Luciferase Advisor: High-Accuracy Model To Flag False Positive Hits in Luciferase HTS Assays

Dipan Ghosh[1], Uwe Koch[2], Kamyar Hadian[3], Michael Sattler[4] and Igor V. Tetko[1,5*]

[1]Institute of Structural Biology, Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany.
[2]Lead Discovery Center GmbH, Otto-Hahn-Straße 15, 44227 Dortmund, Germany
[3]Assay Development and Screening Platform, Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany
[4]Bayerisches NMR-Zentrum, Department of Chemistry, Technical University of Munich, Ernst-Otto-Fischer-Straße 2, 85747 Garching, Germany
[5]BIGCHEM GmbH, Ingolstaedter Landstrasse 1 b. 60w, 85764 Neuherberg, Germany

* Address for correspondence
Dr. Igor V. Tetko,
Institute of Structural Biology, Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany
Email: itetko@vcclab.org
Tel.: +49-89-3187-3575
Fax: +49-89-3187-3585

Keywords: Luciferase, False Positives, Machine Learning, Big Data, OCHEM

Abstract
Firefly Luciferase is an enzyme that has found ubiquitous use in biological assays in high-throughput screening (HTS) campaigns. The inhibition of luciferase in such assays could lead to a false positive result. This issue has been known for a long time and there have been significant efforts to identify luciferase inhibitors, to enhance recognition of false positives in screening assays. However, although a large amount of publicly accessible luciferase counterscreen data is available, so far little effort has been devoted to building a chemoinformatic model that can identify such molecules in a given dataset.
In this study we developed models to identify these molecules using various methods, such as molecular docking, SMARTS screening, pharmacophores, and machine learning methods. Amongst the structure-based methods, the pharmacophore-based method showed promising results, with a balanced accuracy of 74.2%. However, machine-learning approaches using associative neural networks outperformed all the other methods explored, producing a final model with a balanced accuracy of 89.7%. The high predictive accuracy of this model is expected to be useful for advising which compounds are potential luciferase inhibitors present in luciferase HTS assays. The models developed in this work are freely available at the OCHEM platform at http://ochem.eu.

Introduction

With advances in molecular biology and other areas such as engineering and computation, high-throughput assay formats have become routine and are widely used in early stage drug discovery today[1]. For hit detection a large fraction (~20%)[2] of these assays rely on bioluminescence; a technique that reduces background noise and benefits from an excellent signal/noise ratio. Such assays primarily rely on the luciferase enzyme, which is naturally found in various organisms across the animal kingdom, such as the firefly (*Photinus* sp.), larvae of certain beetles known as glow worms, and various marine organisms. Among these, the firefly luciferase (FLuc) obtained from fireflies (*Photinus pyralis*) is the most common and widely used variant. The natural substrate for Luciferase is luciferin. The enzyme catalyzes the production of oxyluciferin and light via a Luciferyl-adenylate intermediate, which is detected and measured in the assay.

It has been known for a long time that ligand molecules tested in luciferase-based assays can inhibit the luciferase protein, and thus affect the assay outcome[3-5]. For this reason, there has been significant interest towards understanding and evaluating luciferase inhibition, especially in the context of a high-throughput assay[1]. In 2008, Auld *et al*[6] published the first comprehensive study, where they tested ~72000 compounds for luciferase inhibition. They also identified important scaffolds for FLuc inhibition[6]. In 2012, the same group published a follow-up study where they tested a much larger set of compounds, and identified a few additional scaffolds[2]. They also published a crystal structure of benzothiol, an inhibitor, bound to FLuc, establishing the binding mode and identifying key interactions[2].

However, despite significant interest and large datasets being publicly available, there has been little to no reported effort towards building a computational model for luciferase inhibitors. Such models could potentially be used to identify and filter out these aberrant and false positive results from a high-throughput assay with good accuracy and relative ease. The goal of our study is to develop a model that can advise against possible luciferase inhibitors present in a HTS dataset. In this study, we analyzed the publicly available data to build such a model using machine-learning methods that can identify luciferase inhibitors. We also analyzed the influence of molecular shape and geometry in luciferase inhibition.

Data

All data used in this study are publicly available in PubChem, as summarized in Table 1. The activity data were downloaded in spreadsheet format and structures in SMILES format from PubChem following the Substance ID. The data were then uploaded to the OCHEM platform, which has established workflows for normalizing and managing the data. The data gathered were processed to look for overlap in the compounds tested, which should give an idea about the coverage and reproducibility of the results. We found significant overlaps between the datasets (Figure 1).

Table 1: Summery of the data used in this study including PubChem Assay ID.

set	Concentration used for testing	Number of compounds tested	Number of compounds after excluding inconclusive	% of Actives	PubChem AID†	Year
1	50 µM	72359	70658	2.17	411	2008
1	11.5 µM	70231	70231	0.72		
2	10 µM	195634	195634	1.52	1006	2010
3	50 µM	364105	326367	6.91	588342	2012
3	11.5 µM	323224	323224	3.25		

† AID stands for Assay ID

Set2 is a complete subset of Set3, and Set1 has some unique compounds with respect to Set3. The union of all sets has a size of 375001 compounds. This size of data is good for building models and performing analysis.
In Set1 and 3 there were a few molecules with inconclusive properties. For these molecules, it was not possible to obtain a concentration-response curve, and therefore the activity was uncertain. We excluded these molecules from our analysis, and because of this, Set2 was no longer a complete subset of Set3.

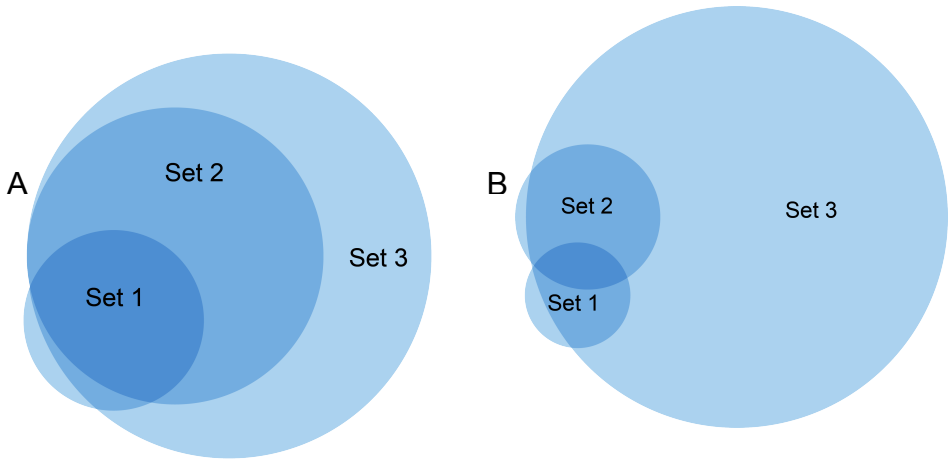


Figure 1: Venn diagram representation of the datasets used. The sizes of the circles reflect the relative sizes of the datasets. **A.** All molecules **B.** Active molecules

We also performed a similar analysis on the active compounds from the three assays. Here we noticed that Set3 has a much larger active compound pool as compared to the others (Figure 1B). This is explained by the fact that there is a significant difference between the highest concentrations tested in the respective assays: Set2 was measured at a maximum concentration of 10 µM, whereas both

Set1 and Set3 were tested at a maximum concentration of 50 μM. Due to the higher concentration, Set1 and Set3 contained larger percentages of active molecules compared to Set2. To compare data at the same concentration for all sets, we extracted and used the inhibition data at 11.5 μM for Set1 and Set3. For a few molecules, there were no data points available at 11.5 μM, hence they were not considered.

We found that the more recent assays had a significantly larger percentage of active molecules when compared at the same concentration (Table 1). This could be either due to difference in the chemical spaces, or that more recently performed assays are more sensitive due to improvements in assay technology. To assess if chemical space plays a role, we analyzed the common molecules in all three assays (N = 61224). We found the same increasing trend (0.7%, 1.0% and 2.4% for Set1, 2 and 3 respectively). Because the chemical space is fixed, this result points to an increase in assay sensitivity. Indeed, to identify potential luciferase inhibition through counterscreening, calibration of the counterscreen assay with known inhibitors is recommended to determine assay sensitivity[2]. Because of this problem the different assays cannot be directly compared.

Methods

Docking studies

For molecular docking, Autodock Vina was used[7]. SMILES of the molecules were downloaded from PubChem, and, using CORINA[8], their optimized 3D structures were obtained. The molecules were prepared for docking using AutoDockTools[9], and were then docked onto the luciferase enzyme with an optimal bounding box enclosing the binding pocket. The binding box was chosen to be large enough to cover the intended docking site, but not too large, in order to minimize calculation time. Default settings were used for the preparation and docking processes.

The resulting docking poses were analyzed using PYMOL[10]. A plane was defined, by choosing three points just outside the binding pocket. This plane denoted the beginning of the binding pocket and for each atom of a ligand, a position vector was calculated with respect to this plane. From this, we calculated which atoms were inside and outside the binding pocket. This information was then averaged over all the docking poses, resulting in the final score that determined how much of a ligand was inside the binding pocket.

Pharmacophore Analysis

Because crystal structure of luciferase bound to an inhibitor was available, we investigated a 3D-structure-based pharmacophore approach to distinguish between the active molecules and the inactives. The pharmacophore development and screening were performed using LigandScout[11]. The detailed procedure for developing the pharmacophores has been described in the results section.

Machine learning methods. Using the freely accessible platform On-line Chemical and Modeling Environment (OCHEM)[12], we built more than 150 models for all three datasets. We used primarily Associative Neural Networks (ASNN)[13, 14] and Support Vector Machine (LibSVM)[15] algorithms for training the models. Associative neural network (ASNN) is an ensemble-based method inspired by the function and structure of neural network correlations in brain. The method operates by

simulating the short- and long-term memory of neural networks and thalmo-cortical organization of brain.[16] These methods on average provided the highest predictive accuracy in comparison to other methods available on the OCHEM web site. The methods were used with default parameters as specified on the OCHEM web site.

Molecular descriptors. A variety of descriptors available within the OCHEM environment were used to train the models.
Adriana.Code[17] comprises a unique combination of topological (2D), spatial (3D) and global molecular descriptors calculated on a sound geometric and physicochemical basis. Adriana offers simple molecular property descriptors such as molecular weight and molecular dipole moment as well as increasingly sophisticated geometric descriptors such as Molecular Radius of Gyration.
ALogPS calculates two descriptors provided by the ALOGPS[18] program, which determine the water/octanol partition coefficient ($logP_{calc}$), and water solubility coefficient ($logS_{calc}$)[19].
CDK (3D) or the Chemistry Development Kit is an open source chemoinformatics project[20]. There are several types of descriptors available from the package, that are integrated into the OCHEM environment. Descriptors calculated with the recently released 2.0 version of CDK were used in this study[21].
ChemAxon Descriptors (3D) are a set of descriptors developed and implemented by the ChemAxon company[22]. The available descriptors are subdivided into seven categories, namely Elemental Analysis, Charge, Geometry, Partitioning, Protonation, Isomers, and Others. Descriptors that return a Boolean or Numerical value were implemented into OCHEM.
Dragon[23] (3D) is a well-known software package, for the calculation of molecular descriptors, developed by the Milano Chemometrics and QSAR Research Group of Prof. R. Todeschini. It comprises perhaps one of the largest and most comprehensive molecular descriptor libraries available, with a total of 5,270 descriptors available. The descriptors are divided into 30 discrete blocks, such as Topological, Constitutional, Drug-like indices, etc. Dragon is built into OCHEM, and for this study, Dragon version 6 was used.
GSFRAG[24] belongs to the category of 2D fragment descriptors. It calculates the occurrence numbers of certain special fragments from k=2 to 10 vertices in a molecular graph G, that can be used as molecular descriptors in quantitative structure-property/activity studies.
ISIDA descriptors are part of the ISIDA project, which stands for *In-SIlico* Design and data Analysis[25]. These fragment-like 2D descriptors are calculated from molecular graphs using three different methods, namely paths, trees, and neighbors. The descriptors are generated from the fragments by using different atom and bond labeling methods[26].
Mera and Mersy[27] (3D) are two related groups of descriptors. Mera provides a group of descriptors that deal with molecular area and surface. Mersy is abbreviated as Mera Symmetry, and the descriptors are calculated using 3D representations of molecules in the framework of the MERA algorithm.
Spectrophores are 1D descriptors that encode the property fields surrounding the molecules. This provides a chemical-class-independent descriptor that can be used to build models.

QNPR or Quantitative Name Property Relationship are 1D descriptors that are directly based on the IUPAC names or SMILES representation of the molecules. The descriptors are calculated by splitting the respective string into all possible continuous substrings[28].
ToxAlert's[29] Extended Functional Group (EFG)[30] category is a descriptor based on classification initially provided by the CheckMol software[31]. The coverage was extended to include new groups, particularly heterocycles[30]. ToxAlert covers total of 583 functional groups.

Statistical coefficients. For internal validation of the generated models, we used 5-fold stratified cross validation. Accuracy is defined as the percentage of correctly classified samples, given by the formula

$$ACC = (TP + TN) / (TP + FP + TN + FN) \quad (1)$$

where TP and TN stand for True Positive and True Negative, and FP and FN stand for False Positive and False Negative respectively. Due to the large size difference between the active and the inactive populations, Balanced Accuracy was used for determining the quality of the models. It is defined as:

$$BA = 0.5*(TP/P + TN/N) \quad (2)$$

where the P=TP + FN and N = TN+FP are number of positive and negative samples, respectively.

Results and Discussion

Molecular Docking

In an effort to directly visualize the interaction of the ligands with Luciferase, we performed high throughput molecular docking using Autodock Vina. Interestingly, through visual inspection, we found that there was a positional difference between the docked population of the inhibitory and non-inhibitory molecules (Figure 2). However, the docking score reported by Vina did not show significant differences between both sets. The optimal score to separate active and inactive compounds (-7.1) using Vina provided a BA of 65.8%. In order to quantify the difference in binding, we calculated the percentage of the ligand that was inside the binding pocket on an atom-by-atom basis, and then averaged over all the ligand poses (Figure 2). Doing this allowed us to quantify the positional difference

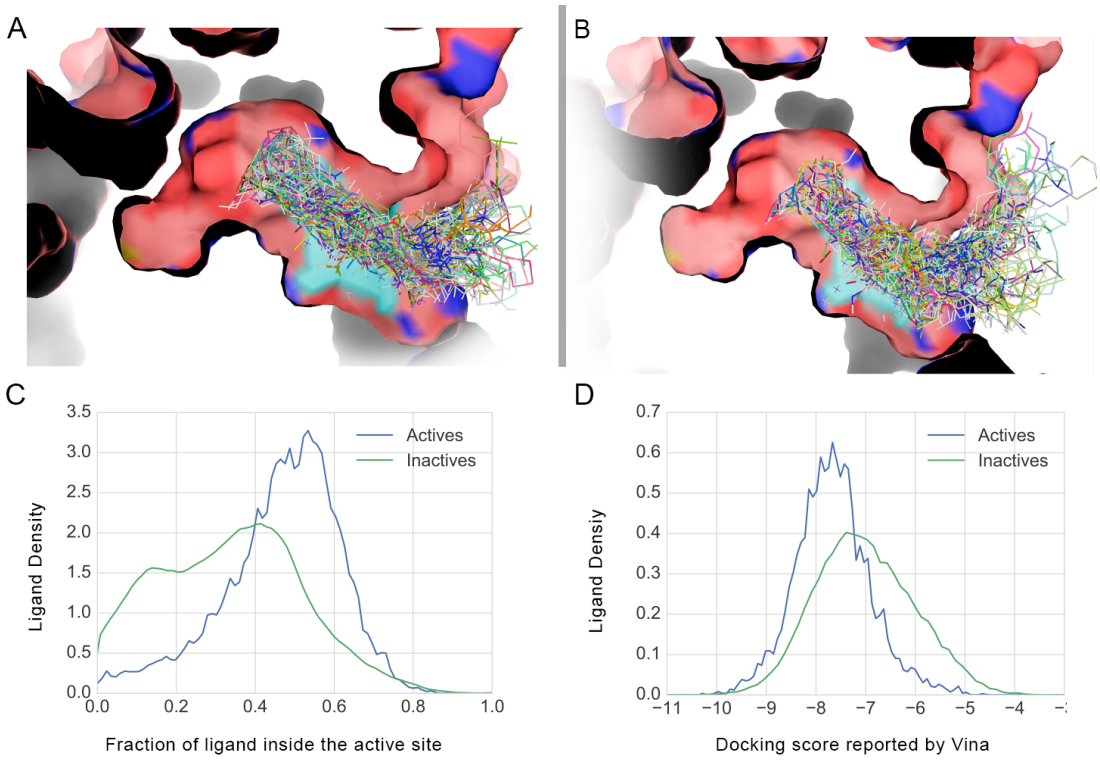


Figure 2: Graphical representation of molecules docked onto Luciferase (top), and histogram of fraction of ligand present inside the active site and Vina docking score **A.** Luciferase Inhibitors **B.** Luciferase non-inhibitors. **C.** Density plot of ligands vs. fraction of ligand inside the active site. **D.** Density plot of ligands vs. docking score reported by Vina. Note that the Vina score is not able to distinguish between the inhibitors and the non-inhibitors as effectively.

that can be seen in Figure 2C, together with a measure of compatibility between the binding pocket and the ligand. From the distribution, one can see that the inhibitory ligands are docked inside the active site significantly more than the non-inhibitory molecules. We applied a threshold of 0.4, and were able to obtain 67.2% balanced accuracy in classifying the two groups. Therefore, calculating the fraction of the ligand inside the active site, one can differentiate between the inhibitors and non-inhibitors with an even better accuracy than using the Vina docking score.

Scaffold Analysis

We were also interested in the chemical nature of the active compounds, so we performed a scaffold tree analysis using Scaffold Hunter[32, 33]. This allowed us to directly visualize the structural hierarchy of the active compounds. It was immediately clear that there is a great deal of variability in the chemical motifs involved; they are not specific to a chemical subtype (Figure 3).

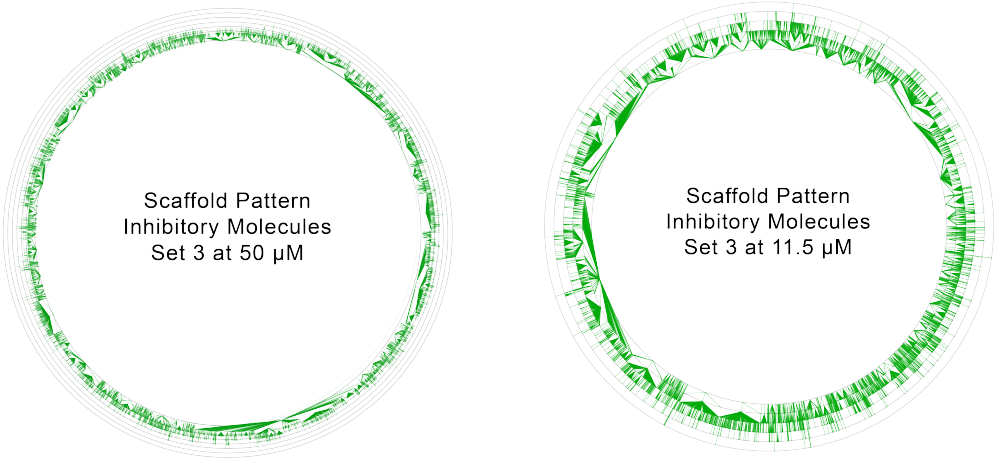
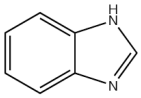
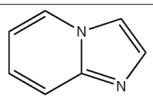
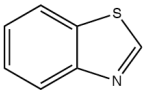
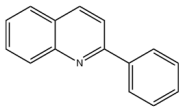
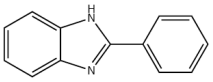
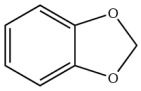
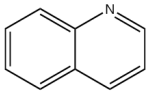
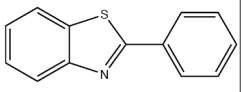


Figure 3: Scaffold tree of Set3 in two different concentrations. The larger size and much higher variability in the chemical space can be clearly seen.
We compared the scaffold structures of Set3 at 50 and 11.5 µM and found that at the lower concentration, the scaffold hierarchy gets simplified considerably due to the reduced number of active molecules (reduction of about 50%, see Table 1). We also noticed some prominent scaffolds emerge.

Upon closer examination, it became apparent that a clear majority of the scaffolds involved, although they belong to different chemical families, have a very flat structure with multiple aromatic rings. Using the SetCompare utility of OCHEM[34], we quantified this observation, and found that such scaffolds are enriched several times in the inhibitor population than in the non-inhibitors (Table 2) : This implies that the presence of particular functional groups is less important than the overall 3D shape and structure of the molecule, when considered from the perspective of luciferase inhibition. This was also corroborated by reported literature[2], where the addition of a non-planar element, such as cyclohexane or a branched motif, to a pre-established motif drastically reduced inhibition. It should also be noted that all the scaffolds have a very limited coverage, therefore indicating a high variability in the chemical space.

Table 2: Scaffold analysis using OCHEM

Scaffold Structure	Inhibitors	Non-inhibitors	Enrichment Factor
	6.5	1.8	3.6
	2.2	0.4	5.5
	4.8	1.1	4.3
	1.4	0.2	7.0
	0.8	0.1	8.0
	4.3	2.0	2.2
	3.5	1.6	2.2
	0.9	0.1	9.0

In order to take the idea of prominent scaffolds one step further, we decided to build a filter using SMARTS to screen active molecules from inactive ones based on the scaffold structure. All the SMARTS were uploaded to ToxAlerts[29] on the OCHEM platform, and can be accessed there online. As can be seen from Table 3, even with a general scaffold such as Benzo-imidazole, only *ca.* 21% of the actives were captured, along with 13% of the inactive molecules. The addition of further groups increases selectivity, but reduces coverage significantly. Due to this, the SMARTS query suffers from exclusivity between selectivity or specificity, and creating an effective filter with this approach proved very difficult due to the large chemical space and variability of the set. Although the scope of such a filter is limited, we gained an understanding of the governing scaffold structure behind the inhibition process: This was useful in designing and refining the pharmacophore during our pharmacophore analysis.

Table 3: Filtering active compounds employing SMARTS.

Scaffolds encoded in SMARTS †	Actives	Inactives	Enrichment Factor

Benzo-imidazole scaffold	21.66	12.93	1.7
Benzyl imidazole scaffold	4.46	1.06	4.2
Biphenyl system with non-aromatic linker	8.85	6.21	1.4
2-(2-(1H-pyrrol-2-yl)ethyl)-1H-benzoimidazole scaffold	0.71	0.07	10.1
6-Phenyl napthyl scaffold	2.87	0.92	3.1
Biphenyl system with non-ring linker	6.83	4.11	1.7
2-Phenyl benzo-imidazole scaffold	5.97	0.78	7.7
2-(2-(naphthalen-2-yl)ethyl)-1H-pyrrole scaffold	0.25	0.13	1.9

†For representation purposes, scaffolds that the SMARTS query represents have been used. All the SMARTS queries can be found in the TOXALERTS section of the OCHEM platform

Pharmacophore analysis

From the scaffold analysis, we saw that the inhibitors are not scaffold specific, but depend on the overall 3D structure of the molecule. Therefore, we investigated a 3D-structure-based pharmacophore approach to distinguish between the active molecules and the inactives. We started with a crystal structure of luciferase bound to a benzothiol inhibitor (PDB ID: 4e5d), and using Ligandscout[11] identified the key interactions between the ligand and the enzyme (Figure 4). This provided the basis of our pharmacophore, which lacks selectivity, but is moderately specific (Table 4). The initial pharmacophore is defined as a combination of three hydrophobic groups and two hydrogen bond acceptors, as can be seen in Table 4. We added aromatic rings to the pharmacophore to increase the selectivity, and further made optional both the hydrogen bond donor to water interactions, and the hydrophobic interactions of the pharmacophore. This significantly increased coverage, but had a negative impact on specificity (Table 4). We then looked at various scaffolds identified in our earlier analysis (Table 2), and found that there are several members of active compounds where two aromatic systems are bound to a linker group.

To cover this possibility during searching, we allowed for one feature to be omitted. This made the pharmacophore much more flexible, as it can accommodate a biphenyl, benzyl or benzo-imidazole, and many other scaffolds, as long as the aromatic groups satisfy the geometry criteria. This is the crucial difference between the pharmacophore and the SMARTS query. For example, in the case of the SMARTS filter that was designed to capture biphenyl systems with a non-aromatic linker, the shape information is irrelevant. If, due to the nature of linker, the structure of the ligand becomes non-planar, the SMARTS would still pick it up. On the other hand, in a pharmacophore query, we are not specifying the motifs

involved; as long as there are two aromatic groups present at the specified 3D position and orientation, it will be picked up. Due to this reason, we were able to get 74.2% balanced accuracy with our designed pharmacophore with our current dataset. This resulting accuracy is higher than any approach based on SMARTS analysis and molecular docking we have explored thus far.

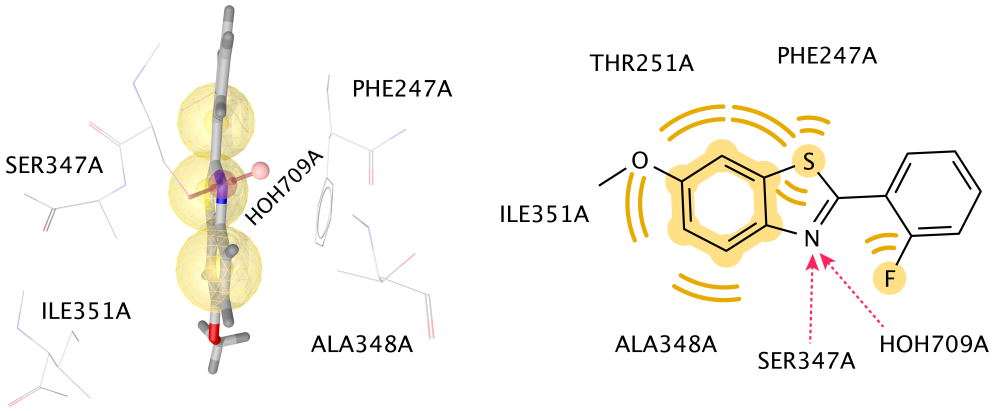
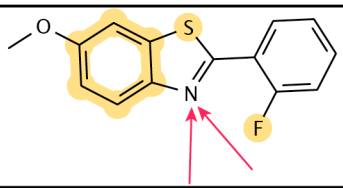
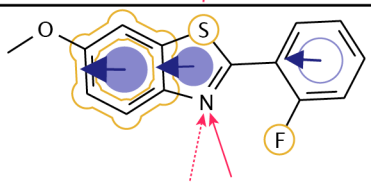
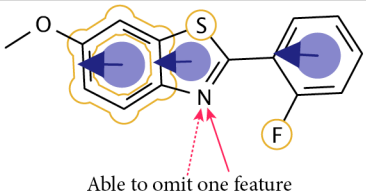


Figure 4: 3D and 2D representation of the interactions of luciferase and benzothiol, its inhibitor (PDB ID 4e5d). The yellow spheres represent hydrophobic interactions, and red arrows show hydrogen bond donor interactions.

Table 4: Filtering luciferase inhibitors using pharmacophores.

Pharmacophore Representation†	Actives (%)	Inactives (%)
	8.2	3.5
	33.5	15.6
	74.2	24.7

†Hydrophobic interactions have been shown in yellow, aromatic groups in purple and hydrogen bonds in red. An outlined shape indicates that the feature was marked as optional.

Machine Learning Models

We built models with various different descriptors that were discussed in the methods section. Across all three datasets, we found that Dragon descriptors, along

with CDK and Adriana provided the highest performance. Dragon6 comprises a total of 5270 descriptors. Many of them capture the shape attributes of the molecules well. The same is true for Chemaxon, CDK and Adriana sets, which also have similar types of descriptors in the package. Thus 3D-based descriptors provided the highest accuracy for prediction of inhibitors of luciferase, which indicates the importance of including 3D structural information when modelling luciferase inhibition.

On the other hand, descriptors based on functional groups, such as Structural Alerts[29], performed poorly throughout. The best results were calculated with the ISIDA descriptors, which provide a comprehensive coverage of different molecular types with automatically generated descriptors. The 2D E-state indices resulted in the second-best models, which had performance that was not statistically different from the performance of models based on ISIDA descriptors.

Table 5: Associative Neural Network analysis

Descriptor	Balanced Accuracy † (%)		
	Set1	Set2	Set3
Dragon6 (3D)	83.7 ± 0.8*	83.6 ± 0.3*	88.1 ± 0.1*
CDK (3D)	83.5 ± 0.9*	84.3 ± 0.3*	88.0 ± 0.1*
ISIDA fragments	81.3 ± 0.8	82.7 ± 0.4*	87.7 ± 0.1*
Adriana (3D)	85.1 ± 0.8*	83.4 ± 0.3*	86.7 ± 0.2*
ALogPS, OEstate	81.3 ± 0.9	81.5 ± 0.3	86.6 ± 0.2
GSFrag	79.5 ± 0.9	80.7 ± 0.4	85.8 ± 0.2
QNPR	79.3 ± 0.9	80.2 ± 0.4	85.4 ± 0.2
Chemaxon Descriptors (3D)	81.2 ± 0.8	81.8 ± 0.3	85.3 ± 0.2
SIRMS	78.1 ± 0.9	81.1 ± 0.4	85.3 ± 0.2
Mera, Mersy (3D)	82.1 ± 0.8	81.8 ± 0.4	84.3 ± 0.2
Inductive Descriptors (3D)	78.1 ± 0.9	78.8 ± 0.4	80.7 ± 0.2
Structural Alerts	73.0 ± 1.0	72.7 ± 0.4	79.1 ± 0.2
Spectrophores (3D)	78.1 ± 0.9	77.4 ± 0.4	78.4 ± 0.2
Consensus Model	86.2 ± 0.7	86.4 ± 0.3	89.3 ± 0.1

†Balanced accuracy for all three datasets obtained using various descriptors and Associative Neural Network algorithm sorted by accuracy of models for set 3.

*Models that are marked with an asterisk were used to create the consensus model.

Consensus Models

Consensus models were built for each dataset. This was performed by averaging the results of the four best-performing models, selected according to the balanced accuracy. As shown in table 5, the consensus models had an accuracy *ca.* 1-3% better than the individual models: All further analysis was performed using these consensus models.

Analysis across datasets:
To observe the effects of the increasing volume of data in the training sets of the models, as well as to determine the performance of the models against new compounds, we used the other two sets as test sets against each trained model:
Since Set1 is the smallest, and also had the least sensitivity amongst the three datasets, models from this set would not be able to effectively predict molecules from Set2 and Set3. As one can see from Table 6, Set1 models show lower accuracy against Set2 or Set3, in comparison to itself. In case of Set2, the sensitivity is higher and training set size is larger than that of Set1, and therefore the model can effectively predict molecules from Set1. However, against Set3 the same model does not perform well, and this can be explained by the same argument as in case of Set1. The model built from Set3 provided the best results, as the training set is the largest and also the sensitivity is the highest, providing the largest number of active molecules in the training set. This makes Set3 the main dataset from which to build our final model.

Table 6: Cross correlation of models between the datasets used in the study.

		Test Set †		
		Set1	Set2	Set3
Training Set	Set1	86.2 ± 0.7 (70,231)	81.2% ± 0.3 (195,546)	81.0% ± 0.2 (323,224)
	Set2	89.8% ± 0.7 (70,231)	86.4 ± 0.3 (195,546)	85.5% ± 0.2 (323,224)
	Set3	90.8 ± 0.5 (70,231)	87.7 ± 0.2 (195,546)	89.3 ± 0.1 (323,224)

†Number inside the parenthesis denotes the number of tested molecules in the respective set.

Analysis of incorrect predictions

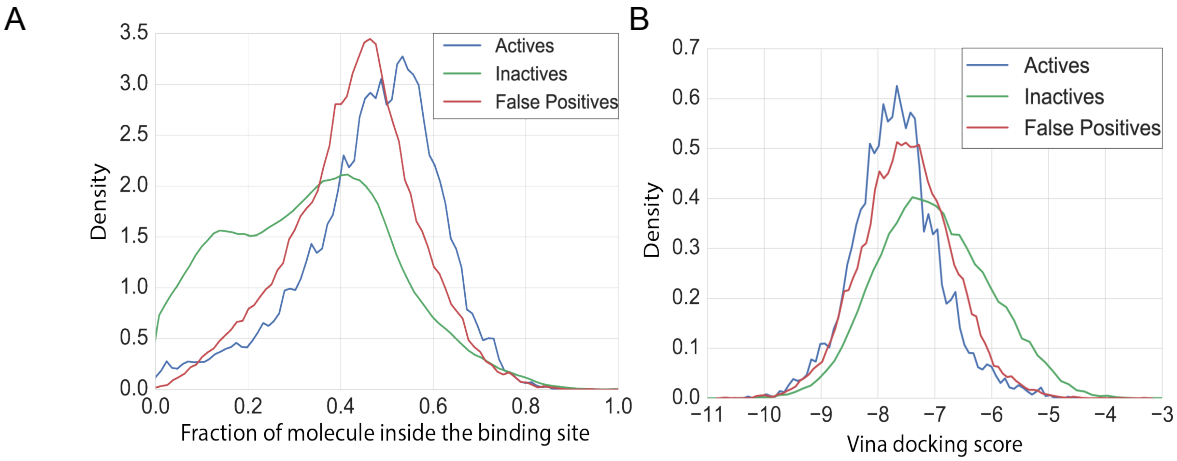


Figure 5: **A.** Density plot of fraction of ligand present inside the active site, for the false positive predictions. Note that the majority of the population lies in between the regular active and inactive molecules. **B.** Density plot of docking score reported by Vina

In order to gain a better understanding of the inaccuracy of the models, we analyzed the compounds that were predicted incorrectly. First, we selected molecules that were predicted incorrectly in at least two consensus models. For the FN (actives that are predicted as inactives), we found 130 molecules, and for the FP (inactives that are predicted as actives), there were 13594 molecules. We attempted to understand the nature of the false positives, by docking them against Luciferase, and performing the analysis described in the docking section. This revealed that FP molecules have the propensity to dock inside the active site of luciferase more than regular inactives (Figure 5), but less than that of regular actives. This means that these molecules have some structural features that are capable of fitting inside the active site of luciferase, but the interactions are not favorable. This is well corroborated by the docking score reported by Vina, where the false positives have more favorable binding energy compared to the inactives, but less favorable compared to the actives. The structural features are being recognized by the machine learning algorithms, and because the machine learning methods do not consider the interactions, the molecules are being marked as inhibitors, when in fact due to unfavorable interactions they do not inhibit luciferase.

Since aggregation is known to play a role in inhibition[35], we decided to investigate whether the activity of some compounds could be due to aggregation. As a property, aggregation is dependent on many variables, and therefore it is very difficult to predict: There has been significant effort in developing this area, and an aggregation advisor (http://advisor.bkslab.org)[36] has been established to address this problem. This on-line server checks new molecules against a database of known aggregators; the database contains compounds that are known to aggregate at concentrations of 10 μM or lower. Because at elevated concentration aggregation is promoted further, this test will identify such molecules in our datasets that were screened at 10 and 50 μM.

We found that 3.2% of the active compounds are known to aggregate, as compared to 2.1% among the inactive molecules. It is also worth mentioning that in Set1 and Set3 assays, 0.01% Tween-20 was used as a detergent, presumably to prevent aggregation. In the case of Set2, compounds were dissolved in DMSO. Therefore, one might expect that in Set2, more aggregators would be present in the active pool. However, due to small number of aggregator molecules, we observed no appreciable difference in percentages of aggregation for active/inactive in Set1, Set3 vs. Set2. The use of detergent could decrease the percentage of aggregators amid active molecules. Still, the fraction of aggregators amid active molecules is 50% large than amid non-active ones. Thus, aggregation plays a significant role in making the molecules change class across experiments, and may have played some role in inhibiting luciferase.

Effect of concentration

Table 7: Effect of concentration reporting balanced accuracy in consensus models for Set1 and Set3

Set	50 μM	11.5 μM
Set1	85.3 ± 0.4	86.2 ± 0.8
Set3	87.2 ± 0.1	89.3 ± 0.1

As mentioned previously, there is a concentration difference in the datasets taken for this study, and the models built are dependent on this concentration because the activity of molecules change based on concentration. We noted that at higher concentrations, models became less accurate (Table 7). To better understand this, we counted the number of molecules (N=2666) that were incorrectly predicted as inactives by the model developed using 50 μM data. We found that 81% of these molecules became inactive upon lowering concentration. Contrary to that only 54% (N=22303) of correctly predicted active molecules (corresponding to average 50% decrease of actives when lowering concentration from 50 to 11.5μM) became inactive. Therefore, at higher concentration, such molecules introduce noise into the data, leading to inaccuracy. The models reported here were built using activity data at 10 μM or 11.5 μM: This must be taken into consideration when applying the model.

Merging datasets to create the final model
To create the final model, we chose Set3 to be our primary set, as reasoned above. We then added to it only the unique active molecules from Set1 and 2, reasoning that since these molecules are active in an assay with lower sensitivity, they have a higher probability to be active and not false positives. We decided not to merge the inactives from three datasets together, as doing so would lead to having inactive molecules that come from experiments with lower sensitivity, which may bring false negatives. This gives us a merged dataset with N = 323443, and with 3.3% of active molecules. Using the same procedure as previously discussed, we obtained the consensus model, which has a balanced accuracy of 89.7%. It can be accessed at http://ochem.eu/article/104546.

Sensitivity of existing filters

As we have explored the inhibition of luciferase and the nature of the inhibitors in this study, we wondered where these identified inhibitors lie in the context of existing frequent hitter and Pan Assay Interference Substance (PAINS[37]) filters: These filters are implemented on OCHEM as part of the ToxAlerts platform[29], and we ran them against our dataset (Table 8). We found that PAINS filters flagged approximately twice as many active compounds as inactive compounds; the AlphaScreen filters to detect promiscuous compounds also provided an approximate threefold enrichment of flagged actives over inactives. However, the promiscuity filter that was designed to identify compounds likely to hit multiple assays[38], provided a much smaller enrichment. The highest enrichment was calculated for the AlphaScreen filter, however, this filter had the lowest coverage. The most prominent alert among the AlphaScreen filter that picked up luciferase inhibitors was the Aminal alert (aminal on a pyridine-based system, Figure S1). This alert picked up several compounds with a planar structure (Figure S2), and provided an enrichment factor of 6.2. It should also be noted that the number of alerts involved in this case is very small, which gets reflected in the poor coverage of this filter. The difference in the number of alerts in each filter contributes to the specificity/selectivity trade-off.

We also noted that most of the compounds were flagged as being reactive, unstable or toxic. This is expected, as the responsible filter is known to pick up drug-like molecules: It is worth mentioning here that the presence of such alerts by itself does not make a molecule toxic in the context of medicinal application, due to dosage and clearance from the body.

Table 8: Luciferase inhibitors tested against a variety of other filters.

Compound Filters †	Actives (%)	Inactives (%)	Enrichment
Pan Assay Interference Substance (PAINS) (480)	9.8	4.9	2.0
Promiscuity (178)	4.7	3.8	1.2
AlphaScreen FH filters (25)	1.7	0.6	2.8
Reactive, Unstable, Toxic (340)	66.9	62.3	1.1

†The numbers in parentheses represent the number of alerts in each respective filter.

Discussion

The developed chemoinformatic model is suitable for providing an early warning against potential inhibitors of luciferase that may interfere with HTS experiments. Since the model does not have 100% accuracy, some compounds can be predicted as luciferase inhibitors when in reality they are not. On the other hand, even if the molecule is indeed a luciferase inhibitor, that does not mean that it cannot be a potential lead. Hence, we strongly advise not to discard the flagged molecules as

false leads but rather to consider them further, to better interpret experimental results.
Thus, the model described here should be used to identify *potential* interference in luciferase-based assay systems. The identified molecules should be re-tested using other assay protocols that do not rely on luciferase. The merit of this study is that one can find potential interference in very large datasets, and only the flagged molecules then need be tested by orthogonal assays. This reduces cost, time and effort in the counterscreening effort.

Conclusions

In this study, we explored various methods of filtering and detecting luciferase inhibitors in a luciferase-based HTS assay. We designed computational models using machine-learning methods on publicly available data from PubChem. We also used molecular docking to understand how inhibitors bind to luciferase, and performed a scaffold analysis to gain a better understanding of the chemical nature of such inhibitors. The machine learning models outperformed other methods of filtering luciferase inhibitors, such as SMARTS or pharmacophore-based filters. We were able to obtain a prediction accuracy of 89.7%, which makes the final model a good tool for filtering potential luciferase inhibitors. Still, the predictions of the model should be considered as advice and flagged compounds can be retested in orthogonal assays. All models and data reported here are publicly accessible at http://ochem.eu/article/104546.

Reproducibility of OCHEM models

All OCHEM models are developed using standardized workflows, which can be used at the OCHEM to produce another model or reproduce previous results. The full specification of details of the workflow are stored in xml file, which can be exported, imported or used as template for model development. This feature provides the reproducibility of OCHEM models. Moreover, OCHEM platform can be installed locally at the commercial or academic premises and be used to apply or reproduce models on local computers of the users. Majority of models available in OCHEM can be also exported and used as standalone versions. For both these applications commercial or academic licenses for some tools, such as descriptors calculation, 3D structure generation, standardization of chemical structures, etc. can be also required. Contrary to that, the predictions of models available in OCHEM can be used under the CC-BY-NC license while the data can be downloaded under the CC-BY license. These features makes OCHEM a powerful public portal for development and sharing of reliable and reproducible chemical information and models on the web.[39]

Acknowledgement

The project leading to this report has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676434, "Big Data in Chemistry". The article reflects only the author's view and neither the European Commission nor the Research Executive Agency (REA) are responsible for any use that may be made of the information it contains. We thank Michael Withnall for English correction. The authors thank

ChemAxon (http://www.chemaxon.com) for Academic license of software tools (Standartizer, ChemAxon plugins) used in this study.

Supporting Information Available

The Aminal alert and the compounds filtered by the alert can be found in the supplementary information for this article.

Conflict of Interest

I.V.T. is CEO of BIGCHEM GmbH, which develops the OCHEM platform (http://ochem.eu) used in this study. Other authors declare no conflicts of interest.

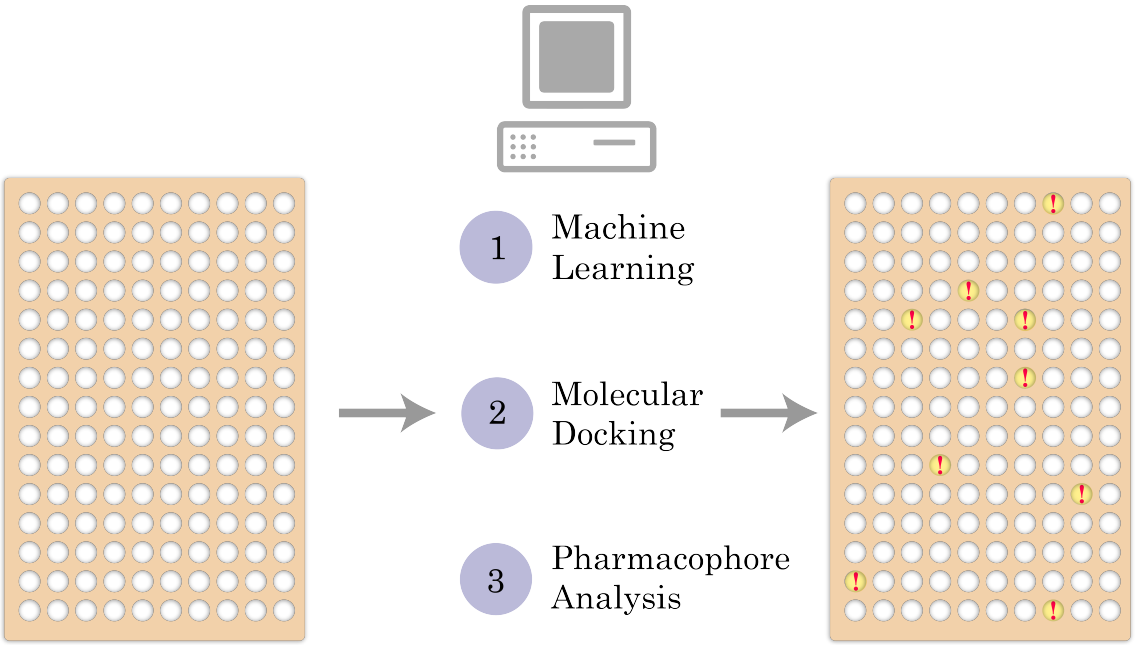
References

1. Thorne, N.; Inglese, J.; Auld, D. S., Illuminating insights into firefly luciferase and other bioluminescent reporters used in chemical biology. *Chem. Biol.* **2010**, 17, 646-657.
2. Thorne, N.; Shen, M.; Lea, W. A.; Simeonov, A.; Lovell, S.; Auld, D. S.; Inglese, J., Firefly luciferase in chemical biology: a compendium of inhibitors, mechanistic evaluation of chemotypes, and suggested use as a reporter. *Chem. Biol.* **2012**, 19, 1060-1072.
3. Wang, T. T. Y., β-Naphthoflavone, an Inducer of Xenobiotic Metabolizing Enzymes, Inhibits Firefly Luciferase Activity. *Anal. Biochem.* **2002**, 304, 122-126.
4. Bakhtiarova, A.; Taslimi, P.; Elliman, S. J.; Kosinski, P. A.; Hubbard, B.; Kavana, M.; Kemp, D. M., Resveratrol inhibits firefly luciferase. *Biochem. Biophys. Res. Commun.* **2006**, 351, 481-484.
5. Leitão, J. M. M.; Esteves da Silva, J. C. G., Firefly luciferase inhibition. *J. Photochem. Photobiol., B* **2010**, 101, 1-8.
6. Auld, D. S.; Southall, N. T.; Jadhav, A.; Johnson, R. L.; Diller, D. J.; Simeonov, A.; Austin, C. P.; Inglese, J., Characterization of chemical libraries for luciferase inhibitory activity. *J. Med. Chem.* **2008**, 51, 2372-2386.
7. Trott, O.; Olson, A. J., AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, 31, 455-461.
8. Sadowski, J.; Gasteiger, J.; Klebe, G., Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 1000-1008.
9. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J., AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, 30, 2785-2791.
10. Schrödinger, L. L. C., In *The PyMOL Molecular Graphics System, Version 1.8.6.0*; 2015.
11. Wolber, G.; Langer, T., LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, 45, 160-169.
12. Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.;

Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V., Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput.-Aided Mol. Des.* **2011**, 25, 533-554.
13. Tetko, I. V., Neural Network Studies. 4. Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 717-728.
14. Tetko, I. V., Associative Neural Network. *Neural Process. Lett.* **2002**, 16, 187-199.
15. Chang, C.-C.; Lin, C.-J., LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, 2, 1-27.
16. Villa, A. E.; Tetko, I. V.; Dutoit, P.; De Ribaupierre, Y.; De Ribaupierre, F., Corticofugal modulation of functional connectivity within the auditory thalamus of rat, guinea pig and cat revealed by cooling deactivation. *J. Neurosci. Methods* **1999**, 86, 161-178.
17. Gasteiger, J., Of molecules and humans. *J. Med. Chem.* **2006**, 49, 6429-6434.
18. Tetko, I. V.; Tanchuk, V. Y., Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 1136-1145.
19. Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P., Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1488-1493.
20. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E., The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 493-500.
21. Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C., The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminf.* **2017**, 9, 33.
22. Myrdal, P. B.; Manka, A. M.; Yalkowsky, S. H., AQUAFAC 3: aqueous functional group activity coefficients; application to the estimation of aqueous solubility. *Chemosphere* **1995**, 30, 1619-1637.
23. Todeschini, R.; Consonni, V. In *Handbook of Molecular Descriptors*; Wiley-VCH Verlag GmbH: 2000.
24. Skvortsova, M. I.; Baskin, I. I.; Skvortsov, L. A.; Palyulin, V. A.; Zefirov, N. S.; Stankevich, I. V., Chemical graphs and their basis invariants. *J. Mol. Struct.: THEOCHEM* **1999**, 466, 211-217.
25. Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D., ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, 29, 855-868.
26. Alexandre, V.; Denis, F.; Dragos, H.; Olga, K.; Cedric, G.; Philippe, V.; Vitaly, S. e.; Frank, H.; Igor, V. T.; Gilles, M., ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, 4, 191-198.
27. Potemkin, V. A.; Grishina, M. A., A new paradigm for pattern recognition of drugs. *J. Comput. Aided. Mol. Des.* **2008**, 22, 489-505.
28. Tetko, I. V.; M. Lowe, D.; Williams, A. J., The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS. *J. Cheminf.* **2016**, 8, 2.

29. 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.
30. Salmina, E. S.; Haider, N.; Tetko, I. V., Extended Functional Groups (EFG): An Efficient Set for Chemical Characterization and Structure-Activity Relationship Studies of Chemical Compounds. *Molecules* **2015**, 21, E1.
31. Haider, N., Functionality Pattern Matching as an Efficient Complementary Structure/Reaction Search Tool: an Open-Source Approach. *Molecules* **2010**, 15.
32. Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H., The scaffold tree—visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **2007**, 47.
33. Schäfer, T.; Kriege, N.; Humbeck, L.; Klein, K.; Koch, O.; Mutzel, P., Scaffold Hunter: a comprehensive visual analytics framework for drug discovery. *J. Cheminf.* **2017**, 9, 28.
34. Vorberg, S.; Tetko, I. V., Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM). *Mol. Inf.* **2014**, 33, 73-85.
35. Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K., High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* **2005**, 1, 146-148.
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. 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.
38. 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.
39. Tetko, I. V.; Maran, U.; Tropsha, A., Public (Q)SAR Services, Integrated Modeling Environments, and Model Repositories on the Web: State of the Art and Perspectives for Future Development. *Mol. Inform.* **2017**, 36.

TOC FIGURE


1
Machine Learning
2
Molecular Docking
3
Pharmacophore Analysis