Hit Dexter 2.0: Machine-Learning Models for the Prediction of Frequent Hitters

Conrad Stork,[1] Ya Chen,[1] Martin Šícho,[1,2] Johannes Kirchmair[1,3,4]*

[1] Center for Bioinformatics (ZBH), Department of Computer Science, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, Hamburg, 20146, Germany

[2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Laboratory of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, 166 28 Prague 6, Czech Republic

[3] Department of Chemistry, University of Bergen, N-5020 Bergen, Norway

[4] Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway

[*]J. Kirchmair. E-mail: johannes.kirchmair@uib.no. Tel.: +47 55 58 34 64.

ABSTRACT

Assay interference caused by small molecules continues to pose a significant challenge for early drug discovery. A number of rule-based and similarity-based approaches have been derived that allow the flagging of potentially “badly behaving compounds”, “bad actors” or “nuisance compounds”. These compounds are typically aggregators, reactive compounds and/or pan-assay

interference compounds (PAINS), and many of them are frequent hitters. Hit Dexter is a recently introduced machine learning approach that predicts frequent hitters independent of the underlying physicochemical mechanisms (including also the binding of compounds based on "privileged scaffolds" to multiple binding sites). Here we report on the development of a second generation of machine learning models which now cover both primary screening assays and confirmatory dose-response assays. Protein sequence clustering was newly introduced to minimize the overrepresentation of structurally and functionally related proteins. The models correctly classified compounds of large independent test sets as (highly) promiscuous or non-promiscuous with Matthews correlation coefficient (MCC) values of up to 0.64 and area under the receiver operating characteristic curve (AUC) values of up to 0.96. The models were also utilized to characterize sets of compounds with specific biological and physicochemical properties, such as dark chemical matter, aggregators, compounds from a high-throughput screening library, drug-like compounds, approved drugs, potential PAINS and natural products. Among the most interesting outcomes is that the new Hit Dexter models predict the presence of large fractions of (highly) promiscuous compounds among approved drugs. Importantly, predictions of the individual Hit Dexter models are generally in good agreement and consistent with those of Badapple, an established statistical model for the prediction of frequent hitters. The new Hit Dexter 2.0 web service, available at http://hitdexter2.zbh.uni-hamburg.de, not only provides user-friendly access to all machine learning models presented in this work but also to similarity-based methods for the prediction of aggregators and dark chemical matter as well as a comprehensive collection of available rule sets for flagging frequent hitters and compounds including undesired substructures.

INTRODUCTION

Biochemical assays are a core component of early drug discovery.[1–3] Some small molecules however can pose significant challenges to biochemical assays as they may trigger false outcomes. Whereas false negative results may lead to a loss of valuable bioactive compounds, false positive outcomes can, if they remain undetected, tie up and consume significant resources and time without prospect of success. In the worst case, these "badly behaving compounds", "bad actors" or "nuisance compounds" get reported as bioactive compounds and pollute the medicinal chemistry and chemical biology literature. Once published, invalid assay outcomes may propagate and trigger follow-up studies based on false grounds, which hampers the global drug discovery effort.

Nuisance compounds (Figure 1) include compounds that can form colloidal aggregates,[4,5] compounds with reactive groups,[6] and pan assay interference compounds (PAINS).[7] Note that the PAINS substructures by design do not cover aggregators because they were derived from the outcomes of high-throughput screening campaigns run in the presence of a detergent and casein in order to minimize phenomena related to aggregate formation. They also do not cover many types of reactive compounds because compounds with reactive functional groups had been removed from the screening library prior to assaying.[7]

The different types of badly behaving compounds discussed so far involve no definite assertions about the frequency by which they cause assay interference. Rather than interfering with all different kinds of assays they trigger false outcomes only under specific (and by far not all) conditions. However, a tendency of badly behaving compounds to have higher hit rates is apparent.

Compounds for which a higher than expected hit rate is recorded in historical assay data are referred to as “frequent hitters”.[8] Many of these compounds are aggregators, reactive compounds or PAINS, but importantly, a significant proportion of frequent hitters are true promiscuous compounds. True promiscuity is often related to “privileged scaffolds”[9] or “master key compounds”,[10] which have the ability to bind to multiple binding sites. True promiscuous compounds are not necessarily nuisance compounds. In fact, they can be valuable in the context of drug repurposing and polypharmacology.[11,12]

Computational methods for predicting nuisance compounds and frequent hitters are still in an early stage of development.[13] The most established approaches for identifying problematic compounds are rule-based methods, which flag compounds containing substructures that have been linked to assay interference. In recent years, the 480 patterns encoding substructures derived from PAINS have become not only one of the best known but also one of the most misused rule sets in medicinal chemistry. All too often, the limitations of the PAINS concept, most of which have been pointed out clearly by its inventors, are not paid the necessary attention.[14,15] Further approaches for the prediction of nuisance compounds and frequent hitters include similarity-based approaches, statistical and machine learning approaches, an overview of which is provided in ref 13.

An important statistical approach for the prediction of frequent hitters is Badapple,[16] which performs a hierarchical scaffold analysis to derive a promiscuity score (“pScore”). The pScore corresponds to the likelihood of a compound based on a specific scaffold being promiscuous. Badapple was derived from a large public data set of more than 430k compounds measured in a total of more than 800 different assays.

We recently reported two machine learning models for the prediction of frequent hitters which are accessible via a free web service called “Hit Dexter”.[17] Hit Dexter was developed with the idea of creating a reliable model for the prediction of frequent hitters independent of the underlying physicochemical mechanisms (including also the binding of compounds based on "privileged scaffolds" to multiple binding sites). Such a model could advice scientists for which compounds to exercise extra caution with positive assay readouts. The initial Hit Dexter models were trained on more than 235k compounds measured in at least 50 different confirmatory dose-response assays (CDRAs). They reached a high level of accuracy on independent test data, with Matthews correlation coefficients[18] (MCCs) of up to 0.67 and area under the receiver operating characteristic curve (AUC) values of up to 0.96.

Here we report on the development of a second generation of machine learning models for the prediction of frequent hitters, which are accessible via the new Hit Dexter 2.0 web service.[19] The models are a result of several major refinements and extensions of the data collection, data processing and modeling procedures. For example, a clustering approach was introduced in order to avoid an overrepresentation of structurally and functionally related proteins such as protein kinases. Hit Dexter 2.0 also includes models trained on data measured with primary screening assays (PSAs). In contrast to CDRAs, PSAs are primarily high-throughput screening assays measuring single-dose inhibition. The inclusion of these models in Hit Dexter 2.0 will allow a better representation of assays employed for primary screening.

In addition to method and model refinement, we also report on comprehensive tests of Hit Dexter 2.0 with various types of compounds, including dark chemical matter (DCM),[20] approved drugs and natural products. Last but not least, we present a direct comparison of Hit Dexter 2.0 with Badapple and introduce the new Hit Dexter 2.0 web service.

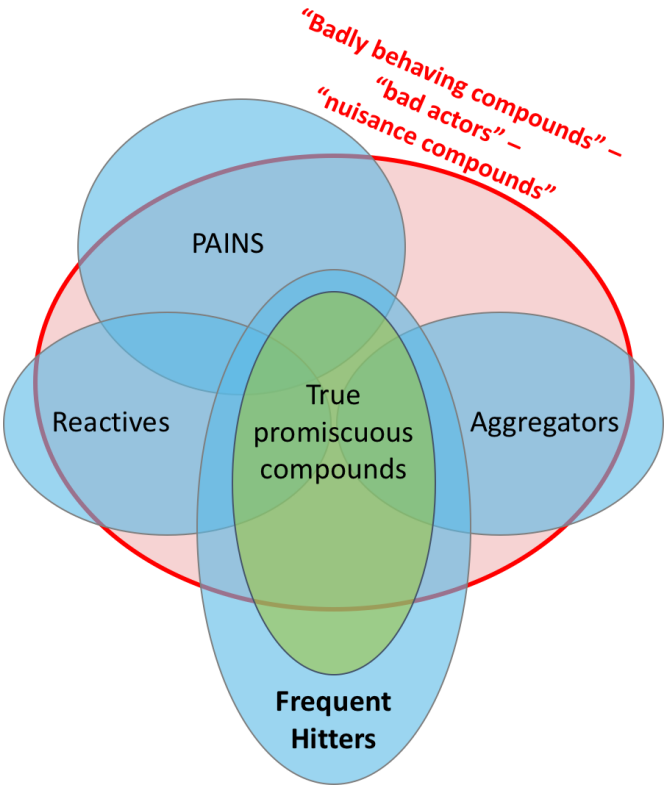

Figure 1. Schematic overview of key concepts and terms used in the context of assay interference and small-molecule drug discovery. Note that by design PAINS exclude aggregators and many types of reactive compounds. However, overlaps between the different types of compounds other than those depicted in this scheme certainly exist.

RESULTS AND DISCUSSION

Data Set Compilation and Analysis

Two large data sets were compiled from PubChem Bioassay,[21,22] one consisting of 803 898 compounds measured in 931 PSAs and the other one consisting of 468 258 compounds measured in 2273 CDRAs. During data preprocessing, filtering in particular, 20 921 and 18 003 compounds were removed from the PSA and CDRA data sets, respectively (Table 1).

Table 1. Number of Compounds Removed During Filtering and Quality Checks.

Reason for removal	PSA data set [cpds]	CDRA data set [cpds]
Invalid SMILES notation	1	3
Salt filter with ambiguous outcome[a]	770	231
Molecular weight outside the range of 200 to 900 Da	11 231	10 815
Elements other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br and I	116	175
Duplicate molecules[b]	8460	5171
Rejected during quality checks[c]	343	1608
Sum	**20 921**	**18 003**

[a] Multi-component compounds for which the main component could not be unequivocally defined (see Methods and ref 17 for detail).

[b] Identified based on canonicalized SMILES. Associated data were merged as outlined in ref 17. This led to the effective reduction of the number of compounds as reported in the table.

[c] Compounds with conflicting data (e.g. activity data). See Methods and ref 17 for detail).

Following this procedure, the proteins covered by the PSA and CDRA data sets were clustered based on sequence similarity: The 429 proteins covered by the PSA data set were assigned to 388 unique protein clusters, and the 712 proteins covered by the CDRA data set were assigned to 537 unique protein clusters (see Methods for detail).

The definition of whether a compound is (highly) promiscuous or not is based on the active-to-tested ratio (ATR), which is calculated according to Equation 1:

$$ATR = \frac{A}{T}, \quad (1)$$

where A is the number of protein clusters for which a compound was measured as active on at least one protein of that cluster, and T is the total number of protein clusters a compound was measured on.

The ideal data set to derive ATRs from would consist of a large number of compounds measured on a large number of protein clusters. Obviously, with the available data a compromise needs to be found between the number of instances available for training and testing the models (i.e. size of the data set in terms of the number of compounds) and the minimum number of protein clusters for which activity data are recorded for the individual compounds.

Figure 2 shows the relationship between data set size and the minimum number of protein clusters for which assay results have been recorded. For example, the processed data set includes 362k compounds which have been measured with PSAs representing at least 50 different protein

clusters. Likewise, 327k compounds are represented by the respective data collected from CDRAs. One of the most obvious differences between the PSA and CDRA data sets is that for the vast majority of compounds of the PSA data set measured data are available for 150 and more protein clusters, whereas for the CDRA data set a steep decline in the number of compounds for which measured data are available for 70 and more protein clusters is observed.

We explored the use of data sets containing all compounds for which bioactivity data has been recorded for at least 20, 50 and (only for PSAs) 100 protein clusters (data not shown). In agreement with previous results,[17] we found the data sets containing all compounds for which bioactivity data has been recorded for at least 50 protein clusters to be highly diverse and most suitable for modeling. We refer to these data sets as the PSA50 and CDRA50 data sets.

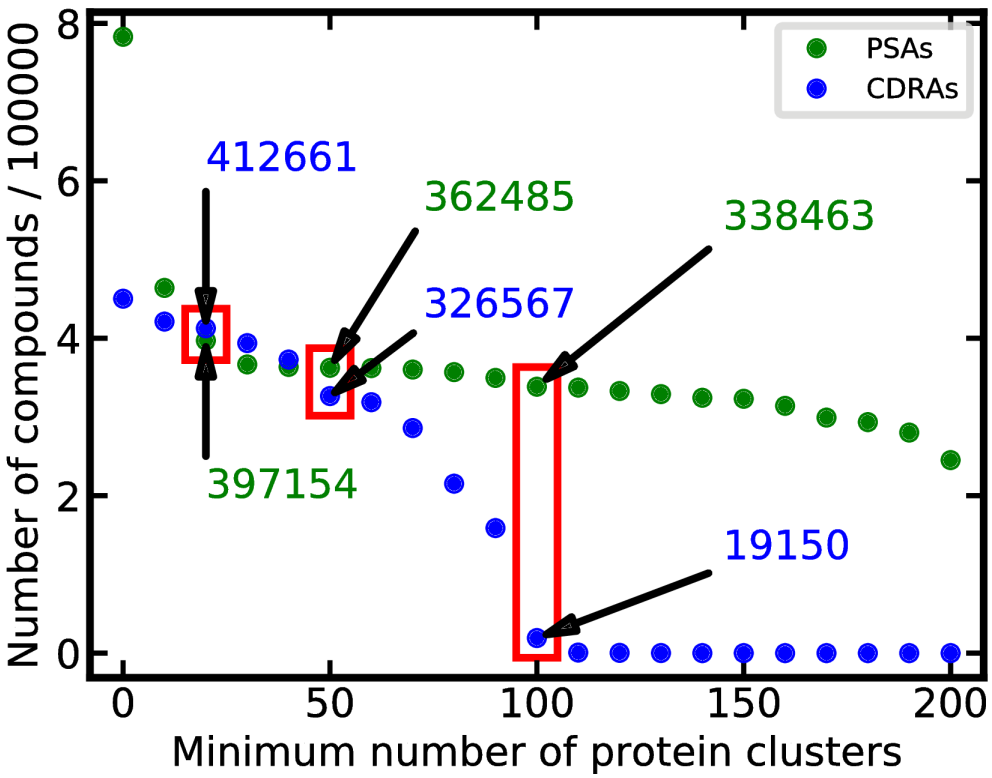

Figure 2. Data set size (number of compounds) as a function of the minimum number of protein clusters for which measured data are available. The rectangles mark the PSA20/CDRA20, PSA50/CDRA50 and PSA100/CDRA100 data sets.

Next, all compounds were assigned a promiscuity label based on their ATR: “NP” for non-promiscuous compounds, “P” for promiscuous compounds and “HP” for highly promiscuous compounds. Note that, according to the definitions of promiscuity summarized in Table 2, highly promiscuous compounds are a subset of promiscuous compounds.

Suitable cutoffs for the assignment of promiscuity labels were calculated for the PSA50 and CDRA50 data sets according to the definitions derived as part of our previous work (recited in Table 2, column “threshold definition”).[17] According to these definitions, any compounds with an ATR greater than 0.024 for PSAs and 0.043 for CDRAs were labeled promiscuous, accounting for 11% and 13% of all compounds, respectively. These proportions of promiscuous compounds are in good agreement e.g. with the findings of a recent study from GlaxoSmithKline, which reported a fraction of 13% of all compounds as "noisy",[23] and higher than the averaged incidence of frequent hitters reported for the AstraZeneca screening library (which is 6%).[24] The mean ATRs for the PSA and CDRA data sets were 0.008 and 0.015, respectively (see Table 2 for more detail). These mean ATRs correspond well to the findings of other studies, such as that on the AstraZeneca compound collection, which reported an overall hit rate of 1.53%.[24]

Table 2. Composition of the Data Sets Used for Model Training and Validation.

Assigned promiscuity class	Number of unique compounds in		Threshold definition[a]	Threshold value	
	PSA50	CDRA50		PSA50[b]	CDRA50[b]
Non-promiscuous (NP)	Total: 247 110 Training set: 222 272 Test set[c]: 24 881	234 811 211 264 23 574	ATR < ATR_{mean}	0.008	0.015
Promiscuous (P)	Total: 29 042 Training set: 26 117 Test set[c]: 2 930	33 982 30 478 3 507	ATR > ATR_{mean} + 1σ[d]	0.024	0.043
Highly promiscuous (HP) - a subset of compounds labeled P	Total: 6 625 Training set: 5 956 Test set[c]: 670	6 246 5 609 637	ATR > ATR_{mean} + 3σ[d]	0.054	0.100

[a] Derived as part of our previous work.[17] Compounds with ATRs between ATR_{mean} and ATR_{mean} + 1σ were not assigned a promiscuity label and removed from all data sets.

[b] ATR threshold values calculated for the individual data sets according to the ATR threshold definition.

[c] Independent test set obtained by random split of the curated data set prior to model development.

[d] Standard deviation.

CDRAs tend to have higher hit rates than PSAs. This is observed in the ATR distributions reported in Figure 3 and also reflected in the higher mean ATR for CDRAs (Table 2). The differences in hit rates can be explained by the fact that CDRAs are often used to measure compounds which have previously been reported as active by a PSA and are hence more likely to also show activity in CDRAs than a random set of screening compounds.

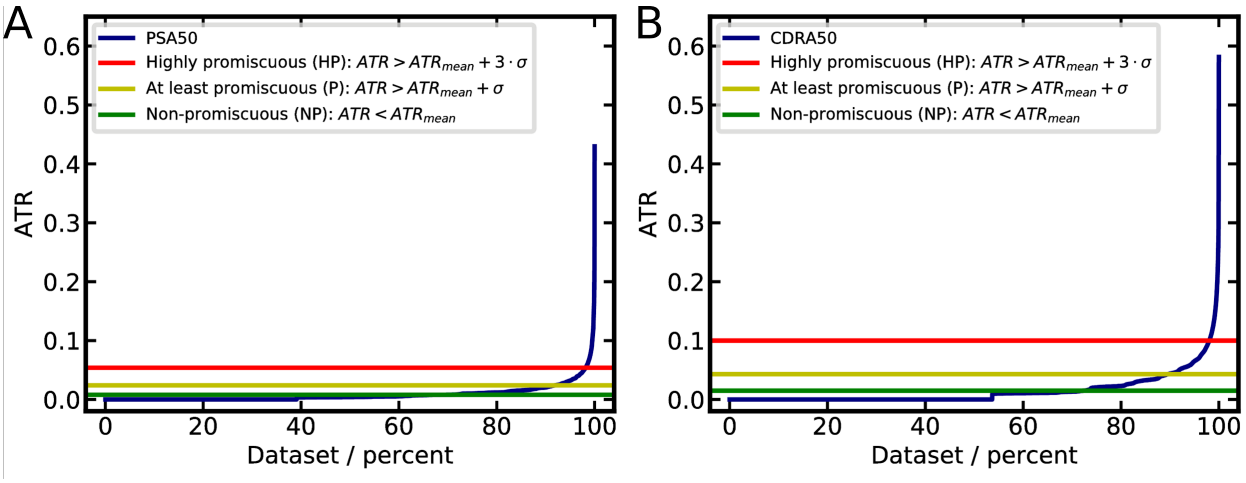

Figure 3. ATR distributions calculated for the (A) PSA50 and (B) CDRA50 data sets. The vertical lines mark the cutoffs applied for the assignment of promiscuity labels. Any compounds below the green line were labeled “NP”, any above the yellow line “P”, and any above the red line “HP”.

Comparison of the Chemical Space of the PSA and CDRA Data Sets

The chemical space of the individual data sets used for modeling was determined and compared using (i) principal component analysis (PCA) on 44 physically meaningful 2D descriptors computed with MOE[25] (listed in Table S1 of ref. [17]) and (ii) pairwise similarity analysis based on the Tanimoto coefficient calculated from Morgan2 fingerprints.[26,27]

In a first experiment, we analyzed whether the reduction of the PSA and CDRA data sets to subsets of compounds annotated with measured data for at least 50 protein clusters leads to a substantial loss of coverage. As shown by the PCA scatter plots and histograms reported in Figure 4, the chemical space of compounds covered by the PSA50 and CDRA50 data sets is–to a large extent–comparable with that of the PSA0 and CDRA0 data sets, respectively. Only about 11% of all compounds of the PSA0 data set and about 9% of all compounds of the CDRA0 data set have a maximum Tanimoto coefficient of less than 0.5 measured against any compounds present in the PSA50 and CDRA50 data sets, respectively. In other words, this means that by constraining the data used for model development to compounds for which measured data has been recorded for at least 50 protein clusters does not lead to a substantial reduction of chemical space coverage as compared to the complete (processed) PubChem Bioassay data sets.

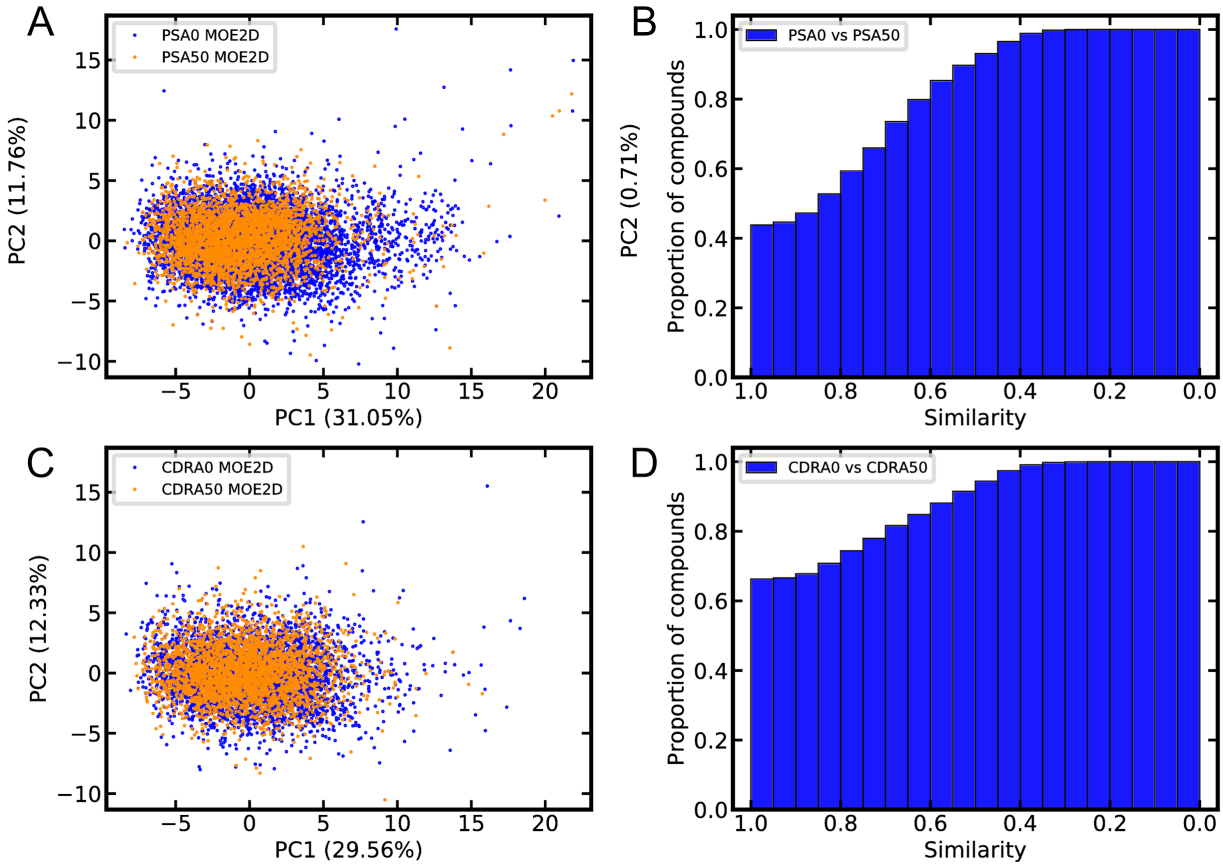

Figure 4. Comparison of the chemical space of the (A, B) PSA0 and PSA50 data sets and (C, D) CDRA0 and CDRA50 data sets. The PCA scatter plots are based on the first against the second component and derived from 44 physicochemically meaningful 2D descriptors calculated with MOE (see Table S1 in ref. [17]). PCA was performed on the full dataset. For the sake of clarity, only a randomly selected 1% of all data points are shown. The axis labels report the percentage of the total variance explained by the respective principal component. The histograms show the proportion of compounds of the (B) PSA0 or (D) CDRA0 data set represented by compounds of the (B) PSA50 or (D) CDRA50 data set at a given minimum similarity (Tanimoto coefficient calculated from Morgan2 fingerprints).

Following the same protocol, we also compared the chemical space covered by the PSA50 and CDRA50 data sets. As shown in Figure 5A and D, no substantial differences in coverage between the two data sets are apparent from the PCA and pairwise similarity analysis. Last but not least we compared the PSA50 and CDRA50 data sets to the complete ChEMBL database.[28,29] The ChEMBL database was prepared following the identical data preprocessing protocol (without considering any biological data) and consists of about 1.5 million compounds. From this comparison it can be seen that chemical space of the ChEMBL database is wider than that of the PSA50 and CDRA50 data sets (Figure 5B, C, E and F). This is an expected result, since the ChEMBL database contains substantially more compounds from a large number of diverse sources. Nevertheless, the plots also show that approximately 50% of all ChEMBL compounds are well represented by the PSA50 and CDRA50 data sets.

Model Development

Prior to model development, the PSA50 and CDRA50 data sets were randomly split into a training and a test set with a ratio of 9:1. In contrast to our previous work,[17] an additional data preprocessing step was implemented which checks for the presence of any compounds with distinct canonicalized SMILES but identical Morgan2 fingerprints (as in the case of stereoisomers, for example) because this can lead to inconsistent predictions. In order to address these issues, any instances with identical Morgan2 fingerprints were merged if their promiscuity labels were identical. If their labels differed, all instances with identical Morgan2 fingerprints were removed from the training data. Table 3 lists the number of compounds removed from the training and test sets as part of this process.

For assays of both screening stages (i.e. PSAs and CDRAs), two binary classifiers were developed: one to distinguish promiscuous from non-promiscuous compounds (P-NP classifier)

and another one to distinguish highly promiscuous from non-promiscuous compounds (HP-NP classifier). An overview of the size of the training and test set is reported in Table 2.

As a first step in the model building process, the most suitable machine learning algorithm and descriptor set were identified. In addition to the extra tree classifiers (ETC) and random forest classifiers (RFC) explored in our previous work,[17] we also tested several meta classifiers such as the AdaBoost Classifier[30,31] and Bagging Classifier,[32,33] both in combination with the ETC and RFC. With respect to descriptors, we explored all 206 2D descriptors available in MOE, Morgan2 fingerprints (1024 bit), and MACCS keys (166 bits).

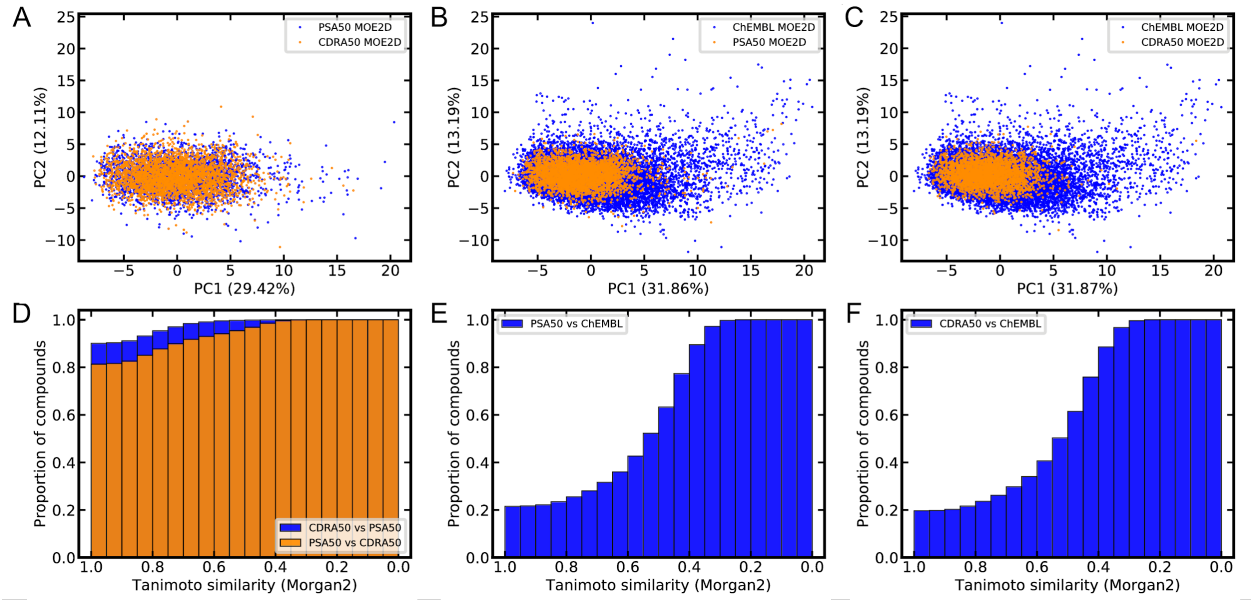

Figure 5. Comparison of the chemical space of the PSA50, CDRA50 and ChEMBL data sets in the (A, B, C) MOE 2D descriptor space and (D, E, F) Morgan2 fingerprint space. The PCAs were performed on the full datasets. For the sake of clarity, only a randomly selected 1% of all data points are shown. The axis labels report the percentage of the total variance explained by the

respective principal component. The histograms show the proportion of compounds of the specified data sets represented by the (D, E) PSA50 and (D, F) CDRA50 data at a given minimum similarity (Tanimoto coefficient calculated from Morgan2 fingerprints).

Table 3. Number of Compounds Filtered Due to Duplicate Morgan2 Fingerprints.

Data set	No. of compounds merged due to identical fingerprints and promiscuity labels	No. of compounds with identical fingerprints removed due to contradicting promiscuity labels
PSA50 training set	2 522	303
PSA50 test set	41	8
CDRA50 training set	1 664	281
CDRA50 test set	26	4

The various combinations of machine learning algorithms and descriptors were tested with 10-fold cross-validation (see Methods for more detail). The performance of the individual classifiers was compared based on the MCC, which quantifies the correlation between the predictions and their true value by taking into account the true positive, false positive, true negative and true positive predictions. MCC values range from -1 to +1, where a value of +1 indicates perfect prediction, a value of 0 a performance equal to random prediction, and a value of -1 total disagreement of the prediction. In addition, we generated receiver operating characteristic (ROC) curves and calculated the area under the ROC curves (AUCs). By considering these three

components, a solid understanding of the goodness of the models can be obtained: Whereas the MCC quantifies the capability of a model to correctly classify a compound of interest, ROC curves and (to some extent also) AUC values quantify a model's ability to identify (highly) promiscuous compounds by assigning them high probabilities as compared to non-promiscuous compounds (i.e., ranking (highly) promiscuous compounds early in a list).

For all data sets the best performance during 10-fold cross-validation was obtained by ETCs in combination with Morgan2 fingerprints (MCC values between 0.56 and 0.58). These observations are consistent with the observations made during our previous work.[17]

Following these experiments, the hyperparameters (Table 4) for the ETCs (in combination with Morgan2 fingerprints) were optimized, again with 10-fold cross-validation and with MCC as performance measure. Optimization of the hyperparameters did not yield substantial improvements of the models. The most suitable settings for the number of estimators and the maximum fraction of features considered per split were 100 and 0.2, respectively. Classifiers using these hyperparameters obtained MCC values between 0.57 and 0.60 and AUC values between 0.91 and 0.96 during 10-fold cross-validation (Figure 6). The final models (with optimized parameters) were built on the complete training sets balanced with the synthetic minority oversampling technique (SMOTE).[34]

Table 4. Hyperparameters Optimized by Grid Search.[a]

Parameter	Option
Number of estimators (estimators)[b]	10[c],50,**100**,150,200,250,300,400,500,600
Maximum fraction of features considered per split (max_features)[b]	sqrt[c], **0.2**, 0.4, 0.6, 0.8, none[d]

[a] Bold numbers indicate settings used for the production of the final models.

[b] Parameter name in the scikit-learn[35] implementation.

[c] Default value.

[d] All features are used.

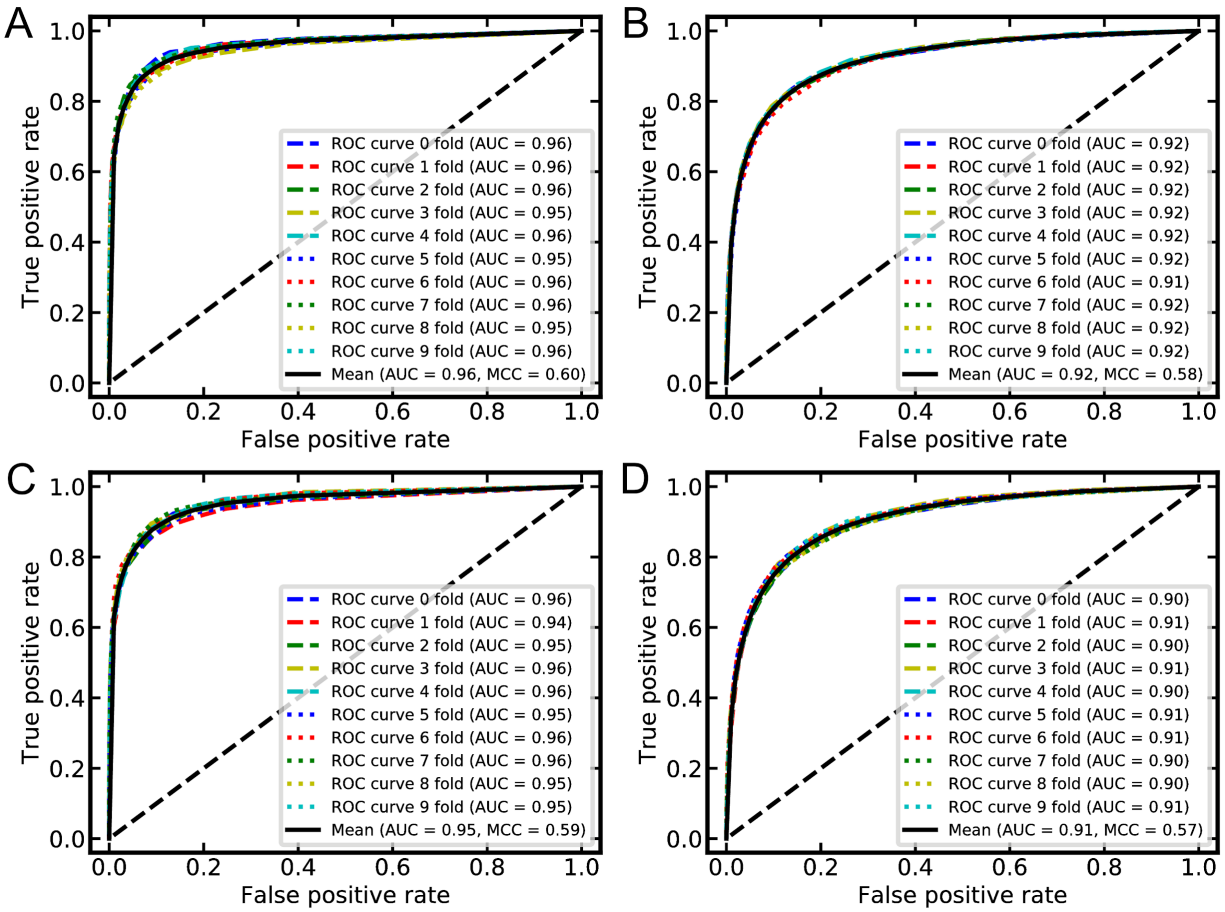

Figure 6. ROC curves obtained during 10-fold cross-validation for the selected models (i.e. ETC in combination with Morgan2 fingerprints; n_estimators = 100; max_features = 0.2). (A) HP-NP classifier for PSAs, (B) P-NP classifier for PSAs, (C) HP-NP classifier for CDRAs, (D) P-NP classifier for CDRAs.

Model Evaluation on Independent Test Data

The final models (ETC; Morgan2 fingerprints; SMOTE for balancing the training data; n_estimators = 100; max_features = 0.2), referred to as Hit Dexter 2.0 models, were tested on an independent test set (Table 2) derived by random split of the preprocessed PSA50/CDRA50 data sets prior to model building. The MCC values obtained by the four classifiers (i.e. HP-NP and P-

NP classifiers, trained on PSA or CDRA data) for the independent test set was between 0.60 and 0.64 (Figure 7A), whereas their AUC values ranged from 0.91 to 0.96. Both HP-NP classifiers performed on average slightly better than the P-NP classifiers. This is expected because the margin between the cutoffs utilized to assign compounds to either of the two promiscuity classes is larger for the HP-NP classifier.

The robustness of the Hit Dexter 2.0 models was further probed by iteratively removing compounds from the test set that are similar to any of the compounds in the training data. More specifically, the maximum allowed similarity between the compounds of the test set and any compounds in the training set (measured as Tanimoto coefficient calculated from Morgan2 fingerprints) was reduced by 0.02 during each iteration (Figure 8). For example, for the subset of test compounds with a maximum Tanimoto coefficient of 0.8, MCC values of 0.55 to 0.58 were obtained, whereas AUC values were between 0.90 and 0.95 (Figure 7B). For the subset of test compounds with a maximum Tanimoto coefficient of 0.7, MCC values were between 0.44 and 0.50, and AUC values between 0.87 and 0.92 (Figure 7C). Decent performance was observed for subsets of test compounds with a maximum Tanimoto coefficient as low as 0.6 (MCC values between 0.34 and 0.42; AUC values between 0.82 and 0.88). Overall, the MCC values obtained for the initial Hit Dexter models[17] are slightly higher than those obtained for Hit Dexter 2.0. This may be a result of the protein clustering procedure, as compounds active on several related proteins (which therefore may contain characteristic structural patterns that can be more easily recognized by machine learning algorithm) may no longer be part of the P (and HP) data set.

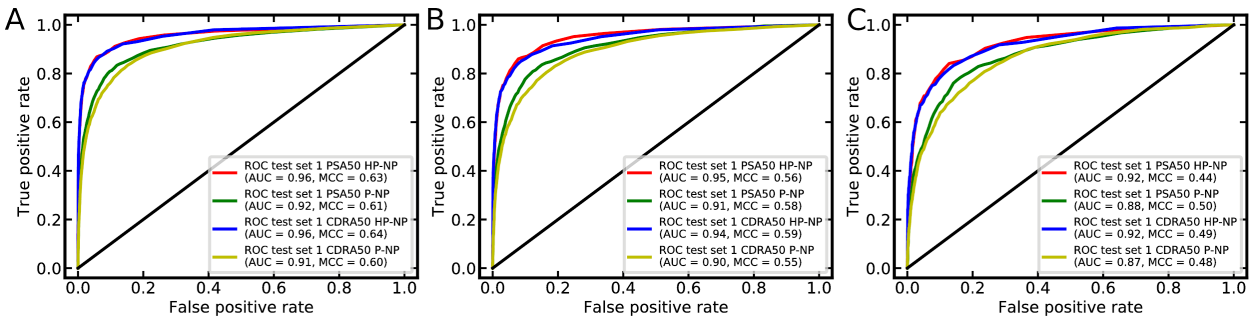

Figure 7. ROC curves obtained for the four classifiers on the independent test set (A) and subsets thereof, consisting of compounds with a maximum Tanimoto coefficient of (B) 0.8 or (C) 0.7 measured against any compound of the training set. MCC and AUC values are also reported.

.

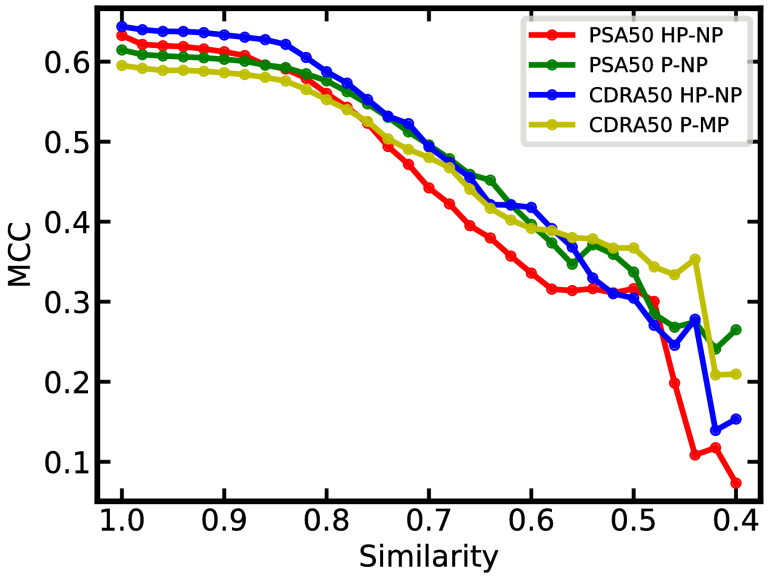

Figure 8. Classification performance (measured as MCC values) as a function of the maximum molecular similarity (Tanimoto coefficient calculated from Morgan2 fingerprints) between any pair of training and test compounds.

Application of Hit Dexter 2.0 to different data sets

In order to obtain a better understanding of the scope and limitations of Hit Dexter 2.0, we analyzed its predictions for a number of data sets with distinct characteristics:

- The dark chemical matter (DCM) data set:[20] a library of compounds which have been tested in at least one hundred different biochemical assays and have never shown activity. This data set originates from Novartis and PubChem assay data collected for 140k compounds (all size indications in this list referring to the unprocessed data sets).
- The aggregators data set:[36] a set of 12.6k compounds known to form colloidal aggregates. This library serves as data resource for Aggregator Advisor.[37]
- The Enamine HTS Collection:[38] Enamine is a leading provider of screening compounds and screening blocks. The Enamine HTS collection was selected as a representative library widely used in high-throughput screening. It contains 1.9 million compounds.
- The ChEMBL database:[29] a curated chemical database of 1.7 million (mainly) drug-like compounds, richly annotated with measured bioactivity data.
- The approved drugs subset of DrugBank:[39] a set of 2158 drugs approved in at least one jurisdiction, at some point in time.
- A "potential PAINS" data set: a subset of 51.7k compounds of the Enamine HTS Collection that match PAINS patterns (see Methods for detail).
- A natural products data set: a comprehensive set of 208k known natural products compiled from 18 different sources as part of our previous work.[40]

The chemical structures included in these data sets were prepared following the identical procedure employed for preprocessing the Hit Dexter 2.0 training data. Any compounds present

in the training data were removed from the individual data sets. The size of the filtered data sets is listed in Table 5.

In the previous section we showed that Hit Dexter 2.0 performs well on compounds represented by at least one instance in the training data with a minimum fingerprint-based Tanimoto coefficient of 0.6 (Figure 8). Considering this threshold, a large proportion of synthetic compounds is covered well by the training data (Figure S1). Taking the PSA50 training set of the P-NP classifier as an example, almost all DCM compounds, more than 80% of all aggregators and approximately 60% of all approved drugs are represented by at least one instance in the training data with a Tanimoto coefficient of 0.6 or higher. The percentage of compounds from ChEMBL and the Enamine HTS collection covered at this level of (minimum) similarity is about 40%. In contrast to synthetic compounds, only for approximately 15% of all natural products the maximum pairwise similarity (measured as Tanimoto coefficient based on Morgan2 fingerprints) with all instances of the training data is 0.6 or higher (Figure 9).

Table 5. Agreement of Predictions of the PSA and CDRA Classifiers.

Data set	No. of cpds in the HP-NP data set	No. of compounds (%) predicted as HP by the PSA HP-NP classifier	No. of compounds (%) predicted as HP by the CDRA HP-NP classifier	Agreement of predictions [%]	No. of compounds in the P-NP data set	No. of compounds (%) predicted P by the PSA P-NP classifier	No. of compounds (%) predicted P by the CDRA P-NP classifier	Agreement of predictions [%]
DCM	11 116	79 (0.7)	69 (0.6)	26.5	10 944	306 (2.8)	361 (3.3)	19.1
Aggregators	5786	225 (3.9)	272 (4.7)	34.0	4183	514 (12.3)	596 (14.3)	37.6
Enamine HTS collection	1 856 964	5883 (0.3)	7961 (0.4)	33.2	1 853 518	31 068 (1.7)	46 190 (2.5)	38.1
ChEMBL	1 194 343	27 643 (2.3)	26 447 (2.2)	37.0	1 166 478	88527 (7.6)	95 031 (8.2)	46.0
Approved Drugs	972	48 (4.9)	58 (6.0)	39.5	813	93 (11.4)	102 (12.6)	36.4
Potential PAINS	49 498	1670 (3.4)	2246 (4.5)	45.9	49 044	4867 (9.9)	6925 (14.1)	50.2

Natural Products	167 873	8010 (4.8)	7919 (4.7)	36.4	167 557	24 046 (14.4)	22 641 (13.5)	57,0
BADAPPLE_NP	110 624	1575 (1.4)	1873 (1.7)	32.0	108 620	7239 (6.7)	9853 (9.1)	42.4
BADAPPLE_P	346	82 (23.7)	87 (25.1)	55.1	330	170 (51.5)	203 (61.5)	69.6
BADAPPLE_HP	104	53 (51.0)	52 (50.0)	75.0	98	66 (67.4)	70 (71.4)	88.9

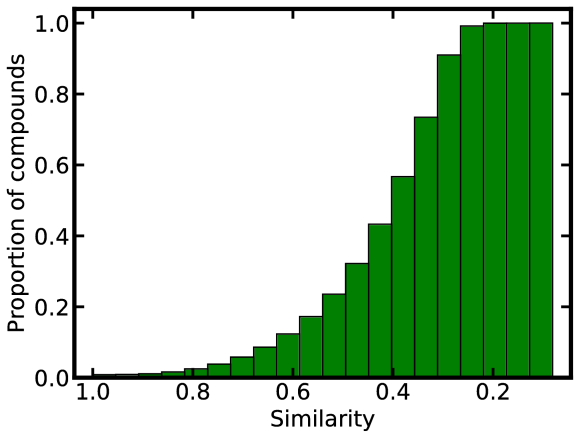

Figure 9. Example of the distribution of maximum pairwise similarities (Tanimoto coefficient calculated from Morgan2 fingerprints) between the training set (in this case, the PSA50 training set of the P-NP classifier) and the natural products data set.

Considering the limited prediction accuracy of the classifiers below a Tanimoto coefficient of 0.6 (Figure 8), the distance to the nearest neighbor(s) in the training data should be closely monitored. Therefore, in the following sections, in addition to the percentages of compounds referring to the complete, processed data sets, we report in brackets the percentages of compounds referring to the respective subsets consisting exclusively of compounds with a minimum Tanimoto coefficient of 0.6. In other words, the percentages reported in brackets are likely more accurate, but they are based on less-representative subsets. Predictions on the data sets listed above are reported in Figures 10 and 11. From the graphs obtained for the DCM data set (Figure 10A) it can be seen that any of the four models (i.e. HP-NP and P-NP classifiers, each for PSAs and CDRAs) classify at least 96% [96%] of all compounds of the DCM data set as non-promiscuous (both HP-NP classifiers obtained 99% [99%] correct classifications). This is an encouraging result since any of these compounds have been tested in a large number of assays and have never shown activity, for which reason they are unlikely frequent hitters (note that all

reported numbers are based on a decision threshold of 0.5, which is the default value). In contrast to the observations made for the DCM data set, a substantial number of compounds of the aggregators data set (approximately 15% [18%]) are predicted as promiscuous and approximately 4% [5%] as highly promiscuous (with classifiers derived from assays of either screening stage (Figure 10B). This again is a plausible result because aggregators are known to cause false positive assay readouts under, importantly, specific assay conditions. It is hence expected that not all known aggregators will be flagged by Hit Dexter 2.0. The distribution of class probabilities among compounds from the Enamine HTS Collection is similar to that of the DCM data set (Figure 10C). This suggests that the Enamine HTS Collection is a well-curated screening library. For the ChEMBL database, the distributions of class probabilities are located somewhere in between those of the DCM data set and the aggregators data set (Figure 10D), meaning that there is a relevant fraction of compounds predicted as promiscuous (approximately 8% [17%]) or highly promiscuous (approximately 2% [6%]). The results obtained for the approved drugs data set may seem surprising: Hit Dexter 2.0 predicts approximately 13% [26%] of approved drugs as promiscuous and 6% [12%] as highly promiscuous (Figure 10E), which is generally even higher than the rates predicted for aggregators. Several approved drugs are known to form colloidal aggregates under specific assay conditions. However, a substantial part of the predicted frequent hitter behavior is likely linked to true promiscuity. Given the challenges involved in designing selective small molecules, this is not only plausible but also forms the basis for drug repurposing and polypharmacology. Note that the percentages of compounds predicted as (highly) promiscuous differ substantially between the (processed) approved drugs data set and the respective subset of compounds well-represented by the training data. These differences are plausible because of substantial differences in the composition of the two data

sets: the processed approved drugs data set consists of around one thousand compounds, and the subset consists of only approximately three hundred compounds.

Interestingly, the distributions of class probabilities for approved drugs are even slightly steeper than those calculated for potential PAINS (Figure 10F). This supports the case that compounds matching PAINS patterns are not necessarily frequent hitters.

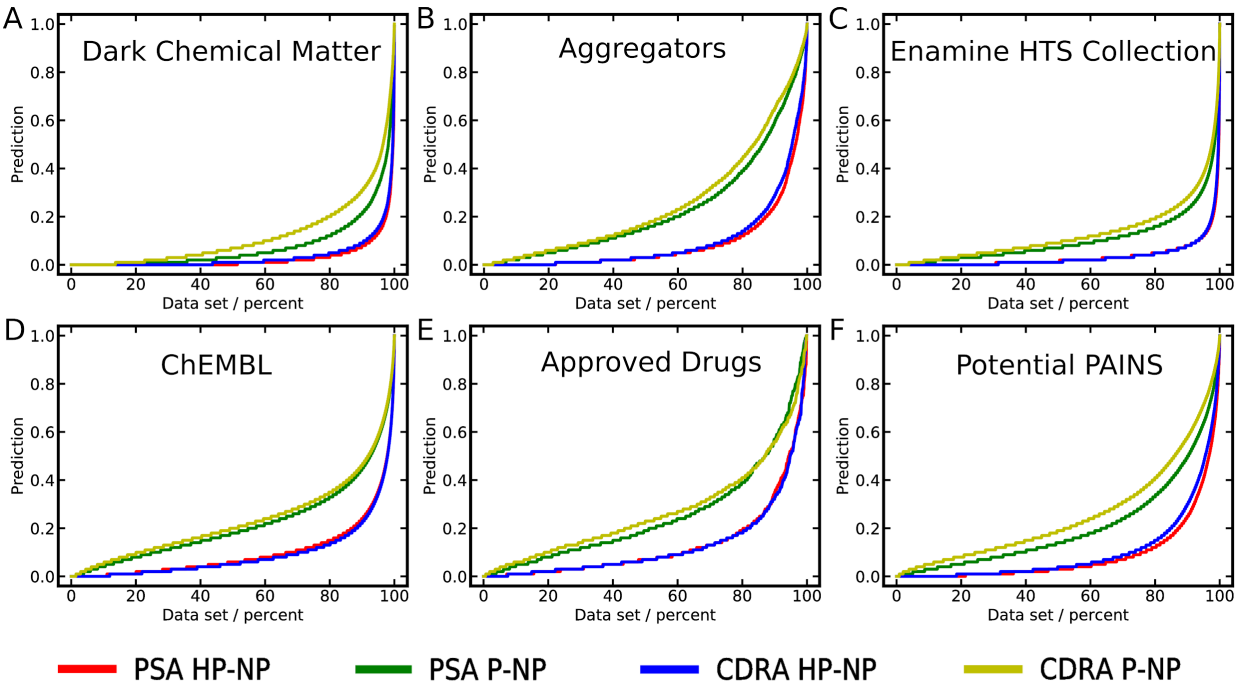

Figure 10. Distribution of class probabilities predicted with classifiers derived from PSA and CDRA data for (A) dark chemical matter, (B) known aggregators, (C) screening compounds from the Enamine HTS Collection, (D) the ChEMBL database, (E) approved drugs from DrugBank and (F) subset of compounds of the Enamine HTS collection that match at least one PAINS pattern.

Last but not least we utilized Hit Dexter 2.0 to predict the promiscuity of 208k natural products. Natural products can be challenging to screen in vitro and it is known that several classes of natural products are prone to interfere with biochemical assays for different reasons.[41] This is reflected by the predictions of Hit Dexter 2.0. The class probability distribution curves (in particular of those of the P-NP classifiers) show a steeper increase than for any other investigated data set (Figure 11A). Particularly noticeable is the high percentage of (highly) promiscuous compounds among flavonoids (the natural product classes were assigned with an automated approach presented previously),[40] with approximately 65% [73%] of all flavonoids predicted as promiscuous and 20% [31%] as highly promiscuous (Figure 10B). Among the investigated flavonoid subclasses (anthocyanidins, chalcones, flavandioles, flavanoles, flavanones, flavanonoles, flavones, flavonoles and isoflavones), chalcones showed the highest rates of highly promiscuous (~42% [50%]) and promiscuous (~85% [86%]) compounds (Figure 11C; note that anthocyanidins are not represented in sufficient numbers in the data set that would allow to draw definite conclusions on their hit rates in assays). In contrast to the observations with flavonoids, Hit Dexter 2.0 reports less than 2% [9%] of all basic alkaloids (see ref 40 for the exact definition) as highly promiscuous and less than 6% [21%] as promiscuous (Figure 10D; note that the subset of compounds covered by the training data according to the above-mentioned criterion is just 10%).

Flavonoids have been reported in the literature to exhibit bioactivity on a large number of different proteins.[41] For example, according to a recent analysis,[41] by the year 2016 more than 680 distinct activities had been reported for quercetin, which is one of the most widely distributed flavonoids in nature but also a known aggregator and PAINS. For this particular flavonoid, PubChem Bioassay currently lists conclusive testing results of more than one

thousand distinct assays, with quercetin reported as active in close to one out of two of these assay outcomes.

Whereas the health-promoting benefits of quercetin and other flavonoids are undisputed, it is reasonable to assume that many of the recorded activities are likely a result of assay interference. It is important to reemphasize at this occasion that the potential of a compound to interfere with assays does not per se lower its value as a bioactive compound, but it may make the rational optimization of its activity a difficult or, in some cases, even impossible task.

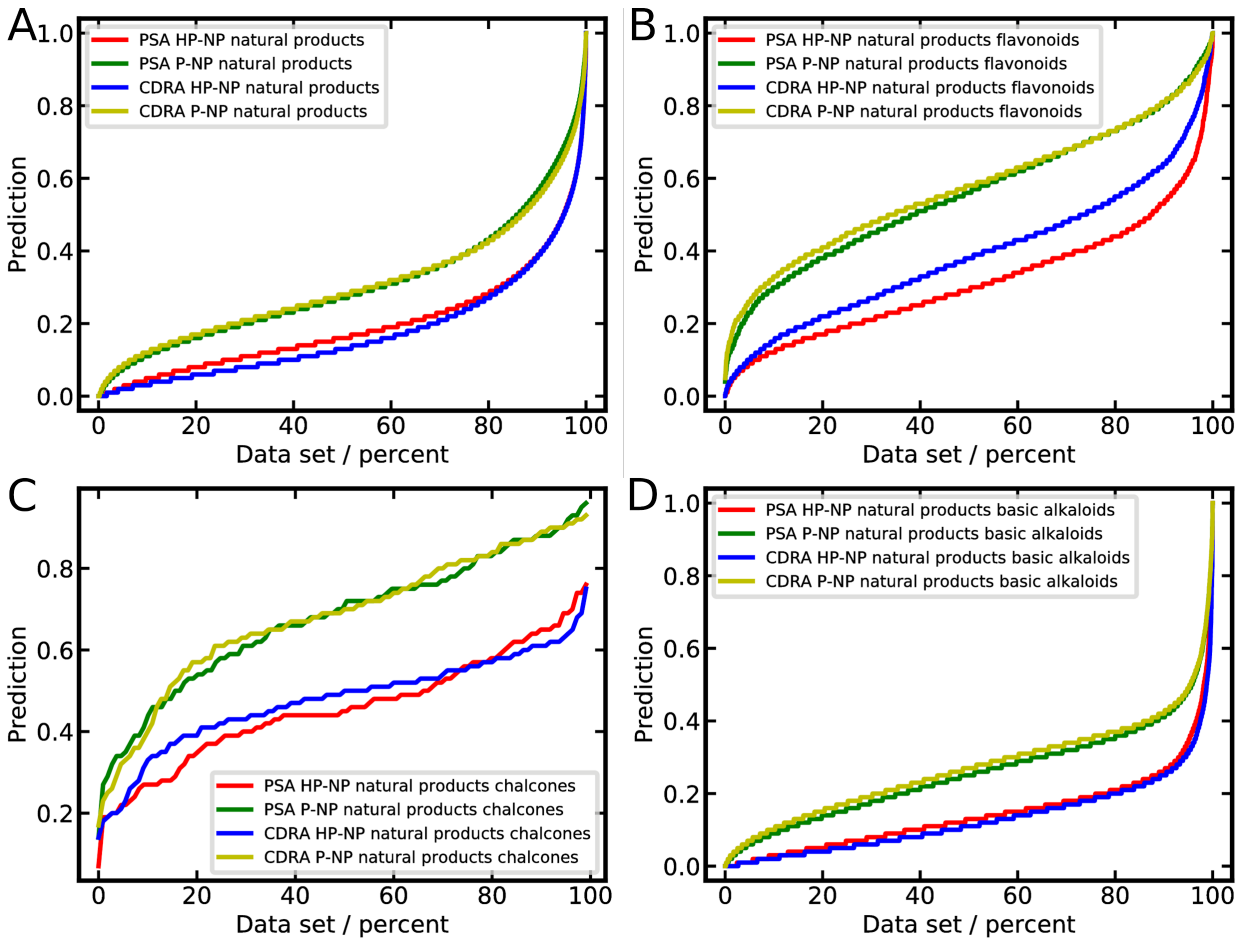

Figure 11. Distribution of class probabilities obtained with classifiers trained on PSA and CDRA data for (A) all natural products, (B) flavonoids, (C) chalcones and (D) basic alkaloids.

Exploration of the Badapple Data Sets with Hit Dexter 2.0

In contrast to Hit Dexter 2.0, which is trained on complete molecular structures, the Badapple model is derived from molecular scaffolds, each of which was assigned a promiscuity score. In order explore to what extent predictions of Hit Dexter 2.0 based on molecular scaffolds are in agreement with Badapple (and the underlying data sets), we compiled sets of 142 468 non-promiscuous scaffolds ("BADAPPLE_NP"), 610 promiscuous scaffolds ("BADAPPLE_P") and 231 highly promiscuous scaffolds ("BADAPPLE_HP") from the Badapple data sets (published in ref 16).

As shown in Figure 12, Hit Dexter 2.0 is able to recognize compound promiscuity based on molecular scaffolds even though it was trained on complete molecular structures. Hit Dexter 2.0 correctly predicted the vast majority (91 to 99% [90 to 99%]) of non-promiscuous scaffolds (Figure 12A). Both P-NP classifiers detected about 57% [72%] of all promiscuous scaffolds as such (Figure 12B), and both HP-NP classifiers predicted around 50% [79%] of the highly promiscuous scaffolds as such (Figure 12C). This can be considered a good agreement for two reasons: First, Hit Dexter 2.0 and Badapple use distinct thresholds for labeling compounds, and second, the BADAPPLE_HP and BADAPPLE_P data sets are small in size and contain only 300 and 100 scaffolds (after preprocessing and removal of duplicates), respectively.

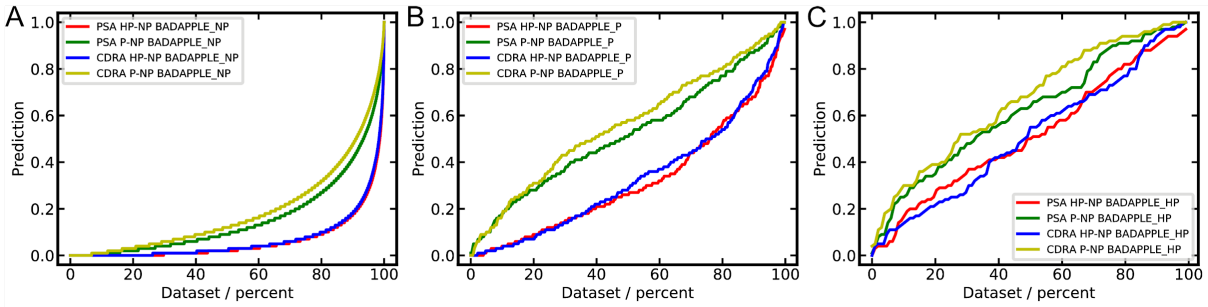

Figure 12. Hit Dexter 2.0 predictions of scaffold promiscuity for the (A) BADAPPLE_NP, (B) BADAPPLE_P and (C) BADAPPLE_HP data sets.

Comparison of the PSA and CDRA Models

As shown in the previous section, the overall behavior and performance of the PSA and CDRA models are comparable. In particular, the numbers of compounds assigned by the different models to one of the three promiscuity classes are similar. One interesting aspect to investigate is the agreement between predictions of the PSA and CDRA classifiers. Table 5 provides an overview of this for each of the above-mentioned data sets. Taking the Approved Drugs subset of DrugBank as an example, the PSA and CDRA models predicted 48 compounds (~5%) and 58 compounds (~6%) of this data set as highly promiscuous. The agreement between both predictions (defined as the fraction of compounds predicted as highly promiscuous by both classifiers as compared to those predicted as highly promiscuous by either classifier) was approximately 40%. Given the fact that only a small number of compounds was predicted as highly promiscuous, this can be considered a good agreement. For the BADAPPLE_HP data set, which consists entirely of highly promiscuous scaffolds, the agreement between the predictions made by the PSA HP-NP and CDRA HP-NP classifiers was 75%. Nevertheless, both classifiers

have different sensitivity and for this reason the use of both predictors in parallel is recommended.

Hit Dexter 2.0 Web Service

The previously introduced Hit Dexter web service[17] was extended substantially. In addition to all models described in this work, we also implemented capabilities to predict aggregators and DCM based on molecular similarity, and to flag nuisance and undesired compounds based on several established collections of SMARTS patterns:

- The "hard filters" rule set developed at Glaxo Wellcome,[42] consisting of 55 patterns of undesired functional groups.
- A rule set developed at the University of Dundee,[43] consisting of 105 patterns of unwanted functional groups and substructures that likely cause interference with HTS assays.
- The "HTS deck filters" rule set developed at Bristol-Meyers Squibb,[44] consisting of 180 patterns of unwanted functional groups derived from intuition and experience.
- The SureChEMBL rule set of ToxAlert,[45] consisting of 166 patterns of toxicophores.
- The "excluded functionality filters" rule set of the NIH Molecular Libraries Small Molecule Repository,[46] consisting of 116 patterns for removing unwanted functional groups.
- The "Lint" rule set developed at Pfizer,[47] consisting 57 patterns of problematic functional groups during drug optimization.
- The PAINS set of substructures linked to assay interference,[7] consisting of 480 patterns. Note that the original PAINS patterns were encoded by Sybyl line notation[7] whereas the

Hit Dexter 2.0 web service utilizes SMARTS patterns in combination with the substructure search implemented in RDKit.[48] This may lead to differing results in some cases.

- A set of 28 substructures derived from undesirable compounds. This is a subset of rules recently introduced by investigators from GlaxoSmithKline. The 28 substructures are listed in Table S2 of ref 23 (value "remove" in column "GSK Recommendation").

Search queries can either be sketched with the JSME Molecule Editor,[49] pasted as individual SMILES or uploaded as a list of SMILES. Predictions are presented as a heat map (Figure 13) and include the results from all the machine learning models, similarity-based and rule-based approaches. Importantly, also the distance to the nearest neighbor in the training data is reported, which gives an indication of the reliability of predictions. A column with comments summarizes the conclusions that may be drawn from the predictions. We believe that these comments will be helpful in particular to occasional users of Hit Dexter 2.0.

The processing of a single compound takes few seconds. Predictions for 1000 compounds take approximately 4 hours. The authors plan to increase the capacity of the web service should the need arise.

Molecule name	SMILES	Comment	Molecular weight (Da)	clogP	Hit Dexter: Probability and prediction confidence of a compound being moderately or highly promiscuous PSA				Hit Dexter: Probability and prediction confidence of a compound being moderately or highly promiscuous CDRA			
					Moderate or high promiscuity	Distance to closest training instance	High promiscuity	Distance to closest training instance	Moderate or high promiscuity	Distance to closest training instance	High promiscuity	Distance to closest training instance
1	CCOC(=O)N1CCN(c	o Predicted as non-promiscuous by the PSA classifier with a probability of 1.0, at high confidence	428.489	0.896	0.000	0.250	0.000	0.250	0.000	0.000	0.000	0.000
3	O=C(C=Cc1cc(O)c(C	o Predicted as promiscuous by the PSA classifier with a probability of 0.99, at moderate confidence	288.255	2.111	0.990	0.420	0.980	0.490	0.940	0.240	0.790	0.240
2	O=c1c(O)c(-c2ccc(O	o Predicted as promiscuous by the PSA classifier with a probability of 1.0, at high confidence	302.238	1.988	1.000	0.220	1.000	0.240	1.000	0.000	0.990	0.000

Showing 1 to 3 of 3 entries

Previous 1 Next

Figure 13. Example of a heat map generated with the Hit Dexter 2.0 web service for three query compounds.

CONCLUSIONS

In this work we report on the second generation of machine learning models for the prediction of frequent hitters independent of the underlying physicochemical mechanisms, including the binding of compounds based on "privileged scaffolds" or of “master key compounds” to multiple binding sites. These models are, among others, accessible via the Hit Dexter 2.0 web service.[19]

In addition to a number of refinements of the data preparation and modeling strategy, substantial improvements presented in this work include the implementation of a protein clustering method in order to avoid an overrepresentation of structurally and functionally related proteins in the training data, and the utilization of PSA data, in addition to CDRA data, for machine learning. During comprehensive tests on independent data, models based on either PSA or CDRA data were shown to predict frequent hitters with high accuracy and robustness. While predictions from both model types were generally in good agreement, the parallel use of both types of classifiers can support the interpretation of results and is recommended.

Hit Dexter 2.0 was used for profiling compounds with specific biological and physicochemical properties, such as dark chemical matter, aggregators, compounds from a high-throughput screening library, drug-like compounds, approved drugs, potential PAINS and natural products. The predictions obtained with Hit Dexter 2.0 confirm common observations and knowledge but also led to some less anticipated observations, such as the high fraction of frequent hitters predicted among approved drugs. A further encouraging observation made was the good

agreement between predictions of Hit Dexter 2.0 and the Badapple data sets of molecular scaffolds and their observed promiscuity.

Since its initial launch in late 2017, the Hit Dexter web service has evolved from a small web presence with rudimentary features into a one-stop shop for the interrogation of compounds regarding their likelihood to exhibit frequent hitter behavior and/or interfere with biochemical assays, and their general desirability in the context of drug discovery. More specifically, the new Hit Dexter 2.0 web service provides user-friendly access to machine learning approaches for the prediction of compound promiscuity, similarity-based methods for the prediction of aggregators and dark chemical matter, and a comprehensive collection of established and new rule sets for flagging frequent hitters and compounds with undesired substructures.

We believe that Hit Dexter 2.0 will enable investigators to make better-informed decisions during hit triage and follow-up. However, the models should not be used as the sole basis for the acceptance or rejection of hits.

METHODS

Data sets

Activity data measured for chemical substances (*substance type* = ”chemical”) on single protein targets (*target* = ”single” and *target type* = ”Protein Targets”) in 932 primary screening assays (*screening stage* = ”primary screening”) and 2266 confirmatory dose-response assays (*screening stage* = ”confirmatory, dose-response”) were separately downloaded from the PubChem Bioassay database[21,22,50] via the PUG REST interface.[51] The download of BioAssay record AID 1224865 failed permanently and was therefore not considered in this work. The SMILES

notations for all 803 898 compounds of the primary screening assays (PSAs) and all 468 260 compounds of the confirmatory dose-response assays (CDRAs) were retrieved via the PubChem Identifier Exchange Service.[52] Salt components, compounds with unsupported elements, and conflicting bioactivity data were identified and removed following the procedure described in ref 17. Also compounds with a molecular weight below 200 Da and above 900 Da were removed. In addition, all molecular structures were neutralized and tautomers merged using the "canonalize" method implemented in the "tautomer" class of MolVS.[53] Subsequently, duplicate compounds were removed based on identical SMILES. In order to ensure the consistency of predictions, compounds with identical Morgan2 fingerprints and differing promiscuity labels (e.g. stereoisomers) were removed from the training sets. This concerned a total of 1945 compounds for the PSA and 2 825 compounds for the CDRA training sets. For any compounds with identical Morgan2 fingerprints only one instance was kept in the training sets.

For each PubChem Bioassay record the unique identifier for genes of the NCBI Protein database[54] ("gene identifier", GI) was obtained via the PubChem PUG REST interface. In total, 429 and 712 unique GIs were retrieved for the PSA and CDRA records, respectively. Subsequently, using these GIs, the protein sequences of all proteins of interest were downloaded in FASTA file format from the NCBI Protein database. Clustering of all protein sequences with *cd-hit* [55] (*sequence identity* = 60%; *tolerance* = 3) resulted in 388 and 537 protein clusters for the PSA and CDRA records, respectively.

For model development, each data set was split randomly into an external test set (10%) and a training set (90%) with the "train_test_split" method of the "model_selection" module of scikit-learn (version: 0.19.1).[35] Only the training set was used for model selection. Initial experiments for selecting the most suitable machine learning algorithm and descriptor sets were performed

with default parameters for ETCs and RFCs, except for the number of estimators, which was set to 50, the class weight, which was set to “balanced”, and bootstrapping, which was disabled. Default parameters were used for all meta classifiers (with ETCs and RFCs parameterized as described above). Stratified splitting was performed as part of cross-validation.

All data sets used to explore and determine the performance of Hit Dexter 2.0 were prepared and filtered according to identical protocol as outlined for the training data.

The Badapple data sets were compiled from the original source[16] by merging scaffolds with identical SMILES and removing any instances with contradicting promiscuity labels. Scaffolds assigned a pScore above 300 were included in the BADAPPLE_HP data set, scaffolds assigned a pScore between 100 and 299 were included in the BADAPPLE_P data set, and scaffolds assigned a pScore between 0 and 99 were included in the BADAPPLE_NP data set.

Hardware and Software

All calculations are performed on Linux workstations running openSUSE 42.2 and equipped with Intel i5 processors (3.2 GHz) and 16GB of main memory.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI:
Additional figures and tables: Maximum pairwise similarity distributions between the PSA50 training set and six chemical data sets.

AUTHOR INFORMATION

Corresponding Author

*J. Kirchmair. E-mail: johannes.kirchmair@uib.no. Tel.: +47 55 58 34 64.

ORCID

Conrad Stork: 0000-0002-5499-742X

Ya Chen: 0000-0001-5273-1815

Martin Šícho: 0000-0002-8771-1731

Johannes Kirchmair: 0000-0003-2667-5877

NOTES

The authors declare no competing financial interest. The Hit Dexter 2.0 web service is available at the following address: http://hitdexter2.zbh.uni-hamburg.de.

ACKNOWLEDGEMENTS

Rainer Fährrolfes, Florian Flachsenberg, Robert Schmidt and Gerd Embruch from the Center of Bioinformatics (ZBH) of the University of Hamburg are thanked for technical support and discussions.

FUNDING

CS and JK are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number KI 2085/1-1. JK is also supported by the Bergen Research Foundation (BFS) - grant no. BFS2017TMT01. YC is supported by the China Scholarship

Council (201606010345). MS is supported by the Ministry of Education of the Czech Republic - project numbers NPU I-LO1220 and LM2015063.

ABBREVIATIONS

ATR, active to tested ratio

AUC, area under the ROC curve

CDRA, confirmatory dose-response assay

DCM, dark chemical matter

ETC, extra tree classifier

HP, highly promiscuous

HTS, high-throughput screening

MCC, Matthews correlation coefficient

MOE, Molecular Operating Environment

NP, non-promiscuous

P, promiscuous

PAINS, pan-assay interference compounds

PCA, principal component analysis

PSA, primary screen assay

RFC, random forest classifier

ROC, receiver operating characteristic

SMARTS, SMILES arbitrary target specification

SMILES, simplified molecular input line entry specification

SMOTE, synthetic minority oversampling technique

REFERENCES

(1) Szymański, P.; Markowicz, M.; Mikiciuk-Olasik, E. Adaptation of High-Throughput Screening in Drug Discovery-Toxicological Screening Tests. *Int. J. Mol. Sci.* **2012**, *13*, 427–452.

(2) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W; Schopfer, U.; Sittampalam, G. S. Impact of High-Throughput Screening in Biomedical Research. *Nat. Rev. Drug Discov.* **2011**, *10*, 188–195.

(3) Janzen, W. P. Screening Technologies for Small Molecule Discovery: The State of the Art. *Chem. Biol.* **2014**, *21*, 1162–1170.

(4) Ganesh, A. N.; Donders, E. N.; Shoichet, B. K.; Shoichet, M. S. Colloidal Aggregation: From Screening Nuisance to Formulation Nuance. *Nano Today* **2018**, *19*, 188–200.

(5) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *J. Med. Chem.* **2002**, *45*, 1712–1722.

(6) Rishton, G. M. Reactive Compounds and in Vitro False Positives in HTS. *Drug Discov. Today* **1997**, *2*, 382–384.

(7) 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.

(8) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.-M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjögren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G. Development of a Virtual Screening Method for Identification of "Frequent Hitters" in Compound Libraries. *J. Med. Chem.* **2002**, *45*, 137–142.

(9) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S. Methods for Drug Discovery: Development of

Potent, Selective, Orally Effective Cholecystokinin Antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246.

(10) Medina-Franco, J. L.; Giulianotti, M. A.; Welmaker, G. S.; Houghten, R. A. Shifting from the Single to the Multitarget Paradigm in Drug Discovery. *Drug Discov. Today* **2013**, *18*, 495–501.

(11) Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery. *J. Med. Chem.* **2014**, *57*, 7874–7887.

(12) Peters, J.-U. Polypharmacology – Foe or Friend? *J. Med. Chem.* **2013**, *56*, 8955–8971.

(13) Stork, C.; Kirchmair, J. PAIN(S) Relievers for Medicinal Chemists: How Computational Methods Can Assist in Hit Evaluation. *Future Med. Chem.* **2018**, *10*, 1533–1535.

(14) Baell, J. B.; Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017-Utility and Limitations. *ACS Chem. Biol.* **2018**, *13*, 36–44.

(15) Kenny, P. W. Comment on The Ecstasy and Agony of Assay Interference Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 2640–2645.

(16) Yang, J. J.; Ursu, O.; Lipinski, C. A.; Sklar, L. A.; Oprea, T. I.; Bologa, C. G. Badapple: Promiscuity Patterns from Noisy Evidence. *J. Cheminform.* **2016**, *8*, 29.

(17) 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, 564-571.

(18) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **1975**, *405*, 442–451.

(19) Hit Dexter 2.0 web service. http://hitdexter2.zbh.uni-hamburg.de (accessed Nov 23, 2018).

(20) Wassermann, A. M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F. J.; Studer, C.; Peltier, J. M.; Grippo, M. L.; Prindle, V.; Tao, J.; Schuffenhauer, A.; Wallace, I. M.; Chen, S.; Krastel, P.; Cobos-Correa, A.; Parker, C. N.; Davies, J. W.; Glick, M. Dark Chemical Matter as a Promising Starting Point for Drug Lead Discovery. *Nat. Chem. Biol.* **2015**, *11*, 958–966.

(21) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound

Databases. *Nucleic Acids Res.* **2016**, *44*, D1202–D1213.

(22) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955–D963.

(23) Chakravorty, S. J.; Chan, J.; Greenwood, M. N.; Popa-Burke, I.; Remlinger, K. S.; Pickett, S. D.; Green, D. V. S.; Fillmore, M. C.; Dean, T. W.; Luengo, J. I.; Macarrón, R. Nuisance Compounds, PAINS Filters, and Dark Chemical Matter in the GSK HTS Collection. *SLAS Discov* **2018**, *23*, 532–544.

(24) M Nissink, J. W.; Blackburn, S. Quantification of Frequent-Hitter Behavior Based on Historical High-Throughput Screening Data. *Future Med. Chem.* **2014**, *6*, 1113–1126.

(25) Molecular Operating Environment (MOE), Version 2016.08; Chemical Computing Group, Montreal, QC.

(26) 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.

(27) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(28) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.

(29) ChEMBL version 23. http://www.ebi.ac.uk/chembl (accessed Dec 8, 2017).

(30) Freund, Y.; Schapire, R. E. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. In *Lecture Notes in Computer Science*; 1995; pp 23–37.

(31) Hastie, T.; Rosset, S.; Zhu, J.; Zou, H. Multi-Class AdaBoost. *Stat. Interface* **2009**, *2*, 349–360.

(32) Breiman, L. Pasting Small Votes for Classification in Large Databases and On-Line. *Mach. Learn.* **1999**, *36*, 85–103.

(33) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140.

(34) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.

(35) 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.

(36) 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.

(37) Aggregator Advisor web service. http://advisor.bkslab.org (accessed Oct 1, 2018).

(38) Enamine HTS collection. https://enamine.net (accessed May 23, 2018).

(39) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.

(40) Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58*, 1518–1532.

(41) Bisson, J.; McAlpine, J. B.; Friesen, J. B.; Chen, S.-N.; Graham, J.; Pauli, G. F. Can Invalid Bioactives Undermine Natural Product-Based Drug Discovery? *J. Med. Chem.* **2016**, *59*, 1671–1690.

(42) 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.

(43) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem* **2008**, *3*, 435–444.

(44) Pearce, B. C.; Sofia, M. J.; Good, A. C.; Drexler, D. M.; Stock, D. A. An Empirical Process for the Design of High-Throughput Screening Deck Filters. *J. Chem. Inf. Model.* **2006**, *46*, 1060–1068.

(45) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2310–2316.

(46) NIH Molecular Libraries Small Molecule Repository. https://grants.nih.gov/grants/guide/notice-files/not-rm-07-005.html (accessed Oct 1, 2018).

(47) Blake, J. Identification and Evaluation of Molecular Properties Related to Preclinical Optimization and Clinical Fate. *McCalls* **2005**, *1*, 649–655.

(48) RDKit version 2016.09.4: Open-Source Cheminformatics Software. http://www.rdkit.org (accessed Nov 23, 2018).

(49) Bienfait, B.; Ertl, P. JSME: A Free Molecule Editor in JavaScript. *J. Cheminform.* **2013**, *5*, 24.

(50) PubChem Bioassay database. https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi (accessed Sep 10, 2018).

(51) Kim, S.; Thiessen, P. A.; Cheng, T.; Yu, B.; Bolton, E. E. An Update on PUG-REST: RESTful Interface for Programmatic Access to PubChem. *Nucleic Acids Res.* **2018**, *46*, W563–W570.

(52) PubChem Identifier Exchange Service. https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi (accessed Sep 11, 2018).

(53) MolVs version 0.1.1. https://github.com/mcs07/MolVS (accessed Jul 12, 2018).

(54) NCBI Resource Coordinators, Nucleic Acids Res. 2016, 44, D7 –D19.

(55) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28*, 3150–3152.

TOC GRAPHICS