Discovering Highly Potent Molecules from an Initial Set of Inactives Using Iterative Screening

Isidro Cortés-Ciriano[1,*], Nicholas C. Firth[2,3], Andreas Bender[1], and Oliver Watson[3]

[1]Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom.

[2] Centre for Medical Image Computing, Department of Computer Science, UCL, London WC1E 6BT, United Kingdom.

[3] Evariste Technologies, London, United Kingdom.

[*]Corresponding author: isidrolauscher@gmail.com

Abstract

The versatility of similarity searching and QSAR to model the activity of compound sets *within* given bioactivity ranges (*i.e.,* interpolation) is well established. However, their relative performance in the common scenario in early-stage drug discovery where lots of inactive data are available, but no active data points (*i.e.,* extrapolation *from the low-activity to the high-activity range*) has not been thoroughly examined yet. To this aim, we have designed an iterative virtual screening strategy which was evaluated on 25 diverse bioactivity data sets from ChEMBL. We benchmark the efficiency of Random Forest (RF), multiple linear regression, Ridge Regression, similarity searching and random selection of compounds to identify a highly active molecule in the test set among a large number of low-potency compounds. We use the number of iterations required to find this active molecule to evaluate the performance of each experimental setup. We show that linear and Ridge Regression often outperform RF and similarity searching, reducing the number of iterations to find an active compound by a factor of two of more. Even simple regression methods seem better able to extrapolate to high-bioactivity ranges than RF, which only provides output values in the range covered by the training set. In addition, examination of the scaffold diversity in the data sets used shows that in some cases similarity searching and RF require two times as many iterations as random selection depending on the chemical space covered in the initial training data. Lastly, we show using bioactivity data for COX-1 and COX-2 that our framework can be extended to multi-target drug discovery, where compounds are selected by concomitantly considering their activity against multiple targets. Overall, this study provides an approach for iterative screening where only inactive data are present in early stages of drug discovery in order to discover highly potent compounds, and the best experimental set up to do so.

Introduction

The wealth of bioactivity data accumulated over the last years in high-throughput screening campaigns enables the application of artificial intelligence (AI) in drug discovery to model compound activity[1–4]. AI enable researchers to work with high-dimensional data that escape human intuition, *e.g.,* the many variables that govern the modulation of protein activity by small molecules, thus making it an unparalleled approach to better understand and model complex systems. Machine learning models are possibly the most common AI approach used in drug development[2–4]. The tasks where machine learning has enabled data-driven decision making, and contributed to unravel fundamental biological aspects of pharmacology, include the prediction of drugs' side effects, computational ADMET profiling, toxicity prediction, the derivation of structural alerts, as well as target-based and ligand-based virtual screening[5–15].

A major goal in early-stage drug discovery campaigns is to discover hits as efficiently as possible. Today, high-throughput screening technologies enable cherry picking of compounds at sufficient speed to perform iterative screening on a large scale (thousands of compounds tested per day)[16–18]. However, *what compounds should be picked* at each iteration still remains a largely unresolved question. Hence, learning patterns in the screening data using artificial intelligence to increase hit rates, and hence discover active molecules faster and more efficiently, is gaining increasing attention[19–22].

Virtual screening (VS) refers to techniques that capitalize on compound and/or target information to discover novel bioactive molecules more efficiently than random selection. These techniques are most often applied in combination with high-throughput screening to prioritize compounds with higher chances of being active for further experimental testing[14,23–28]. Discovering highly selective and potent molecules from a set of initial hits is fundamental to drug discovery, and hence, considerable effort has been invested in the development of VS approaches in both the public and private sectors.

The simplest, and perhaps most widely used VS approach, is similarity searching[29–33]. This approach consists of computing the similarity between a set of active molecules, and a (usually large) collection of structurally diverse compounds of unknown activity. Although the versatility of similarity searching is well established, chemical information across compounds is not

integrated, thus limiting its power to learn chemical patterns predictive of bioactivity across active molecules. Also, what makes two molecules similar, and how to efficiently compute this similarity is still the subject of intense debate, and it depends strongly on the compound descriptors used[32,34,35].

Chemical information and bioactivity data on a target of interest can be integrated using Quantitative Structure-Activity Relationship (QSAR)[12,36,37]. QSAR embraces those mathematical approaches that regress compound activity on compound descriptors. This permits to model (often non-linear) relationships between chemical features and bioactivity across compounds, and hence, interpolate compound activity to the extent the training data allows[38].

Both similarity searching and QSAR rely on the *similarity principle*, which states that the bioactivities of similar molecules (where similar can refer to structural similarity, or similarity in bioactivity space[39–43]) tend to be more correlated than those of dissimilar ones[32,44]. This principle thus implies that a given QSAR model is not likely to generate meaningful predictions for molecules that are dissimilar in descriptor space to those in the training data. Similarly, a model will not accurately predict activity values outside the range covered in the training data (*e.g.,* predict that a compound is active using a model trained on only inactive compounds). Hence the interest in developing reliable techniques to estimate errors in prediction for individual instances[38,45,46].

The applicability domain of a model refers to the regions of descriptor space for which it generates reliable predictions[38]. Commonly used machine learning algorithms in QSAR, such as RF (RF), show high interpolation power (*i.e.,* they perform accurately within their applicability domain). However, their performance in extrapolation (*i.e.,* when applied to molecules outside their applicability domain) is limited, due to the method of prediction used[47]. That is, the predicted value is given as the average value of data from the training set at each leaf.

The virtual screening algorithms published in the literature have mostly focused on the prediction of compound activity *within* the bioactivity range of the training data (*i.e.,* interpolation), but did not consider how to move from inactives to actives (*i.e.,* extrapolation). The question then arises whether, and to what extent, QSAR models can provide meaningful information to guide drug development in cases where only bioactivity data about compounds

with low activity is available, a common scenario in early-stage drug discovery. The interpolation power of diverse QSAR methods has been extensively benchmarked[48,49]. Nevertheless, to the best of our knowledge, a systematic assessment of the efficiency of QSAR models, and similarity searching, to discover highly-active molecules from an initial set of low-active hits is still missing. To fill this shortage, we have benchmarked the efficiency of diverse algorithms using an iterative virtual screening framework and public IC_{50} data for 25 targets from ChEMBL (Figure 1). To quantify their relative performance we use the number of iterations required to discover an active molecule in the test set among a large number of less active molecules. As a starting point, the models are only trained on compounds showing a IC_{50} value several orders of magnitude higher than the active to be found (from 1 to 4 pCI_{50} units; Figure 1). To find general trends in the data, we have designed a factorial experiment with interactions to control for confounding factors affecting model performance. Overall, we show that similarity searching does not perform better than random picking of compounds in many occasions, and that algorithmically simple algorithms, including Ridge and multiple linear regression outperform RF, especially when the bioactivities of the initial training molecules and the active to be found differ by 3-4 pIC_{50} units. From this analysis, we provide guidelines on which method might be more suitable on the basis of the chemical diversity of the available training data, which could ultimately prove valuable to guide and design prospective virtual screening campaigns more efficiently.

Methods

Data Collection and Curation

We gathered IC_{50} data for 25 diverse protein targets from ChEMBL database version 23 using the chembl_webresource_client python module[50–52]. To assemble high-quality data sets, we only kept IC_{50} values for small molecules that satisfied the following filtering criteria: (i) an activity unit equal to "nM", (ii) activity relationship equal to '=', (iii) target type equal to "SINGLE PROTEIN", and (iv) organism equal to *Homo sapiens*. IC_{50} values were modeled in a logarithmic scale ($pIC_{50} = -\log_{10} IC_{50}$). The average pIC_{50} value was calculated when multiple pIC_{50} values were available for the same compound. Further information about the data sets is given in Table 1. All data sets are provided in the Supporting Information.

Table 1. Data sets used in this study.

Target preferred name	Target abbreviation	Uniprot ID	ChEMBL ID	Number of bioactivity data points
Alpha-2a adrenergic receptor	A2a	P08913	CHEMBL1867	203
Tyrosine-protein kinase ABL	ABL1	P00519	CHEMBL1862	773
Acetylcholinesterase	Acetylcholinesterase	P22303	CHEMBL220	3,159
Androgen Receptor	Androgen	P10275	CHEMBL1871	1,290
Serine/threonine-protein kinase Aurora-A	Aurora-A	O14965	CHEMBL4722	2,125
Serine/threonine-protein kinase B-raf	B-raf	P15056	CHEMBL5145	1,730
Cannabinoid CB1 receptor	Cannabinoid	P21554	CHEMBL218	1,116
Carbonic anhydrase II	Carbonic	P00918	CHEMBL205	603
Caspase-3	Caspase	P42574	CHEMBL2334	1,606
Thrombin	Coagulation	P00734	CHEMBL204	1,700
Cyclooxygenase-1	COX-1	P23219	CHEMBL221	1,343
Cyclooxygenase-2	COX-2	P35354	CHEMBL230	2,855
Dihydrofolate reductase	Dihydrofolate	P00374	CHEMBL202	584
Dopamine D2 receptor	Dopamine	P14416	CHEMBL217	479
Norepinephrine transporter	Ephrin	P23975	CHEMBL222	1,740
Epidermal growth factor receptor erbB1	erbB1	P00533	CHEMBL203	4,868
Estrogen receptor alpha	Estrogen	P03372	CHEMBL206	1,705
Glucocorticoid receptor	Glucocorticoid	P04150	CHEMBL2034	1,447
Glycogen synthase kinase-3 beta	Glycogen	P49841	CHEMBL262	1,757
HERG	HERG	Q12809	CHEMBL240	5,207
Tyrosine-protein kinase JAK2	JAK2	O60674	CHEMBL2971	2,655
Tyrosine-protein kinase LCK	LCK	P06239	CHEMBL258	1,352
Monoamine oxidase A	Monoamine	P21397	CHEMBL1951	1,379
Mu opioid receptor	Opioid	P35372	CHEMBL233	840
Vanilloid receptor	Vanilloid	Q8NER1	CHEMBL4794	1,923

Molecular Representation

The python module *standardizer* (https://github.com/flatkinson/standardiser) was used to standardize all chemical structures. Inorganic molecules were removed, and the largest fragment was kept in order to filter out counterions.

We computed circular Morgan fingerprints [53] using RDkit (release version 2013.03.02) [54]. Morgan fingerprints encode compound structures by considering radial atom neighborhoods. The choice of Morgan fingerprints was motivated by the high retrieval rates obtained with these fingerprints in benchmarking studies of compound descriptors [34,55]. The radius was set to 2 and the fingerprint length to 128 to reduce computing time, as we did not obtain significantly higher predictive power when increasing the fingerprint size when modelling these data sets.

Model Training

All models reported here were built using the python library scikit learn[56]. RF models were trained using the default parameter values except for the number of trees, which was set to 100. This value was chosen because using more than 100 trees does not generally lead to increased model performance when modelling bioactivity data sets[57–59], and thus permits to reduce the training times. Default parameter values were used to train Ridge Regression (α=0.1) and multiple linear regression models.

Simulation of Prospective Iterative Screening

- *Data set split*

To simulate a prospective screening scenario, we initially split the data sets into three subsets in the following manner (Figure 1A):

(i) **Inactive molecules** (shown in green in Figure 1A)**:** compounds annotated with an activity value lower than a given bioactivity threshold (parameter 'Max inactives' $\in$ {5,6,7}) are used as the training set. For instance, if the value of the parameter 'Max inactives' is set to 6, all compounds exhibiting a pIC_{50} value equal to or smaller than 6 are selected for training. The goal is to train the models at iteration zero using only inactive or moderately active compounds. This is a common scenario in drug discovery campaigns where the goal is to find highly active molecules starting from a set of low to moderately active compounds.

(ii) **Active molecules** (shown in red in Figure 1A): in each prospective screening simulation, one highly active compound (*i.e.,* IC_{50} larger than the value for the parameter "Min actives"; Figure 1) is kept in the test set, whereas the other active compounds are dismissed from both the training and test sets, and hence, no longer considered in a given simulation. Active compounds are those showing an activity value equal to or larger than the bioactivity threshold defined by the value of the parameter 'Min actives' $\in \{7,8,9\}$.

(iii) **Remaining molecules** (shown in orange in Figure 1A): this subset is composed of those compounds with an activity value equal to or larger than the value of the parameter 'Max inactives' and smaller than the value of the parameter 'Min actives'. All these are kept in the test set (see below).

- *Iterative screening*

We developed a modelling workflow to simulate a prospective drug discovery scenario where, given an initial set of inactive compounds, a RF, linear, or Ridge Regression model is trained on these inactives to predict the activity for a set of molecules (*i.e.,* test set) containing only one active molecule (Figure 1B). The goal is to then test how many screening iterations are required by each method to find this active molecule. Once the predicted values for the test set are calculated, the *C* molecules with the highest predicted activity are selected (where *C* corresponds to the value of the parameter 'number of choices', set to 1 in the current study), and experimentally (or virtually in our framework) tested on the target under consideration. If the active molecule kept in the test set is not identified in one given iteration, it is incorporated into the training set and the model is regenerated. The steps above are repeated until the active molecule is identified and the number of iterations needed to reach this goal is recorded.

We implemented a random picking and a similarity searching approach based on the Tanimoto coefficient[60,61] to serve as baseline methods for comparison. The underlying idea is to evaluate whether a commonly used algorithm in virtual screening, *i.e.* a RF, provides higher extrapolation capabilities than (i) just picking *C* molecules at random from the test set until an active molecule is found, or (ii) picking the *C* molecules showing the highest average Tanimoto similarity to the top *N* active molecules in the training set, where $N \in \{1, 5, 10, 20, 50\}$. The similarity searching

runs are referred to in the Figures and in the main text as “Tanimoto 1”, “Tanimoto 5”, “Tanimoto 10”, “Tanimoto 20”, “Tanimoto 50”, depending on the value of *N*.

The steps of the iterative screening workflow can be summarized as follows (Figure 1B):

1. Train:
    - In the case of RF, linear, and Ridge Regression, train a model on the training set (initially the training set corresponds to the inactive molecules, green box in Figure 1A). Next, use these models to predict the activity for the molecules in the test set (initially composed of one active molecule and those highlighted in orange in Figure 1A).
    - In the case of similarity searching, compute the average Tanimoto similarity between the molecules in the test set and the *N* most active molecules in the training set.
2. Select:
    - The molecule from the test set with the highest predicted activity in the case of RF, linear regression, and Ridge Regression.
    - The molecule showing the highest average Tanimoto similarity to the *N* most active molecules in the training set when using the similarity searching method, or
    - 1 molecule at random from the test set in the case of the random picking approach.
3. Evaluate:
    - If the active molecule is among the selected ones, stop and record the number of iterations needed to identify the target molecule. If not, add this molecule to the training set and repeat steps 1-3 till the active molecule is found.

Experimental Design

The *discovery power* of the models was quantified using the variable “Number of molecules tested”, which corresponds to the number of molecules virtually tested until the active molecule in the test set was found.

To benchmark the *discovery power* of the algorithms described above, we designed a balanced fixed-effect full-factorial experiment with replications[62]. We considered the following 3 factors:

(i) *Data set*: 25 data sets (Table 1).

(ii) *Algorithm*: RF, Ridge Regression, linear regression, Tanimoto similarity searching, and random compound selection.

(iii) pIC_{50} cut-off: this corresponds to the combinations for the values of '*Max inactives*' (cutoff pIC_{50} value to include a molecule in the training set; see Figure 1A), and '*Min actives*' (minimum pIC_{50} value required to consider a molecule active; Figure 1A). We considered the following pairs of '*Max inactives*' and '*Min actives*' values: 5-7, 5-8, 5-9, 6-7, 6-8, 6-9, 7-8, 7-9, and 8-9.

This factorial design was studied with the following linear model:

$$\begin{aligned} &Number\ of\ molecules\ tested \\ &\quad = Data\ set_i + Algorithm_j + pIC50\ cutoff_k + (Data\ set * Algorithm)_{i,j} \\ &\quad + (Data\ set * pIC50\ cutoff)_{i,k} + (Algorithm * pIC50\ cutoff)_{j,k} + \mu_0 + \varepsilon_{i,j,k,l} \\ &(i \in \{1, \dots, N_{data\ sets} = 25\};\ j \in \{1, \dots, N_{algorithms} = 9\};\ k \in \{1, \dots, N_{pIC50\ cutoff} = 8\}; \\ &\quad l \in \{1, \dots, N_{repetitions} = 250\};) \end{aligned}$$

where the response variable, "Number of molecules tested', corresponds to the number of molecules required in a given instance of the simulated virtual screening approach to identify the active molecule contained in the test set. *Data set*$_i$, *Algorithm*$_j$, *pIC*$_{50}$ *cutoff*$_k$ are the main effects considered in the model, while the terms Data set * *Algorithm*, *Data set* $*$*pIC*$_{50}$ *cutoff*, and *Algorithm* $*$*pIC*$_{50}$ *cutoff* correspond to the interaction terms.

The levels "A2a" (*Data set*), "random" (*Algorithm*), and "5-7" (*pIC*$_{50}$ *cut-off*) were used as reference factor levels to calculate the intercept term of the linear model, μ_0, which corresponds to the mean value of 'Number of molecules tested' for this combination of factor levels. The coefficients (slopes) for the other combinations of factor levels correspond to the difference between their mean 'Number of molecules tested' value and the intercept. The error term, $\epsilon_{i,j,k,l,m}$, corresponds to the random error of each 'Number of molecules tested' value, defined as $\epsilon_{i,j,k,l} = \text{Nb. molecules tested}_{i,j,k,l} - \text{mean}(\text{Nb. molecules tested}_{i,j,k})$ These errors are assumed to (i) be mutually independent, (ii) have zero expectation value, and (iii) have constant variance.

We ran 250 prospective screening simulations for each combination of factor levels to ensure a balanced experimental design. We found that this was necessary to obtain robust statistics, as the number of iterations required to find the active molecule varied by two orders of magnitude across replicates. In each of the 250 simulations run for each combination of parameter values, the active molecule to be identified was different. For each combination of factor levels, we chose a different random number as seed for each of the 250 replicates.
This experimental design corresponds to a total of 450,000 simulations (25 data sets × 9 algorithms × 8 pIC_{50} cutoffs × 250 repetitions). The normality and homoscedasticity assumptions of the linear model were respectively assessed with (i) quantile–quantile (Q-Q) plots and (ii) by visual inspection of the distribution of "Nb. molecules tested" values, and by plotting the fitted values against the residuals[62]. Homoscedasticity means that the residuals are evenly distributed (*i.e.,* equally dispersed) across the range of the dependent variable used in the linear model. It is necessary to test this condition to guarantee that the modeling errors (*i.e.,* residuals) and the dependent variable are not correlated. A systematic bias of the residuals would indicate that the errors are not random and that they contain predictive information that should be included in the model[63,64].

Multi-target virtual screening experiments

We used the 1,070 compounds present in both the COX-1 and COX-2 data sets (Table 1) to test the discovery power of RF and Ridge Regression using multiparameter optimization, *i.e.* considering the bioactivity against a set of protein targets in parallel (here, two). In this scenario, the goal is to find a specific target compound in the test set that satisfies two criteria based on its activity on COX-1 and COX-2. The three target compounds were defined as those satisfying the following activity cut-off values: (i) pIC_{50} on COX-1 > 9 and pIC_{50} on COX-2 > 10; (ii) pIC_{50} on COX-1 > 8 and pIC_{50} on COX-2 < 5; and (iii) pIC_{50} on COX-1 < 5 and pIC_{50} on COX-2 > 8.7, corresponding to dual ligands, selective ligands for COX-1 over COX-2, and selective ligands for COX-2 over COX-1.

To compare the efficiency of RF and Ridge Regression to find one of these three target compounds, we trained either two RF or two Ridge Regression models on 200 randomly selected compounds from the whole activity range. These models were then used to predict the activity for the remaining 870 compounds on both COX-1 and COX-2. These sets of predictions were then combined using one of three metrics (see below), the compounds were ranked, and

the top-ranked compound was selected. The simulation was stopped if the selected compound was the target compound. Otherwise, the selected compound was added to the training sets, the two models were regenerated, and the process started again until the target compound was found. Each simulation was run 100 times, each time using a different set of 200 compounds as initial training set.

The three metrics used to rank the compounds in the test set are:

- "Euc": Euclidean distance for the predicted bioactivities for each compound in the test set on COX-1 and COX-2, and the threshold bioactivity values on these two proteins. The compound in the test set displaying the lowest distance was selected.

- "Cumulative Distribution Function (CDF)": we iteratively fit two Ridge or two RF models, one to predict activity on COX-1 and the other to predict activity on COX-2. At each iteration, we calculated the mean squared error for the predictions on the test set for both models, namely $\varepsilon_{COX\text{-}1}$ and $\varepsilon_{COX\text{-}2}$. We next calculated the probability (P_{COX-1}) that the predicted activity for each molecule in the test set on COX-1 ($\hat{y}_{COX-1}$) is higher than the activity cut-off value for COX-1 (T_{COX-1}; *e.g.,* pIC_{50} 9) as:

$$P_{COX-1} = \varphi\left(\frac{\hat{y}_{COX-1} - T_{COX-1}}{\varepsilon_{COX-1}}\right)$$

where $\varphi(x)$ is the normal CDF. Similarly, for COX-2 we calculated P_{COX-2} as:

$$P_{COX-2} = \varphi\left(\frac{\hat{y}_{COX-2} - T_{COX-2}}{\varepsilon_{COX-2}}\right)$$

The probability that the two predicted activities for a given molecule, $\hat{y}_{COX-1}$ and $\hat{y}_{COX-2}$, are in the region of interest (*i.e.,* $\hat{y}_{COX-1} > T_{COX-1}$ and $\hat{y}_{COX-2} > T_{COX-2}$) was

considered to be the product of the two probabilities P_{COX-1} and P_{COX-2}. The molecule with the highest combined probability was selected at each iteration.

- 'Addition': a RF or Ridge Regression model was trained to predict the sum of the activity values, for both dual and selective inhibitors, on both COX-1 and COX-2. This model was then applied to the test set and the compound with the highest predicted value was selected.

Conformal prediction

Cross-conformal predictors were built as previously reported[65,66]. In brief, RF models were trained on 70% of the training data randomly selected using 10-fold cross validation. The cross-validation predictions served to calculate a list of non-conformity values for the molecules in the training set, using the standard deviation across the forest as a scaling factor[58,67]. The validity was assessed on the remaining 30% of the data.

We note that a plethora of methods to compute errors in prediction have been developed[38]. We decided to consider conformal prediction for this analysis as the state of the art given that (i) the calculated confidence intervals are always guaranteed to be valid if the exchangeability principle holds, and (ii) they have proved a versatile technique in diverse early-stage drug discovery applications[20, 58, 74, 65,67–73].

Results and Discussion

Data set modelability

We first evaluated the modelability of the 25 data sets considered using RF (Table 1) in order to test whether our descriptor choice permits to model compound activity with high predictive power. To this end, we trained a RF model on 70% of the data set selected at random (training set), and used the resulting models to predict the activity for the remaining 30% (test set). The mean R^2 values (averaged across 5 replicates) for the observed against the predicted pIC_{50} values on the set were above 0.5 for all data sets (see Figure S1 for details), indicating that our choice of descriptors provides a molecular representation that captures aspects of the chemical structures related to bioactivity. The average $RMSE_{test}$ values were in the 0.5-0.9 range, consistent with the expected modelling errors for heterogeneous IC_{50} data from ChEMBL[75,76].

Extrapolation power of RF and Ridge regression

We next sought to investigate the extrapolation power of linear models and RF. To this end, we used as a training set all compounds in each data set with a pIC_{50} value <7, and as test set the remaining (*i.e.*, higher-activity) data. RF models did not predict values higher than 7 pIC_{50} units for the molecules in the test set in none of the 25 data sets considered (Figure S2). This is consistent with the formulation of RF, as RF predictions are the average value of those similar instances in the training data. Hence, RF models never generate predictions outside the range of activities comprised in the training data[76]. By contrast, Ridge Regression models often extrapolated compound activity to values outside those present in the training set, generating predictions higher than the maximum activity value in the training set (Figure S3). Although the correlation between observed and predicted values for this low-activity to high-activity evaluation is generally poor for all data sets, there are three cases where the errors in prediction for molecules with a pIC_{50} value of around 8 were often smaller than 0.5 (see the performance for the A2a, Carbonic, Dopamine datasets in Figure S3). Overall, these results indicate that less algorithmically complex models, in this case Ridge Regression, extrapolate compound activities to values different from the training set better than RF.

Although inaccurate, extrapolated predictions might be informative if they were accompanied by a quantitative measure of their reliability. Thus, we next evaluated whether the confidence

intervals calculated using conformal prediction[72] serve to identify those predictions that are unreliable because the available training data is not representative of the molecules in the test set, and thus, force the models to extrapolate.

To this aim, we challenged conformal prediction in intrapolation using the workflow previously described[72] across the 25 data sets considered (Methods). As expected, the errors in prediction correlate with the confidence level ($R^2 > 0.95$, $P < 0.01$; Figure S4). This indicates that the confidence intervals provide valuable information about the reliability of individual predictions, which can then serve to prioritize compounds for further experimental testing[20, 69,74]. Next, when the models were trained on compounds with pIC_{50} values below 7 pIC_{50} units and applied to the remaining data. In this case the generated conformal predictors were found to be not valid (Figure S5). The lack of validity of these conformal predictors indicates that the exchangeability principle does not hold; *i.e.,* the molecules in the training data are not representative of those to which the models are to be applied, which is true here given that the training data are inactive or marginally active, while the test data is taken from the highly active range. Hence the obtained confidence intervals do not provide reliable information about the errors of individual predictions.

These results are of great importance for the current study in that they highlight that conformal prediction, as a method that has been shown to generate valid confidence intervals[77], does not permit to generate reliable confidence intervals in cases where the available training data only encompasses inactive molecules and the goal is to find highly-potent molecules with an activity several orders of magnitude higher than the highest active molecule in the training set.

Benchmarking the prospective discovery power of different virtual screening approaches

To evaluate the prospective discovery power of the different virtual screening approaches considering all combinations of model parameters (factor levels), we designed a factorial design that we evaluated using a linear model (see Methods for details). The fitted linear model displayed an R^2 value adjusted for the number of parameters of 0.52 ($P < 10^{-15}$), and a standard error for the residuals of 269.2. This indicates that a substantial fraction of the variability in performance across the simulated iterative screening scenarios can be explained by the factors considered in the linear model, and hence, we can use it to better understand their relative performance in a statistically robust manner. Figures 2 and 3 show the average values of the 'Nb. molecules tested' for all data sets across the levels of the factors *Algorithm* and *pIC_{50} cut-*

offs. The values for the coefficients, namely slopes and intercept, and their *P* values are reported in Table S1. Overall, the residuals are randomly scattered for the region with higher density of "Nb. Molecules tested" values, *i.e.,* 0-500 iterations, and become more dispersed for higher values (see Figure 4A). The distribution of the residuals around zero and the linear trend observed in the Q-Q plot indicates that the assumptions of normality and homoscedasticity of the linear model are fulfilled (Figure 4B-C). Thus, the choice of a linear model is adequate to study the relationship between the discovery power of the models and the parameters considered in the iterative virtual screening framework.

Analysis of the interaction terms in the factorial analysis revealed a significant interaction between the factors *Data set* and *Algorithm* ($P < 10^{-15}$), *Data set* and *pIC_{50} cut-offs* ($P < 10^{-15}$), and *Algorithm* and *pIC_{50} cut-offs* ($P < 10^{-15}$). The presence of interactions is illustrated by the non-parallel lines in Figures 2 and 3, and indicates that the discovery power of the algorithms explored here depends on *both* the data set modelled and the values of the factor *pCI_{50} cut-off*. Thus, the main effects alone cannot explain the variability in discovery power across algorithms. For instance, Ridge Regression permits to find the active molecule in the test set in fewer iterations than RF for data set COX-1 when the value of *pCI_{50} cut-offs* is '5; 8' (green line; Figure 2), whereas it requires more iterations when the value of *pCI_{50} cut-offs* is '5; 9' (purple line; Figure 2).

The difference in performance across algorithms shrinks for most data sets as the difference between the value of "Max inactives" and "Min inactives" decreases (factor *pCI_{50} cut-offs)*. In fact, the performance of the algorithms is overall comparable when the difference between these two parameters is 1 pIC_{50} unit (magenta and light blue lines in Figures 2 and 3), and often comparable to random picking. We observe stronger variability across algorithms when the difference is 3 or 4 pIC_{50} units (purple, green, and grey lines in Figures 2 and 3). Notably, the number of iterations required to find the active molecule in the test set is significantly lower when using Ridge Regression for data sets Acetylcholinesterase, Androgen, and Dopamine (Figure 2), and data sets Opioid and Vanilloid (Figure 3). The similarity searching approach based on the Tanimoto similarity leads to comparable results across the values of *N* for most data sets, except for few cases, *e.g.,* data set A2a (Tanimoto 5; purple line in Figure 2) or data set Caspase (Tanimoto 50; purple line in Figure 2).

More relevant is the fact that random selection of compounds requires a significantly smaller number of iterations than similarity searching and RF in multiple cases. For instance, similarity searching requires about two times more iterations than random picking, Ridge Regression and RF for data sets COX-1 (Figure 2) and Vanilloid (Figure 3), and about the same number of iterations as random picking for data sets Androgen, B-raf, and Dopamine (Figure 2). For some data sets, similarity searching required twice as many iterations than Ridge and linear regression (Figures 2 and 3). The active molecule in the case of the Vanilloid data set (Figure 3) was found after ~600 iterations on average, whereas similarity searching required ~900 iterations. In summary, three trends are apparent from the data: (i) similarity searching and RF tend to perform a biased exploration in a similar area, either in chemical or activity space (or both), so they perform worst, (ii) random picking doesn't extrapolate, but the sampling is not biased to any particular region in chemical or activity space, so this can be considered as the 'neutral option' for selection; and (iii) regression methods (to an extent) can extrapolate, so they perform best.

Influence of chemical diversity on iterative screening

We next sought to determine whether the chemical similarity between the molecules in the training and test sets could explain the notable differences in performance of the studied virtual screening approaches across data sets. To this aim, we computed the Tanimoto similarity for each compound with a pIC_{50} value < 6 against all other compounds in the data set with an activity > 6 pIC_{50} units; Figures S6-10). This analysis revealed that RF performs worse than random where the molecules with pIC_{50} values < 4 and those with values in the 6-8 range are highly similar (see data sets COX-2 and Dihydrofolate; Figure S9). As a more general trend, it is important to highlight that the chemical space of inactives (*e.g.,* pIC_{50} < 6) and medium actives (*e.g.,* pIC_{50} in the 6-8 range) may or not be similar, but the chemical space of highly actives usually is. Medium actives can bridge to the highly actives in cases where they are midway in chemical space from both actives and inactives. However, in cases where there is a gap in chemical space between actives and inactives, only regression might provide a bridge due to their extrapolation power.

Further analysis of specific examples revealed that low scaffold diversity in the training set often underlies the lower performance of similarity searching and RF as compared to random

selection. The enrichment for specific chemical moieties in the training data might lead to overfitting, as the presence of few chemotypes favours the selection of highly similar compounds to these, which in turn, often display similar activities to the low-active molecules in the training set. This might represent a trap for the algorithms, hampering the discovery of other scaffolds displaying high activity. A clear example of this phenomenon occurs in the Vanilloid data set (Figure 5) where 7 out of the most active 18 molecules in the training set (defined as compounds with pIC_{50} values < 6) share a common 2-(3-fluoro-4-methylsulfonylaminophenyl) propenamide scaffold (Figure 5A). Notably, this scaffold is not present in the most active molecules in the data set (Figure 5B). Figure 5C shows that the top ranked molecules in the test set by similarity searching (*i.e.,* molecules with pIC_{50} value between 6 and 8 in this example) contain that scaffold, although with different substituents. These series of compounds were reported in a study of 2-(3-fluoro-4-methylsulfonylaminophenyl) propenamide derivatives[79]. Hence, the abundance of 2-(3-fluoro-4-methylsulfonylaminophenyl) propenamide derivatives in the training data makes similarity searching select compounds from a confined area in chemical space, thus hampering the discovery of the active compounds, and leading to lower performance than random picking (purple line in Figure 3).

The overrepresentation of a scaffold in the training data might be however beneficial when activity cliffs are present, *i.e.,* small structural modifications lead to marked changes in activity[79–83]. In such cases, the highly active molecules in the data set would be very similar to the inactives in the training set, and hence, it would be easy to discover them based on the similarity of their fingerprints. To evaluate this in our data, we computed the average difference in activity for molecules with identical fingerprints, and for those with fingerprints differing in an increasingly higher number of bits (Figure 6). We find compounds leading to exactly the same fingerprint and whose activities differ by > 4 pIC_{50} units (Figures 6 and 7). These generally correspond to molecules that differ in the length of an alkyl chain linking two rings. For instance, the activity of compound CHEMBL3098275 (pIC_{50}: 4.0) is 4 orders of magnitude lower than that of compound CHEMBL3586191 (pIC_{50}: 8.44) in the Estrogen data set. The only difference between these two is the length of the alkyl chain. We note that the fingerprints for these two compounds are the same because the substructures that form the alkyl link map to the same position in the fingerprint. We observe that the presence of activity cliffs affects the relative performance of the virtual screening approaches considered. For instance, RF and similarity searching (Tanimoto 1) outperform other methods in the case of the Estrogen data set,

especially when the initial training data encompasses compounds with pIC_{50} values < 5 (purple squares in Figure 3; Estrogen data set).

Overall, these examples illustrate that low chemical diversity in the initial set of training molecules might lead to overfitting in cases where the active molecules are structurally dissimilar, thus requiring additional iterations to escape the local minimum in chemical space, and find structurally novel chemistry. However, in cases where activity cliffs are present similarity searching and RF permit to reach active molecules faster. Given that ~58% of data points in ChEMBL were obtained from the literature (967,242 unique compounds; 5,635,084 bioactivities in ChEMBL 22[84]), it is thus advisable to investigate the presence of analogue series on a per data set basis to account for the potential overfitting of similarity searching if the goal is to find novel scaffolds. Nevertheless, in cases where a small change might lead to increased activity, the overrepresentation of a given chemical scaffold in the training data might be beneficial. It is important to note that the presence of highly similar compounds, even if they show diverse bioactivities depending on small substituents (see the examples in Figure 7), might be a convenient strategy in optimization, but not in early drug discovery phases, where a wide sampling of the chemical space to find novel scaffolds is sought after.

Although we have considered a wide range of data sets and explored an ample set of parameter settings, we note that our study has several limitations. The data sets used here, despite being diverse in terms of target biology and size, combine heterogeneous IC_{50} data obtained under uncontrolled experimental settings, and cover a biased set of the chemical space, as medicinal chemistry publications often report SAR studies on analogue series. This lack of structural diversity underlies the difference in performance across algorithms in specific cases. For instance, careful analysis of the Vanilloid data set revealed that the low scaffold diversity in the training set, due to the presence of tens of 2-(3-fluoro-4-methylsulfonylaminophenyl) derivatives, significantly decreases the efficiency of RF and similarity searching to discover the active molecule in the test set (Figures 3 and 5). Hence, the results obtained here might not perfectly translate to an industrial setting, as higher-quality proprietary bioactivity data are generally obtained using normalized experimental conditions, chemical libraries are often less biased towards particular chemical families than ChEMBL data, and encompass more chemically diverse molecules. One should also note that there are fundamental biological differences across druggable targets, and hence, these should also

guide whenever possible the design of the most suitable iterative screening approach. Despite these limitations, we believe that the trends reported here are strong enough to challenge the role of similarity searching as a baseline method for the discovery of highly-potent compounds from an initial set of inactives. Our results suggest that using linear models might help reduce resources by 50% in some well-defined scenarios, and thus, make the drug discovery process more efficient.

Multi-target iterative virtual screening

So far, we have considered the activity on one protein target as the only criterion to find an active compound. However, in real-world drug discovery settings it is essential to consider compound activity on other targets to avoid unwanted side-effects, to the extent the *in vitro* data recapitulate *in vivo* effects[5]. Hence, we designed three multiparameter optimization virtual screening experiments, each based on a different metric (see Methods for details), using the data sets COX-1 and COX-2, as these are the only two data sets that share a large number of compounds in ChEMBL (version 23), namely 1,070[85]. COX-1 is constitutively expressed serving as the source of housekeeping prostaglandins, whereas the expression of COX-2 increases in pathophysiological situations such as acute pain, inflammation or cancer[86–88]. Hence, selective COX-2 inhibitors are used to treat pain and inflammation, while avoiding COX-1 mediated side effects (*e.g.*, stomach bleeding)[89].

Here, we used two parameters, activity on COX-1 and COX-2, to concomitantly rank the compounds in the test set. Instead of finding one highly active molecule in the test set, the aim is to find one of the three specific target compounds we selected, each of which satisfies one of the following sets of criteria (indicated with red dots in Figure 8): (i) pIC_{50} on COX-1 > 9 and pIC_{50} on COX-2 > 10 (*i.e.*, dual activity against both isoforms of the COX enzyme); (ii) pIC_{50} on COX-1 > 8 and pIC_{50} on COX-2 < 5; and (iii) pIC_{50} on COX-1 < 9 and pIC_{50} on COX-2 > 8.7. We found that for each of the three target compounds we could find a combination of metric and algorithm that performed significantly better than random picking (P <0.05, t-test). Specifically, the difference in the number of iterations between random picking and the best algorithm-metric combination was 148, 212, and 187 iterations for target compounds 1, 2 and 3, respectively (Table 2). By averaging the results across targets we find that the “Euclidean” and “CDF” metrics outperform “Addition” (P < 0.01), although the effect size is low, namely 22 iterations.

Overall, these results indicate that the multi-parameter optimization methods proposed here can also be used for multi-target drug design. Future studies on data sets encompassing compounds with annotated activities across multiple targets will be needed to more comprehensively evaluate the performance of these metrics and algorithms to perform multiparameter compound optimization.

Table 2 Number of iterations (mean +/- standard error; *n*=100) required to find one of the three target compounds using the multiparameter optimization strategy designed for COX-1 and COX-2.

	Ridge Euc.	**Ridge Addition**	**Ridge CDF**	**Random**
Target compound 1	353.1 +/- 9.0	427.5+/-8.2	362.4+/-8.9	435
Target compound 2	277.1 +/- 9.3	303.9+/-9.0	278.5+/-9.3	435
Target compound 3	407.1 +/- 8.0	370.1+/-8.5	396.8+/-8.1	435
	RF Euc.	**RF Addition**	**RF CDF**	435
Target compound 1	360.4+/-28.9	287.2+/-29.4	372.7+/-29.4	435
Target compound 2	222.6+/-29.0	250.9+/-27.5	223.5+/-28.9	435
Target compound 3	261.3+/-25.1	248.7+/-26.0	251.5+/-25.5	435

Conclusions

In this contribution, we have shown that RF and similarity searching often show comparable performance to random picking in the discovery of novel bioactive molecules. In addition, we show that linear and Ridge Regression often enable the discovery of highly-potent molecules ~2-3 faster than RF and similarity searching. In summary, (i) similarity searching and RF are confined to the original training space and do not extrapolate compound activities; (ii) regression methods can extrapolate, even if not always very predictive, into the desired direction, and enable the discovery of active molecules faster, and (iii) random picking is the baseline method and is not biased either way. The importance of these results are enormous given the strong reliance on similarity searching of a number of drug discovery campaigns to tackle the task of finding active molecules from an initial set of low-potent compounds.

Figures

Figure 1 Framework for the simulation of a prospective screening campaign to find highly-active molecules from a set of inactive/low-active compounds. (**A**) For each data set, an initial training set is assembled using the compounds with a pIC_{50} value smaller than the value of the parameter "Max inactives". The initial test set is composed of compounds with a pIC_{50} value higher than the value of the parameter "Max inactives" and smaller than that of the parameter "Min actives", and one molecule randomly chosen with a pIC_{50} value greater than the value of "Min actives". (**B**) First, the initial training data is used to select *C* molecules (1 in this study) from the test set. If this molecule has an activity higher than that of the parameter "Min actives", the process is stopped. If not, the selected molecule is added to the training set and the process is repeated till the active molecule is found.

Figure 2 Mean Nb. molecules tested for 15 out of the 25 data sets considered averaged across 250 simulations. Each colored line corresponds to the combination of *Max inactive* and *Min inactive* values used in the virtual iterative screening framework shown in Figure 1. The shape of the points indicates the difference between the values of the parameters *Max inactives* and *Min inactives*. The *y*-axis indicates the average number of iterations (n=250) required to find the active molecule in the test set. Overall, the most efficient algorithm corresponds to that requiring the lowest average number of iterations to find the active molecule in the test set. As can be observed by the presence of non-parallel lines (*i.e.,* interactions), the most efficient algorithm varies depending on the parameter values used.

Figure 3 Mean Nb. molecules tested for 10 out of the 25 data sets considered averaged across 250 simulations. Similar to Figure 2.

Figure 4 Verification of the assumptions of the linear model. (A) Heteroscedasticity of the residuals. Fitted "Nb molecules tested" values against the residuals. Overall, the residuals are centered around zero and, roughly, present a comparable dispersion across the range of the dependent variable, indicating that the assumption of the heteroscedasticity of the residuals is fulfilled. The assumption of the normality of the residuals, assessed with the distribution of the residuals (B) and a quantile−quantile (Q-Q) plot (C). The residuals follow a Gaussian distribution with zero mean. This indicates that the assumption of the normality of the residuals is fulfilled.

Figure 5 Example of the effect of the chemical diversity in the training data on the discovery power of similarity searching. (Top panel) The 20 molecules in the training set (Vanilloid data set) with the highest pIC_{50} values are shown. The value for the parameter 'Max inactives' was set to 6 pIC_{50} units. (Middle panel) Top 18 most active molecules in the Vanilloid data set. (Bottom panel) The 18 molecules in the test set with the highest similarity to the top molecules in the training set (top panel) are shown. The red circles highlight the substructure that is overrepresented in the training data set and that leads to overfitting in this case.

Figure 6 Average activity difference for compounds whose fingerprints differ in the number of bits indicated in the *x*-axis. It can be seen that for some data sets the average pIC_{50} difference for compounds leading to the same fingerprint is > 1 pIC_{50} units. This high pIC_{50} difference for structurally similar compounds indicates the presence of activity cliffs in the data sets.

Figure 7 Examples of compound pairs with identical fingerprints but showing markedly different pIC_{50} values on the same target are shown (activity cliffs). These examples illustrate that small structural differences can lead to differences in activity of more than 4 pIC_{50} units. The presence of activity cliffs determines which algorithm performs best to find active molecules: regression methods when activity cliffs are present, and RF and similarity searching in the absence of activity cliffs.

Figure 8 Bioactivities for the 1,070 compounds present in both the COX-1 and COX-2 data sets. The three structures correspond to the three target compounds used in the multi-target iterative screening simulations.

Author Contributions

I.C.-C. and O.W. designed research, trained the models, and analyzed the results. I.C.-C. generated the figures and wrote the paper with substantial input from O.W., A.B., and N.C.F.

Acknowledgements

This project has received funding from the European Union's Framework Programme For Research and Innovation Horizon 2020 (2014-2020) under the Marie Curie Sklodowska-Curie Grant Agreement No. 703543 (I.C.C.). A.B. thanks the European Research Commission (Starting Grant ERC-2013-StG 336159 MIXTURE) for funding. N.C.F is funded by EPSRC (EP/M006093/1).

Conflicts of Interest

I.C.-C., O.W., and N.C.F. hold equity interest in Evariste technologies.

Supporting Information Available: The Supporting Information consists of (i) the 25 data sets used in this study, (ii) Figures S1-10, and (iii) Table S1. The Supporting Information is available free of charge on the ACS Publications website.

References

(1) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250.

(2) Lavecchia, A. Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discov. Today* **2015**, *20*, 318–331.

(3) Lima, A. N.; Philot, E. A.; Trossini, G. H. G.; Scott, L. P. B.; Maltarollo, V. G.; Honorio, K. M. Use of Machine Learning Approaches for Novel Drug Discovery. *Expert Opin. Drug Discov.* **2016**, *11*, 225–239.

(4) Wale, N. Machine Learning in Drug Discovery and Development. *Drug Dev. Res.* **2011**, *72*, 112–119.

(5) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-Scale Prediction and Testing of Drug Activity on Side-Effect Targets. *Nature* **2012**, *486*, 361–367.

(6) Cheng, F.; Li, W.; Liu, G.; Tang, Y. In Silico ADMET Prediction: Recent Advances, Current Challenges and Future Trends. *Curr. Top. Med. Chem.* **2013**, *13*, 1273–1289.

(7) Lavecchia, A.; Di Giovanni, C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem.* **2013**, *20*, 2839–2860.

(8) Moroy, G.; Martiny, V. Y.; Vayer, P.; Villoutreix, B. O.; Miteva, M. A. Toward in Silico Structure-Based ADMET Prediction in Drug Discovery. *Drug Discov. Today* **2012**, *17*, 44–55.

(9) Bulusu, K. C.; Guha, R.; Mason, D. J.; Lewis, R. P. I.; Muratov, E.; Kalantar Motamedi, Y.; Cokol, M.; Bender, A. Modelling of Compound Combination Effects and Applications to Efficacy and Toxicity: State-of-the-Art, Challenges and Perspectives. *Drug Discov. Today* **2015**, *21*, 225–238.

(10) Alves, V. M.; Muratov, E. N.; Capuzzi, S. J.; Politi, R.; Low, Y.; Braga, R. C.; Zakharov, A. V.; Sedykh, A.; Mokshyna, E.; Farag, S.; Andrade, C. H.; Kuz'min, V. E.; Fourches, D.; Tropsha, A. Alarms about Structural Alerts. *Green Chem.* **2016**, *18*, 4348–4360.

(11) Yang, H.; Li, J.; Wu, Z.; Li, W.; Liu, G.; Tang, Y. Evaluation of Different Methods for Identification of Structural Alerts Using Chemical Ames Mutagenicity Data Set as a Benchmark. *Chem. Res. Toxicol.* **2017**, *30*, 1355–1364.

(12) Cortes-Ciriano, I. Bioalerts: A Python Library for the Derivation of Structural Alerts from Bioactivity and Toxicity Data Sets. *J. Cheminform.* **2016**, *8*, 13.

(13) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010.

(14) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, *52*, 867–881.

(15) Schneider, G. Virtual Screening: An Endless Staircase? *Nat. Rev. Drug Discov.* **2010**, *9*, 273–276.

(16) Crouch, S. P. .; Slater, K. J. High-Throughput Cytotoxicity Screening: Hit and Miss. *Drug Discov. Today* **2001**, *6*, 48–53.

(17) Martis, E. A.; Radhakrishnan, R. High-Throughput Screening : The Hits and Leads of Drug Discovery- An Overview. *J. Appl. Pharm. Sci.* **2011**, *01*, 2–10.

(18) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discov.*

2002, *1*, 882–894.
(19) Paricharak, S.; IJzerman, A. P.; Bender, A.; Nigsch, F. Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-House HTS Data. *ACS Chem. Biol.* **2016**, *11*, 1255–1264.
(20) Ahmed, L.; Georgiev, V.; Capuccini, M.; Toor, S.; Schaal, W.; Laure, E.; Spjuth, O. Efficient Iterative Virtual Screening with Apache Spark and Conformal Prediction. *J. Cheminform.* **2018**, *10*, 8.
(21) Svensson, F.; Norinder, U.; Bender, A. Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 439–444.
(22) Garnett, R.; Gärtner, T.; Vogt, M.; Bajorath, J. Introducing the 'Active Search' Method for Iterative Virtual Screening. *J. Comput. Aided. Mol. Des.* **2015**, *29*, 305–314.
(23) Klebe, G. Virtual Ligand Screening: Strategies, Perspectives and Limitations. *Drug Discov. Today* **2006**, *11*, 580–594.
(24) Köppen, H. Virtual Screening - What Does It Give Us? *Curr. Opin. Drug Discov. Devel.* **2009**, *12*, 397–407.
(25) Song, C. M.; Lim, S. J.; Tong, J. C. Recent Advances in Computer-Aided Drug Design. *Brief. Bioinform.* **2009**, *10*, 579–591.
(26) Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H. Structure-Based Virtual Screening for Drug Discovery: A Problem-Centric Review. *AAPS J.* **2012**, *14*, 133–141.
(27) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W.; Jr. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334–395.
(28) Zhang, S. Computer-Aided Drug Discovery and Development. In *Methods in molecular biology (Clifton, N.J.)*; 2011; vol. 716, pp 23–38.
(29) Willett, P.; And, J. M. B.; Downs, G. M. Chemical Similarity Searching. **1998**.
(30) Stumpfe, D.; Bajorath, J. Similarity Searching. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 260–282.
(31) Willett, P. Similarity Searching Using 2D Structural Fingerprints. In *Methods in molecular biology (Clifton, N.J.)*; 2010; vol. 672, pp 133–158.
(32) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.
(33) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.* **2015**, *7*, 20.
(34) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49*, 108–119.
(35) Bender, A. How Similar Are Those Molecules after All? Use Two Descriptors and You Will Have Three Different Answers. *Expert Opin. Drug Discov.* **2010**, *5*, 1141–1151.
(36) Hansch, C. Quantitative Approach to Biochemical Structure-Activity Relationships. *Acc. Chem. Res.* **1969**, *2*, 232–239.
(37) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178–180.
(38) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Mark, T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Carol, A.; Myatt, G.; Nikolova-jeliazkova, N.; Patlewicz, G. Y.; Perkins, R. *Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure – Activity Relationships*; 2005; vol. 2.
(39) Nguyen, H. P.; Koutsoukas, A.; Mohd Fauzi, F.; Drakakis, G.; Maciejewski, M.; Glen, R. C.; Bender, A. Diversity Selection of Compounds Based on "Protein Affinity Fingerprints" Improves Sampling of Bioactive Chemical Space. *Chem. Biol. Drug Des.* **2013**, 252–266.
(40) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.;

Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting Ligand Binding to Proteins by Affinity Fingerprinting. *Chem. Biol.* **1995**, *2*, 107–118.
(41) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. a. Biological Spectra Analysis: Linking Biological Activity Profiles to Molecular Structure. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 261–266.
(42) Paull, K. D.; Shoemaker, R. H.; Hodes, L.; Monks, A.; Scudiero, D. A.; Rubinstein, L.; Plowman, J.; Boyd, M. R. Display and Analysis of Patterns of Differential Activity of Drugs Against Human Tumor Cell Lines: Development of Mean Graph and COMPARE Algorithm. *J. Natl. Cancer Inst.* **1989**, *81*, 1088–1092.
(43) Zaharevitz, D. W.; Holbeck, S. L.; Bowerman, C.; Svetlik, P. A. COMPARE: A Web Accessible Tool for Investigating Mechanisms of Cell Growth Inhibition. *J. Mol. Graph. Model.* **2002**, *20*, 297–303.
(44) Willett, P. Similarity Methods in Chemoinformatics. *Annu. Rev. Inform. Sci.* **2009**, *43*, 3–71.
(45) Toplak, M.; Močnik, R.; Polajnar, M.; Bosnić, Z.; Carlsson, L.; Hasselgren, C.; Demšar, J.; Boyer, S.; Zupan, B.; Ståhring, J. Assessment of Machine Learning Reliability Methods for Quantifying the Applicability Domain of QSAR Regression Models. *J. Chem. Inf. Model.* **2014**, *54*, 431–441.
(46) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.
(47) Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC_{50} s for Realistically Novel Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 2077–2088.
(48) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504.
(49) Tetko, I. V; Solov'ev, V. P.; Antonov, A. V; Yao, X.; Doucet, J. P.; Fan, B.; Hoonakker, F.; Fourches, D.; Jost, P.; Lachiche, N.; Varnek, A. Benchmarking of Linear and Nonlinear Approaches for Quantitative Structure-Property Relationship Studies of Metal Complexation with Ionophores. *J. Chem. Inf. Model.* **2006**, *46*, 808–819.
(50) Nowotka, M.; Papadatos, G.; Davies, M.; Dedman, N.; Hersey, A. Want Drugs? Use Python. **2016**.
(51) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic Acids Res.* **2015**, *43*, W612-20.
(52) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2011**, *40*, 1100–1107.
(53) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
(54) Landrum, G. RDKit: Open-Source Cheminformatics.
(55) Koutsoukas, A.; Paricharak, S.; Galloway, W. R. J. D.; Spring, D. R.; IJzerman, A. P.; Glen, R. C.; Marcus, D.; Bender, A. How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2013**, *54*, 230–242.
(56) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
(57) Sheridan, R. P. Three Useful Dimensions for Domain Applicability in QSAR Models Using Random Forest. *J. Chem. Inf. Model.* **2012**, *52*, 814–823.

(58) Cortés-Ciriano, I.; van Westen, G. J. P.; Bouvier, G.; Nilges, M.; Overington, J. P.; Bender, A.; Malliavin, T. E. Improved Large-Scale Prediction of Growth Inhibition Patterns on the NCI60 Cancer Cell-Line Panel. *Bioinformatics* **2016**, *32*, 85–95.
(59) Firth, N. C.; Atrash, B.; Brown, N.; Blagg, J. MOARF, an Integrated Workflow for Multiobjective Optimization: Implementation, Synthesis, and Biological Evaluation. *J. Chem. Inf. Model.* **2015**, *55*, 1169–1180.
(60) Salim, N.; Holliday, J.; Willett, P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435–442.
(61) Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci. 42*, 1407–1414.
(62) Winer, B.; Brown, D.; Michels, K. *Statistical Principles in Experimental Design*, 3rd ed.; Psychology, M.-H. series in, Ed.; McGraw-Hill: New York, NY, USA, 1991.
(63) Roy, K.; Ambure, P.; Aher, R. B. How Important Is to Detect Systematic Error in Predictions and Understand Statistical Applicability Domain of QSAR Models? *Chemom. Intell. Lab. Syst.* **2017**, *162*, 44–54.
(64) Roy, K.; Das, R. N.; Ambure, P.; Aher, R. B. Be Aware of Error Measures. Further Studies on Validation of Predictive QSAR Models. *Chemom. Intell. Lab. Syst.* **2016**, *152*, 18–33.
(65) Sun, J.; Carlsson, L.; Ahlberg, E.; Norinder, U.; Engkvist, O.; Chen, H. Applying Mondrian Cross-Conformal Prediction To Estimate Prediction Confidence on Large Imbalanced Bioactivity Data Sets. *J. Chem. Inf. Model.* **2017**, *57*, 1591–1598.
(66) Vovk, V. Cross-Conformal Predictors. *Ann. Math. Artif. Intell.* **2015**, *74*, 9–28.
(67) Svensson, F.; Aniceto, N.; Norinder, U.; Cortes-Ciriano, I.; Spjuth, O.; Carlsson, L.; Bender, A. Conformal Regression for Quantitative Structure–Activity Relationship Modeling—Quantifying Prediction Uncertainty. *J. Chem. Inf. Model.* **2018**, acs.jcim.8b00054.
(68) Svensson, F.; Norinder, U.; Bender, A. Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 439–444.
(69) Norinder, U.; Boyer, S. Conformal Prediction Classification of a Large Data Set of Environmental Chemicals from ToxCast and Tox21 Estrogen Receptor Assays. *Chem. Res. Toxicol.* **2016**, *29*, 1003–1010.
(70) Norinder, U.; Boyer, S. Conformal Prediction Classification of a Large Data Set of Environmental Chemicals from ToxCast and Tox21 Estrogen Receptor Assays. *Chem. Res. Toxicol.* **2016**, *29*, 1003–1010.
(71) Cortés-Ciriano, I.; Bender, A.; Malliavin, T. Prediction of PARP Inhibition with Proteochemometric Modelling and Conformal Prediction. *Mol. Inform.* **2015**, *34*, 357–366.
(72) Svensson, F.; Norinder, U.; Bender, A. Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 439–444.
(73) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014**, *54*, 1596–1603.
(74) Johansson, U.; Ahlberg, E.; Boström, H.; Carlsson, L.; Linusson, H.; Sönströd, C. Handling Small Calibration Sets in Mondrian Inductive Conformal Regressors; Springer, Cham, 2015; pp 271–280.
(75) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC_{50} Data - a Statistical Analysis. *PLoS One* **2013**, *8*, e61007.
(76) Cortés-Ciriano, I.; Bender, A. How Consistent Are Publicly Reported Cytotoxicity Data? Large-Scale Statistical Analysis of the Concordance of Public Independent Cytotoxicity

Measurements. *ChemMedChem* **2015**, *11*, 57–71.
(77) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
(78) Shafer, G.; Vovk, V. A Tutorial on Conformal Prediction. *J. Mach. Learn. Res.* **2008**, *9*, 371–421.
(79) Ha, T.-H.; Ryu, H.; Kim, S.-E.; Kim, H. S.; Ann, J.; Tran, P.-T.; Hoang, V.-H.; Son, K.; Cui, M.; Choi, S.; Blumberg, P. M.; Frank, R.; Bahrenberg, G.; Schiene, K.; Christoph, T.; Frormann, S.; Lee, J. TRPV1 Antagonist with High Analgesic Efficacy: 2-Thio Pyridine C-Region Analogues of 2-(3-Fluoro-4-Methylsulfonylaminophenyl)Propanamides. *Bioorg. Med. Chem.* **2013**, *21*, 6657–6664.
(80) Bajorath, J. Representation and Identification of Activity Cliffs. *Expert Opin. Drug Discov.* **2017**, *12*, 879–883.
(81) Hu, Y.; Bajorath, J. Extending the Activity Cliff Concept: Structural Categorization of Activity Cliffs and Systematic Identification of Different Types of Cliffs in the ChEMBL Database. *J. Chem. Inf. Model.* **2012**, *52*, 1806–1811.
(82) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932–2942.
(83) Maggiora, G. M. On Outliers and Activity Cliffs--Why QSAR Often Disappoints. *J. Chem. Inf. Model. 46*, 1535.
(84) and, R. G.; John H. Van Drie*, †. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. **2008**.
(85) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.
(86) Cortes-Ciriano, I.; Murrell, D. S.; van Westen, G. J. P.; Bender, A.; Malliavin, T. Prediction of the Potency of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling. *J. Cheminf.* **2014**, *7*, 1.
(87) Luo, C.; He, M.; Bohlin, L. Is COX-2 a Perpetrator or a Protector? Selective COX-2 Inhibitors Remain Controversial. *Acta Pharmacol. Sin.* **2005**, *26*, 926–933.
(88) Moore, B. C.; Simmons, D. L. COX-2 Inhibition, Apoptosis, and Chemoprevention by Nonsteroidal Anti-Inflammatory Drugs. *Curr. Med. Chem.* **2000**, *7*, 1131–1144.
(89) Kismet, K.; Akay, M. T.; Abbasoglu, O.; Ercan, A. Celecoxib: A Potent Cyclooxygenase-2 Inhibitor in Cancer Prevention. *Cancer Detect. Prev.* **2004**, *28*, 127–142.
(90) Fine, M. Quantifying the Impact of NSAID-Associated Adverse Events. *Am. J. Manag. Care* **2013**, *19*, s267--272.