Fr-PPIChem: An Academic Compound Library Dedicated to Protein-Protein Interactions

Nicolas Bosc, Christophe Muller, Laurent Hoffer, David Lagorce, Stéphane Bourg, Carine Derviaux, Marie-Edith Gourdel, Jean-Christophe Rain, Thomas W Miller, Bruno O. Villoutreix, et al.

Fr-PPIChem: An Academic Compound Library Dedicated to Protein-Protein Interactions

Nicolas Bosc[a,b,‡], Christophe Muller[c,‡], Laurent Hoffer[d], David Lagorce[e], Stéphane Bourg[f], Carine Derviaux[c], Marie-Edith Gourdel[g], Jean-Christophe Rain[g], Thomas W. Miller[c], Bruno O. Villoutreix[h], Maria A. Miteva[i], Pascal Bonnet[f], Xavier Morelli[c,d], Olivier Sperandio,[a,b,], Philippe Roche[d,*]*

[a]Inserm U973 MTi, 25 rue Hélène Brion 75013 Paris

[b]Institut Pasteur, Unité de Bioinformatique Structurale, CNRS UMR3528, 28 rue du Dr Roux 75015 Paris

[c]IPC Drug Discovery Platform, Institut Paoli-Calmettes, 232 Boulevard de Sainte-Marguerite, 13009, Marseille, France

[d]CRCM, CNRS, INSERM, Institut Paoli-Calmettes, Aix-Marseille Univ, 13009 Marseille, France

[e]Université de Paris, INSERM US14, Plateforme Maladies Rares - Orphanet, 75014 Paris, France

[f]Institut de Chimie Organique et Analytique (ICOA), Université d'Orléans, UMR CNRS 7311, BP 6759, 45067 Orléans. France

[g]Hybrigenics Services SAS, 1 rue Pierre Fontaine, 91000 Evry Courcouronnes, France

[h]Université de Lille, INSERM, Institut Pasteur de Lille, U1177 - Drugs and Molecules for living Systems, 59000 Lille, France

[i]Inserm U1268 MCTR, CNRS UMR 8038 CiTCoM – Univ. De Paris, Faculté de Pharmacie de Paris, 75006 Paris, France

[‡]These authors contributed equally to this work

[*]**Corresponding authors**: E-mail: *olivier.sperandio@inserm.fr*, *philippe.roche@inserm.fr*

running title : Fr-PPIChem, a PPI-oriented chemical library

Abstract

Protein-protein interactions (PPIs) mediate nearly every cellular process and represent attractive targets for modulating disease states but are challenging to target with small molecules. Despite this, several PPI inhibitors (iPPIs) have entered clinical trials, and a growing number of PPIs have become validated drug targets. However, high-throughput screening efforts still endure low hit rates mainly because of the use of unsuitable screening libraries. Here, we describe the collective effort of a French consortium to build, select and store in plates a unique chemical library dedicated to the inhibition of PPIs. Using two independent predictive models–and two updated databases of experimentally confirmed PPI inhibitors developed by members of the consortium, we built models based on different training sets, molecular descriptors and machine learning methods. Independent statistical models were used to select putative PPI inhibitors from large commercial compound collections showing great complementarity. Medicinal chemistry filters were applied to remove undesirable structures from this set (such as PAINS, frequent hitters and toxic compounds) and to improve drug likeness. The remaining compounds were subjected to a clustering procedure to reduce the final size of the library while maintaining its chemical diversity. In practice, the library showed a 46-fold activity rate enhancement when compared to a non-iPPI-enriched diversity library in high-throughput screening against the CD47-SIRPα PPI. The Fr-PPIChem library is plated in 384-well plates and will be distributed on demand to the scientific community as a powerful tool for discovering new chemical probes and early hits for the development of potential therapeutic drugs.

KEYWORDS chemical libraries, focused library, PPI, protein-protein interaction, inhibitors, QSAR, diversity, drugs, screening, ADMET, CD47, SIRPα, HTS, chemical biology

INTRODUCTION

Protein-protein interactions (PPIs) play a major role in most biological processes and a variety of cellular disorders and disease states[1, 2]. As a consequence, PPIs have emerged as an important new class of therapeutic targets[3]. However, they are considered challenging to target with small molecules mainly due to the nature of their interfaces, which have not evolved to interact with physiological small molecules such as second messengers and metabolites[4, 5].

The successful development of PPI inhibitors (iPPIs) has increased over the last decade, and several compounds are now moving toward advanced-phase clinical trials[6-8]. However, despite these encouraging results, targeting PPIs remains highly challenging, and specific tools are still needed to improve the effectiveness of drug discovery campaigns. One of the most popular approaches for discovering iPPIs, especially when no structural data are available for the target, is by high-throughput screening (HTS). However, HTS often results in poor hit rates and high false-positive rates for this class of targets, which can be attributed to the poorly suited screening libraries that are either not enriched in iPPIs or are too small to contain sufficient diversity[9, 10]. Therefore, developing innovative chemical libraries dedicated to discovering active small molecules targeting PPIs is of great importance to the scientific community to improve both the hit rates of those screenings and the quality of the identified chemical probes.

Several databases such as TIMBAL[11], iPPI-DB[12] and $2P2I_{DB}$[13] have been developed to store experimentally validated iPPIs and determine their physicochemical profile of PPI modulators. Analysis of the compounds present in these databases revealed that, on average, iPPIs are heavier, more hydrophobic, contain more aromatic rings and have different 3D shapes than conventional drugs[14-17]. These characteristics have been used to build statistical models capable of

filtering chemical databases to search for putative iPPIs[14, 15, 18-20] or to extract PPI-specific chemical information.

Our teams were responsible for building the PPI-HitProfiler[14] and $2P2I_{HUNTER}$[19] models leading to the design a PPI-oriented chemical library, which has been validated on several structurally diverse targets[21-23]. Following these encouraging results, a French consortium was organized to build and share a unique chemical library dedicated to the inhibition of PPIs. A new, optimized, combined version of the two predictive models was built considering the two updated chemical databases iPPI-DB[12] and $2P2I_{DB}$[13]. For this, different models were built based on different training sets, molecular descriptors and machine learning methods. Both statistical models were independently used to retrieve 'iPPI-like molecules' from large commercial compound collections and the two models showed great complementarity. Medicinal chemistry filters were applied to remove undesirable structures from this set and to improve drug-likeness. The remaining compounds were subjected to a clustering procedure to reduce the final size of the library while maintaining chemical diversity. Finally, we describe a single case study comparing the rate of obtaining active compounds by screening the Fr-PPIChem library to the rate with a traditional diversity library for a therapeutically relevant PPI target.

MATERIALS AND METHODS

I- 2P2I$_{HUNTER}$ models

Dataset Preparation

197 orthosteric iPPIs, corresponding to 21 protein-protein complexes, were extracted from 2P2I$_{DB}$ version 03-2016 (http://2p2idb.cnrs-mrs.fr/)[13]. Bromodomain inhibitors were not considered in this work. Additionally, an SD file containing 1,826 approved drugs from DrugBank (*version 4.5.0*) was extracted from https://www.drugbank.ca/ and considered as the negative dataset (decoy set)[24].

In an effort to work with consistent datasets, a curation/standardization procedure was applied for the positive and negative sets using ChemAxon (http://www.chemaxon.com). A diversity set was selected to reduce structural redundancy in the positive set (because some chemical series are overrepresented in 2P2I$_{DB}$) and to reduce the number of compounds in the decoy set. The ChemMineR library from R was used for this purpose[25]. The full procedure of standardization and diversity selection is described in given as supplementary information (Figure S1). Finally, 85 nonredundant iPPIs, corresponding to 21 protein-protein complexes, were selected in the positive dataset and 734 in the decoy dataset.

Chemical Libraries

Screening collection of compounds from MolPort (https://www.molport.com), Ambinter (http://www.ambinter.com), and Zinc "All Purchasable" (http://zinc.docking.org) were retrieved corresponding to 6.3 (version 2016), 5.7 (version 2015) and 16 (version 2012) million compounds, respectively.

Molecular Descriptors

a: Dragon molecular descriptors

Dragon v.6 (http://www.talete.mi.it) was used to calculate values for 9 molecular descriptors[26] (molecular weight, number of multiple bonds, number of rings, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, unsaturation count, hydrophilic factor and topological surface area) as previously described[19]. The value of cLogP was predicted using cxcalc from ChemAxon. The 10 descriptors were normalized between modeling and test sets using the same normalization parameters.

b: MOE 2D molecular descriptors

MOE 2D descriptors were computed using Molecular Operating Environment from Chemical Computing Group (https://www.chemcomp.com). Chirality descriptors were not taken into account because this information was not available for all molecules. All MOE 2D descriptors were normalized between modeling and test sets using the same normalization parameters.

c: ISIDA fragment descriptors

Two classes of the ISIDA fragment descriptors[27] were used: "sequences" (I) and "augmented atoms" (II). For each class, three sub-types are defined AB, A and B. Sequences represent the shortest paths from one atom to another one. (AB) represent sequences of atoms and bonds, (A) sequences of atoms only and (B) sequences of bonds only[28]. Sequences can also be represented as Atom Pairs (AP) where only terminal atoms and the topological distance between them are explicitly represented. An "augmented atom" represents a selected atom with its environment including sequences of AP, AB, A and B types issued from this atom. Descriptors of length from 2 to 15 atoms were calculated for sequences and from 2 to 10 atoms for augmented atoms. For

sake of clarity, sequences of atoms and bonds of length from x to y atoms were annotated as IAB(x-y).

Construction and validation of models

Support Vector Machine (SVM) was used as the machine learning method to build predictive models using the LIBSVM classifier[29]. SVM models giving a probability for a compound to be a PPI inhibitor or not were constructed[30]. The polynomial method was chosen as the kernel function. The default degree value of 3 was kept. Cost and gamma hyperparameters were optimized using a 5-fold cross-validation procedure to get the best averaged ROC AUC value (Figure S2). All inhibitors of a given protein-protein family were in the same set (see supplementary information for details).

The area under the curve (AUC) of the ROC was chosen as the statistical criteria to evaluate model performance. As additional statistical criterion, balanced accuracy (BA) was calculated to offer another view of model performance. BA is calculated as a function of elements of the confusion matrix : true positive (TP), false positive (FP), true negative (TN) and false negative (FN) as follows : $BA=\frac{1}{2}(\frac{TP}{TP+FN}+\frac{TN}{TN+FP})$.

For the best pool of descriptors determined in 5-fold cross-validation and for each modeling set, a procedure of y-scrambling was performed to estimate the part of models related to chance[31].

II- PPI-HitProfiler models

Dataset preparation

The molecules were selected as described previously[20]. Briefly, 3,051 nonredundant iPPIs were obtained from iPPI-DB[12] (version 2016) and TIMBAL database[11] (version May 2015). In addition, a diversity set was derived from this dataset. Compounds were first described using FCFP4 fingerprint computed by Pipeline Pilot[32] and a hierarchical clustering was performed

using the Ward method with the Tanimoto distance. Only the cluster centers were selected using a similarity threshold of 0.7 resulting in a chemically diverse set of 719 iPPIs.

Putative non-iPPIs were selected from two different sources: BindingDB[33] and BDM as described previously[20]. In total, 83,572 non-iPPIs were identified (44,228 from BindindDB and 39,344 from BDM). A diversity selection was performed as described previously resulting in a set of 35,686 diverse compounds used as decoy.

The molecules were standardized and filtered using an in house protocol developed in Pipeline Pilot[32] to exclude compounds containing non organic atoms, to keep only small molecules (peptides and macrocycles were not considered in this study) and to remove potential solvent molecules or counter-ions.

Molecular descriptors

MOE descriptors.

MOE 2D and available 3D descriptors were computed. The latter descriptors require to have the active conformation of each molecule. In the absence of such information for the vast majority of the molecules used in this study, up to 50 conformations per molecule were generated using a stochastic method and the MMFF94x force field as described previously[20]. The 3D descriptors were calculated on all the conformations for a given compound and the mean value for each descriptor was kept as the final value. A total of 298 descriptors (186 2D and 112 3D) were calculated for each molecule of each data set.

Dragon molecular descriptors.

Dragon V7 (https://chm.kode-solutions.net/products_dragon.php) was used to calculate 2,600 2D descriptors.

RDKit descriptors.

A total of 117 2D descriptors were calculated using RDKit (version 2016_03_1, http://www.rdkit.org).

It was impossible to calculate several descriptors for a significant number of molecules in these data sets. Hence, these descriptors were deleted resulting in a total of 2,990 descriptors.

Descriptor selection

To build models with relevant descriptors, the most discriminative descriptors between iPPIs and putative non-iPPIs (decoy set) were identified prior to modeling using the protocol described previously[34]. Briefly, it allows to remove invariant or correlated descriptors and those whose absolute correction exceeds 0.6 with molecular weight, clogP and TPSA. It also excludes descriptors whose distribution is not significantly different between iPPIs and non-iPPIs. Finally, to eliminate the descriptors whose significance was due to over-represented chemotypes, this protocol was applied on the PPI/non-PPI inhibitor data sets and their corresponding diverse subsets (Figure 1). Only the descriptors that were found in both cases were kept. Finally, the last selection step consisted in ensuring that the descriptors were also significantly discriminant for at least 60% of the PPI targets against the non-iPPIs. Eventually, 167 molecular descriptors (Dragon: 124, RDKit: 7 and MOE: 36) were identified and are detailed in the supplementary information (Table S1).

Model training and validation

Four different machine learning algorithms were used in this protocol to build classification models. (i) Decision trees as implemented in the J48 function of Weka (version 3.7)[35]; (ii) Random Forests (RF) as implemented in Weka; (iii) JRip which is the implementation of the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) in Weka; (iv) SVM using

the C-SVM function with the radial basis function as presented in the R statistical software package Caret[36, 37]. J48 and JRip were used with their default parameters but a parametrization of RF and SVM was performed doing a grid search using a 5-fold cross validation. The optimal numbers of trees and of variables selected by the trees were both tuned for RF, while the optimal hyper-parameters cost and gamma were found for SVM.

The four classification methods were applied on iPPIs and non-iPPIs described with the 167 selected descriptors. Normalization was performed on the variables to be centered to zero mean and scaled to unit variance. The data set was split into a training set (70%) and a test set (30%) conserving the same proportion of iPPIs and non-iPPIs. Note that as described in the previous section "Descriptor selection", both the training and the test sets were used for the selection of the 167 descriptors.

Once the models were trained, they were first evaluated on the training set. Then, a 5-fold cross validation was performed on the training set and the performance were averaged on the five iterations. Finally, the performance of the models was evaluated by predicting the test set (i.e. the 30% remaining data).

At each step, the performance of the models was assessed using the sensitivity, specificity, and a specific definition of the enrichment factor $EF = \frac{TP}{TP+FP} \times \frac{TP+FP+TN+FN}{TP+FN}$. It highlights the ratio between the proportion of true actives before and after filtering. Matthews correlation coefficient (MCC) was also estimated as follows : $\mathrm{MCC} = (\frac{(\mathrm{TP}*\mathrm{TN})-(\mathrm{FP}*\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}})$.

III- Combining, Filtering and clustering

ADMET properties and PAINS.

The 103,656 predicted iPPI compounds resulting from the two models were annotated and filtered with the last release of our Free ADME-Tox package named FAF-Drugs[38]. The overall

procedure involved a data curation step with removal of salts, counterions and mixtures, standardization via the ChemAxon Standardizer routine and removal of duplicates using simple CANSMILES strings generated with OpenBabel. Various physicochemical property thresholds to filter out unwanted chemicals can be selected in FAF-Drugs. As we are interested in the design of iPPI chemical probes, these thresholds were tuned such as to also explore regions of the chemical space in the extended rule of 5 region and beyond the rule of 5 area[39]. The following rules were used: aromatic rings ≤ 6, fused aromatic rings ≤ 3, rotatable bonds ≤ 20, heteroatoms ≤ 12, logP (XlogP3) and logD between -7 and 8, molecular weight ≥ 300 g.mol-1, number of halogens (Br, I, Cl) ≤ 5 and less than 5 consecutive CH_2 units. The *in silico* filtering approach continued with the annotation of molecules containing toxicophores. Different types of documented toxicophores were investigated, and among a list of 139 structures or substructures potentially toxic (Table S2), molecules containing one occurrence of some chemical groups reported in the literature as toxic were rejected (68 such groups are flagged including anhydride, aldehyde, isocyanate, triflate, etc. and molecules that contain these chemical groups are downloaded in file called rejected compounds see: http://fafdrugs4.mti.univ-paris-diderot.fr/groups.html). Another list of 57 toxicophores was analyzed, and molecules that contain less than 2 to 4 occurrences of these chemical groups (depending on the type of chemical group, data reported in the literature about these groups and our analysis of approved drugs, withdrawn drugs and drugs with a blackbox warning)[38] were flagged and accepted. This category of structural alerts involves, for instance, catechol, nitro, imidazole, adamantane, etc. Similarly, another set of 12 potential toxicophores were flagged. These, for example, include hydrazine, sulfoxide and thiocarbamate. In addition, molecules containing a PAINS substructure were flagged. However, only molecules flagged by the 22 PAINS filter-A were rejected[40]. This filtering stage resulted in 78,243 compounds.

Chemical diversity analysis.

The 78,243 remaining compounds were first prepared using VSPrep with default parameters[41]. Duplicates were then removed based on their canonical smiles. A multi-step clustering method implemented in Pipeline Pilot[32] was then performed to gradually select the 10,000-targeted molecules for plating. Different thresholds varying between 0.2, 0.3 and 0.4 for the Tanimoto distance (Tc) based on maximal dissimilarity was used with the Euclidian metric and FCFP6 fingerprints. A minimum of 3 molecules was required in each cluster to ensure further SAR analysis. First, the clustering was performed using Tc of 0.3 leading to 8,851 clusters. Then, for the highest populated clusters (n≥20), a second clustering was performed using a Tc of 0.4 leading to 549 clusters. Finally, on the 15,856 singletons identified in the first clustering step, a third clustering was applied using a Tc of 0.2 providing 1,523 clusters. A total of 10,923 diverse molecules was thus retrieved.

CD47/SIRPα screening

CD47-SIRPα reagents were prepared and NCATS Genesis chemical library screening was conducted as described previously[42, 43]. Fr-PPIChem chemical library screening was performed using the same reagents and homogeneous time-resolved fluorescence (HTRF) optimized protocol with the following alterations: Compounds were dispensed acoustically from 1 mM DMSO stocks using an Echo 550 and assay plates were read on a Pherastar FS multimodal plate reader all of which was incorporated into an automated Labcyte Echo Workstation. The Fr-PPIChem library was first screened using the CD47-SIRPα HTRF assay at a single fixed concentration (20 μM) followed by confirmatory testing in a 7-point concentration response. Concentration response data were fitted using the NCGC CurveFit and filtered by curve classification and potency. Compounds meeting the activity criteria (curve class -1.1, -1.2 and >

90% inhibition) were evaluated in a combination counter and orthogonal screen using the thermal shift assay. HTRF active compounds were incubated at 20 μM with CD47 or SIRPα and subjected to a thermal ramp from 20 to 95 °C in the presence of 2.5X SYPRO Dye (Invitrogen) to measure it melting point temperature via a BioRad CFX384 RTPCR instrument. The variability of the thermal shift assay was ±0.17 °C (standard deviation, n=16). The activity cutoff of Tm ≥ 0.5 °C represents 3*SD. Compounds that induced an increase in the Tm ≥ 0.5 °C uniquely for CD47 or SIRPα were considered active.

RESULTS AND DISCUSSION

The compounds included in the Fr-PPIChem library were selected from commercially available libraries based on their predicted abilities to inhibit protein-protein complexes. Two complementary statistical models were built to select iPPIs.

Development and validation of the $2P2I_{HUNTER}$ models.

A positive training set of diverse nonredundant iPPIs was compiled from $2P2I_{DB}$. The negative dataset representing the putative non-iPPIs was constructed by selecting compounds from DrugBank. This selection was based on the fact that approved drugs are often developed to be highly specific to a given target and are unlikely to interact with PPIs with high affinity.

Three types of descriptors were used to build three distinct statistical models: ISIDA substructural fragments, Dragon and MOE 2D molecular descriptors. SVM was used to differentiate between the positive dataset and the decoy. Other methods, such as random forest, were tested (data not shown), but SVM gave the best results for each type of descriptor. To build and validate the models, the initial training set was randomly divided into 5 modeling/test sets, and the inhibitors of the same PPI family were placed in the same set. Using this strategy, a 5-

fold cross-validation (CV) was undertaken for each modeling set to optimize the SVM hyperparameters. The performances of the best models constructed on the three types of descriptors were very similar based on their balanced accuracy (Table 1).

When filtering a large database of molecules, one must keep in mind that the goal is to predict inhibitors with the highest possible precision. In this context, the models built using ISIDA descriptors IAB(2-2) appear to be the best models. Moreover, the scrambling procedure involving these descriptors led to models with lower AUC average values (0.5) for the 5 tests. This demonstrates that the ISIDA-based models are robust and not the result of random correlations. Therefore, these models were used to filter the collections of screening compounds from MolPort and Ambinter, leading to 36,447 iPPI-like molecules from Ambinter and 37,927 from MolPort.

Development and validation of the PPI-HitProfiler models

Similar to our previous study[20], several machine learning models were built (see material and methods for details). The results show that all 4 new PPI-HitProfiler models have good predictive power, and all managed to positively discriminate the iPPIs (Table 2). All the models have excellent specificities and high sensitivities ranging from 0.76 to 0.91 on the test set and good values for BA and MCC. The application of a Y-scrambling procedure to the 4 models confirmed that the predictions were not due to chance in every case, as significant drops were observed in the sensitivities and in the capacity of enriching libraries (EF approximately 1).

Given the performances of the 4 PPI-HitProfiler models and their excellent capacity to discard non-iPPIs, the choice was made to combine all these models to retrieve iPPI-like molecules from the commercially purchasable catalogs from the ZINC database resulting in a list of 143,967 compounds.

Combination, filtering and clustering of the predicted inhibitors

The two sets of compounds selected by $2P2I_{HUNTER}$ and PPI-HitProfiler were combined, leading to a collection of 103,656 molecules. We optimized in-house-designed FAF-Drugs software for this project to predict and remove PAINS and undesirable compounds[38]. Several physicochemical filters were applied based on analyses of known chemical probes leading to 78,243 remaining compounds (see Material and Methods for details).

A final clustering step was conducted to generate a diverse PPI-oriented chemical library composed of 10,923 compounds (see Material and Methods for details). Molecules that met the quality control requirements and were commercially available were purchased from Ambinter, leading to a 10,314-member compound chemical library (Fr-PPIChem), of which 5,177 (50.2%) were selected by PPI-HitProfiler models and 5,245 (50.8%) were selected by $2P2I_{HUNTER}$ models, while 108 compounds were selected by both models. The Fr-PPIChem chemical library was stored in 384-well plates.

Properties of the Fr-PPIChem Library

The PPI-oriented library has been analyzed and compared to other chemical libraries using visual representations and analytical methods. We used maps of normalized principal moments representation in which low range of the [0,1] interval corresponds to rod- and disk-like shapes and high range of the [0,1] interval to sphere-like shapes[44]. This allows us to differentiate three-dimensional compounds from what is usually referred to as flatland[45]. The Y-axis allows differentiation between rod- and disk-like shapes (Figure 2). The densities along each axis and each population along with the corresponding mean values have been added to highlight the similarities or dissimilarities of the shapes of the compounds in the different datasets.

When examining the molecular shape distributions of the different datasets, one can note several characteristics. DrugBank compounds tend to be the most elongated and the least three-dimensional. PPI-HitProfiler and $2P2I_{HUNTER}$ compounds have very similar distributions of molecular shapes, and both sets of compounds are more three-dimensional and less elongated than DrugBank compounds, while they are more elongated than 2P2I compounds but of comparable three-dimensionality.

iPPIs exhibit different properties than conventional drugs[14, 15]. The eleven descriptors employed by Hamon *et al*. were used to perform PCA[19]. The first two components, representing 53.5% and 22.2% of the total variance for the $2P2I_{DB}$ and DrugBank compounds, respectively, were used for plotting purposes (Figure 3). Distributions of some physicochemical properties of the different sets are given in Supplementary Material (Figure S3). Interestingly, 88.4% of the compounds in Fr-PPIChem follow Veber's rule, which states that compounds with 10 or fewer rotatable bonds and polar surface area equal to or less than 140 $Å^2$ have a high probability of good oral bioavailability[46]. On a similar note, 77.8% of the library complies with Lipinski's rule of five[47].

Molecules selected by the PPI-HitProfiler and $2P2I_{HUNTER}$ models cover different chemical spaces (Figure 3). Notably, combining the space covered by these two models allows all $2P2I_{DB}$ space to be covered even though the individual models cover only part of this space. This clearly demonstrates the high complementarity of the two approaches for searching for iPPI-like molecules.

Comparison of Fr-PPIChem to commercially available PPI-focused libraries, namely Asinex, ChemDiv, Life Chemicals and Otava, show a similar profile in terms of physiochemical properties (Figure S4). Indeed, boxplots calculated for LogP, molecular weight, number of

hydrogen bond donors and acceptors, TPSA, number of rotatable bonds and number of rings, demonstrate that Fr-PPIChem compounds share a similar chemical space.

Experimental evaluation of the Fr-PPIChem library

Finally, we compared the performance of the Fr-PPIChem library to a non-iPPI enriched library in a screening for small molecule iPPIs between the immune checkpoint CD47 and its counter receptor SIRPα. As described in Miller *et al.,* 2019[42], a robust set of HTS-capable biochemical assays (HTRF and AlphaScreen) were designed and optimized based on the interactions of the soluble domains of CD47 and SIRPα. The HTRF assay was used to screen a large, diverse, drug-like library at the National Center for Advancing Translational Studies (NCATS, a division of the NIH). This library (the Genesis collection) was composed of 94,965 (now over 120,000) molecules designed to be diverse (1,000 scaffolds that vary in representation from 20 to 100 compounds per chemotype), highly curated (PAINS and Lipinski), and lead-like (sp^3-enriched, spirocycles, and novel chemotypes). The screening strategy consisted of an HTRF-based primary screen in qHTS format (6 or 7-point concentration response) with activity defined as > 25% inhibition (due to the low overall activity rate). Active compounds were retested and confirmed using the HTRF assay followed by a counter screening (HTRF) and an orthogonal screening (AlphaScreen) to remove false positive compounds. This resulted in 12 active compounds after primary screening and 5 after counter and orthogonal screenings (as described in Miller *et al*. 2019[42]) for an overall activity rate of < 0.01% (**Table 3**). Notably, none of the active compounds were complete inhibitors showing more than 50% activity. We then compared the activity rate of this typical diversity library to that of the focused Fr-PPIChem library using the same HTRF assay and reagents with a slightly modified process to remove false positives. The Fr-PPIChem library (10,314 compounds) was first screened at a single concentration (20

μM) followed by confirmatory testing in a 7-point concentration response and combination counter and orthogonal screenings via a thermal shift assay. The number of active compounds following the primary and confirmatory screening was substantially higher for the Fr-PPIChem library using the same criteria as the NIH screen (779 vs. 5). To focus on higher quality active compounds, we increased the stringency of the activity criteria to > 90% inhibition and IC50 < 12 μM resulting in 173 active compounds. Following thermal shift assay screening to remove false positives and promiscuous compounds as well as to define CD47 or SIRPα interaction, 25 active compounds remained for an overall activity rate of 0.24% representing a 46-fold enrichment for the Fr-PPIChem library vs. the NCATS Genesis library.

Distribution of the Fr-PPIChem library.

The Fr-PPIChem library can be requested to Xavier Morelli (xavier.morelli@inserm.fr, coordinator of the national consortium). A material transfer agreement (MTA) has to be signed by both parties before the chemical library can be delivered at cost price (including the cost of fluids, shipping and handling). Confidentiality will be guaranteed by this MTA and no intellectual property (IP) will be claimed by the national consortium. The library is provided in 384 well plates (33 plates in total) at a concentration of 1mM and a volume of 15μl per well. People requesting the Fr-PPIChem library should provide evidence that a miniaturized and optimized test is available to screen the library.

CONCLUSION

Here, we present a national consortium effort that has produced a unique focused library enriched with iPPI-like compounds. For this purpose, two models based on complementary approaches have been constructed. These models were validated by 5-fold cross-validation and achieved high performances.

The application of these models to large chemical libraries resulted in 124,589 structures. Compounds with undesirable moieties (toxic, PAINS, etc.) and those exhibiting undesirable physicochemical properties were removed leading to a medicinal chemistry-driven collection of 78,243 molecules. Finally, this set was refined to a chemically diverse library of 10,314 compounds (Fr-PPIChem) using a clustering procedure.

The shapes of the molecules in Fr-PPIChem are similar to those of known PPI modulators; therefore, they should be able to bind PPI cavities. Interestingly, selections from PPI-HitProfiler and $2P2I_{HUNTER}$ are equally represented in the final library, and PCA showed that they cover a complementary chemical space.

Fr-PPIChem was evaluated against the CD47/SIRPα PPI target for which no small molecule inhibitor is currently available. An almost 50-fold enrichment in hit rate performance was observed compared to the recently reported screening of a large non-PPI library on the same target[42].

This academic PPI-oriented library constitutes a unique tool for improving the hit rate of screening campaigns. In this effort, this new collection will be accessible to the scientific community, and its reasonable size makes it practical for screening by small biotech companies as well as in the academic environment against numerous PPI targets. It is anticipated that the distribution of the library will accelerate the identification of bioactive molecules targeting challenging protein-protein interfaces and should result in the development of new chemical probes and drug candidates for clinical applications in a time- and cost-effective manner.

Acknowledgment

This work was supported by research funding from the French National Research Agency (ANR-15-CE18-0023, ANR-18-CE11-0023 and ANR-20-COVI-0047), Canceropole PACA

Prematuration and U01 CA218259/CA/NCI NIH HHS (USA). N.B. and L.H. were supported by fellowships from ANR-15-CE18-0023. S.B. and P.B. wish to thank the Région Centre Val de Loire for financial support. We acknowledge the Datacentre IT and Scientific Computing plateform of the CRCM.

Abbreviations

AP, Atom pairs; CV, cross-validation; FN, false negative; FP false positive; HTRF, homogeneous time-resolved fluorescence; MCC, Matthews correlation coefficient; PCA, principal component analysis; PPI, protein-protein interaction; RF, random forests; SVM, support vector machine; TN, true negative; TP, true positive; TPSA, Topological polar surface area.

REFERENCES

[1] Li, Z., Ivanov, A. A., Su, R., Gonzalez-Pecchi, V., Qi, Q., Liu, S., Webber, P., McMillan, E., Rusnak, L., Pham, C., Chen, X., Mo, X., Revennaugh, B., Zhou, W., Marcus, A., Harati, S., Johns, M. A., White, M. A., Moreno, C., Cooper, L. A., Du, Y., Khuri, F. R., and Fu, H. (2017) The OncoPPi network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies, *Nat Commun 8*, 14356.

[2] Donev, R. (2018) Preface, *Adv Protein Chem Struct Biol 112*, xiii-xiv.

[3] Mullard, A. (2012) Protein–protein interaction inhibitors get into the groove, *Nat Rev Drug Discov 11*, 173-175.

[4] Arkin, M. R., Tang, Y., and Wells, J. A. (2014) Small-molecule inhibitors of protein-protein interactions: progressing toward the reality, *Chem Biol 21*, 1102-1114.

[5] Kuenemann, M. A., Sperandio, O., Labbé, C. M., Lagorce, D., Miteva, M. A., and Villoutreix, B. O. (2015) In silico design of low molecular weight protein–protein interaction inhibitors: Overall concept and recent advances, *Prog Biophys Mol Biol 119*, 20-32.

[6] Sheng, C., Dong, G., Miao, Z., Zhang, W., and Wang, W. (2015) State-of-the-art strategies for targeting protein-protein interactions by small-molecule inhibitors, Chem Soc Rev 44, 8238-8259.

[7] Scott, D. E., Bayly, A. R., Abell, C., and Skidmore, J. (2016) Small molecules, big targets: drug discovery faces the protein-protein interaction challenge, *Nat Rev Drug Discov 15*, 533-550.

[8] Gentile, M., Petrungaro, A., Uccello, G., Vigna, E., Recchia, A. G., Caruso, N., Bossio, S., De Stefano, L., Palummo, A., Storino, F., Martino, M., and Morabito, F. (2017) Venetoclax for the treatment of chronic lymphocytic leukemia, *Expert Opin Investig Drugs 26*, 1307-1316.

[9] Gupta, S. (2011) New drug development, JayPee Brothers Medical Publishers Ltd, Delhi, India.

[10] Barker, A., Kettle, J. G., Nowak, T., and Pease, J. E. (2013) Expanding medicinal chemistry space, *Drug Discov Today 18*, 298-304.

[11] Higueruelo, A. P., Jubb, H., and Blundell, T. L. (2013) TIMBAL v2: update of a database holding small molecules modulating protein-protein interactions, *Database (Oxford) 2013*, bat039. doi: 10.1093/database/bat039.

[12] Labbe, C., Kuenemann, M., Zarzycka, B., Vriend, G., Nicolaes, G., Lagorce, D., Miteva, M., Villoutreix, B., and Sperandio, O. (2016) iPPI-DB: an online database of modulators of protein-protein interactions, *Nucleic Acids Res 44*, D542-D547.

[13] Basse, M. J., Betzi, S., Morelli, X., and Roche, P. (2016) 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein-protein interactions, *Database (Oxford). 2016*, baw007. doi: 010.1093/database/baw1007.

[14] Sperandio, O., Reynès, C., Camproux, A., and Villoutreix, B. (2010) Rationalizing the chemical space of protein-protein interaction inhibitors., *Drug Discov Today 15*, 220-229.

[15] Morelli, X., Bourgeas, R., and Roche, P. (2011) Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I), *Curr Opin Chem Biol 15*, 475-481.

[16] Villoutreix, B. O., Labbé, C. M., Lagorce, D., Laconde, G., and Sperandio, O. (2012) A leap into the chemical space of protein-protein interaction inhibitors, *Curr Pharm Des 18*, 4648-4667.

[17] Zhang, X., Betzi, S., Morelli, X., and Roche, P. (2014) Focused chemical libraries--design and enrichment: an example of protein-protein interaction chemical space, *Future Med Chem 6*, 1291-1307.

[18] Neugebauer, A., Hartmann, R. W., and Klein, C. D. (2007) Prediction of Protein–Protein Interaction Inhibitors by Chemoinformatics and Machine Learning Methods, *J Med Chem 50*, 4665-4668.

[19] Hamon, V., Bourgeas, R., Ducrot, P., Theret, I., Xuereb, L., Basse, M. J., Brunel, J. M., Combes, S., Morelli, X., and Roche, P. (2014) 2P2I HUNTER: a tool for filtering orthosteric protein-protein interaction modulators via a dedicated support vector machine*, J R Soc Interface. 11*, 20130860. doi: 20130810.20131098/rsif.20132013.20130860.

[20] Bosc, N., Kuenemann, M. A., Bécot, J., Vavrusa, M., Cerdan, A. H., and Sperandio, O. (2017) Privileged Substructures to Modulate Protein-Protein Interactions, *J Chem Inf Model 57*, 2448-2462.

[21] Hamon, V., Brunel, J. M., Combes, S., Basse, M. J., Roche, P., and Morelli, X. (2013) 2P2Ichem: Focused Chemical Libraries Dedicated to Orthosteric Modulation of Protein-Protein Interactions, *MedChemComm 4, 797-809*.

[22] Milhas, S., Raux, B., Betzi, S., Derviaux, C., Roche, P., Restouin, A., Basse, M. J., Rebuffet, E., Lugari, A., Badol, M., Kashyap, R., Lissitzky, J. C., Eydoux, C., Hamon, V., Gourdel, M. E., Combes, S., Zimmermann, P., Aurrand-Lions, M., Roux, T., Rogers, C., Muller, S., Knapp, S., Trinquet, E., Collette, Y., Guillemot, J. C., and Morelli, X. (2016) Protein-Protein Interaction Inhibition (2P2I)-Oriented Chemical Library Accelerates Hit Discovery, *ACS Chem Biol. 11*, 2140-2148.

[23] Raux, B., Voitovich, Y., Derviaux, C., Lugari, A., Rebuffet, E., Milhas, S., Priet, S., Roux, T., Trinquet, E., Guillemot, J. C., Knapp, S., Brunel, J. M., Fedorov, A. Y., Collette, Y., Roche, P., Betzi, S., Combes, S., and Morelli, X. (2016) Exploring Selective Inhibition of the First

Bromodomain of the Human Bromodomain and Extra-terminal Domain (BET) Proteins, *J Med Chem 59*, 1634-1641.

[24] Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration, *Nucleic Acids Res 34*, D668-D672.

[25] Cao, Y., Charisi, A., Cheng, L.-C., Jiang, T., and Girke, T. (2008) ChemmineR: a compound mining framework for R, *Bioinformatics 24*, 1733-1734.

[26] Todeschini, R., Consonni, V., Mannhold, R., Kubinyi, H., and Timmerman, H. (2000) Handbook of Molecular Descriptors, Wiley-VCH. doi:10.1002/9783527613106.

[27] Ruggiu, F., Marcou, G., Varnek, A., and Horvath, D. (2010) ISIDA Property-Labelled Fragment Descriptors, *Mol. Inf. 29*, 855-868.

[28] Varnek, A., Fourches, D., Hoonakker, F., and Solov'ev, V. P. (2005) Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures, *J Comput Aided Mol Des 19*, 693-703.

[29] Chang, C.-C., and Lin, C.-J. (2011) LIBSVM : a library for support vector machines, *ACM T Intel Syst Tec 2*, 1-27.

[30] Platt, J. (1999) Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*, In Advances in Large Margin Classifiers*, pp 61-74.

[31] Rücker, C., Rücker, G., and Meringer, M. (2007) y-Randomization and its variants in QSPR/QSAR, *J Chem Inf Model 47*, 2345-2357.

[32] Dassault Systèmes BIOVIA. Pipeline Pilot, version 9.0.2 (2019), San Diego: Dassault Systèmes.

[33] Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities, *Nucleic Acids Res 35*, D198-201.

[34] Kuenemann, M., Labbe, C., Cerdan, A., and Sperandio, O. (2016) Imbalance in chemical space: How to facilitate the identification of protein-protein interaction inhibitors, *Sci. Rep. 6*, 23815. doi: 10.1038/srep23815.

[35] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009) The WEKA data mining software: an update, *SIGKDD Explor. Newsl. 11*, 10-18.

[36] Team, R. C. (2018) R: A Language and Environment for Statistical Computing., R Foundation for Statistical Computing, Vienna. Austria.

[37] Kuhn, M. (2008) Building Predictive Models in R Using the caret Package, *J. Stat. Softw. 28*, 1-26. doi:10.18637/jss.v028.i05.

[38] Lagorce, D., Bouslama, L., Becot, J., Miteva, M. A., and Villoutreix, B. O. (2017) FAF-Drugs4: free ADME-tox filtering computations for chemical biology and early stages drug discovery, *Bioinformatics 33*, 3658-3660.

[39] Doak, B. C., Over, B., Giordanetto, F., and Kihlberg, J. (2014) Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates, *Chem Biol 21*, 1115-1142.

[40] Baell, J. B., and Holloway, G. A. (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays, *J Med Chem 53*, 2719-2740.

[41] Gally, J. M., Bourg, S., Fogha, J., Do, Q. T., Aci-Sèche, S., and Bonnet, P. (2019) VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening, *Curr Med Chem. 26*, 1-15. doi: 10.2174/0929867326666190614160451.

[42] Miller, T. W., Amason, J. D., Garcin, E. D., Lamy, L., Dranchak, P. K., Macarthur, R., Braisted, J., Rubin, J. S., Burgess, T. L., Farrell, C. L., Roberts, D. D., and Inglese, J. (2019) Quantitative high-throughput screening assays for the discovery and development of SIRPα-CD47 interaction inhibitors, *PLoS One 14*, e0218897. doi: 10.1371/journal.pone.0218897.

[43] Burgess, T. L., Amason, J. D., Rubin, J. S., Duveau, D. Y., Lamy, L., Roberts, D. D., Farrell, C. L., Inglese, J., Thomas, C. J., and Miller, T. W. (2020) A homogeneous SIRPα-CD47 cell-based, ligand-binding assay: Utility for small molecule drug development in immuno-oncology, *PLoS One 15*, e0226661. doi: 10.1371/journal.pone.0226661.

[44] Kuenemann, M. A., Bourbon, L. M., Labbé, C. M., Villoutreix, B. O., and Sperandio, O. (2014) Which three-dimensional characteristics make efficient inhibitors of protein-protein interactions?, *J Chem Inf Model 54*, 3067-3079.

[45] Lovering, F., Bikker, J., and Humblet, C. (2009) Escape from flatland: increasing saturation as an approach to improving clinical success, *J Med Chem 52*, 6752-6756.

[46] Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002) Molecular Properties That Influence the Oral Bioavailability of Drug Candidates, *J Med Chem 45*, 2615-2623.

[47] Lipinski, C., Lombardo, F., Dominy, B., and Feeney, P. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings., *Adv Drug Deliv Rev 46*, 3-26.

Table 1

Descriptors	Validation	Specificity[a]	Sensitivity[b]	BA[c]	AUC[d]	MCC[e]
Dragon	5-CV	0.686	0.738	0.712	0.753	0.270
	test	0.782	0.694	0.738	0.766	0.328
MOE 2D	5-CV	0.730	0.706	0.718	0.786	0.286
	test	0.827	0.565	0.696	0.778	0.291
ISIDA-IAB(2-2)	5-CV	0.882	0.570	0.726	0.824	0.371
	test	0.891	0.518	0.704	0.807	0.348

Table 1. Averaged statistics of the $2P2I_{HUNTER}$ model, computed for the best SVM models obtained with each type of descriptors in 5-fold cross-validation (5-CV) and for the 5 test sets (test). [a]*TN/(TN + FP)*; [b]*TP/(TP + FN)*, [c] 0.5*(specificity + sensitivity); [d]Area under the curve. [e] Matthews correlation coefficient.

Table 2

Models	Datasets	Specificity[a]	Sensitivity[b]	BA[c]	EF[d]	MCC[e]
RF	RF-5-CV	1.00	0.73	0.87	27.93	0.85
	RF-Test	1.00	0.76	0.88	28.04	0.86
SVM	SVM-5-CV	1.00	0.89	0.95	27.91	0.94
	SVM-Test	1.00	0.91	0.95	27.89	0.95
JRip	JRip-5-CV	1.00	0.72	0.86	24.00	0.78
	JRip-Test	1.00	0.77	0.89	24.88	0.82
J48	J48-5-CV	0.99	0.72	0.86	23.68	0.77
	J48-Test	0.99	0.76	0.88	23.53	0.79

Table 2. Averaged statistics computed for the 4 new optimized PPI-HitProfiler models on the training set (Train), the 5-fold cross validation (5-CV), and the test set (Test). [a]*TN/(TN+FP)*; [b]*TP/(TP+FN)*; [c] 0.5*(specificity + sensitivity); [d]enrichment factor (EF) and [e]Matthews correlation coefficient (MCC) are described in the methods section.

Table 3

Genesis Library Size (# of compounds)	94,965	
	# of actives[a]	activity rate[b]
Primary Screen (HRTF)	12	0.013%
Confirmation (HTRF)	8	0.008%
Counter Screen (HTRF)	5	0.005%
Orthogonal Screen (AlphaScreen)	5	0.005%

Fr-PPIChem Library Size (# of compounds)	10,314	
	# of actives[a]	activity rate[b]
Primary Screen (HRTF)	1623	15.736%
Confirmation (HTRF)	173	1.677%
Orthogonal/Counter Screen (TSA)	25	0.242%

Overall activity rate enrichment[c]	46

Table 3. Results from HTS screens for small molecule inhibitors of the CD47-SIRPα PPI. [a]Active compounds defined as in Results. [b]Activity rate calculated as the number of actives divided by the library size. [c]Overall activity rate enrichment calculated from the final activity rate for the Fr-PPIChem library divided by the final activity rate for the Genesis library.

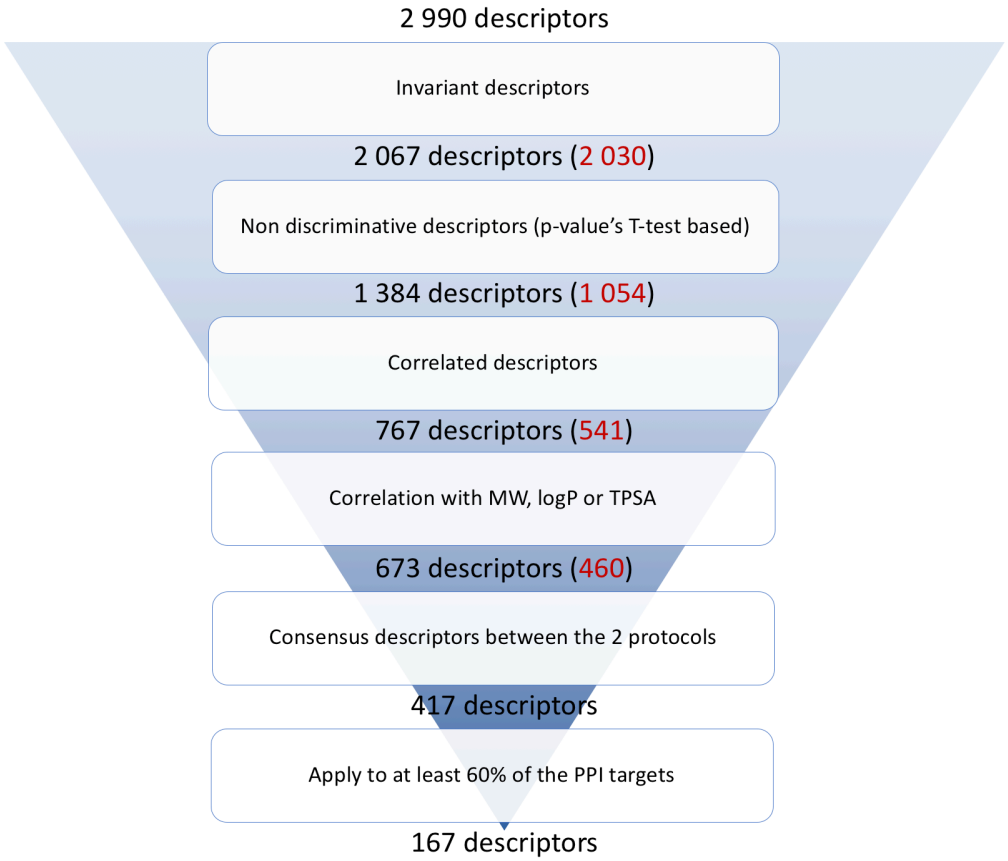

Figure 1: Variable selection procedure prior to the construction of the PPI-HitProfiler models.

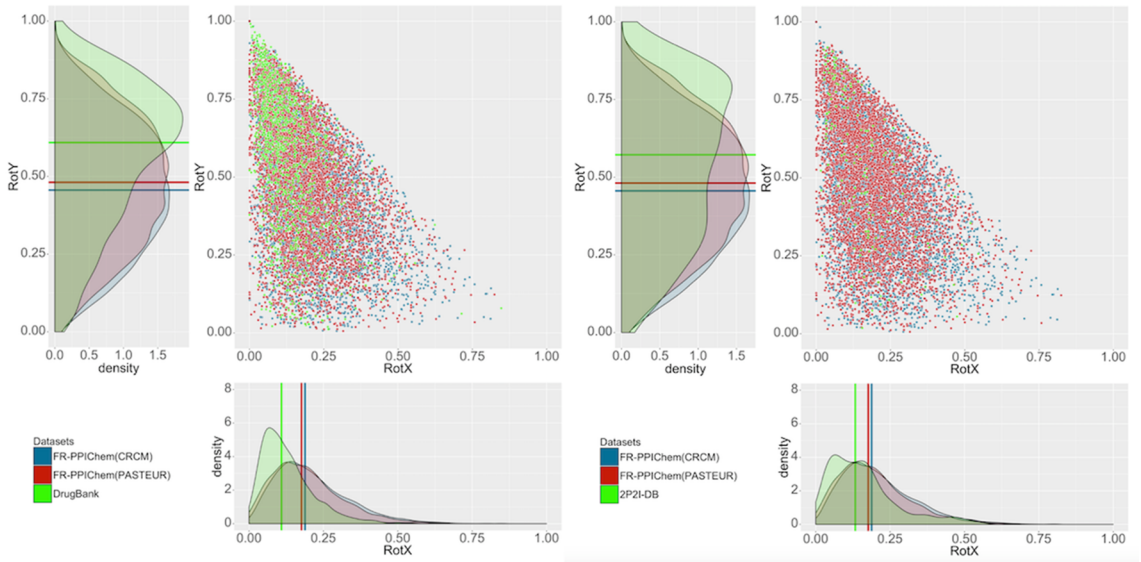

Figure 2. Rotated NPR1 and NPR2 plots of the Fr-PPIChem and reference databases. Comparison of the Fr-PPIChem contributions ($2P2I_{HUNTER}$ and PPI-HitProfiler) with DrugBank (left) and with $2P2I_{DB}$ (right).

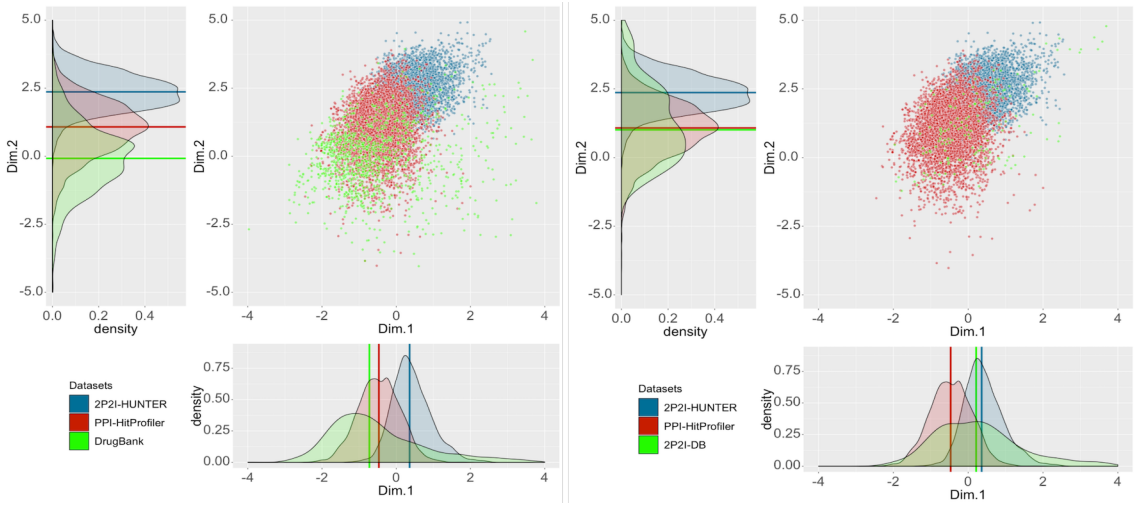

Figure 3. 2D PCA for Fr-PPIChem ($2P2I_{HUNTER}$ and PPI-HitProfiler contributions), DrugBank (left) and $2P2I_{DB}$ (right) using 11 molecular descriptors. Density curves along principal components and for each datasets are also provided.

Supplementary information

Fr-PPIChem: An academic compound library dedicated to protein-protein interactions

Nicolas Bosc[a,b,‡], Christophe Muller[c,‡], Laurent Hoffer[d], David Lagorce[e], Stéphane Bourg[f], Carine Derviaux[c], Marie-Edith Gourdel[g], Jean-Christophe Rain[g], Thomas Miller[c], Bruno O. Villoutreix[h], Maria A. Miteva[i], Pascal Bonnet[f], Xavier Morelli[c,d], Olivier Sperandio,[a,b,], Philippe Roche[d,*]*

[a]*Inserm U973 MTi, 25 rue Hélène Brion 75013 Paris*

[b]*Institut Pasteur, Unité de Bioinformatique Structurale, CNRS UMR3528, 28 rue du Dr Roux 75015 Paris*

[c]*IPC Drug Discovery Platform, Institut Paoli-Calmettes, 232 Boulevard de Sainte-Marguerite, 13009, Marseille, France*

[d]*CRCM, CNRS, INSERM, Institut Paoli-Calmettes, Aix-Marseille Univ, Marseille, France*

[e]*Université de Paris, INSERM US14, Plateforme Maladies Rares - Orphanet, Paris, France*

[f]*Institut de Chimie Organique et Analytique (ICOA), Université d'Orléans, UMR CNRS 7311, BP 6759, 45067 Orléans. France*

[g]*Hybrigenics* Services *SAS,* 1 rue Pierre Fontaine, 91000 Evry Courcouronnes, France

[h]*Université de Lille, INSERM, Institut Pasteur de Lille, U1177 - Drugs and Molecules for living Systems, Lille, France*

[i]*Inserm U1268 MCTR, CNRS UMR 8038 CiTCoM – Univ. De Paris, Faculté de Pharmacie de Paris, Paris, France*

[‡]These authors contributed equally to this work

[*]To whom correspondence should be addressed. E-mail: *olivier.sperandio@inserm.fr* , *philippe.roche@inserm.fr*

Figure S1

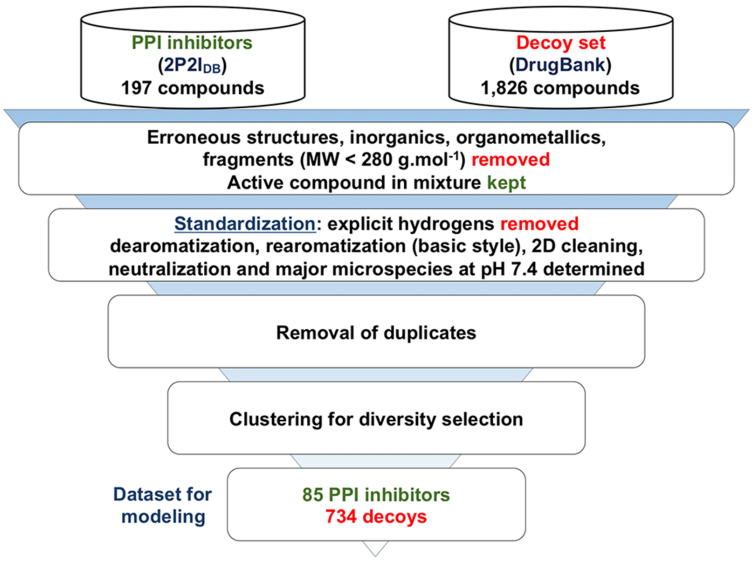

Figure S1 : Training set Preparation. In an effort to work with consistent datasets, the following procedure was applied using ChemAxon (http://www.chemaxon.com). Erroneous molecules were detected using Structure Checker (*version 15.5.4.0*) and discarded from the dataset. Next, inorganics, organometallics and fragments were also removed. Concerning mixtures, only the molecule with the largest number of atoms was kept when it was clear that the experimental biological activity of the mixture was due to this compound. When no clear distinction could be made the whole mixture was discarded. This step was performed using Standardizer (*version 15.5.4.0*). In an effort to represent all molecules in the same manner, further standardization steps were performed: removal of explicit hydrogens, dearomatization, rearomatization (basic style), 2D cleaning and neutralization. Then, the major microspecies at pH 7.4 was determined for each molecule using cxcalc from ChemAxon. Next, erroneous cases due to standardization (detected with Structure Checker) as well as duplicates were removed from all datasets. As a result, 170 iPPIs and 1,193 approved drugs were conserved.

The ChemMineR library from R was used for diversity selection. In brief, atom pair descriptors were calculated for each molecule. Then, a fingerprint was generated for each molecule based on these atom pair descriptors. Finally, a single-linkage binning clustering was performed using the generated fingerprints and for a given similarity threshold. Several thresholds were tested to obtain the most coherent clustering of molecules. Visual observation of clustering led to choose a Tanimoto threshold of 0.8. For each cluster, the molecule with the lowest distance to the cluster center was selected. When only two molecules were present in a cluster, one molecule was randomly chosen. Finally, 85 iPPIs were selected in the positive dataset and 734 in the decoy dataset.

Figure S2

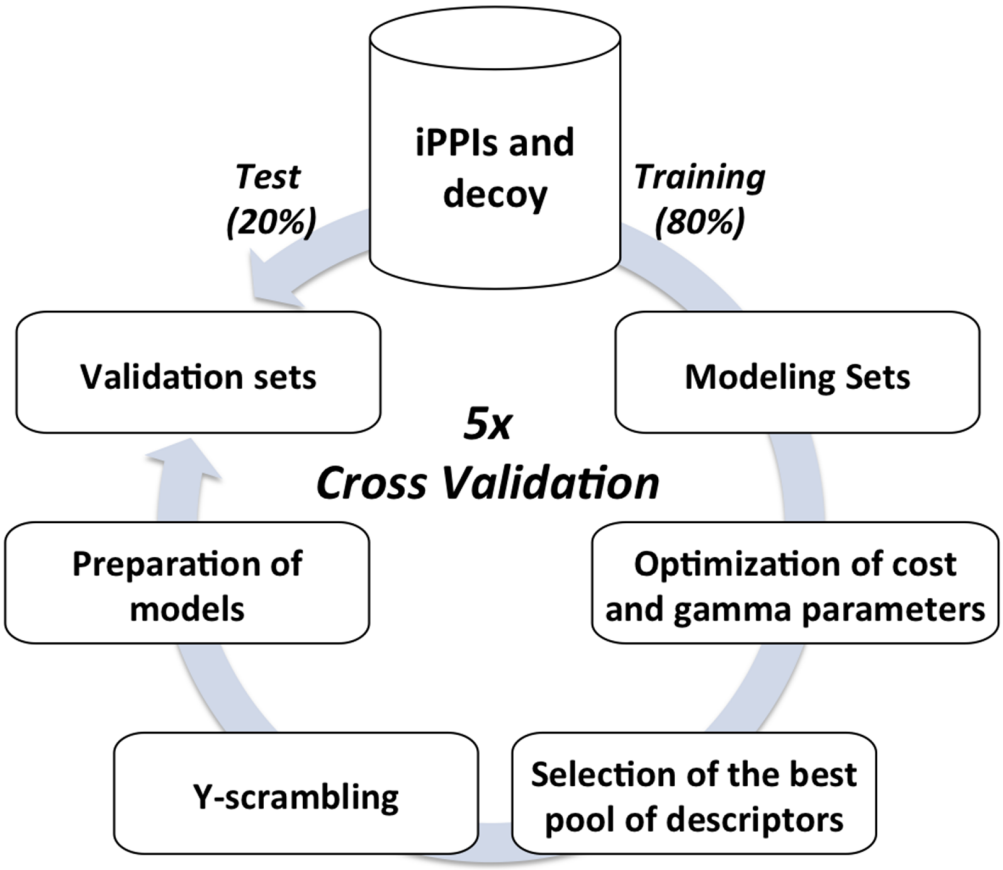

Figure S2. Workflow applied for 2P2I$_{HUNTER}$ model optimisation and validation. First, initial dataset (85 iPPIs and 734 compounds as decoy) was randomly split into five modeling and tests sets. Importantly, special attention was paid to place all inhibitors of a given protein-protein family in the same dataset. Then, for each modeling set another 5-fold cross-validation was undertaken to determine the best pool of descriptors and to optimize cost and gamma parameters. For this purpose, cost ranged from 1 to 151 incrementing by 10, and gamma varied from 10^{-4} to 10^{-3} incrementing by 10^{-4}, from 10^{-3} to 10^{-2} incrementing by 10^{-3} and from 10^{-2} to 1 incrementing by 10^{-2}. Instances were weighted according to the proportion of their class. For the best pool of descriptors and for each modeling set, a procedure of y-scrambling was performed to estimate the part of models related to chance. Finally, validated models were applied on the test set. The whole procedure was repeated five times so that all compounds of the initial dataset is predicted once.

Figure S3

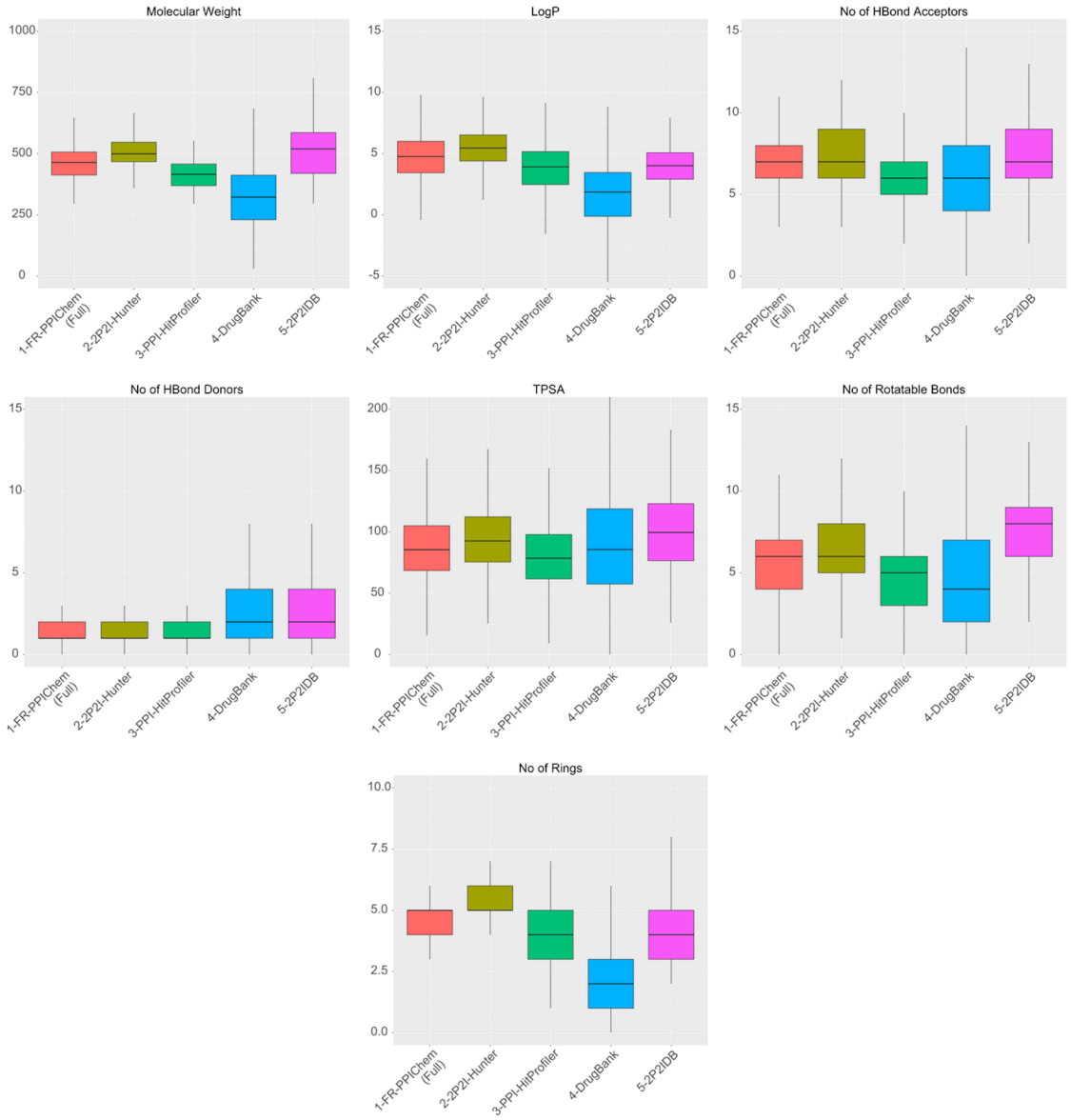

Figure S4. Box plots of molecular weight, logP, number of hydrogen bond acceptors and donors, TPSA, number of rotatable bonds, and number of rings for compounds in Fr-PPIChem, as full (1) or for individual contributions selected by $2P2I_{HUNTER}$ (2) and PPI-HitProfiler (3) modelsas well as DrugBank « approved » and $2P2I_{DB}$.

Figure S4

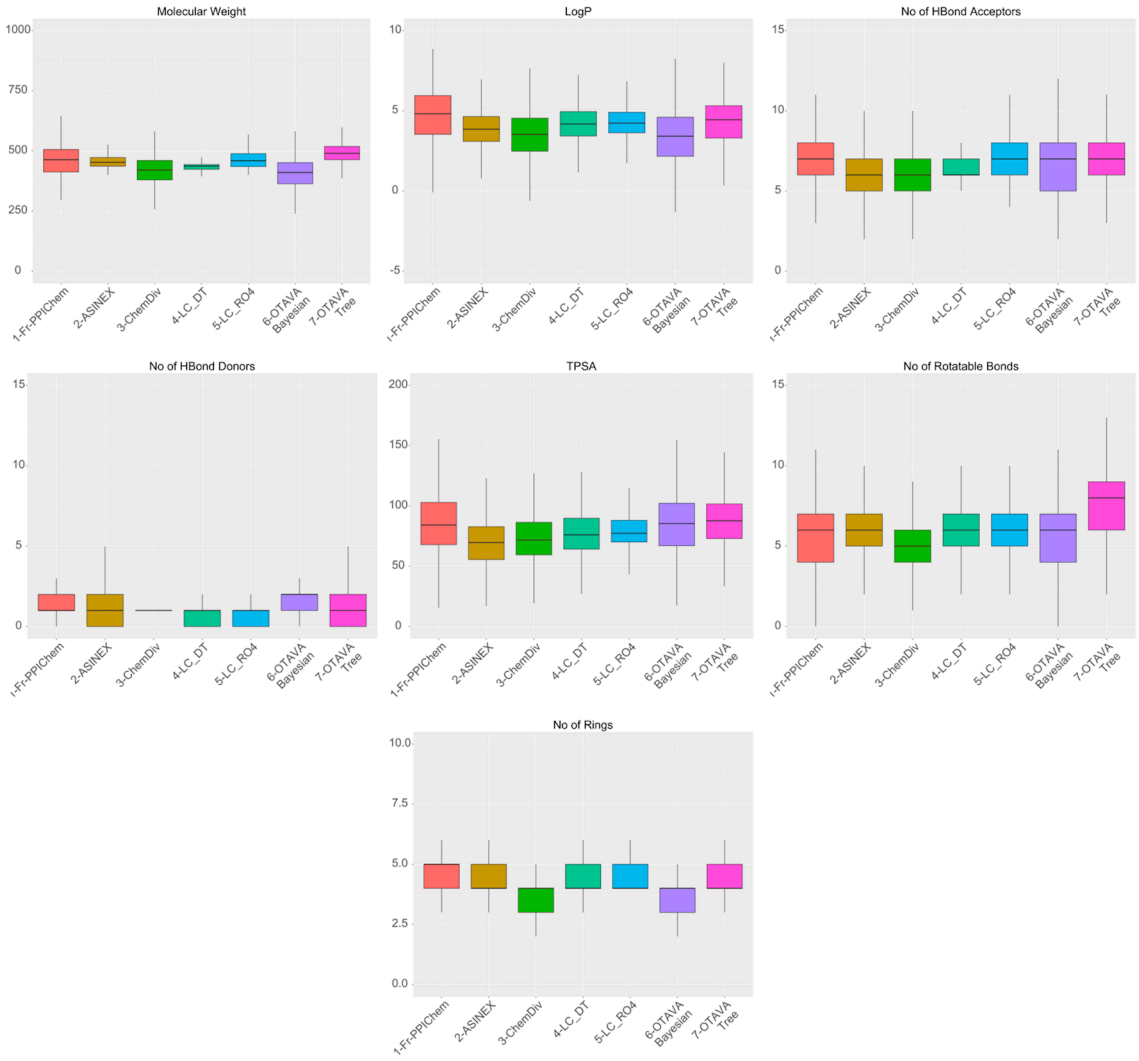

Figure S5. Box plots of molecular weight, logP, number of hydrogen bond acceptors and donors, TPSA, number of rotatable bonds, and number of rings for compounds in Fr-PPIChem and commercially available PPI-Focused libraries: Asinex, ChemDiv, LifeChemicals (Decision Tree), LifeChemicals (Rule of 4), Otava (Bayesian) and Otava (Decision Tree).

Table S1
List of molecular descriptors

Name	Origin	Description	Dragon Block
Mi	Dragon	mean first ionization potential (scaled on Carbon atom)	Constitutional indices
nTA	Dragon	number of terminal atoms	Constitutional indices
C%	Dragon	percentage of C atoms	Constitutional indices
nR03	Dragon	number of 3-membered rings	Ring descriptors
PW3	Dragon	path/walk 3 - Randic shape index	Topological indices
PW4	Dragon	path/walk 4 - Randic shape index	Topological indices
PW5	Dragon	path/walk 5 - Randic shape index	Topological indices
MAXDP	Dragon	maximal electrotopological positive variation	Topological indices
Psi_i_t	Dragon	intrinsic state pseudoconnectivity index - type T	Topological indices
Psi_e_A	Dragon	electrotopological state pseudoconnectivity index - type S average	Topological indices
SRW09	Dragon	self-returning walk count of order 9	Walk and path counts
X1A	Dragon	average connectivity index of order 1	Connectivity indices
X2A	Dragon	average connectivity index of order 2	Connectivity indices
X4A	Dragon	average connectivity index of order 4	Connectivity indices
X0Av	Dragon	average valence connectivity index of order 0	Connectivity indices
X4Av	Dragon	average valence connectivity index of order 4	Connectivity indices
IVDE	Dragon	mean information content on the vertex degree equality	Information indices
IC1	Dragon	Information Content index (neighborhood symmetry of 1-order)	Information indices
IC2	Dragon	Information Content index (neighborhood symmetry of 2-order)	Information indices
BIC5	Dragon	Bond Information Content index (neighborhood symmetry of 5-order)	Information indices
P_VSA_LogP_3	Dragon	P_VSA-like on LogP bin 3	P_VSA-like descriptor
P_VSA_ppp_con	Dragon	P_VSA-like on potential pharmacophore points con - conjugated atoms	P_VSA-like descriptor
Eta_F_A	Dragon	eta average functionality index	ETA indices
Eta_sh_y	Dragon	eta y shape index	ETA indices
nCrq	Dragon	number of ring quaternary C(sp3)	Functional group counts
nCb-	Dragon	number of substituted benzene C(sp2)	Functional group counts
nR=Cp	Dragon	number of terminal primary C(sp2)	Functional group counts
nRCOOR	Dragon	number of esters (aliphatic)	Functional group counts

Table S1
List of molecular descriptors

nArCOOR	Dragon	number of esters (aromatic)	Functional group counts
nArCONHR	Dragon	number of secondary amides (aromatic)	Functional group counts
nROCON	Dragon	number of (thio-) carbamates (aliphatic)	Functional group counts
nArOCON	Dragon	number of (thio-) carbamates (aromatic)	Functional group counts
nRCO	Dragon	number of ketones (aliphatic)	Functional group counts
nCONN	Dragon	number of urea (-thio) derivatives	Functional group counts
nN=C-N<	Dragon	number of amidine derivatives	Functional group counts
nArCNO	Dragon	number of oximes (aromatic)	Functional group counts
nRCN	Dragon	number of nitriles (aliphatic)	Functional group counts
nN+	Dragon	number of positively charged N	Functional group counts
nRNHO	Dragon	number of hydroxylamines (aliphatic)	Functional group counts
nN(CO)2	Dragon	number of imides (-thio)	Functional group counts
nArOR	Dragon	number of ethers (aromatic)	Functional group counts
nImidazoles	Dragon	number of Imidazoles	Functional group counts
nThiophenes	Dragon	number of Thiophenes	Functional group counts
nOxazoles	Dragon	number of Oxazoles	Functional group counts
nTriazoles	Dragon	number of Triazoles	Functional group counts
nPyridazines	Dragon	number of Pyridazines	Functional group counts
nPyrazines	Dragon	number of Pyrazines	Functional group counts
C-007	Dragon	CH2X2	Atom-centred fragments
C-009	Dragon	CHRX2	Atom-centred fragments
C-025	Dragon	R--CR--R	Atom-centred fragments
C-031	Dragon	X--CR--X	Atom-centred fragments
C-032	Dragon	X--CX--X	Atom-centred fragments
C-035	Dragon	R--CX..X	Atom-centred fragments
C-038	Dragon	Al-C(=X)-Al	Atom-centred fragments
C-041	Dragon	X-C(=X)-X	Atom-centred fragments
C-043	Dragon	X--CR..X	Atom-centred fragments
C-044	Dragon	X--CX..X	Atom-centred fragments

H-051	Dragon	H attached to alpha-C	Atom-centred fragments
O-060	Dragon	Al-O-Ar / Ar-O-Ar / R..O..R / R-O-C=X	Atom-centred fragments
N-074	Dragon	R#N / R=N-	Atom-centred fragments
N-079	Dragon	N+ (positively charged)	Atom-centred fragments
S-106	Dragon	R-SH	Atom-centred fragments
SdsCH	Dragon	Sum of dsCH E-states	Atom-type E-state indices
SdssC	Dragon	Sum of dssC E-states	Atom-type E-state indices
SdNH	Dragon	Sum of dNH E-states	Atom-type E-state indices
SsOH	Dragon	Sum of sOH E-states	Atom-type E-state indices
NaasC	Dragon	Number of atoms of type aasC	Atom-type E-state indices
NssO	Dragon	Number of atoms of type ssO	Atom-type E-state indices
NaaS	Dragon	Number of atoms of type aaS	Atom-type E-state indices
CATS2D_01_DD	Dragon	CATS2D Donor-Donor at lag 01	CATS 2D
CATS2D_08_DA	Dragon	CATS2D Donor-Acceptor at lag 08	CATS 2D
CATS2D_04_DP	Dragon	CATS2D Donor-Positive at lag 04	CATS 2D
CATS2D_06_DP	Dragon	CATS2D Donor-Positive at lag 06	CATS 2D
CATS2D_04_DL	Dragon	CATS2D Donor-Lipophilic at lag 04	CATS 2D
CATS2D_07_DL	Dragon	CATS2D Donor-Lipophilic at lag 07	CATS 2D
CATS2D_08_DL	Dragon	CATS2D Donor-Lipophilic at lag 08	CATS 2D
CATS2D_09_DL	Dragon	CATS2D Donor-Lipophilic at lag 09	CATS 2D
CATS2D_03_AA	Dragon	CATS2D Acceptor-Acceptor at lag 03	CATS 2D
CATS2D_04_AP	Dragon	CATS2D Acceptor-Positive at lag 04	CATS 2D
CATS2D_03_AL	Dragon	CATS2D Acceptor-Lipophilic at lag 03	CATS 2D
CATS2D_04_AL	Dragon	CATS2D Acceptor-Lipophilic at lag 04	CATS 2D
CATS2D_04_PL	Dragon	CATS2D Positive-Lipophilic at lag 04	CATS 2D
CATS2D_09_LL	Dragon	CATS2D Lipophilic-Lipophilic at lag 09	CATS 2D
B01[C-O]	Dragon	Presence/absence of C - O at topological distance 1	2D Atom Pairs
B02[C-O]	Dragon	Presence/absence of C - O at topological distance 2	2D Atom Pairs
B02[N-S]	Dragon	Presence/absence of N - S at topological distance 2	2D Atom Pairs

Table S1
List of molecular descriptors

B02[S-S]	Dragon	Presence/absence of S - S at topological distance 2	2D Atom Pairs
B03[C-O]	Dragon	Presence/absence of C - O at topological distance 3	2D Atom Pairs
B03[O-O]	Dragon	Presence/absence of O - O at topological distance 3	2D Atom Pairs
B03[F-F]	Dragon	Presence/absence of F - F at topological distance 3	2D Atom Pairs
B04[C-O]	Dragon	Presence/absence of C - O at topological distance 4	2D Atom Pairs
B04[O-S]	Dragon	Presence/absence of O - S at topological distance 4	2D Atom Pairs
B04[F-F]	Dragon	Presence/absence of F - F at topological distance 4	2D Atom Pairs
B05[C-O]	Dragon	Presence/absence of C - O at topological distance 5	2D Atom Pairs
B05[S-S]	Dragon	Presence/absence of S - S at topological distance 5	2D Atom Pairs
B05[S-Cl]	Dragon	Presence/absence of S - Cl at topological distance 5	2D Atom Pairs
B06[C-C]	Dragon	Presence/absence of C - C at topological distance 6	2D Atom Pairs
B06[C-O]	Dragon	Presence/absence of C - O at topological distance 6	2D Atom Pairs
B06[N-O]	Dragon	Presence/absence of N - O at topological distance 6	2D Atom Pairs
B06[O-S]	Dragon	Presence/absence of O - S at topological distance 6	2D Atom Pairs
B07[C-C]	Dragon	Presence/absence of C - C at topological distance 7	2D Atom Pairs
B07[C-N]	Dragon	Presence/absence of C - N at topological distance 7	2D Atom Pairs
B07[C-O]	Dragon	Presence/absence of C - O at topological distance 7	2D Atom Pairs
B08[C-C]	Dragon	Presence/absence of C - C at topological distance 8	2D Atom Pairs
B08[C-N]	Dragon	Presence/absence of C - N at topological distance 8	2D Atom Pairs
B08[C-O]	Dragon	Presence/absence of C - O at topological distance 8	2D Atom Pairs
B09[C-C]	Dragon	Presence/absence of C - C at topological distance 9	2D Atom Pairs
B09[C-O]	Dragon	Presence/absence of C - O at topological distance 9	2D Atom Pairs
B10[C-C]	Dragon	Presence/absence of C - C at topological distance 10	2D Atom Pairs
B10[C-O]	Dragon	Presence/absence of C - O at topological distance 10	2D Atom Pairs
F02[C-O]	Dragon	Frequency of C - O at topological distance 2	2D Atom Pairs
F02[S-Cl]	Dragon	Frequency of S - Cl at topological distance 2	2D Atom Pairs
F03[C-O]	Dragon	Frequency of C - O at topological distance 3	2D Atom Pairs
F03[O-O]	Dragon	Frequency of O - O at topological distance 3	2D Atom Pairs
F03[S-S]	Dragon	Frequency of S - S at topological distance 3	2D Atom Pairs

Table S1
List of molecular descriptors

F03[F-F]	Dragon	Frequency of F - F at topological distance 3	2D Atom Pairs
F04[N-S]	Dragon	Frequency of N - S at topological distance 4	2D Atom Pairs
F04[O-S]	Dragon	Frequency of O - S at topological distance 4	2D Atom Pairs
F04[F-F]	Dragon	Frequency of F - F at topological distance 4	2D Atom Pairs
F05[C-O]	Dragon	Frequency of C - O at topological distance 5	2D Atom Pairs
F06[N-O]	Dragon	Frequency of N - O at topological distance 6	2D Atom Pairs
F06[O-S]	Dragon	Frequency of O - S at topological distance 6	2D Atom Pairs
DLS_04	Dragon	modified drug-like score from Chen et al. (7 rules)	Drug-like indices
DLS_cons	Dragon	DRAGON consensus drug-like score	Drug-like indices
NumSaturatedCarbocycles	RDKit		
FractionCSP3	RDKit		
slogp_VSA11	RDKit		
smr_VSA2	RDKit		
smr_VSA10	RDKit		
peoe_VSA1	RDKit		
peoe_VSA5	RDKit		
BCUT_PEOE_1	MOE		
BCUT_PEOE_2	MOE		
BCUT_SLOGP_1	MOE		
BCUT_SMR_1	MOE		
lip_violation	MOE		
BCUT_SMR_2	MOE		
b_double	MOE		
b_rotR	MOE		
chiral_u	MOE		
GCUT_PEOE_1	MOE		
GCUT_PEOE_2	MOE		
GCUT_SLOGP_1	MOE		
GCUT_SLOGP_2	MOE		

Table S1
List of molecular descriptors

GCUT_SMR_1	MOE
GCUT_SMR_2	MOE
PEOE_RPCplus	MOE
PEOE_VSAplus1	MOE
PEOE_VSAplus3	MOE
PEOE_VSAmoins2	MOE
PEOE_VSAmoins4	MOE
PEOE_VSA_FNEG	MOE
PEOE_VSA_FPNEG	MOE
Q_RPCplus	MOE
Q_VSA_PPOS	MOE
ASAplus	MOE
DASA	MOE
DCASA	MOE
dipole	MOE
FASAplus	MOE
FASAmoins	MOE
pmiX	MOE
vsurf_CP	MOE
vsurf_EDmin1	MOE
vsurf_EWmin3	MOE
vsurf_IW6	MOE
vsurf_Wp3	MOE

Table S2
substructures used to filter potentially toxic compounds

1_2_aminothiazole	[$(s1c([NX3H1&!R][#6])nc([#6])[cH]1),$(s1c([NX3H1&!R][#6])n[cH]c([#6])1)]
1_2_dicarbonyl_oxalyl	[#1,#6][CX3&!R](=[OX1])[CX3&!R](=[OX1])[#1,#6]
1_2_thiazol_3_one	[S,s]1[NH1,nH1][C,c](=[OX1])[C,c]~[C,c]1
1_aminobenzotriazole	Nn1nnc2ccccc12
2_phenylbenzimidazole	c2ccc3nc(c4ccccc4)[nX3H1]c3c2
3_amino_9_ethylcarbazoles	CCn1c2c(cc(cc2)N)c3ccccc31
4_subst_n_alkyltetrahydropyridines	[a]C1=CCN([A])CC1
4_vinyl_pyridine	c1cnccc1[CX3&!R]=[CX3&!R]
6_membered_aromatic_sulfur_NSN	C1(=[OX1&!R])C=CSC=C1
6_membered_aromatic_sulfur_NSC	C1(=[OX1&!R])C=CSN=C1
6_membered_aromatic_sulfur_CSN	C1(=[OX1&!R])C=NSC=C1
6_membered_aromatic_sulfur_CSC	C1(=[OX1&!R])C=NSN=C1
9_aminoacridine	c1cccc2c([NX3&!RH2])c3ccccc3nc12
acetal_1_in_ring	C1([SX2&!R][#6&!R])=NCCS1
acetal_both_in_ring	C1(=[NX2&!R][CX2&!R]#[NX1&!R])SCCS1
acetylene_alkyne	[$([CX2H1&!R]),$([CX2&!R][#6])]#[#6&!R][#6]
acrylamide	[$([NX3H2&!R][#6]),$([NX3H1&!R]([c,C])[#6]),$([NX3&!R]([c,C])([c,C])[#6])][CX3&!R](=[OX1])[CX3&!R]=[$([CX3H2&!R]),$([CX3H1&!R][c,C]),$([CX3H0&!R]([c,C])[c,C])]
acyclic_acetal	[#6][OX2&!R][$([CX4&!RH2]),$([CX4&!RH1][#6]),$([CX4&!R]([#6])[#6])][OX2&!R][#6]
acyclic_acid_halide_acyl_halide	*[CD3&!R]([F,Cl,I,Br])=[OX1]
acyl_amide	[C,c][C;!R](=O)[N;!R][C;!R](=O)[C,c]
acyl_cyanides	[NX1]#[CX2&!R][CX3&!R](=[OX1])[#6]
acyl_isoamide_aromatic	n1coc(=[OX1])cc1
adamantane	C1C2CC3CC1CC(C2)C3
aldehyde	[CX3&!RH1]=[OD1]
aliphatic_ketone	[#6&!R][CD3&!R](=[OD1&!R])[#6&!R]
alkyl_halide_I	[CX4&!R][Br,Cl]
alkyl_halide_Cl_Br	[CX4&!R][I]
alphahalo_ketone_carbonyl	[OX1&!R]=[CX3&!RH0]([#6])[#6&!R][Cl,F,Br,I]
anhydride	[OX1]=[CD3]([*])[OD2][CD3](=[OX1])[*]
anthracene	a1aa2aa3aaaaa3aa2aa1
azide	[$(N#[N+]-[N-]),$([N-]=[N+]=N),$([#7]=[#7+]=[#7-]),$([#7]=[#7]=[#7])]
aziridine	[NX3]1[CX4][CX4]1

Table S2
substructures used to filter potentially toxic compounds

azo	[#6][NX2]=[NX2][#6]
azocyanamide	[N;R0]=[N;R0]C#N
beta_heterosubstituted_carbonyl	[CH3][CX4&!RH2][CX4&!RH1]([F,Br,I,Cl])[CX4&!RH2][CX3&!R](=[OD1&!R])
betalactams	C1C(=[OX1])N([#6])C1
carbamic_acid	[$([NX4+]),$([ND1&!R]),$([NX3&!H2]),$([NX3&!RH1][#6]),$([NX3&!R]([#6])[#6])][CD3&!R](=[OD1&!R])[OX2&!RH1,OX1-]
carbazide	O=*N=[N+]=[N-]
carbodiimide	[#6][NX2]=C=[NX2][#6]
catechol	[OH]c1c([OH])cccc1
chloramidine	[Cl]C([C&R0])=N
coumarines	[OX1&!R]=c1ccc2ccccc2o1
crown_2_2	[C&!R][#8&!R,#7&!R,#16&!R][C&!R][C&!R][#8&!R,#7&!R,#16&!R][C&!R][C&!R][#8&!R,#7&!R,#16&!R][C&!R]
crown_2_3	[C&!R][#8&!R,#7&!R,#16&!R][C&!R][C&!R][C&!R][#8&!R,#7&!R,#16&!R][C&!R][C&!R][#8&!R,#7&!R,#16&!R][C&!R]
crown_3_3	[C&!R][#8&!R,#7&!R,#16&!R][C&!R][C&!R][C&!R][#8&!R,#7&!R,#16&!R][C&!R][C&!R][C&!R][#8&!R,#7&!R,#16&!R][C&!R]
cyanohydrins	[ND1&!R]#[CD2&!R][CX4&!R][OD2H1]
cyanophosphonate	P(OC)(OC)(=O)C#N
cyclic_crown_2_2	[c,C&R][#8&R,#7&R,#16&R][c,C&R][c,C&R][#8&R,#7&R,#16&R][c,C&R][c,C&R][#8&R,#7&R,#16&R][c,C&R]
cyclic_crown_2_3	[c,C&R][#8&R,#7&R,#16&R][c,C&R][c,C&R][#8&R,#7&R,#16&R][c,C&R][c,C&R][c,C&R][#8&R,#7&R,#16&R][c,C&R]
cyclic_crown_3_3	[c,C&R][#8&R,#7&R,#16&R][c,C&R][c,C&R][c,C&R][#8&R,#7&R,#16&R][c,C&R][c,C&R][c,C&R][#8&R,#7&R,#16&R][c,C&R]
diazonium	[c,C][N+&!R]#[N&!R]
ellipticine	Cc1c2[nH]c3ccccc3c2c(C)c2cnccc12
enamine	[$([#6][NX3&!RH1][#6&!RH1]=[#6&!RH1]),$([#6][NX3&!R]([#6])[#6&!RH1]=[#6&!RH1]),$([NX3&!RH2][#6&!RH1]=[#6&!RH1])]
epoxide	[OX2]1[CX4][CX4]1
fmoc	c12c(cccc1)c1c([CX4H1]2CO[CX3H0](=[OX1]))cccc1
formic_acid_esters	[CX3H1&!R](=[OX1])[OX2&!R][#6]
furocoumarines	[OX1&!R]=c1ccc2cc3ccoc3cc2o1
halo_alkene	[$([#6&!RH2]),$([#6&!RH1][C]),$([#6&!R]([C])[C])]=[$([#6&!RH2]),$([#6&!RH1][Br,F,I,Cl]),$([#6&!R]([C])[Br,F,I,Cl])]
halo_amine	[#6][NX3]([#6])[Cl,Br,I,F]
halopyrimidine	c1cnc([F,Cl,Br,I])nc1
hemiaminal	[OX2&!RH1][$([CX4&!RH2]),$([CX4&!RH1][#6]),$([CX4&!R]([#6])[#6])][$([NX3&!RH2]),$([NX3&!RH1][#6]),$([NX3&!R]([#6])[#6])]
hemiketal	[OX2&!RH1][$([CX4&!RH2]),$([CX4&!RH1][#6]),$([CX4&!R]([#6])[#6])][$([OX2&!RH1]),$([OX2&!R][#6])]
heteroatom_heteroatom_N_N	[$([#6][NH1;!R]-[NH1;!R][#6]),$([#6][NH1;!R]-[N;!R]([#6])[#6]),$([#6][N;!R]([#6])-[N;!R]([#6])[#6])]
heteroatom_heteroatom_N_S	[#6]-[$([NH1;!R]),$([N;!R][#6])]-[$([SX2;!R]-[#6]),$([SX4H1;!R](-[#6])-[#6]),$([SX4;!R](-[#6])(-[#6])-[#6])]

Table S2
substructures used to filter potentially toxic compounds

heteroatom_heteroatom_O_N	[$([#6][#8;!R][NH1;!R][#6]),$([#6][#8;!R][N;!R]([#6])[#6])]
heteroatom_heteroatom_S_O	[#6][SD2;!R]-[#8;!R][#6]
heteroatom_heteroatom_S_S	[#6][SD2;!R][SD2;!R][#6]
hydantoin	[OX1]=C1[NX3]C(=[OX1])C(=C)[NX3]1
hydralazine	n1ncc2ccccc2c1([NX3&!RH1][NX3&!RH2])
hydrazide	[NX3&!R][NX3&!R][CD3&!R]=[OX1]
hydrazine	[#6][NX3&!RH1][NX3&!RH2]
hydrazone	[CX3&!R]=[NX2&!R][NX3&!R]
hydroxamic_acid	[c][CX3&!R](=[OX1&!R])[NX3!RH1][OX2&!RH1]
hydroxylamine	[c;!$([#6]=[#8])][$([NX3&!RH1]),$([NX3&!R][#6])][OX2&!RH1]
imidazole	[#6]n1[cH1][cH1]n[cH1]1
imide	[C,c][C;!R](=O)[N;!R][C;!R](=O)[C,c]
imidoyl_halide	[$([CX3&!RH1]),$([CX3&!RH0][#6])](=[NX2&!R])[F,Br,I,Cl]
imine_C	[H,C][NX2&!R]=[CX3&!R]([C])[H,C]
imine_c_arom	[H,c][NX2&!R]=[CX3&!R]([c])[H,c]
isocyanate	[OD1&!R]=[CD2&!R]=[ND2&!R][#6]
isocyanide_isonitrile	[#6][#7+]#[#6-]
isothiocyanate	[SD1&!R]=[CD2&!R]=[ND2&!R][#6]
lawesson_reagent_derivative	[$([#6]P1(=S)SP(S1)(=S)[#6]),$([#6,Cl,F,#8]P(=[#16])([#16])[#16])]
maleimide	[CH]1=[CH]C(=[OX1])NC1=[OX1]
meta_aminophenol	c1([NX3&!R])cc([OX2&!RH1])ccc1
michael_acceptors	[C&!R]=[$([C&!R][CX3](=[O])[O]),$([C&!R][CX3](=[O])[C]),$([C&!R][CX3](=[O])[O][NX3]),$([C&!R][CX2]#[NX1]),$([C&!R][S](=[O])(=[O])[C]),$([C&!R][S](=[O])(=[O])[NX3]),$([C&!R][CX4]1[OX2][CX4]1)]
mustard_gas	[Cl]CCSCC[Cl]
nitramine	[$([NX3][NX3&!R](=O)=O),$([NX3][NX3+&!R](=O)[O-]),$([NX3][NX3+&!R]([O-])[O-])][!#8]
nitro	[$([NX3&!R](=O)=O),$([NX3+&!R](=O)[O-]),$([NX3+&!R]([O-])[O-])]
nitroso	[OD1&!R]=[$([ND2&!R][#6]),$([ND2&!R][NX3])]
ortho_aminophenol	c1([NX3&!R])c([OX2&!RH1])cccc1
ortho_aniline	[NX3H2&!R]c1[cH][cH]c(*)[cH][cH]1
ortho_hydroxyanilines	c1cccc([OX2&!RH1])c1[$([NX3&!RH2]),$([NX3&!RH1][#6]),$([NX3&!R]([#6])[#6])]
orthonitrophenyl_ester	[$(C(=O)Oc1c([$([NX3&!R](=O)=O),$([NX3+&!R](=O)[O-]),$([NX3+&!R]([O-])[O-])])cccc1),$(C(=O)Oc1ccccc1([$([NX3&!R](=O)=O),$([NX3+&!R](=O)[O-]),$([NX3+&!R]([O-])[O-])]))]
orthoquinone	[OD1&!R]=C1C(=[OD1&!R])C=CC=C1
oxime	[CX3]=[NX2][OX2H1]

Table S2
substructures used to filter potentially toxic compounds

oxonium	[O+]
para_aminophenol	c1([NX3&!R])[cH][cH]c([OX2&!RH1])[cH][cH]1
para_hydroquinone	[OH]c1ccc([OH])cc1
para_hydroxyanilines	c1cc([OX2&!RH1])ccc1[$([NX3&!RH2]),$([NX3&!RH1][#6]),$([NX3&!R]([#6])[#6])]
para_para_dihydroxybiphenyl	[OH]c1ccc(c2ccc([OH])cc2)cc1
para_para_dihydroxystilbene	[OH]c1ccc([#6&!R]=[#6&!R]c2ccc([OH])cc2)cc1
paranitrophenyl_ester	C(=O)Oc1ccc([$([NX3&!R](=O)=O),$([NX3+&!R](=O)[O-]),$([NX3+&!R]([O-])[O-])])cc1
pentafluorophen_ester	C(=O)([CH3])Oc1c(F)c(F)c(F)c(F)c1(F)
perhaloketone	[#6][CD3&!R](=[OD1&!R])[CD4&!R]([F,Br,I,Cl])([F,Br,I,Cl])[F,Br,I,Cl]
peroxide	[#6][#8][#8][#6,Cl,F,I,Br]
phenanthrene_het_N=C	c1cccc2c1c3c(nc2)cccc3
phenanthrene_het_C=N	c1cccc2c1c3c(cn2)cccc3
phenanthrene_het_N=N	c1cccc2c1c3c(nn2)cccc3
phenol	[OX2H]c1[cH1][cH1][cH1][cH1][cH1]1
phosphorane	[#6][OX2&!R][PX4&!R](=[OX1])([OX2&!R][#6])[OX2&!R][#6]
phosphonic_acid	[$(P([#6,Cl,I,Br,F,#1])([#6,Cl,I,Br,F,#1])([#6,Cl,I,Br,F,#1])[#6,Cl,I,Br,F,#1])]
polyenes	[#6]=[#6][#6]=[#6][#6]=[#6]
propiolactone	[OX1&!R]=C1OCC1
propiosultone	[OX1&!R]=[SX4]1(=[OX1&!R])OCCC1
quinone	O=[#6]1[#6]:,=[#6][#6](=O)[#6]:,=[#6]1
sulfonate_ester	[C][OX2&!R][$([SX4&!RH1]),$([SX4&!R][C])](=[OX1])(=[OX1])
sulfonic_acid	[#6][SX4&!R](=[OX1])(=[OX1])[$([OX2H]),$([OX1-])]
sulfonic_acid_ester	[#6][SX4&!R](=[OX1])(=[OX1])[$([OX2&!R][#6,#1]),$([OX1-])]
sulfonium	[S+;X3]([#6])([*])[*]
sulfonyl_cyanide	[#6][SD4&!R,SD4+2&!R](=[OD1&!R])(=[OD1&!R])[CD2&!R]#[NX1]
sulfonyl_halide	[#6][#16]([F,Cl,I,Br])(=[#8])=[#8]
sulfonyl_urea	[#6][NX3H1][CX3](=[OX1])[NX3H1][SX4](=[OX1])(=[OX1])[#6]
sulfoxide	[$([#16X3](=[OX1])([#6])[#6]),$([#16X3+]([OX1-])([#6])[#6])]
sulphanylamino	[#6][NX3H1][SX2H1]
thiazole	[$(s1nccc1),$(s1ccnc1)]
thiazolidinedione	S1C(=[OX1])[NX3H1]C(=[OX1])C1(=C)
thioacetal	[#6][CX4&!R]([#16&!R][#6])[#16&!R][#6]

Table S2
substructures used to filter potentially toxic compounds

thiocarbamate	[$([NX3H2&!R][CX3&!R](=[SX1&!R])[#8&!R][#6]),$([NX3H1&!R]([#6])[CX3&!R](=[SX1&!R])[#8&!R][#6]),$([NX3&!R]([#6])([#6])[CX3&!R](=[SX1&!R])[#8&!R][#6])]
thioester	[$([OX1&!R]=[c,C][SX2]),$([#8][c,C]=[#16])]
thioic_acid	[$([CX3]([OX2H1])=S),$([CX3]([SX2H1])=O),$([CX3]([SX2H1])=S)]
thioketone	[SD1&!R]=[#6]
thiol	[#6][SX2H1&!R]
thiophene	[$(s1c([#6])c([#6])cc1),$(s1cc([#6])c([#6])c1),$(s1c([!#1])[cH][cH][cH]1),$(s1[cH]c([!#1])[cH][cH]1),$(s1[cH][cH]c([!#1])[cH]1),$(s1[cH][cH][cH]c([!#1])1)]
toxoflavins	CN1C2=NC(=O)N(C(=O)C2=NC=N1)C
triacyloxime	C(=O)N(C(=O))OC(=O)
triazenes	[NX3&!R][NX2&!R]=[NX2&!R]
triflate	OS(=O)(=O)C(F)(F)F
triphenyl	c1ccccc1[CX4&!R](c1ccccc1)c1ccccc1