM are a subset of the compounds labeled as PROM.

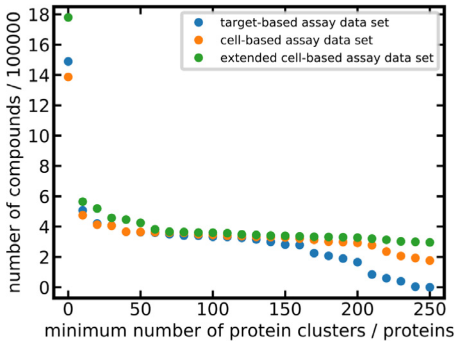

Fig. 2. Data set size (number of compounds) as a function of the minimum number of protein clusters (in the case of target-based assays) or proteins (in the case of cell-based assays) for which measured data are available.

Analysis of compound hit rates and assignment of promiscuity class labels

The ATR (Eq. (1)) can be used to assign categorical promiscuity values to compounds, such as “non-promiscuous”, “promiscuous” or “highly promiscuous”. The significance and robustness of the ATR depends on the quality and quantity of the underlying data: the higher the value of *T* (i.e. the total number of assays a compound was tested in), the more robust the ATR. The main advantage of the ATR over alternative metrics is its interpretability as it reflects the hit rate of a compound.

In this work, we set the minimum threshold of *T* for a compound to be included in the data sets used for model development to 100, which represents a good balance between ATR quality and coverage (Fig. 2). This filtering procedure resulted in a set of 332,653 compounds measured in target-based assays, 345,743 compounds measured in cell-based assays designed to measure a specific bioactivity, and 360,094 compounds measured in an extended set of cell-based assays.

Based on the ATR thresholds reported in Table 5, all compounds were assigned a promiscuity label: highly promiscuous (HPROM), promiscuous (PROM) or non-promiscuous (NPROM). According to these definitions, roughly 2% of the compounds are labeled HPROM across the three assay data sets. Likewise, the percentages of compounds labeled PROM were around 9% across the three assay data sets (note that all HPROM compounds are also part of the PROM subset). The percentages of compounds labeled NPROM are approximately 90% across the three assay data sets (Table 5 and Fig. 3).

To obtain a training set and a test set (separately for all three data sets), a stratified random split was performed to obtain 90% training data and 10% test (hold out) data. Following a fingerprint-based data merging procedure (i.e. merging of instances having identical fingerprints and identical class labels, and removal of any instances having identical fingerprints but conflicting class labels; see Materials and Methods for details) the target-based, cell-based and extended cell-based training sets contain 243,949, 256,873 and 265,049 compounds, respectively (Table 5).

As shown in Table 5, the average ATR across the extended set of cell-based assays is higher than for the cell-based and the target-based assay sets, suggesting that non-specific interactions are likely to play an important role in the assays exclusive to the extended set of cell-based assays (i.e. cell-based assays not designed to measure specific biological processes but to capture properties such as cell-viability and cytotoxicity).

Analysis of the chemical space covered by the training sets

The chemical space covered by the training set is a decisive factor for the applicability domain of a model. In order to obtain an understanding of the relevance of our three training sets to early drug discovery we run a pairwise comparison of the molecular structures included in these training sets and all molecular structures included in the ChEMBL database. Fig. 4 shows the distributions of the pairwise, maximum Tanimoto coefficients based on Morgan2 fingerprints (with a length of 1024 bits) for the three data sets vs. the ChEMBL database. The distributions are similar for the three data sets, with approximately

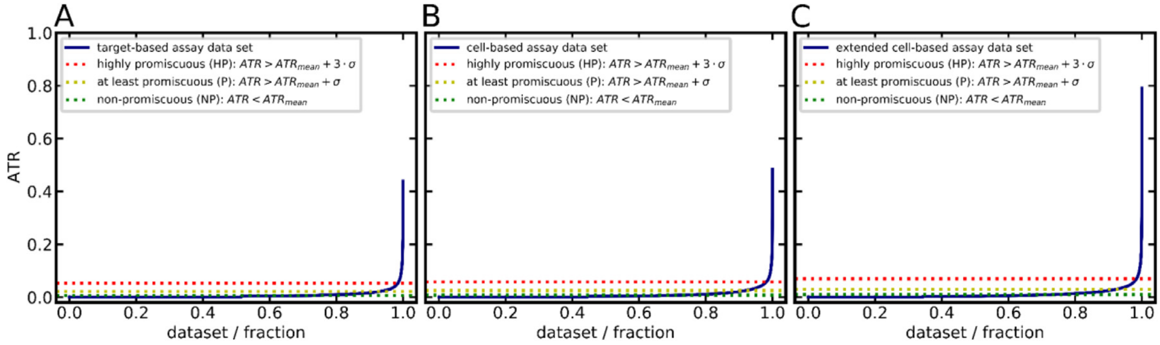

Fig. 3. ATR distribution among compounds of the (A) target-based assay data set, (B) cell-based assay data set, and (C) extended cell-based assay data set.

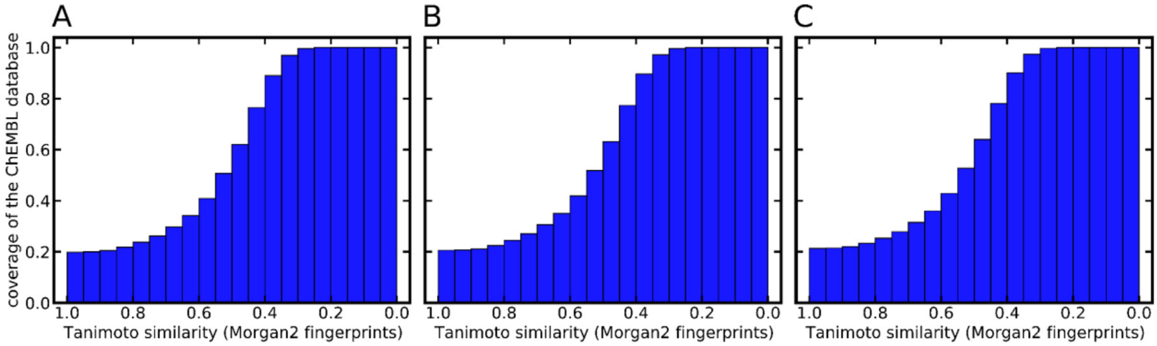

Fig. 4. Cumulative coverage of the compounds included in the ChEMBL database by the compounds included in the (A) target-based assay data set (B) cell-based assay data set, and (C) extended cell-based assay data set.

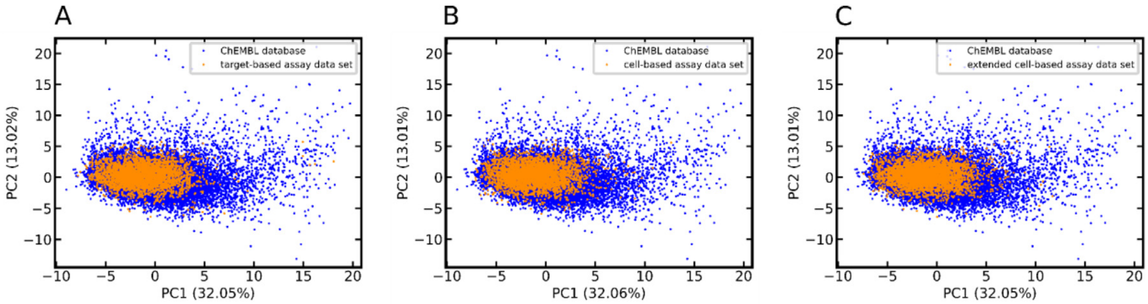

Fig. 5. PCA of the ChEMBL database and the (A) target-based assay set, (B) cell-based assay set, and (C) extended cell-based assay set. The PCA is derived from the 44 2D molecular property descriptors (see Table S1 in Ref. [46]) implemented in MOE. For the sake of clarity, only 1% of the data points (randomly selected) are visualized. The numbers in parentheses report the variance explained by the respective principal component (PC).

50% of the compounds in the ChEMBL database represented by at least one compound in the respective training set with a Tanimoto coefficient of 0.5 or higher.

The Principal Component Analysis (PCA) scatter plots presented in Fig. 5 show that the areas in chemical space that are most densely populated with the compounds from the ChEMBL23 database are also well represented by the assay data sets used for model training. However, there are a significant number of compounds included in the ChEMBL database that are chemically distinct from those represented by the training sets. These are in particular compounds with PC1 values greater than 10, which account for 2.5% of the total number of compounds of the ChEMBL database. Visual inspection of these compounds reveals that they are unusually large, with molecular weight between 575 and 900 Da.

The target-based and the cell-based assay data sets (training data only) have an overlap of 180,278 compounds (representing 75% of the target-based and 72% of the cell-based assay data set, respectively). Only 13,045 (7%) of these compounds have contradicting promiscuity labels (with HPROM treated as a subset of PROM). At first sight the level of agreement between readouts from target-based and cell-based assays seems surprisingly high. However, a closer look reveals that the agreement stems primarily from compounds consistently labeled as NPROM. Among the 20,481 compounds present in both data sets and labeled as PROM in at least one of them, only 6616 (32%) have identical class labels. This indicates that target-based and cell-based assays perform indeed differently and that they should be represented by dedicated models.

Development of machine learning models for compound promiscuity prediction

Two types of classifiers were generated for the target-based, cell-based and extended cell-based assay data sets: classifiers discriminating HPROM from NPROM compounds, and classifiers discriminating PROM from NPROM compounds.

Identification of the best setup for model generation

In order to identify the best setup for model generation we tested all possible combinations of four machine learning algorithms (i.e. KNN,

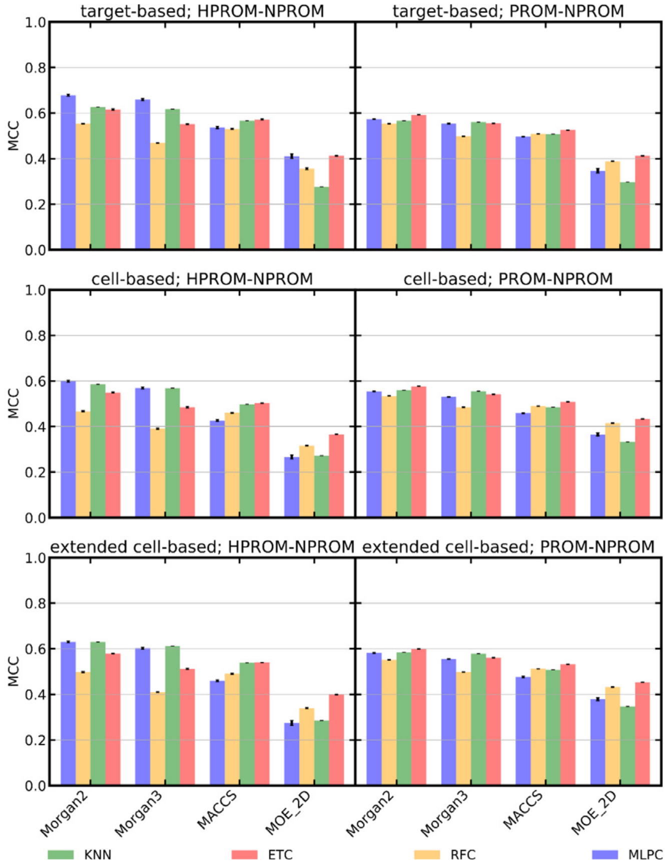

Fig. 6. Performance (quantified as MCC) of machine learning models trained on different types of descriptors. The variance of the ten experiments (each using a distinct random seed between 42 to 51; see Materials and Methods for details) is indicated by error bars.

ET, RF, MLP) and four sets of descriptors (i.e. Morgan2 and Morgan3 fingerprints, each of 1024 bits in length, MACCS keys, and the complete set of 206 2D physicochemical property descriptors implemented in MOE, referred to as "MOE_2D") within a 10-fold cross-validation framework. For each setup ten of these cross-validation experiments were performed using distinct random seeds. This allowed, for each setup, the calculation of a standard deviation that is independent of the cross-validation.

As shown in Fig. 6, the task of discriminating HPROM from NPROM compounds (MCCs of up to 0.679) is simpler than that of discriminating PROM from NPROM compounds across the three assay data sets (the MCC of the best HPROM-NPROM classifier, 0.679, is significantly higher than that of the best PROM-NPROM classifier, 0.599; p-value 2.48×10^{-12}). This is expected because of the larger ATR margin between the HPROM and the NPROM class (margin of 3σ) than between the PROM and the NPROM class (margin of 1σ). No substantial differences in model performance were observed with respect to the type of assay modeled: the best setups yielded comparable MCCs for the target-based assay set (MCCs 0.679 and 0.592 for HPROM-NPROM and PROM-NPROM classification, respectively), cell-based assay set (MCCs 0.602 and 0.577, respectively), and extended cell-based assay set (MCCs 0.631 and 0.599, respectively).

The differences in model performance that can be attributed to the model algorithms are rather small, on average 0.104 in MCC. The maximum difference in MCC observed for any model trained on identical input (i.e. same data set and same descriptor set) was 0.224. Overall, the MLP classifiers performed best in HPROM-NPROM classification (the MCC of the best MLP classifier, 0.679, is significantly higher than that of the second-best model, a KNN model that obtained an MCC of 0.630; p-value of 8.81×10^{-11}), and the ET classifiers performed best in PROM-NPROM classification (the MCC of the best ET classifier, 0.599, is significantly higher than that of the second-based model, a KNN model that obtained an MCC of 0.585; p-value of 7.79×10^{-11}). Interestingly, in this cross-validation scenario the simple one-nearest neighbor approach performed almost as well as the more complex machine learning algorithms (MCCs of up to 0.587; p-value of 7.79×10^{-11} against the best MLP classifier).

In contrast to what we observed for the machine learning algorithms, the differences in model performance that can be attributed to the molecular descriptors were, in part, substantial. On average, the best performance was obtained by models trained on Morgan2 fingerprints (MCC averaged over all models trained on Morgan2 fingerprints: 0.679). They were closely followed by the models based on Morgan3 fingerprints (MCC averaged over all models trained on Morgan3 fingerprints: 0.659; the difference in the average MCC of models trained on Morgan2 and Morgan3 fingerprints is significant, with a p-value of 1.10×10^{-79}). The MOE_2D physicochemical property descriptors and the MACCS keys yielded models that are clearly inferior, with MCCs not exceeding 0.453 and 0.572, respectively.

The highest MCC during model optimization (0.679) was obtained by the HPROM-NPROM MLP classifier in combination with Morgan2 fingerprints (the MCC of the second-best classifier, which is the respective model trained on Morgan3 fingerprints, was 0.659; the difference in the MCCs is significant, with a p-value of 3.98×10^{-5}).

Hyperparameter optimization

Focusing now on Morgan2 fingerprints, in the next phase of model development we optimized the hyperparameters of the individual algorithms (i.e. KNN, ET, RF and MLP). More specifically, we conducted a grid search within a 10-fold cross-validation framework to identify the hyperparameters yielding the best performing models for a particular combination of machine learning algorithm and descriptors in terms of MCC (averaged over the respective HPROM-NPROM and PROM-NPROM classifiers for the three assay data sets). An overview of the explored hyperparameters and value ranges, as well as the selected hyperparameter values, is provided in Table 6.

The impact of individual hyperparameter settings on model performance is generally small (Table SI_3). The largest improvement in MCC observed during hyperparameter optimization was 0.037 (for the PROM-NPROM MLP classifier trained on the cell-based assay data set; the optimized classifier performed significantly better than the classifier using default hyperparameters; p-value of 1.01×10^{-10}). The AUC values improved consistently with the MCCs (Table SI_3), except for KNN, for which the MCCs increased with fewer numbers of neighbors while the AUC values decreased. In the case of the RF and ET classifiers, gains in performance beyond 200 estimators were marginal and do not justify the additional demands in computational power and memory. The same is true for the MLP classifier, for which we identified 250 as the most suitable number of perceptrons for our purposes.

The best of all models (an HPROM-NPROM MLP classifier for target-based assays; single hidden layer with 250 perceptrons; activation function relu) yielded an MCC of 0.686 (the optimized classifier performed significantly better than the classifier using default hyperparameters; p-value of 3.79×10^{-3}). The models chosen from hyperparameter optimization are listed in Table 7.

Model performance as a function of the size of the training set

In order to determine the impact of the size of the training set on model performance we trained and tested the optimized HPROM-NPROM and PROM-NPROM MLP classifiers on fractions of 0.01 to 1.00 of the full training sets (within a 10-fold cross-validation framework). From Fig. 7 it is observed that models built on just 20% of the data already achieve good performance (MCCs between 0.434 and 0.524). Larger data sets may add significant value but primarily if they cover

Table 6
Overview of hyperparameters optimized during grid search within a 10-fold cross-validation framework.[1]

Classifier	Parameter	Values
KNN	n_neighbors (number of neighbors considered)	1, **3**, 5, 10
RF,	n_estimators (number of trees)	50, 100, **200**, 300, 400, 500
ET	max_features (features taken into account for best split search)	'sqrt', 'none', **'0.2'**, '0.4', '0.6', '0.8'
MLP	hidden_layer_sizes (number of perceptrons per layer)[2]	50, 100, **250**, 500
	hidden_layer_sizes (number hidden layer)[2]	**1**, 2, 3, 4, 5
	activation (activation function)	**'relu'**, 'tanh', 'logistic'

[1] The hyperparameter values indicated in bold are those we identified as most suitable for model building. These values were used for the generation of the final models.

[2] hidden_layer_sizes accepts two values: one for the number of perceptrons per layer and one for the number of hidden layers.

Table 7
Cross-validation and test set performance of the best models of different types.

Data	Classification	Machine learning algorithm	Cross-validation performance[1]					Test set performance				
			MCC[2]	AUC[2]	Balanced accuracy	Sensitivity	Specificity	MCC	AUC	Balanced accuracy	Sensitivity	Specificity
target-based assay data set	HPROM-NPROM	KNN	0.624	0.843	0.733	0.469	0.998	0.376	0.909	0.871	0.818	0.924
		ET	0.630	0.964	0.734	0.469	0.998	0.508	0.966	0.677	0.357	0.997
		RF	0.588	0.964	0.695	0.392	0.999	0.484	0.965	0.677	0.358	0.996
		MLP	0.686	0.946	0.796	0.595	0.997	0.648	0.949	0.798	0.601	0.995
	PROM-NPROM	KNN	0.587	0.844	0.745	0.506	0.984	0.412	0.864	0.816	0.822	0.809
		ET	0.597	0.928	0.746	0.504	0.986	0.518	0.910	0.721	0.464	0.977
		RF	0.578	0.929	0.718	0.445	0.991	0.518	0.908	0.731	0.489	0.973
		MLP	0.599	0.907	0.777	0.578	0.975	0.580	0.899	0.768	0.562	0.974
cell-based assay data set	HPROM-NPROM	KNN	0.571	0.827	0.704	0.41	0.998	0.338	0.899	0.857	0.812	0.902
		ET	0.572	0.950	0.697	0.395	0.998	0.531	0.940	0.692	0.387	0.997
		RF	0.514	0.947	0.651	0.303	0.999	0.520	0.932	0.692	0.387	0.996
		MLP	0.611	0.929	0.754	0.512	0.996	0.576	0.915	0.767	0.541	0.992
	PROM-NPROM	KNN	0.566	0.845	0.74	0.501	0.979	0.413	0.860	0.809	0.834	0.783
		ET	0.593	0.925	0.747	0.511	0.983	0.551	0.911	0.743	0.513	0.973
		RF	0.572	0.925	0.717	0.445	0.989	0.543	0.910	0.747	0.525	0.968
		MLP	0.579	0.910	0.77	0.571	0.969	0.561	0.901	0.764	0.562	0.965
extended cell-based assay data set	HPROM-NPROM	KNN	0.600	0.842	0.721	0.443	0.998	0.340	0.895	0.858	0.798	0.919
		ET	0.599	0.956	0.708	0.417	0.999	0.527	0.944	0.683	0.368	0.998
		RF	0.537	0.956	0.662	0.325	0.999	0.519	0.943	0.686	0.374	0.997
		MLP	0.639	0.939	0.764	0.532	0.997	0.567	0.921	0.753	0.511	0.994
	PROM-NPROM	KNN	0.586	0.854	0.749	0.516	0.981	0.429	0.871	0.819	0.835	0.804
		ET	0.618	0.935	0.757	0.529	0.985	0.565	0.923	0.742	0.506	0.978
		RF	0.590	0.934	0.725	0.459	0.990	0.554	0.920	0.748	0.523	0.973
		MLP	0.607	0.921	0.783	0.593	0.972	0.587	0.910	0.781	0.596	0.967

[1] The optimized hyperparameters are reported in Table 6.

[2] The variance is reported in Table SI_3.

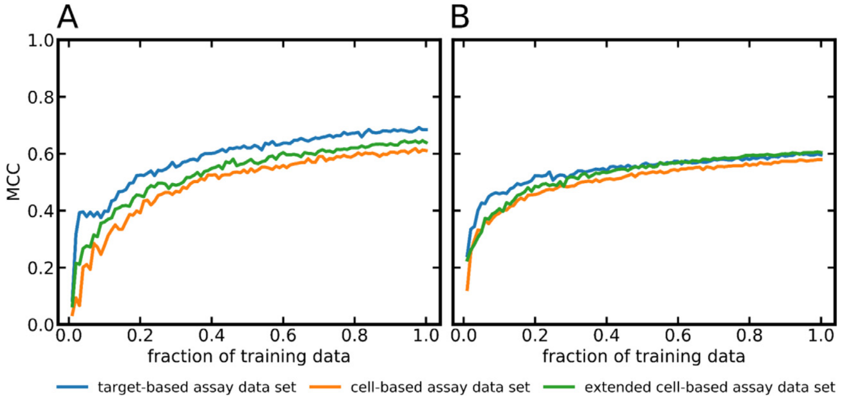

Fig. 7. Performance (quantified as MCC) of the hyperparameter-optimized MLP classifiers trained on the target-based, cell-based and extended cell-based assay data set as a function of training set size. (A) HPROM-NPROM classifiers, (B) PROM-NPROM classifiers. For each data point the variance was calculated from ten calculations (with different random seeds; see Materials and Methods for details). Because the variance values were within the range of 1.4 × 10^{-6} to 5.9 × 10^{-4} they are not visualized in these graphs.

distinct areas in the chemical space and hence contribute to the extension of the applicability domain of the model.

Evaluation of the final machine learning models

A total of 24 final models of different types (i.e. models trained on the full training set, balanced with SMOTE; see Materials and Methods for details) were tested on the holdout data (i.e. 10% of the data that was set aside prior to model building). The 24 models result from the combination of three different training sets (i.e. target-based, cell-based and extended cell-based assay data set), four machine learning algorithms (KNN, ET, RF, MLP), and two different types of classification (i.e. HPROM-NPROM and PROM-NPROM). All of these models are built on Morgan2 fingerprints and utilize the

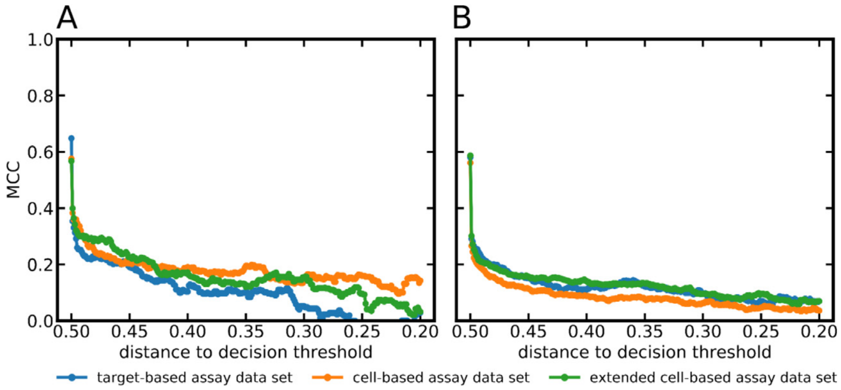

Fig. 8. Performance (quantified as MCC) of (A) the HPROM-NPROM MLP classifiers and (B) the PROM-NPROM MLP classifiers as a function of the distance of the predicted class probability to the decision threshold (the decision threshold applied to all models in this study is 0.5). For each data point the variance was estimated by running the models on ten randomly selected subsets of the test data (see Materials and Methods for details). Because the variance values were within the range of 9.6×10^{-6} to 4.5×10^{-3} they are not visualized in these graphs.

hyperparameters sets optimized during the previous cross-validation experiments.

Model performance on the test set

The average MCC obtained by the 24 models on the respective test sets was 0.507, which is 0.087 lower than in the cross-validation scenario (Table 7). Overall, the decrease in performance (on the test set compared to cross-validation) was more pronounced for the HPROM-NPROM classifiers (average decline in MCC 0.103) than the PROM-NPROM classifiers (average decline 0.070). The steeper drop in performance observed for the HPROM-NPROM classifiers is likely related to the fact that the number of compounds representing the active class is much lower for the HPROM-NPROM training set (approximately 5000 compounds) than for the PROM-NPROM training set (approximately 23,000 compounds). The best MCC among all HPROM-NPROM classifiers was obtained by the MLP classifier trained on the target-based assay data set (MCC 0.648). The best-performing PROM-NPROM classifier was the MLP classifier trained on the extended cell-based assay data set (MCC 0.587).

Importantly, a substantial decrease in performance was observed for the HPROM-NPROM KNN classifiers for all three assay data sets (for example, the KNN classifier of the target-based assay data set; three nearest neighbors; cross-validation MCC 0.624; test set MCC 0.376) and also the PROM-NPROM KNN classifiers for all three assay data sets (for example, the KNN classifier of the target-based assay data set; three nearest neighbors; cross-validation MCC 0.587; test set MCC 0.421). This decline in the performance may be related to model overfitting.

In contrast to the observations made for the KNN, the MCC values obtained by the RF and ET classifiers remained stable. The maximum decline in MCC observed for these models was 0.122. The MLP classifiers showed the most robust performance across the three data sets and the two types of classifications (i.e. HPROM-NPROM and PROM-NPROM), with a maximum decline in MCC of 0.072. For this reason these six MLP classifiers were selected to form the Hit Dexter 3 set of machine learning models and they were investigated further regarding their applicability domains.

Prediction success as a function of the distance of the predicted class probability from the decision threshold

Commonly, a directly proportional relationship is observed between the reliability of class assignments and the distances between the predicted class probabilities and the decision threshold. This holds true also for the Hit Dexter 3 models. Fig. 8 shows that class assignments based on predicted probabilities close to 0 or close to 1.0 (this corresponds to a distance to the decision threshold of approximately 0.5 as we apply a decision threshold of 0.5 in all cases) are particularly reliable (MCC values of up to 0.648) for the Hit Dexter 3 models. The MLP classifiers differentiating PROM and NPROM compounds for the three data sets report predicted class probabilities greater than 0.95 or smaller than 0.05 for on average 97% of the compounds in the test set.

Prediction success as a function of the distance of the test compounds to the training set

It is expected that test compounds that are structurally dissimilar from those represented by the training data pose greater challenges to the model than those that are structurally related. Fig. 9 shows that the Hit Dexter 3 models perform well for compounds represented by at least one molecule in the training set that is structurally related (i.e. having at least one compound in the training set for which the pairwise Tanimoto coefficient based on Morgan2 fingerprints is at least 0.7). Predictions for compounds that are more dissimilar to those represented in the training data are less reliable and should be considered with the necessary caution.

Prediction success as a function of the applied decision threshold

In the current context, the decision threshold applied to a classifier decides on when a compound is classified as a frequent hitter or as a non-promiscuous compound. The default value for the decision threshold is 0.5. There are some use cases where a different decision threshold may be preferred. For example, in cases where the detection of frequent hitters is a priority (i.e. prioritization of sensitivity over specificity), a lower decision threshold may result in better predictions. Fig. 10 visualizes the effects of changes in the decision threshold on the MCC, balanced accuracy, sensitivity and specificity. The fact that the curves remain fairly stable until the decision threshold approaches extreme values (i.e. values close to 0.0 or 1.0) indicates that the classifiers produce clear predictions for most compounds. In cases where sensitivity is of primary importance, users are advised to consider any compounds with predicted probabilities greater than 0.0 as potential frequent hitters.

Predicting frequent hitters in cell-based assays with models trained on data from target-based assays and vice versa

Compounds may behave differently in target-based and cell-based assays, in particular also with regard to their assay interference and frequent hitter behavior. In order to obtain a better understanding of the relevance and value of dedicated models for the prediction of frequent hitters in target-based and cell-based assays, we compared the performance of the MLP classifiers on test data of the same assay domain to their performance on test data of the other assay domain (i.e. classifiers trained on target-based assay data were tested on cell-based assay test set and vice versa).

As shown in Fig. 11, the PROM-NPROM MLP classifiers trained and tested on data from the same assay domain clearly outperformed those trained on the other domain. The graphs also indicate that the difference in performance is not the result of differences in the chemical space covered by the individual data sets: even for test compounds that are structurally closely related to those represented by the training set, models trained on target-based assay data do not perform well on cell-based assay data and vice versa.

For the cell-based assay test data, the MCC of the PROM-NPROM MLP classifier trained on target-based assay data was just 0.189 (vs.

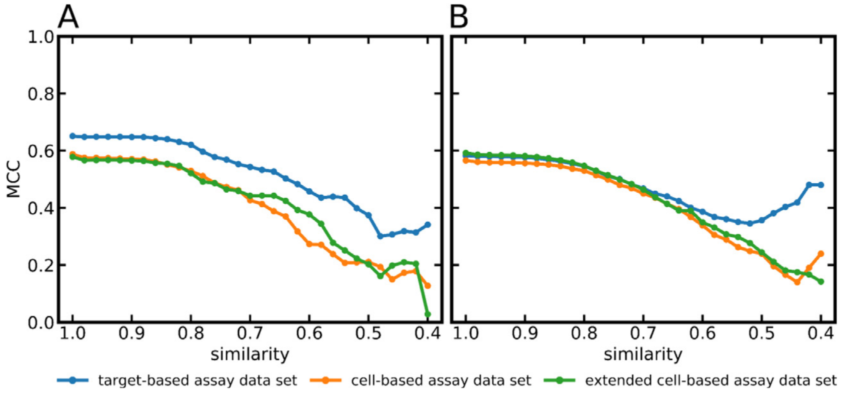

Fig. 9. Performance (quantified as MCC) of (A) the HPROM-NPROM MLP classifiers and (B) the PROM-NPROM MLP classifiers as a function of the structural similarity (measured as Tanimoto similarity on Morgan2 fingerprints) between the compounds in the test and the training sets. For each data point the variance was estimated by running the models on ten randomly selected subsets of the test data (see Materials and Methods for details). Because the variance values were within the range of 1.9×10^{-6} to 2.9×10^{-3} they are not visualized in these graphs.

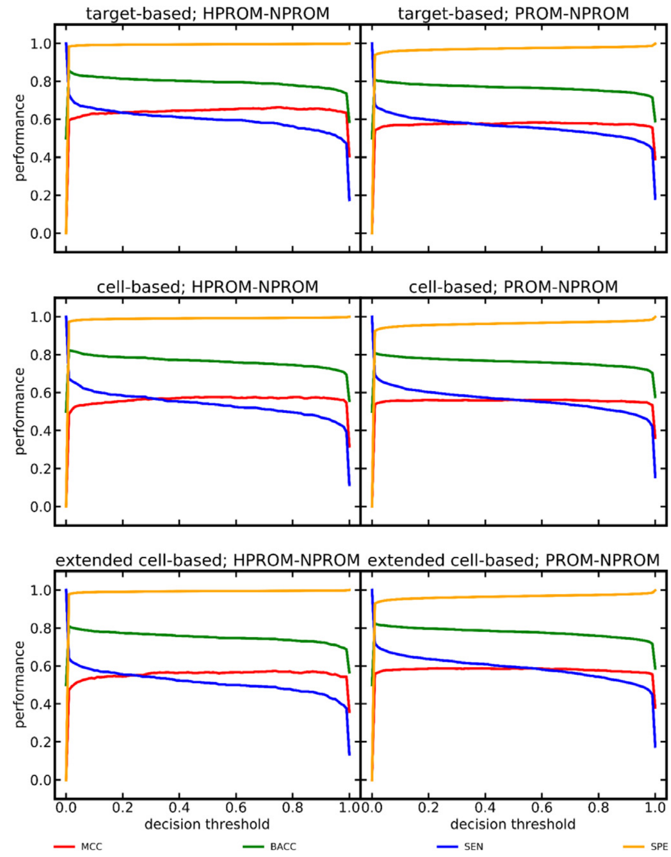

Fig. 10. Performance of the Hit Dexter 3 models as a function of the selected decision threshold.

0.561 for the classifier trained on the cell-based assay data; any compounds present in the training and in the test set are disregarded in the calculation of these MCC values). Likewise, for the target-based assay test data the MCC for the PROM-NPROM MLP classifier trained on cell-based assay data was just 0.235 (vs. 0.580 for the classifier trained on the target-based assay data). These results show that target-based and cell-based assays clearly behave differently and that dedicated models are required to adequately predict their behavior.

Model performance on dark chemical matter

We tested the Hit Dexter 3 models also on the dark chemical matter (DCM) data set compiled by Wassermann et al. The DCM data set consists of 135,489 compounds which have been tested in at least 100 target-based and cell-based assays without a single positive assay outcome. These compounds are not necessarily without activity on any protein but they are unlikely frequent hitters.

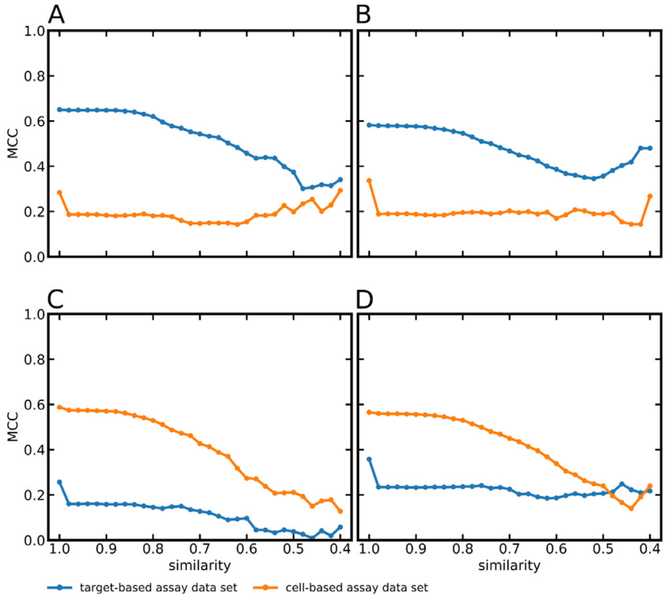

Fig. 11. Performance (quantified as MCC) of MLP classifiers as a function of the pairwise similarity between the test compound and its nearest neighbor in the training set (measured as Tanimoto coefficient derived from Morgan2 fingerprints). (A) HPROM-NPROM MLP classifier trained on the target-based assay data set, (B) PROM-NPROM MLP classifier trained on the target-based assay data set, (C) HPROM-NPROM MLP classifier trained on the cell-based assay data set, (D) PROM-NPROM classifier trained on the cell-based assay data set. For each data point the variance was estimated by running the models on ten randomly selected subsets of the test data (see Materials and Methods for details). Because the variance values were within the range of 9.2×10^{-6} to 6.2×10^{-3} they are not visualized in these graphs.

In the test of the Hit Dexter 3 models on the DCM data set, any test compounds also present in the training set of the respective models were disregarded (leaving 24,111 to 37,711 DCM compounds for testing, depending on the individual training set). The target-based, cell-based and extended cell-based HPROM-NPROM MLP models correctly assigned 99.0%, 98.6% and 98.7% of the DCM compounds to the NPROM class. In comparison, the percentage of correct assignments of the PROM-NPROM models were 95.4%, 93.7% and 93.6%, respectively. This result corroborates the validity (in particular the specificity) of the models.

Model performance on known bad actors

To test the capacity of the six Hit Dexter 3 models to identify bad actors in biological assays, we ran the models on two recently published

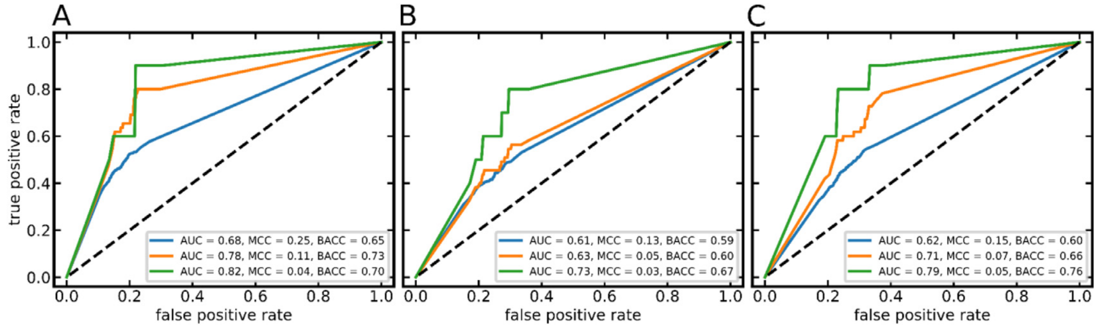

Fig. 12. ROC curves obtained with the Hit Dexter 3 PROM-NPROM classifiers trained on (A) target-based assay data, (B) cell-based assay data, and (C) extended cell-based assay data, and tested on the data set of Borrel et al. The compounds of the test set were annotated as frequent hitters according to Definition 1 (blue curves), Definition 2 (orange curves) and Definition 3 (green curves).

data sets containing experimentally confirmed bad actors (cave: bad actors are not necessarily frequent hitters; Hit Dexter 3 is designed to identify frequent hitters). As in all previous experiments, we disregarded all compounds present in these test sets that are also part of the training data of the individual models.

The first data set is from the work of Dahlin et al. [37]. This data set consists of 1139 compounds that are known to cause false readouts in various types of biological assays. For the 891 to 1002 test compounds not represented in the training set of the individual models, the target-based, cell-based and extended cell-based HPROM-NPROM MLP classifiers assigned 24.1%, 25.5% and 23.0% of all compounds to the HPROM class. The models distinguishing PROM and NPROM compounds flagged 40.3%, 39.3% and 40.3% as promiscuous, respectively. Because bad actors are not necessarily frequent hitters (and vice versa), the percentages of compounds reported by our models as PROM or HPROM are within the expected range.

The second data set is from the work of Borrel et al. [22]. This data set contains 8947 compounds, 891 of which have been observed to cause false positive readouts in bioluminescence assays due to luciferase inhibition (in one out of one assay) or autofluorescence (in one or several out of 24 assays), and 8056 compounds that are confirmed to behave benign in these assays. We explored three ways of translating the measurements recorded with these interference assays into “frequent hitter data"”: Compounds were labeled as frequent hitter if they produced

Definition 1. a false-positive signal in at least one assay (luciferase assay or assay to test for autofluorescence).

Definition 2. a false-positive signal in the luciferase assay AND at least one of the (24) assay setups to test for autofluorescence.

Definition 3. a false-positive signal in the luciferase assay AND at least nine of the (24) assay setups to test for autofluorescence.

All other compounds were labeled as non-promiscuous.

As shown in Fig. 12, the Hit Dexter 3 models reached AUC values of up to 0.82 (PROM-NPROM MLP classifier trained on the target-based assay data set, in combination with Definition 3), which confirms the ability of the models to rank bad actors early in a rank-ordered list of compounds. The MCC and balanced accuracy indicate moderate perfor-

Table 8
Performance of the Hit Dexter 2.0 and Hit Dexter 3 machine learning models on the DCM data sets and the known bad actors data set of Dhalin et al.

Hit Dexter version	training set	test set	classification	number of compounds in test set[1]	number of compounds	fraction of compounds
					correctly classified as DCM or bad actors[4]	
Hit Dexter 2.0	PSA data[2]	DCM	HPROM-NPROM	20,894	20,806	0.996
Hit Dexter 2.0	CDRA data[3]	DCM	HPROM-NPROM	42,567	42,341	0.995
Hit Dexter 3	target-based assay data	DCM	HPROM-NPROM	37,711	37,317	0.990
Hit Dexter 3	cell-based assay data	DCM	HPROM-NPROM	30,967	30,529	0.986
Hit Dexter 3	extended cell-based assay data	DCM	HPROM-NPROM	24,327	24,015	0.987
Hit Dexter 2.0	PSA data[2]	DCM	PROM-NPROM	20,872	20,472	0.981
Hit Dexter 2.0	CDRA data[3]	DCM	PROM-NPROM	41,587	40,080	0.964
Hit Dexter 3	target-based assay data	DCM	PROM-NPROM	37,421	35,695	0.954
Hit Dexter 3	cell-based assay data	DCM	PROM-NPROM	30,875	28,942	0.937
Hit Dexter 3	extended cell-based assay data	DCM	PROM-NPROM	24,111	22,572	0.936
Hit Dexter 2.0	PSA data[2]	Known Bad Actors [37]	HPROM-NPROM	974	140	0.144
Hit Dexter 2.0	CDRA data[3]	Known Bad Actors [37]	HPROM-NPROM	963	140	0.145
Hit Dexter 3	target-based assay data	Known Bad Actors [37]	HPROM-NPROM	1002	241	0.241
Hit Dexter 3	cell-based assay data	Known Bad Actors [37]	HPROM-NPROM	987	252	0.255
Hit Dexter 3	extended cell-based assay data	Known Bad Actors [37]	HPROM-NPROM	965	222	0.230
Hit Dexter 2.0	PSA data[2]	Known Bad Actors [37]	PROM-NPROM	910	304	0.334
Hit Dexter 2.0	CDRA data[3]	Known Bad Actors [37]	PROM-NPROM	896	330	0.368
Hit Dexter 3	target-based assay data	Known Bad Actors [37]	PROM-NPROM	941	379	0.403
Hit Dexter 3	cell-based assay data	Known Bad Actors [37]	PROM-NPROM	906	356	0.393
Hit Dexter 3	extended cell-based assay data	Known Bad Actors [37]	PROM-NPROM	891	359	0.403

[1] Any compounds present in the training set were removed from the test set.
[2] Primary screening assay data.
[3] Confirmatory dose-response assay data.
[4] Note that Hit Dexter models are not designed to identify all different kinds of bad actors but rather to identify frequent hitters (of which a significant portion are in fact bad actors).

mance but, again, Hit Dexter 3 is designed to predict frequent hitters, and it can be expected that a substantial proportion of the compounds observed to cause false-positive readouts in these interference assays will behave benign in other assay types and setups.

Model performance compared to Hit Dexter 2.0

The set of machine learning models developed in this work to form Hit Dexter 3 differ from the Hit Dexter 2.0 models in several ways. For Hit Dexter 3,

- dedicated models for target-based and cell-based assays were developed whereas the previous set of models only cover target-based assays.
- four different machine learning algorithms (KNN, ET, RF and MLP) instead of just ET were explored. This led to the finding that MLP classifiers perform best.
- the minimum number of data points required to calculate the ATR has been increased from 50 to 100. This results in more robust ATRs (based on which the class labels, i.e. HPROM, PROM and NPROM, are assigned).

Given the fact that the training and test sets utilized for the development and validation of the Hit Dexter 3 and Hit Dexter 2.0 models differ, a 1:1 comparison of model performance is difficult. For models of the same type (e.g. HPROM-NPROM classifier), differences in MCCs on the test data were in the range of -0.035 to +0.015 (cell-based models not included as they are not available in Hit Dexter 2.0). Also on the DCM data sets (the DCM data sets used for testing differ in their composition because of the removal of any compounds that are also present in the training set of the respective model), the models behave similarly, with the Hit Dexter 3 PROM-NPROM classifier (for target-based assays) assigning 5% to the PROM class and the respective Hit Dexter 2.0 models assigning 2% to 4% of the DCM compounds to the PROM class (Table 8). On the set of known bad actors [37], the percentage of compounds predicted as frequent hitters is 40% for Hit Dexter 3 (PROM-NPROM classifier trained on target-based assay data) and 33% to 37% for Hit Dexter 2.0 (PROM-NPROM classifiers; Table 8).

Overall, these results indicate that the performance of the Hit Dexter 3 and Hit Dexter 2.0 machine learning models is comparable. The Hit Dexter 3 models perform a bit better on the set of known bad actors. Finally, the addition of dedicated models for predicting a compound's behavior in cell-based assays is an important advantage of Hit Dexter 3 over Hit Dexter 2.0.

Conclusions

In this work we present the development, refinement and validation of new models for the prediction of frequent hitters in biological assays. The models are trained on a manually curated assay data set that we extracted from the PubChem Bioassay database and, for the first time, these models cover cell-based assays in addition to target-based assays. Further additions include the exploration of four sets of descriptors with additional machine learning algorithms such as KNN and MLP, and the use of more robust ATRs (calculated now on a minimum of 100 distinct assays compared to 50 previously).

The MLP classifiers turned out to obtain the best classification performance and robustness in most cases, with MCCs of up to 0.648 in discriminating HPROM from NPROM compounds, and MCCs of up to 0.580 in discriminating PROM from NPROM compounds. Use cases that require models with high sensitivity or high specificity can be approached by adjusting the decision threshold applied in classification.

Tests of the MLP classifiers on DCM compounds and sets of known bad actors corroborate good performance of the models: the models correctly identified 94 to 99% of all compounds of the DCM data set as non-promiscuous and flagged up to 40% of the known bad actors as frequent hitters (because bad actors are not necessarily frequent hitters this number is in line with the expectations for a good model).

We found that it is indeed important to use dedicated models for predicting the behavior of compounds in target-based and cell-based assays as assays from the different domains can behave very differently. At the same time it is clear that for the further development of this and similar computational methods it will be important to consider assay types and conditions, which poses fundamental challenges related to the scarcity and heterogeneity of the available data.

The best models presented in this work are available via a refined, free web service at https://nerdd.univie.ac.at/hitdexter3/. This web service offers many additional features, encrypted communication via HTTPS and the possibility for users to immediately and permanently delete their data from the web server.

We hope that the new Hit Dexter models, in particular the new models for cell-based assays, will be of high value to the scientific community to tackle the challenge of hit prioritization and the identification of problematic compounds in biological screens.

Funding

C.S. and J.K. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number KI 2085/1-1. N.M. and J.K. are supported by the Trond Mohn Foundation (BFS2017TMT01).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2021.100007.

References

[1] Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, et al. Impact of high-throughput screening in biomedical research. Nat Rev Drug Discov 2011;10:188–95.

[2] Bajorath J. Evolution of assay interference concepts in drug discovery. Expert Opin Drug Discov 2021:1–3.

[3] Evans BE, Rittle KE, Bock MG, DiPardo RM, Freidinger RM, Whitter WL, et al. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. J Med Chem 1988;31:2235–46.

[4] In: Auld DS, Inglese J. Interferences with Luciferase reporter enzymes. Assay Guidance Manual. Markossian S, Grossman A, Brimacombe K, Arkin M, Auld D, Austin CP, editors. et al., editors. Eli Lilly & Company and the National Center for Advancing Translational Sciences, Bethesda (MD); 2016.

[5] Reker D, Bernardes GJL, Rodrigues T. Computational advances in combating colloidal aggregation in drug discovery. Nat Chem 2019;11:402–18.

[6] Yang Z-Y, He J-H, Lu A-P, Hou T-J, Cao D-S. Frequent hitters: nuisance artifacts in high-throughput screening. Drug Discov Today 2020;25:657–67.

[7] Dantas RF, Evangelista TCS, Neves BJ, Senger MR, Andrade CH, Ferreira SB, et al. Dealing with frequent hitters in drug discovery: a multidisciplinary view on the issue of filtering compounds on biological screenings. Expert Opin Drug Discov 2019;14:1269–82.

[8] Ferreira LLG, Andricopulo AD. ADMET modeling approaches in drug discovery. Drug Discov Today 2019;24:1157–65. doi:10.1016/j.drudis.2019.03.015.

[9] Kar S, Leszczynski J. Open access in silico tools to predict the ADMET profiling of drug candidates. Expert Opin Drug Discov 2020;15:1473–87.

[10] Feldmann C, Bajorath J. Machine learning reveals that structural features distinguishing promiscuous and non-promiscuous compounds depend on target combinations. Sci Rep 2021;11:7863.

[11] Irwin JJ, Duan D, Torosyan H, Doak AK, Ziebart KT, Sterling T, et al. An aggregation advisor for ligand discovery. J Med Chem 2015;58:7076–87.

[12] Yang Z-Y, Yang Z-J, Dong J, Wang L-L, Zhang L-X, Ding J-J, et al. Structural analysis and identification of colloidal aggregators in drug discovery. J Chem Inf Model 2019;59:3714–26.

[13] Alves VM, Capuzzi SJ, Braga RC, Korn D, Hochuli JE, Bowler KH, et al. SCAM detective: accurate predictor of small, colloidally aggregating molecules. J Chem Inf Model 2020;60:4056–63.

[14] Reactive compounds and in vitro false positives in HTS. Drug Discov Today 1997;2:382–4.

[15] Hann M, Hudson B, Lewell X, Lifely R, Miller L, Ramsden N. Strategic pooling of compounds for high-throughput screening. J Chem Inf Comput Sci 1999;39:897–902.

[16] Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 2010;53:2719–40.

[17] Baell J, Walters MA. Chemistry: chemical con artists foil drug discovery. Nature 2014;513:481–3.

[18] Baell JB, Nissink JWM. Seven year itch: pan-assay interference compounds (PAINS) in 2017-utility and limitations. ACS Chem Biol 2018;13:36–44.

[19] M. Koptelov, A. Zimmermann, P. Bonnet, R. Bureau, B. Crémilleux. PrePeP: a tool for the identification and characterization of pan assay interference compounds. Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, p. 462–71.

[20] Ghosh D, Koch U, Hadian K, Sattler M, Tetko IV. Luciferase advisor: high-accuracy model to flag false positive hits in Luciferase HTS assays. J Chem Inf Model 2018;58:933–42.

[21] Yang Z-Y, Dong J, Yang Z-J, Lu A-P, Hou T-J, Cao D-S. Structural analysis and identification of false positive hits in Luciferase-based assays. J Chem Inf Model 2020;60:2031–43.

[22] Borrel A, Huang R, Sakamuru S, Xia M, Simeonov A, Mansouri K, et al. High-throughput screening to predict chemical-assay interference. Sci Rep 2020;10:3986.

[23] M Nissink JW, Blackburn S. Quantification of frequent-hitter behavior based on historical high-throughput screening data. Future Med Chem 2014;6:1113–26.

[24] Yang JJ, Ursu O, Lipinski CA, Sklar LA, Oprea TI, Bologa CG. Badapple: promiscuity patterns from noisy evidence. J Cheminform 2016;8:29.

[25] Stork C, Chen Y, Šícho M, Kirchmair J. Hit Dexter 2.0: machine-learning models for the prediction of frequent hitters. J Chem Inf Model 2019;59:1030–43.

[26] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 2021;49:D1388–95.

[27] Feldmann C, Yonchev D, Stumpfe D, Bajorath J. Systematic data analysis and diagnostic machine learning reveal differences between compounds with single- and multitarget activity. Mol Pharm 2020;17:4652–66.

[28] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic Acids Res 2016;44:D1202–13.

[29] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47:D1102–9.

[30] PubChem. https://pubchem.ncbi.nlm.nih.gov/ (accessed June 3, 2021).

[31] NCBI Resource CoordinatorsDatabase resources of the National Center for Biotechnology Information. Nucleic Acids Res 2016;44:D7–19.

[32] Kim S, Thiessen PA, Cheng T, Yu B, Bolton EE. An update on PUG-REST: RESTful interface for programmatic access to PubChem. Nucleic Acids Res 2018;46:W563–70.

[33] Patrícia Bento A, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, et al. An open source chemical structure curation pipeline using RDKit. J Cheminform 2020;12:1–16.

[34] RDKit: Open-source cheminformatics; http://www.rdkit.org/ (accessed June 3, 2021).

[35] ChEMBL 23. http://www.ebi.ac.uk/chembl/ (accessed June 3, 2021).

[36] Wassermann AM, Lounkine E, Hoepfner D, Le Goff G, King FJ, Studer C, et al. Dark chemical matter as a promising starting point for drug lead discovery. Nat Chem Biol 2015;11:958–66.

[37] Dahlin JL, Auld DS, Rothenaigner I, Haney S, Sexton JZ, Nissink JWM, et al. Nuisance compounds in cellular assays. Cell Chem Biol 2021;28:356–70.

[38] NCICADD Group, National Cancer Institute. Chemical Identifier Resolver. https://cactus.nci.nih.gov/chemical/structure (accessed July 26, 2021).

[39] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50:742–54.

[40] Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J Doc 1965;5:107–13. doi:10.1021/c160017a018.

[41] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422–3.

[42] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28:3150–2.

[43] Garreta R, Moncecchi G. Learning scikit-learn: machine learning in Python. Packt Publishing Ltd; 2013.

[44] https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm.

[45] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority oversampling technique. J Artif Intell Res 2002;16:321–57. doi:10.1613/jair.953.

[46] Stork C, Wagner J, Friedrich N-O, de Bruyn Kops C, Šícho M, Kirchmair J. Hit Dexter: a machine-learning model for the prediction of frequent hitters. ChemMedChem 2018;13:564–71.