JOURNAL OF THE ROYAL SOCIETY Interface

2P2I$_{HUNTER}$: a tool for filtering orthosteric protein–protein interaction modulators via a dedicated support vector machine

Véronique Hamon, Raphael Bourgeas, Pierre Ducrot, Isabelle Theret, Laura Xuereb, Marie Jeanne Basse, Jean Michel Brunel, Sebastien Combes, Xavier Morelli and Philippe Roche

J. R. Soc. Interface 2014 **11**, 20130860, published 6 November 2013

JOURNAL OF THE ROYAL SOCIETY Interface

rsif.royalsocietypublishing.org

Research

Cite this article: Hamon V *et al*. 2014 2P2I$_{HUNTER}$: a tool for filtering orthosteric protein–protein interaction modulators via a dedicated support vector machine. *J. R. Soc. Interface* **11**: 20130860.
http://dx.doi.org/10.1098/rsif.2013.0860

Received: 20 September 2013
Accepted: 14 October 2013

Subject Areas:
computational biology, chemical biology, bioinformatics

Keywords:
focused chemical library, protein–protein interactions, small molecule inhibitors, drug design, support vector machine, filtering algorithm

Authors for correspondence:
Xavier Morelli
e-mail: xavier.morelli@inserm.fr
Philippe Roche
e-mail: philippe.roche@inserm.fr

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2013.0860 or via http://rsif.royalsocietypublishing.org.

Royal Society Publishing

2P2I$_{HUNTER}$: a tool for filtering orthosteric protein–protein interaction modulators via a dedicated support vector machine

Véronique Hamon[1], Raphael Bourgeas[1], Pierre Ducrot[2], Isabelle Theret[2], Laura Xuereb[2], Marie Jeanne Basse[1], Jean Michel Brunel[1], Sebastien Combes[1], Xavier Morelli[1] and Philippe Roche[1]

[1]Laboratory of integrative Structural and Chemical Biology (iSCB), Centre de Recherche en Cancérologie de Marseille (CRCM); CNRS UMR 7258, INSERM U 1068, Institut Paoli-Calmettes; and Aix-Marseille Universités, Marseille 13009, France
[2]Institut de Recherches Servier (IdRS), 11 rue des Moulineaux, Suresnes 92150, France

Over the last 10 years, protein–protein interactions (PPIs) have shown increasing potential as new therapeutic targets. As a consequence, PPIs are today the most screened target class in high-throughput screening (HTS). The development of broad chemical libraries dedicated to these particular targets is essential; however, the chemical space associated with this 'high-hanging fruit' is still under debate. Here, we analyse the properties of 40 non-redundant small molecules present in the 2P2I database (http://2p2idb.cnrs-mrs.fr/) to define a general profile of orthosteric inhibitors and propose an original protocol to filter general screening libraries using a support vector machine (SVM) with 11 standard DRAGON molecular descriptors. The filtering protocol has been validated using external datasets from PubChem BioAssay and results from in-house screening campaigns. This external blind validation demonstrated the ability of the SVM model to reduce the size of the filtered chemical library by eliminating up to 96% of the compounds as well as enhancing the proportion of active compounds by up to a factor of 8. We believe that the resulting chemical space identified in this paper will provide the scientific community with a concrete support to search for PPI inhibitors during HTS campaigns.

1. Introduction

Protein–protein interactions (PPIs) play a central role in many cellular processes, including signal transduction, cell adhesion, cell proliferation, growth, differentiation, viral self-assembly and programmed cell death (see reviews [1–3]). These interactions are involved in numerous pathways, including diseases, different stages of cancer development and host–pathogen interactions. Therefore, the modulation of these networks of transient interactions is a promising therapeutic strategy. PPIs are becoming more accepted as potential drug targets despite the long-held assumption that they were poorly druggable [4]. Modulators of these original interactions are likely to lead to the next generation of highly innovative drugs that should reach the market in the next decade. However, the *in silico* design of such compounds remains challenging [5–11]. PPI modulators (PPIMs) can be activators or inhibitors of the interaction, in this work the term 'modulators' only refers to PPI inhibitors.

The identification of hot spots at the interface of PPIs [12] has given a rationale for the possible disruption of protein–protein complexes with small molecules. Since then, there have been an increasing number of studies reporting the disruption of PPIs by small molecules [13–20]. Consequently, these successes have opened the way to the development of strategies to assess the druggability (or more appropriately ligandability) of protein–protein complexes [21–33].

A number of strategies have been used to conceive non-peptidic PPI inhibitors (for reviews, see [2,9,10,15,19]). Two of the main approaches involve the use of

small molecule chemical libraries through high-throughput screening (HTS) and fragment libraries using fragment-based drug discovery (FBDD) [34–39]. There are usually few small hydrophobic pockets at the protein–protein interface [21] that can each be filled with fragments, therefore FBDD is a very promising and efficient approach in the case of PPIs. However, one the major hurdle/challenge of this approach remains how to combine the low-affinity fragments to conceive high-affinity drug leads. In this research article, we only focus on the conception of small molecule chemical libraries dedicated to PPIs and we do not consider fragments.

Despite the progress in PPI drug discovery in the last decade, the success rate of finding hit compounds in HTS campaigns using small molecule compounds remains generally very low [40]. This low success rate suggests that most of the available chemical libraries are not appropriate for screening PPI targets. The poor suitability of commercial libraries demonstrates the need to design targeted chemical libraries that are dedicated to this particular chemical space [41]. These targeted libraries would accelerate and reduce the cost of screening campaigns by enhancing the number of hits while reducing the number of compounds tested which could help in bringing pharmaceutical companies *back on target* [42]. One way to achieve this goal is to design filtering algorithms for large chemical libraries that remove compounds that are unlikely to disrupt PPI interfaces while preserving a large number of potential disruptors in the selected subset. Several studies have focused on the chemical properties of known PPI inhibitors [1,43,44]. A general profile has been defined for these PPI inhibitors by compiling a collection of known PPI inhibitors and comparing them to other drugs. The authors found that PPI inhibitors are generally larger and more hydrophobic compared with other small molecule–protein complexes. They tend to form fewer hydrogen bonds and present more aromatic and hydrophobic interactions at the protein–ligand interface.

Decision tree methods have also been used to design PPI-inhibitor-focused libraries [45–49]. However, these studies focused on a set of validated drug-like PPI inhibitors, regardless of their modes of inhibition. Small molecule PPI inhibitors can be classified as orthosteric or allosteric modulators, depending upon their modes of interaction [50]. The former compete directly with hot spots at the interface [12], while the latter bind to a cavity away from the interface, which usually prevents the conformational changes necessary for binding to the protein partner. In addition, small molecules *in vivo* can prevent the formation of a protein–protein complex through non-direct mechanisms. To target the PPI inhibitors that directly interfere with the interface of protein–protein complexes, we focused on cases where the three-dimensional structures of both the protein–protein and protein–ligand complexes have been characterized. This work resulted in the freely accessible $2P2I_{DB}$ structural database (structural database dedicated to the inhibition of protein–protein interactions; http://2p2idb.cnrs-mrs.fr) [23].

In this research article, we analysed the properties of 40 non-redundant small molecules found in $2P2I_{DB}$ to define a general profile of orthosteric inhibitors. We propose an original protocol, $2P2I_{HUNTER}$, to filter general screening libraries with a machine-learning approach. The models were built using a support vector machine (SVM) with 11 standard DRAGON molecular descriptors. The best models were externally tested on the only two representative PPI bioassays from the publicly available PubChem Bioassay database for biological activities of small molecules (http://pubchem.ncbi.nlm.nih.gov/). This external blind validation demonstrated the ability of the SVM model to reduce the size of the filtered chemical library by eliminating up to 96% of the compounds as well as enhancing the proportion of active compounds by up to a factor of 8.

2. Results and discussion

2.1. Selection of the training dataset

We have developed the 2P2I structural database of all known protein–protein and protein–inhibitor complexes for which both three-dimensional structures are available [23]. Since the first release of the database in 2010, $2P2I_{DB}$ has been updated to v. 1.0 (http://2p2idb.cnrs-mrs.fr/2p2i_db.html).

The database currently contains 14 protein–protein complexes corresponding to various types of interfaces, including helix-based, beta-strand, mixed-folding (helix/beta strand) and loop-binding groove domains [51]. We have proposed a classification of these complexes into two main classes based on the type of interface; Class I, corresponding to protein–peptide complexes and Class II, corresponding to more globular protein–protein complexes contain [33]. Both classes are composed of seven protein–protein families. As $2P2I_{DB}$ grows, our goal is to classify protein–protein complexes based on the characteristics of their interface, to develop specific scoring functions and to associate a dedicated chemical library to each type of interface. However, data currently available do not allow this type of approach yet owing to the small number of protein–protein and protein–inhibitor complexes with a known three-dimensional structure. Although, the 14 complexes in $2P2I_{DB}$ possess different types of interfaces, they still exhibit some common features that are reflected in the properties of the small molecule inhibitors, for example the presence of small hydrophobic pockets, that are not linearly distributed. It is important to note that all other published studies dealing with the analysis of the properties of PPIMs and the conception of PPI-targeted chemical libraries have used all available PPI inhibitors regardless of their mode of action [20,44–48]. Especially, they did not differentiate between orthosteric and allosteric inhibitors which are likely to possess quite different properties.

There are currently 52 small molecule orthosteric inhibitors in the 2P2I database [51]. To guarantee the structural diversity of the compounds, the small molecule orthosteric inhibitors in $2P2I_{DB}$ were clustered with a Tanimoto similarity criterion of 0.8 (OptiSim algorithm implemented in the Tripos package) leading to 40 non-redundant molecules, which were used as the positive set in our learning approach (see the two-dimensional representation of molecules in the electronic supplementary material, figure S1). The selection of decoy molecules (presumed to be inactive against PPI targets) is more complicated owing to the impossibility of testing the compounds against all known PPI targets. The human 'interactome' is estimated to number between approximately 130 000 [52] and 650 000 [53] PPIs. To be definitively validated as a non-PPI inhibitor, a compound would have to be tested against a significant number of these protein–protein complexes. The aim of this work is to develop a filtration tool that is able to increase the hit rate in a screening campaign (i.e. to accelerate and reduce the cost of hit finding in a screening campaign). This aim can

Table 1. Parameters and overall internal performances of best optimized SVM model on the training dataset composed of 40 PPI inhibitors from the 2P2I database as a positive set and 1018 compounds as decoy.

	cost	sigma	accuracy	sensitivity	specificity	AUC of ROC
optimized SVM model	16	0.006	0.98	0.63	1.0	0.98

be achieved by either selecting the maximum number of true PPI inhibitors in a screening library or removing the maximum number of non-PPI inhibitors (or a combination of both). The next step involved the selection of a decoy dataset consisting of small drug-like molecules representative of commercial chemical libraries. We thus compared the molecular properties of the validated PPI inhibitors with several diverse chemical libraries (such as NCI Diversity set II, DiverSet from ChemBridge, LifeChemical and FDA approved) and we analysed the ability to discriminate between compounds from the positive training dataset and the different libraries. We selected NCI Diversity set II (1364 compounds) as a decoy because it allowed a better separation of the two datasets in *t*-test validations and preliminary SVM models (data not shown). As for the positive dataset, decoy compounds were clustered to remove redundancy with Tanimoto similarity comparisons of the UNITY fingerprints. Molecules with reactive groups and detergent compounds as well as those for which it was not possible to calculate all molecular descriptors were removed from the decoy dataset (see Material and methods). The final training dataset selected for the SVM approach contained 40 active (orthosteric PPI inhibitors) and 1018 inactive (non-PPI inhibitors) compounds (see electronic supplementary material, table S1). The chemical diversity and overlapping of the training dataset were assessed using a large set of two-dimensional BCUT metrics (see electronic supplementary material, figure S2). As expected, this analysis demonstrated the lack of redundancy within each dataset. In addition, compounds from the positive dataset covered most of the chemical space defined by the decoy.

2.2. $2P2I_{HUNTER}$: a filtering protocol to design protein–protein interaction libraries

To gain further insight into the selection of PPI inhibitors, SVM models were built and optimized based on the positive dataset and decoy described above. We used a classification SVM with a radial Gaussian basis (RGB) kernel and fivefold internal cross-validation. After comparing the spread of values and various molecular properties between the positive dataset and decoy, we performed *t*-tests to select standard molecular descriptors for the partial separation of the positive dataset and decoy. The final 11 molecular descriptors selected are listed in the electronic supplementary material, table S2. The SVM models were optimized using a receiver operating characteristic (ROC) curve value to counter the imbalance ratio of active/inactive (1/25) [54]. This choice (a compromise between specificity and sensitivity) was also guided by the characteristics of the desired filtration tool.

The entire dataset was randomly split into five folds; four folds were used for training, and the fifth fold was used for testing. This process was repeated 30 times, and the final average performance was calculated. The best selected model resulted in 98% accuracy, 100% specificity and 62.5% sensitivity (tables 1 and 2). To validate the robustness of the model, the PPI classification properties (active or not) of the 40 orthosteric PPI inhibitors from the positive dataset were randomly reordered (by scrambling). The average ROC AUC, obtained over 20 successive Y-scrambling runs, decreased from 0.96 with the actual activity labels to approximately 0.6 (the random threshold value was 0.5).

Table 2. Matrix of confusion of training set of best optimized SVM model.

actives		non-actives	
TP	FN	TN	FP
25	15	1017	1

2.3. PCA analyses of the training dataset

Separation of the positive and decoy compounds in the training dataset using the 11 molecular descriptors was assessed using PCA analyses (figure 1*a*). Unlike the BCUT metrics, the 11 DRAGON descriptors separated the true PPIMs and decoy molecules (figure 1*a* and electronic supplementary material, figure S1). The first two principal components accounted for more than 72% of the entire data variance. The most significant descriptors in the first PCA dimension corresponded to the number of hydrogen bond donors and acceptors (figure 1*b*). Topological polar surface area and molecular weight (MW) also had important contributions. The second dimension of the PCA analysis was dominated by ALogP, the level of unsaturation and the number of rings. Altogether, these principal components represented the four parameters that define rule of 4 ('Ro4') as defined previously [5,55]. MW, ALogP, number of rings and number of hydrogen bond acceptors all significantly contributed to the first or second principal component, as shown in figure 1*b*.

Most compounds from the decoy were clustered together (black ellipse in figure 1). The 40 orthosteric PPI inhibitors selected as the positive dataset occupied three cluster regions in the PCA representation. Interestingly, a correlation could be observed between compound clustering and the protein complexes targeted by the inhibitors. Eight PPI compounds were located in the main region occupied by the decoy. They mainly corresponded to low MW compounds targeting ZipA/FtsZ and HIV-1 integrase LEDGF/p75 complexes (average MW $= 370.9 \pm 44.4$ g mol^{-1}). It is worth pointing out that these inhibitors bind in a non-conventional way to their protein target. ZipA/FtsZ inhibitors exhibit low binding affinity constants (K_d 10–100 µM) and they appear to sit on the protein surface in the X-ray structure [21,44,56]. Their binding is achieved through hydrophobic contacts without any polar interaction between the ligand and the protein

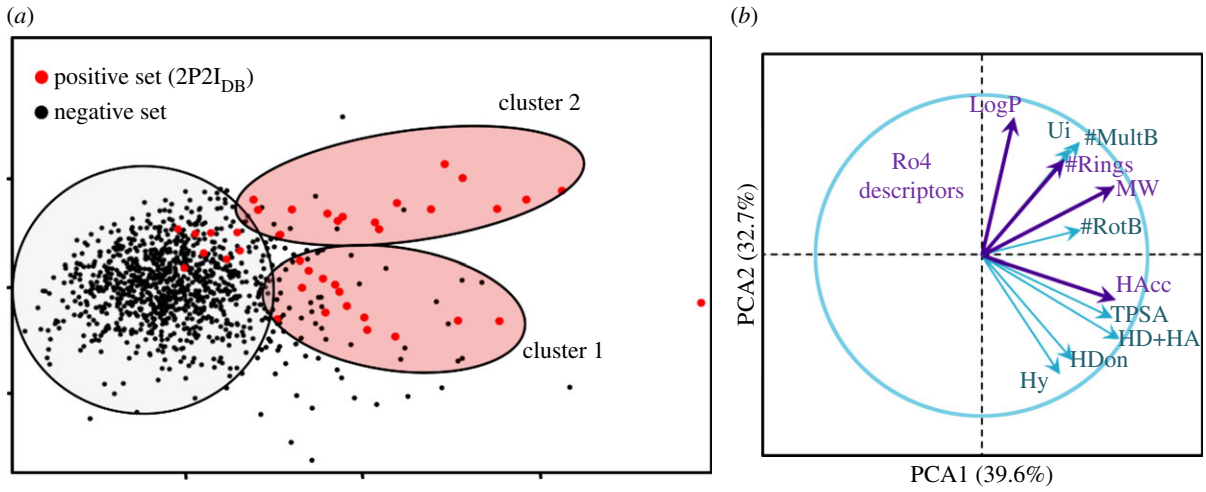


Figure 1. Two-dimensional PCA (*a*) and variable contributions (*b*) for the training dataset. (*a*) The 40 non-redundant orthosteric PPI inhibitors from 2P2I$_{DB}$ (positive set) and the 1018 compounds from Diversity are indicated as grey and black spheres, respectively. Most compounds from the decoy are clustered in the centre left of the representation (black ellipse). (*b*) Distribution and contribution of the 11 DRAGON molecular descriptors in the first two principal components. The first two principal axes represent more than 72% of the data variance. Hydrogen bonds (acceptor and donor), TPSA and MW play a major contribution in the first principal component, whereas ALogP, the level of unsaturation and the number of rings are the major contributors of the second principal component. Ro4 descriptors are highlighted; interestingly, #HBA and MW are among the major contributors in the first principal component and ALogP and the number of rings are important in the second dimension. (Online version in colour.)

target. PPI inhibitors usually bind few (4 ± 1) nearby small pockets [21], whereas HIV-1 integrase LEDGF/p75 inhibitors are targeting only one deep pocket at the integrase dimer interface and do not bind the centre of the interface as conventional orthosteric inhibitors. Each monomer subunit of the integrase forms approximately half of the binding pocket [57,58]. We decided not to remove these non-conventional compounds from the learning dataset although it would have improved the global statistics of the models.

The remaining orthosteric PPI compounds were mainly split between two similarly populated regions (red ellipses in figure 1) that showed little overlap with the decoy. The 14 compounds from cluster 1 (bottom red ellipse in figure 1) corresponded almost exclusively to inhibitors of the XIAP and IL2 families. These compounds are characterized by lower average values for ALogP (2.3 ± 1.0), number of multiple bonds (16.4 ± 3.2), increased average values for hydrogen bond donors (4.5 ± 1.6) and hydrophilic index (1.9 ± 1.1). The 17 compounds from the second cluster (top red ellipse in figure 1) are disruptors of the Bcl-XL, MDM2, MDM4, HPVE2/E1, TNF and TNFR1A families. They are characterized by higher average values for ALogP (6.1 ± 1.5), number of multiple bonds (28.4 ± 6.5), slightly higher MWs (654.0 ± 136.7 g mol^{-1}), number of rings (5.5 ± 0.9) and fewer hydrogen bond donors (1.5 ± 1.1). Only one of these 17 compounds was not 'Ro4' compliant. Finally, the XIAP/SMAC bivalent inhibitor (CZ3) lies outside these clusters mainly owing to its high MW (971.4 g mol^{-1}) and number of hydrogen bond acceptors [16]. It is worth mentioning that the monomeric compound (MW = 486.7) is also an inhibitor of the interaction, although slightly less efficient [59].

2.4. External validation of the support vector machine model: bioassay selection

The freely accessible PubChem BioAssay database (http://pubchem.ncbi.nlm.nih.gov) contains comprehensive information about small molecules and their biological activities [54,60,61]. It contains experimental descriptions and biological test results for more than 600 000 bioassays and is therefore a great resource for academic researchers in the fields of chemical biology, medicinal chemistry and chemoinformatics. We used an advanced query to retrieve relevant protein–protein bioassays to evaluate our SVM model with external data (see electronic supplementary material, figure S3). We first selected (using a keyword search) the series of bioassays grouped by AID summaries that correspond to PPIs. This search found 28 different bioassays (see electronic supplementary material, figure S3). Among these assays, we selected hits for which both primary and secondary (dose–response) assays were available (17 bioassays). At this stage, the bioassays were manually inspected and only those corresponding to PPIs were selected (five bioassays). Only *in vitro* bioassays were selected (three bioassays) to ensure that the inactivity was not owing to membrane barrier crossing, cytotoxicity effects or metabolic conversion of the compounds in the cell. The SVM model is a general model; therefore, we did not use bioassays dedicated to the development of target-specific inhibitors. For example, several bioassays correspond to the development of selective inhibitors of myeloid cell leukaemia sequence 1 (MCL1). In this series of bioassays, the authors selected compounds that are not active on other Bcl-2 proteins, for example Bcl-XL. Because our model was trained with modulators of the Bcl-XL family, it is not expected to perform well for these types of bioassays. Overall, two sets of bioassays were selected (AID 1645 and 1683; electronic supplementary material, figure S3 and table S3). AID 1496 and 1438 correspond to the primary and secondary HTS identification of compounds inhibiting the binding between the RUNX1 Runt domain and the Core-binding factor, beta subunit. AID 1531 (primary screen) and 1892/1897 (secondary screens) correspond to the HTS Assay for modulators of MEK Kinase PB1 Domain interactions via MEKK5.

2.5. Blind validation of 2P2I$_{HUNTER}$ with selected bioassays

The ability to enhance hit rates in screening campaigns with the 'Ro4' and the best SVM model as filtering tools was tested

(a)

SVM				
	actives		non-actives	
	TP	FN	TN	FP
AID1496	4	41	207 048	8 547
AID1531	35	62	276 429	12 913

	accuracy	sensitivity	specificity
AID1496	0.96	0.09	0.96
AID1531	0.96	0.36	0.96

(b)

Ro4				
	actives		non-actives	
	TP	FN	TN	FP
AID1496	24	21	171 642	43 989
AID1531	53	44	231 043	58 335

	accuracy	sensitivity	specificity
AID1496	0.80	0.53	0.80
AID1531	0.80	0.55	0.80

(c)

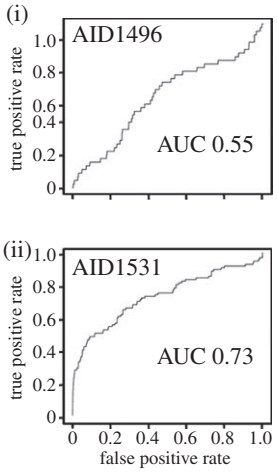


Figure 2. Confusion matrices, performances and ROC curves of the SVM model for the two external validation datasets. (*a*) Repartition of the molecules according to the SVM model prediction (TP, true positive; FN, false negative; TN, true negative; FP, false positive). (*b*) Overall performances of the 'Ro4' model in terms of accuracy, sensitivity and specificity. (*c*) ROC curves and AUC are shown for AID 1496 (top) and 1531 (bottom).

Table 3. Overall performances of 'Ro4' and SVM models for the two available PPI bioassays. EF, enrichment factor.

AID	tested	active	hit rate[a] (%)	model	selected[b]	TP[c]	hit rate[d] (%)	EF	% selected[e]
1531	289 475	97	0.033	'Ro4'	58 388	53	0.09	2.7	20
				SVM	12 948	35	0.27	8.0	4
				'Ro4' + SVM	8436	24	0.28	8.5	3
1496	215 676	45	0.021	'Ro4'	44 009	24	0.055	2.6	20
				SVM	8551	4	0.047	2.2	4
				'Ro4' + SVM	5900	4	0.068	3.2	3

[a]Experimental hit rate.
[b]Number of molecules selected by the model at a probability threshold of 0.5.
[c]Number of true-positive compounds selected.
[d]Hit rate of the model.
[e]Percentage of molecules selected from the initial library.

on the two HTS compound bioassays selected from PubChem. The overall performances for the different models are summarized in table 3. Confusion matrices, ROC curves and overall performances (in terms of accuracy, sensitivity and specificity) are shown in figure 2. To quantify the performance of the SVM approach, we used the enrichment factor metric, which compares the performance of the model to the performance expected if the compounds were randomly selected. The enrichment factor is given by the ratio between the hit rate obtained with the entire chemical library and the hit rate for the filtered library. Based on this metric, we observed a significant improvement of the HTS performances for both bioassays.

With both external validation datasets, the application of the 'Ro4' allows approximately 20% of the compounds from the starting chemical library to be selected, representing a significantly improved hit rate (enrichment factor of approx. 3). Therefore, the 'Ro4' constitutes a simple and fast method of designing chemical libraries enriched in PPI inhibitors. The SVM model showed high accuracy and specificity with values of 0.96 for both validation datasets, compared to 0.8 for the 'Ro4' filtering (figure 2). Although these high values are probably related to the relatively imbalanced ratio of active and inactive compounds in the training dataset, they indicate that the SVM model can be used as a highly efficient tool for removing non-PPI compounds from large screening libraries. The true positive rate (as measured by the sensitivity) was lower for the SVM, with values of 9% and 36% for AID 1496 and 1531, respectively. The 'Ro4' exhibited a higher sensitivity than the SVM model (53% and 55% for AID 1496 and 1531, respectively), which shows that it is able to retrieve more true active compounds in absolute numbers. However, the SVM model is much more stringent. This model was able to eliminate a large number of non-PPI compounds, leading to an increase in the hit rate of a factor up to 8. It is important to point out that none of the active molecules were present in the training dataset. The distributions of MW, number of rings (#Rings) and number of hydrogen bond acceptors (#HBA) in the positive dataset and the two validation datasets were compared to further analyse the performance of the SVM

tool (see electronic supplementary material, figure S4). As expected, the active molecules with low MW were poorly predicted by the SVM model. It should be noted that although molecules in validation bioassays are classified in a binary manner as active or inactive, the activity described in these bioassays is covering a large binding affinity range (from 0.14 to more than 50 μM). A significant amount of the active molecules in both bioassays could almost be considered as fragments (MW 300 g mol^{-1}) and are outside the scope of this SVM model. A similar distribution was observed for the number of rings in the active molecules and those selected by the SVM model for both validation bioassays. However, in the case of AID1531, a series of 12 molecules were poorly predicted owing to their high MWs (average MW = 1063.1 ± 42.0) and high number of rings (average #Rings = 8.1 ± 1.2). Finally, the distribution of #HBA for most active molecules in AID1531 was downshifted to lower values compared with the distribution of selected molecules which could account for the non-selection of some active compounds.

Interestingly, most true positive compounds selected by the SVM model followed the 'Ro4' (69% in the case of AID1531 and 100% in the case of AID1496). This finding prompted us to combine the two approaches to refine the filtering process. We applied the 'Ro4' filtering tool to the compounds selected by the SVM model. The results in table 3 demonstrated a slight increase in the hit rate compared with the SVM model owing to a smaller number of selected compounds and almost the same number of true positive PPI inhibitors predicted. Molecules selected by the SVM model for both validation bioassays overlapped with the region defined by the positive dataset as shown by PCA analysis (see electronic supplementary material, figure S5).

AID1531 led to better overall enrichment factors, indicating that the active compounds in this bioassay were closer to the 2P2I chemical space, as observed in the representation of the two principal components (see electronic supplementary material, figure S5). X-ray structures were available for the unbound Runx1/Runt domain (PDB code 1EAN) and the core-binding factor beta (PDB code 2JHB), which corresponds to the protein targets in bioassay AID1496. However, no structural information was available for the protein targets in AID1531 or the protein–protein complexes. Therefore, we were not able to compare the MEKKK2/MEKK5 and RuntX/CBFb complexes to those present in $2P2I_{DB}$ to determine whether the difference between the two bioassays was related to the structure of the protein–protein interfaces.

2.6. Chemical diversity of selected molecules

The structural diversity of the molecules selected with the SVM model was checked (using a Tanimoto index of 0.8), and on average they did not share the same scaffold. In the case of AID1531, the 35 true-positive compounds corresponded to 26 different classes after the similarity clustering (comparable to the diversity of the active molecules); whereas all four true-positive compounds in AID1496 were dissimilar, indicating that this SVM protocol did not retrieve a series of compounds from the same family.

2.7. Blind validation with in-house PPI bioassays

The overall performance of $2P2I_{HUNTER}$ was also assessed using results from in-house PPI bioassays. Prior to this study, a high-throughput screen against a PPI target was performed by the industrial partner involved in this study to identify PPIMs using a FRET assay in the primary screening and IC50 for the secondary dose–response confirmatory screening, which led to a low success rate of 0.004% (table 4). Application of the SVM protocol to the entire collection of compounds used in the HTS campaign selected 8.5% of the molecules, corresponding to the filtered library (tables 4 and 5). A significant improvement was observed in the hit rate for this resulting filtered chemical library (0.025%), which corresponds to an enrichment factor of 5.9. It is especially noteworthy that 50% of the active compounds found in the HTS campaign were present in the filtered library. In a parallel approach, results from a large series of in-house protein–protein and protein–peptide-related primary or dose–response screenings were pooled together, which allowed us to search for PPIMs in the filtered library (table 5). In particular, we looked for evidence of activity in PPI modulation bioassays for each compound in the filtered library. We were not able to calculate enrichment factors because the bioassays were not conducted on the entire in-house non-filtered collection of compounds for comparison. However, we estimated the percentage of true PPIMs in the filtered library (table 5), which ranged from 0.6 to 4.1%. Not all of the compounds in the filtered library have been tested in the various bioassays, so the percentages are underestimated. A higher percentage of predicted PPIMs was observed for protein–peptide-related bioassays, which could indicate that the SVM protocol performs better for these families of targets.

Table 4. Overall performances of SVM model for in-house bioassays. High-throughput screening of PPI target. EF, enrichment factor.

assay	hit rate[a] (%)	SVM[b] (%)	hit rate SVM[c] (%)	EF
HTS	0.004	8.5	0.025	5.9

[a]Experimental hit rate.
[b]Percentage of molecules selected by the model at a probability threshold of 0.5.
[c]Hit rate of the model.

Table 5. Overall performances of SVM model for in-house bioassays. Cumulative percentage of true PPIMs found on a set of in-house PPI targets.

assay	PPI modulators (%)
protein/peptide (primary screen)	1.6
protein/protein (primary screen)	0.6
protein/peptide (dose–response)	4.1
protein/protein (dose–response)	1.6

2.8. Applicability domain of $2P2I_{HUNTER}$ support vector machine filtration tool

The SVM filtration tool was applied to a set of 25 chemical libraries representing more the 8.3 M compounds and its applicability domain was discussed in details previously [55]. Interestingly, compared with all other chemical libraries tested, the percentage of selected compounds was higher for the PPI library from ChemDiv (figure 3*a*). Selected compounds

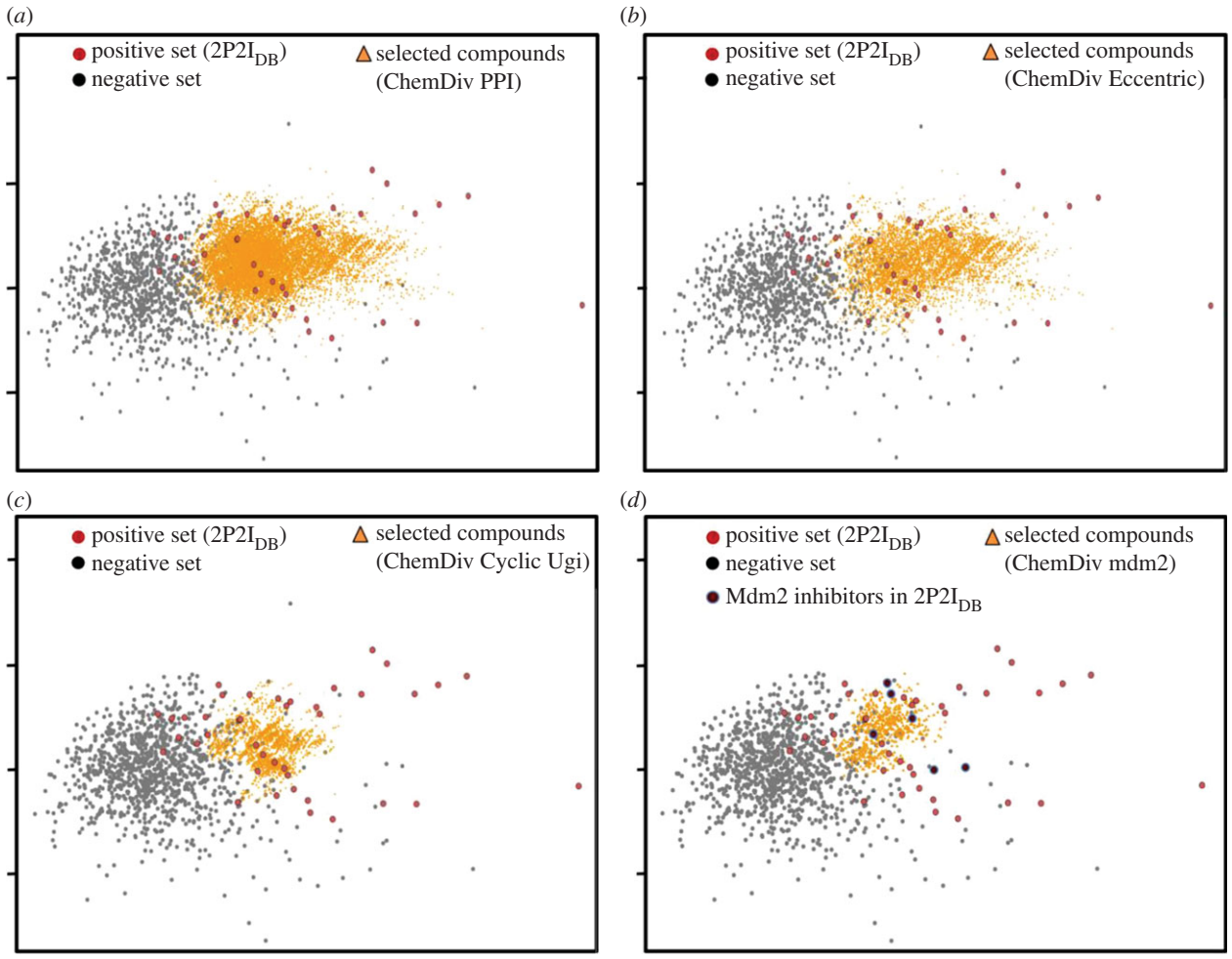


Figure 3. Two-dimensional representation of ChemDiv PPI subsets in the first two principal components of the training set chemical space. (*a*) Total set of 123 001 compounds, (*b*) Eccentric, (*c*) Cyclic Ugi and (*d*) mdm2-focused subsets. Mdm2 inhibitors found in the 2P2I database are indicated as unfilled darker circles. (Online version in colour.)

from the Eccentric subset of ChemDiv (83.1%) cover almost the entire PPI chemical space described by our SVM model (figure 3*b*). This subset contains approximately 7000 compounds that are grouped together based on their original scaffolds outside of the field of (hetero)aromatic systems and their low toxicity [62]. The compounds from the Cyclic Ugi subset overlap with 50% of the PPI inhibitors in 2P2I database (figure 3*c*). The compounds selected from the mdm2-focused subset cover a smaller chemical space that encompasses most of the mdm2/p53 inhibitors found in 2P2I$_{DB}$ (figure 3*d*). Thus, the 2P2I chemical space that we identified is able to select new and/or unusual aromatic scaffolds that exhibit low toxicity and lie outside the field of extensively used benzenoid and heteroaromatic ring systems [62].

The distributions of MW, #Ring, #HBA have been compared for ChemDiv PPI, Eccentric and three other representative chemical libraries (see electronic supplementary material, figure S6). Although, libraries rich in high MW compounds also corresponded to libraries with a higher percentage of selected molecules by the SVM tool, the correlation was not linear. There was a clear shift towards high MW in the case of the Eccentric subset for which 94.5% of the molecules have an MW $\geq$ 400 g mol^{-1} and 83.1% of the molecules are selected. ChemDivPPI and IBScreen libraries contain 64.1% and 48.8% of molecules with MW $\geq$ 400 g mol^{-1}, respectively. However, the number of selected molecules for these two libraries is similar and much lower than that for the Eccentric subset (10.8% and 11.2%, respectively). Not surprisingly, libraries rich in low MW compounds also corresponded to a lower percentage of selected molecules. Maybridge and Otava collections of compounds contain 19.4% and 9.6% compounds with MW $\geq$ 400, respectively, however a similar percentage of molecules is selected by the SVM model (1.1% and 1.2%, respectively). Concerning the number of rings, a weaker correlation was observed. Overall, libraries containing molecules with a higher number of rings correspond to the libraries with high percentages of selected compounds by the SVM tool. However, when comparing among themselves libraries containing molecules with a high number of rings such as Eccentric, ChemDivPPI and IBScreen, no correlation was observed between the high number of rings and the percentage of selected molecules (see electronic supplementary material, figure S6). Finally, the distribution of HBA seemed slightly shifted towards lower values for libraries with lower percentage of selected molecules.

2.9. Conclusion

In the postgenomic era, the identification of complete networks of PPIs within a cell has contributed to major breakthroughs in understanding biological pathways, host–pathogen interactions and cancer development. This work has also led to the identification of novel therapeutic drug targets [63]. PPIs represent a promising new class of attractive therapeutic targets and correspond to the most screened

target in HTS today (93% screening) followed by kinases (82% screening) according to a recent survey [64]. However, this class of targets is still considered to be extremely difficult for targeting by small molecules, owing to the structural characteristics of the interfaces.

Specific strategies are needed to tackle this particularly challenging class of drug targets. Successes in drug discovery developments against PPI targets face two major issues, druggability assessment (or target selection) and adequacy of the chemical libraries used for screening. Our strategy to address these issues was to focus on orthosteric PPIMs. Therefore, we only considered PPIs for which structural information was available to validate that the inhibitor was interacting at the interface. We have recently developed the hand-curated structural 2P2I database by collecting information about protein–protein interfaces for which both the protein–protein and protein–inhibitor complexes have been structurally characterized. We identified key descriptors of PPIs with known inhibitors [23]. Although this database is currently rather small, a recent update has resulted in an increase of almost 30% of small molecules orthosteric inhibitors. As the database is growing, our goal is to cluster protein–protein complexes into different structural families based on the properties of the interface and to associate a dedicated chemical library with each type of interface.

Here, we report the analysis and characterization of the chemical space of PPI orthosteric inhibitors present in the 2P2I database (http://2p2idb.cnrs-mrs.fr) to provide tools to build focused chemical libraries dedicated to PPI targets from large screening libraries. PPI inhibitors are sometimes not considered to be potential drug-like compounds owing to their high MWs and hydrophobicity. However, more than 60% of the active compounds in $2P2I_{DB}$ are compliant with Lipinski 'rule-of-five' [65] and 65% follow Veber rules for bioavability [66] demonstrating that they could possess ADME properties that are compatible with further developments as oral drugs and that rules defined in this article to characterize PPI inhibitors and drug-likeness are not mutually exclusive.

We have developed a general protocol to filter chemical libraries using SVM approaches. This $2P2I_{HUNTER}$ filtration tool has been validated using external bioassays from PubChem and in-house screening results. Our tool demonstrated the potential to significantly increase success rates, leading to enrichment factors of up to 8. The protocol was applied to a set of 25 commercial libraries from the major providers [55]. In all cases, the tool dramatically decreased the size of the original libraries compared with the resulting focused libraries (on average, 94% of the compounds were removed from the original libraries). However, more than 80% of the 7076 compounds from the ChemDiv Eccentric PPI library were selected as putative PPIMs by the $2P2I_{HUNTER}$ filtration tool (see [55] for a detailed analysis of the chemical space).

In general, only a limited number of molecules can be tested in screening campaigns in academic institutions (and increasingly in pharmaceutical companies as well). The $2P2I_{HUNTER}$ protocol has been used to build in-house chemical libraries of small size [55] that will be tested on PPI targets, covering a wide range of biological and chemical spaces including the well-studied mdm2/p53. $2P2I_{HUNTER}$ represents a useful tool for both academics and pharmaceutical companies to expand the scope of chemical libraries dedicated to protein–protein targets and to enhance hit rates in high-throughput experiments, thereby reducing the cost of screening campaigns.

3. Material and methods

3.1. Training dataset collection and preparation

3.1.1. Ligand preparation

A standard ligand preparation protocol was applied using Chemaxon tools (JChem 5.10, 2012, http://www.chemaxon.com). Briefly, molecules were first checked for errors in valence, coordination, aromaticity and covalently bound counter ions. Molecules were then standardized as follows: the largest fragment was kept, explicit hydrogens were removed, molecules were dearomatized, aromatized and neutralized, explicit hydrogens were added and finally the structures were cleaned in two dimensions. Major species at physiological pH 7.4 were then determined using cxcalc module.

3.1.2. Training set preparation

In order to reduce structural redundancy, the 52 inhibitors present in $2P2I_{DB}$ were submitted to a diversity selection using the OptiSim module from Tripos that compares UNITY fingerprints of compounds. A Tanimoto coefficient of 80% was chosen as the similarity threshold, leading to a final active training dataset of 40 compounds. The same OptiSim procedure was applied to select a diverse set of 1089 compounds from the NCI Diversity set II (1364 compounds). Compounds with reactive groups and detergent compounds were removed manually to provide a final decoy dataset composed of 1018 molecules. The positive training set and decoy are available at http://2p2idb.cnrs-mrs.fr/2p2i_hunter.html.

3.1.3. Molecular descriptors

The 10 molecular descriptors used for machine-learning process (MW; number of multiple bonds; number of rings; number of rotatable bonds; number of donor atoms for H-bonds; number of acceptor atoms for H-bonds; unsaturation count; hydrophilic factor; topological polar surface area and ALogP) were computed with DRAGON v. 6 (http://www.talete.mi.it) on the charged compounds prepared with ChemAxon as described above. The 11th descriptor corresponded to the sum of hydrogen bond donors and acceptors. Molecules that generated errors in descriptors computation were discarded.

3.1.4. BCUT descriptors

The DiverseSolution module from Tripos was used to compute BCUT-based metrics [67]. The program performs a partitioning of the BCUT chemical space and determines which molecules from the studied set occupy each mapped cell. Cells coordinates were retrieved and submitted to a principal components analysis.

3.1.5. 'Rule of 4' definition

The 'Ro4' is a rule of thumb that was derived from the distribution of molecular properties of PPI inhibitors in the 2P2I database [5,55]. Small molecules follow the 'Ro4' if they obey three of the following properties: MW $\geq$ 400 Da, ALogP $\geq$ 4, number of rings more than or equal to 4 and number of hydrogen bond acceptors more than or equal to 4. It is worth pointing out that contrary to other rules, compounds need to exceed criteria to be compliant.

3.2. Bioassay selection

Bioassays were retrieved from the PubChem Bioassay website that provides searchable descriptions of more than 600 000 bioassays (http://www.ncbi.nlm.nih.gov/pcassay). The detailed query is shown in the electronic supplementary material, figure S3. Briefly, the keyword 'protein protein' was searched in AID corresponding to bioassay type summary. Then bioassays with both confirmatory and primary screenings were selected. Finally, bioassays were grouped by type summary. The 17 remaining bioassays were checked manually. Bioassays corresponding to enzymatic assays or cell-based were discarded leaving only two AID.

The name of the targets and the detailed description of assay methods for the blind validation with the in-house PPI bioassays could not be disclosed owing to confidentiality issues regarding the pharmaceutical partner.

3.3. Support vector machine model construction and validation

SVM algorithm implemented in the statistical software package R was used for model training [68]. Original data were mapped through a kernel function on to a higher dimensional space where the two sets are more easily separable with a linear classifier. The Caret library in R environment was used to perform SVM modelling. The positive or negative activity of our training set was typed as a factor variable to specify the classification mode of the problem. The RGB method was chosen as kernel function. As a preprocessing treatment of data, each descriptor is centred around a mean of 0 and scaled to a variance equal to 1. The training is performed through a fivefold cross-validation procedure repeated 30 times with a random selection of the training and test sets at each time. The final optimal model was selected according to the average best performance of the ROC AUC (area under curve) statistical metric. The model training in RGB method consists of tuning two hyperparameters cost C and scaling function sigma along a grid of candidate values. The best final model was found with $C = 16$ and $\sigma = 0.006$.

3.4. Enrichment factor

The enrichment factor which corresponds to the ratio of hit rates after and before filtration with the SVM model is given by the following formula:

$$\mathrm{EF} = \frac{\mathrm{Hit}_{\mathrm{selected}}/N_{\mathrm{selected}}}{\mathrm{Hit}_{\mathrm{total}}/N_{\mathrm{total}}} = \frac{\mathrm{Hit\,Rate}_{\mathrm{selected}}}{\mathrm{Hit\,Rate}_{\mathrm{total}}},$$

where EF is the enrichment factor obtained with the SVM filtered chemical library, $\mathrm{Hit}_{\mathrm{selected}}$ is the number of active compounds in the selected library, $\mathrm{Hit}_{\mathrm{total}}$ is the total number of active molecules in the library, N_{selected} is the number of compounds selected by the SVM model and N_{total} is the total number of compounds in the original (non-filtered) library.

3.5. Model validation

Statistical metrics used to assess the quality of the model are defined below.

$$\mathrm{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}}, \quad \mathrm{specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \quad \text{and} \quad \mathrm{sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

where TP, true positive; TN, true negative; FN, false negative; FP, false positive.

The accuracy measures the percentage of correctly predicted molecules altogether. Specificity and sensitivity reflect the ability to correctly predict inactive and active compounds, respectively.

3.6. Y-scrambling

In order to assess the reliability of the model, activity labels of the training set were randomly re-ordered by keeping positive/negative ratio. The two parameters cost and sigma remained unchanged and the same training configuration was applied, fivefold cross-validation repeated 30 times.

3.7. PCA analyses

PCA was performed using the library FactoMineR in R package [69]. All the 11 molecular descriptors were used to derive the principle components that defined the 2P2I chemical space, calculated on the whole training dataset. For the PCA of the external validation datasets and commercial libraries, properties of molecules were projected onto the first two principal components of the training set.

References

1. Wells J, McClendon C. 2007 Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* **450**, 1001–1009. (doi:10.1038/nature06526)
2. Fry DC. 2012 Small-molecule inhibitors of protein–protein interactions: how to mimic a protein partner. *Curr. Pharm. Des.* **18**, 4679–4684. (doi:10.2174/138161212802651634)
3. Morelli X, Hupp T. 2012 Searching for the Holy Grail; protein–protein interaction analysis and modulation. *EMBO Rep.* **13**, 877–879. (doi:10.1038/embor.2012.137)
4. Mullard A. 2012 Protein–protein interaction inhibitors get into the groove. *Nat. Rev. Drug Discov.* **11**, 173–175. (doi:10.1038/nrd3680)
5. Morelli X, Bourgeas R, Roche P. 2011 Chemical and structural lessons from recent successes in protein–protein interaction inhibition (2P2I). *Curr. Opin. Chem. Biol.* **15**, 475–481. (doi:10.1016/j.cbpa.2011.05.024)
6. Wilson CG, Arkin MR. 2011 Small-molecule inhibitors of IL-2/IL-2R: lessons learned and applied. *Curr. Top Microbiol. Immunol.* **348**, 25–59. (doi:10.1007/82_2010_93)
7. Khoury K, Dömling A. 2012 P53 mdm2 inhibitors. *Curr. Pharm. Des.* **18**, 4668–4678. (doi:10.2174/138161212802651580)
8. Cesa LC, Patury S, Komiyama T, Ahmad A, Zuiderweg ER, Gestwicki JE. 2013 Inhibitors of difficult protein–protein interactions identified by high-throughput screening of multiprotein complexes. *ACS Chem. Biol.* **8**, 1988–1997 (doi:10.1021/cb400356m)
9. Arkin MR, Whitty A. 2009 The road less traveled: modulating signal transduction enzymes by inhibiting their protein–protein interactions. *Curr. Opin. Chem. Biol.* **13**, 284–290. (doi:10.1016/j.cbpa.2009.05.125)
10. Azzarito V, Long K, Murphy NS, Wilson AJ. 2013 Inhibition of α-helix-mediated protein–protein interactions using designed molecules. *Nat. Chem.* **5**, 161–173. (doi:10.1038/nchem.1568)
11. Fletcher S, Hamilton AD. 2006 Targeting protein–protein interactions by rational design: mimicry of protein surfaces. *J. R. Soc. Interface* **3**, 215–233. (doi:10.1098/rsif.2006.0115)
12. Wells JA. 1995 Structural and functional epitopes in the growth hormone receptor complex. *Nat. Biotechnol.* **13**, 647–651. (doi:10.1038/nbt0795-647)
13. Blazer LL, Neubig RR. 2009 Small molecule protein–protein interaction inhibitors as CNS therapeutic agents: current progress and future hurdles. *Neuropsychopharmacology* **34**, 126–141. (doi:10.1038/npp.2008.151)

14. Azmi AS, Wang Z, Philip PA, Mohammad RM, Sarkar FH. 2011 Emerging Bcl-2 inhibitors for the treatment of cancer. *Expert Opin. Emerg. Drugs* **16**, 59–70. (doi:10.1517/14728214.2010.515210)
15. Fry D. 2011 Small-molecule inhibitors of protein–protein interactions. In *Nutley* (eds L Vassilev, D Fry), pp. 184. New Jersey: Springer.
16. Koes DR, Camacho CJ. 2011 Small-molecule inhibitor starting points learned from protein–protein interaction inhibitor structure. *Bioinformatics* **28**, 784–791. (doi:10.1093/bioinformatics/btr717)
17. Vu BT, Vassilev L. 2011 Small-molecule inhibitors of the p53-MDM2 interaction. *Curr. Top. Microbiol. Immunol.* **348**, 151–172. (doi:10.1007/82_2010_110)
18. Sakuma Y, Tsunezumi J, Nakamura Y, Yoshihara M, Matsukuma S, Koizume S, Miyagi Y. 2011 ABT-263, a Bcl-2 inhibitor, enhances the susceptibility of lung adenocarcinoma cells treated with Src inhibitors to anoikis. *Oncol. Rep.* **25**, 661–667. (doi:10.3892/or.2010.1123)
19. Higueruelo AP, Jubb H, Blundell TL. 2013 Protein–protein interactions as druggable targets: recent technological advances. *Curr. Opin. Pharmacol.* **13**, 791–796. (doi:10.1016/j.coph.2013.05.009)
20. Higueruelo AP, Jubb H, Blundell TL. 2013 TIMBAL v2: update of a database holding small molecules modulating protein–protein interactions. *Database* **2013**, bat039. (doi:10.1093/database/bat039)
21. Fuller JC, Burgoyne NJ, Jackson RM. 2009 Predicting druggable binding sites at the protein–protein interface. *Drug Discov. Today* **14**, 155–161. (doi:10.1016/j.drudis.2008.10.009)
22. Sugaya N, Ikeda K. 2009 Assessing the druggability of protein–protein interactions by a supervised machine-learning method. *BMC Bioinform.* **10**, 263. (doi:10.1186/1471-2105-10-263)
23. Bourgeas R, Basse M-J, Morelli X, Roche P. 2010 Atomic analysis of protein–protein interfaces with known inhibitors: the 2P2I database. *PLoS ONE* **5**, e9598. (doi:10.1371/journal.pone.0009598)
24. Geppert T, Hoy B, Wessler S, Schneider G. 2011 Context-based identification of protein–protein interfaces and 'hot-spot' residues. *Chem. Biol.* **18**, 344–353. (doi:10.1016/j.chembiol.2011.01.005)
25. Wanner J, Fry DC, Peng Z, Roberts J. 2011 Druggability assessment of protein–protein interfaces. *Future Med. Chem.* **3**, 2021–2038. (doi:10.4155/fmc.11.156)
26. Sugaya N, Furuya T. 2011 Dr. PIAS: an integrative system for assessing the druggability of protein–protein interactions. *BMC Bioinform.* **12**, 50. (doi:10.1186/1471-2105-12-50)
27. Kozakov D *et al*. 2011 Structural conservation of druggable hot spots in protein–protein interfaces. *Proc. Natl Acad. Sci. USA* **108**, 13 528–13 533. (doi:10.1073/pnas.1101835108)
28. Koes DR, Camacho CJ. 2012 PocketQuery: protein–protein interaction inhibitor starting points from protein–protein interaction structure. *Nucleic Acids Res.* **40**, W387–W392. (doi:10.1093/nar/gks336)
29. Metz A, Ciglia E, Gohlke H. 2012 Modulating protein–protein interactions: from structural determinants of binding to druggability prediction to application. *Curr. Pharm. Des.* **18**, 4630–4647. (doi:10.2174/138161212802651553)
30. Thangudu RR, Bryant SH, Panchenko AR, Madej T. 2012 Modulating protein–protein interactions with small molecules: the importance of binding hotspots. *J. Mol. Biol.* **415**, 443–453. (doi:10.1016/j.jmb.2011.12.026)
31. Ulucan O, Eyrisch S, Helms V. 2012 Druggability of dynamic protein–protein interfaces. *Curr. Pharm. Des.* **18**, 4599–4606. (doi:10.2174/138161212802651652)
32. Johnson DK, Karanicolas J. 2013 Druggable protein interaction sites are more predisposed to surface pocket formation than the rest of the protein surface. *PLoS Comput. Biol.* **9**, e1002951. (doi:10.1371/journal.pcbi.1002951)
33. Hamon V, Morelli X. 2013 Druggability of protein–protein interactions. In *Understanding and exploiting protein–protein interactions as drug targets* (ed. G. Zinzalla), pp. 18–31. London, UK: Future Science Ltd. (doi:10.4155/9781909453463)
34. Hajduk PJ, Huth JR, Tse C. 2005 Predicting protein druggability. *Drug Discov. Today* **10**, 1675–1682. (doi:10.1016/S1359-6446(05)03624-X)
35. Edfeldt FN, Folmer RH, Breeze AL. 2011 Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discov. Today* **16**, 284–287. (doi:10.1016/j.drudis.2011.02.002)
36. Valkov E, Sharpe T, Marsh M, Greive S, Hyvönen M. 2012 Targeting protein–protein interactions and fragment-based drug discovery. *Top. Curr. Chem.* **317**, 145–179.
37. Winter A, Higueruelo AP, Marsh M, Sigurdardottir A, Pitt WR, Blundell TL. 2012 Biophysical and computational fragment-based approaches to targeting protein–protein interactions: applications in structure-guided drug discovery. *Q. Rev. Biophys.* **45**, 1–44. (doi:10.1017/S0033583512000108)
38. Scott DE, Ehebauer MT, Pukala T, Marsh M, Blundell TL, Venkitaraman AR, Abell C, Hyvönen M. 2013 Using a fragment-based approach to target protein–protein interactions. *Chembiochem* **14**, 332–342. (doi:10.1002/cbic.201200521)
39. Wu B, Zhang Z, Noberini R, Barile E, Giulianotti M, Pinilla C, Houghten R, Pasquale E, Pellecchia M. 2013 HTS by NMR of combinatorial libraries: a fragment-based approach to ligand discovery. *Chem. Biol.* **20**, 19–33. (doi:10.1016/j.chembiol.2012.10.015)
40. Barker A, Kettle JG, Nowak T, Pease JE. 2013 Expanding medicinal chemistry space. *Drug Discov. Today* **18**, 298–304. (doi:10.1016/j.drudis.2012.10.008)
41. Fry D *et al*. 2013 Design of libraries targeting protein–protein interfaces. *ChemMedChem* **8**, 726–732. (doi:10.1002/cmdc.201200540)
42. Bunnage ME. 2011 Getting pharmaceutical R&D back on target. *Nat. Chem. Biol.* **7**, 335–339. (doi:10.1038/nchembio.581)
43. Pagliaro L, Felding J, Audouze K, Nielsen SJ, Terry RB, Krog-Jensen C, Butcher S. 2004 Emerging classes of protein–protein interaction inhibitors and new tools for their development. *Curr. Opin. Chem. Biol.* **8**, 442–449. (doi:10.1016/j.cbpa.2004.06.006)
44. Higueruelo AP, Schreyer A, Bickerton GRJ, Pitt WR, Groom CR, Blundell TL. 2009 Atomic interactions and profile of small molecules disrupting protein–protein interfaces: the TIMBAL database. *Chem. Biol. Drug Des.* **74**, 457–467. (doi:10.1111/j.1747-0285.2009.00889.x)
45. Neugebauer A, Hartmann RW, Klein CD. 2007 Prediction of protein–protein interaction inhibitors by chemoinformatics and machine learning methods. *J. Med. Chem.* **50**, 4665–4668. (doi:10.1021/jm070533j)
46. Reynes C *et al*. 2010 Designing focused chemical libraries enriched in protein–protein interaction inhibitors using machine-learning methods. *PLoS Comput. Biol.* **6**, e1000695. (doi:10.1371/journal.pcbi.1000695)
47. Sperandio O, Reynès C, Camproux A, Villoutreix B. 2010 Rationalizing the chemical space of protein–protein interaction inhibitors. *Drug Discov. Today* **15**, 220–229. (doi:10.1016/j.drudis.2009.11.007)
48. Villoutreix BO, Labbé CM, Lagorce D, Laconde G, Sperandio O. 2012 A leap into the chemical space of protein–protein interaction inhibitors. *Curr. Pharm. Des.* **18**, 4648–4667. (doi:10.2174/138161212802651571)
49. Labbé CM, Laconde G, Kuenemann MA, Villoutreix BO, Sperandio O. 2013 iPPI-DB: a manually curated and interactive database of small non-peptide inhibitors of protein–protein interactions. *Drug Discov. Today* **18**, 958–968. (doi:10.1016/j.drudis.2013.05.003)
50. Buchwald P. 2010 Small-molecule protein–protein interaction inhibitors: therapeutic potential in light of molecular size, chemical space, and ligand binding efficiency considerations. *IUBMB Life* **62**, 724–731. (doi:10.1002/iub.383)
51. Basse MJ, Betzi S, Bourgeas R, Bouzidi S, Chetrit B, Hamon V, Morelli X, Roche P. 2013 2P2Idb: a structural database dedicated to orthosteric modulation of protein–protein interactions. *Nucleic Acids Res.* **41**, D824–D827. (doi:10.1093/nar/gks1002)
52. Venkatesan K *et al*. 2009 An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90. (doi:10.1038/nmeth.1280)
53. Stumpf M, Thorne T, de Silva E, Stewart R, An H, Lappe M, Wiuf C. 2008 Estimating the size of the human interactome. *Proc. Natl Acad. Sci. USA* **105**, 6959–6964. (doi:10.1073/pnas.0708078105)
54. Li Q, Wang Y, Bryant SH. 2009 A novel method for mining highly imbalanced high-throughput screening data in PubChem. *Bioinformatics* **25**, 3310–3316. (doi:10.1093/bioinformatics/btp589)
55. Hamon V, Brunel JM, Combes S, Basse MJ, Roche P, Morelli X. 2013 2P2Ichem: focused chemical libraries dedicated to orthosteric modulation of

protein–protein interactions. *MedChemComm* **4**, 797–809. (doi:10.1039/c3md00018d)

56. Tsao DHH *et al*. 2006 Discovery of novel inhibitors of the ZipA/FtsZ complex by NMR fragment screening coupled with structure-based design. *Bioorg. Med. Chem.* **14**, 7953–7961. (doi:10.1016/j.bmc.2006.07.050)
57. Tsiang M *et al*. 2012 New class of HIV-1 integrase (IN) inhibitors with a dual mode of action. *J. Biol. Chem.* **287**, 21 189–21 203. (doi:10.1074/jbc.M112.347534)
58. Christ F *et al*. 2010 Rational design of small-molecule inhibitors of the LEDGF/p75-integrase interaction and HIV replication. *Nat. Chem. Biol.* **6**, 442–448. (doi:10.1038/nchembio.370)
59. Cossu F *et al*. 2009 Designing Smac-mimetics as antagonists of XIAP, cIAP1, and cIAP2. *Biochem. Biophys. Res. Commun.* **378**, 162–167. (doi:10.1016/j.bbrc.2008.10.139)
60. Wang Y *et al*. 2010 An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* **38**, D255–D266. (doi:10.1093/nar/gkp965)
61. Xie XQ. 2010 Exploiting PubChem for virtual screening. *Expert Opin. Drug Discov.* **5**, 1205–1220. (doi:10.1517/17460441.2010.524924)
62. Marson CM. 2011 New and unusual scaffolds in medicinal chemistry. *Chem. Soc. Rev.* **40**, 5514–5533. (doi:10.1039/c1cs15119c)
63. Lievens S, Eyckerman S, Lemmens I, Tavernier J. 2010 Large-scale protein interactome mapping: strategies and opportunities. *Expert Rev. Proteomics* **7**, 679–690. (doi:10.1586/epr.10.30)
64. HTStec. 2012 Future Directions of HTS Trends. Cambridge, UK: HTStec. See http://www.htstec.com/.
65. Lipinski C, Lombardo F, Dominy B, 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. (doi:10.1016/S0169-409X(00)00129-0)
66. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. 2002 Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45**, 2615–2623. (doi:10.1021/jm020017n)
67. Pearlman RS, Smith KM. 1999 Metric validation and the receptor-relevant subspace concept. *J. Chem. Info. Comput. Sci.* **39**, 28–35. (doi:10.1021/ci980137x)
68. Cortes C, Vapnik V. 1995 Support-vector networks. *Mach. Learn.* **20**, 273–297. (doi:10.1007/BF00994018)
69. Josse J. 2008 FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* **25**, 1–18.

$2P2I_{HUNTER}$: A Tool for Filtering Orthosteric Protein-Protein Interaction Modulators via a Dedicated Support Vector Machine

Supplementary Material

Table S1

	MW	#Ring	#HBA	LogP	'Ro4'[a]	'Ro5'[b]	Veber[c]
[a]PPI inhibitors	87.5	80	90	52.5	80	62.5	65
SVM decoy	5	16	63	19	<7	99	98

Table S1: Percentage of compounds that satisfy the different parameters of 'Ro4' in $2P2I_{DB}$ and in the decoy selected for the SVM. [a]Percentage of compounds that follow 'Ro4'. [b]Percentage of compounds 'Ro5' compliant. [c]Percentage of compounds following Veber rule.

Table S2

Dragon Molecular descriptor	Description
ALOGP	Ghose-Crippen octanol-water partition coefficient
Hy	Hydrophilic factor
MW	Molecular weight
nBM	Number of multiple bonds
nCIC	Number of rings
nHAcc	Number of acceptor atoms for H-bonds (N,O,F)
nHDon	Number of donor atoms for H-bonds (N and O)
nHDHA	Sum of acceptors and donors (nHAcc+ nHDon)
RBN	Number of rotatable bonds
TPSA	Topological polar surface area using N,O,S,P polar contributions
Uc	Unsaturation count

Table S2: List of Dragon molecular descriptors used in the SVM model.

Table S3

PPI	AID	Assay
Core-binding factor, beta subunit isoform 1/Runt-related transcription factor 1 (AID 1645)	1496	Primary Screen (FRET)
	1438	Dose Response Confirmation
MEKKK2/MEKK5 (AID 1683)	1531	Primary Screen (FRET)
	1892	Single Concentration Confirmation Screen
	1897	Dose Response Confirmation

Table S3: Description of the two high-throughput screening PPI bioassays selected from Pubchem for external validation of the SVM model.

Figure S1: 2D representation of the 40 non-redundant orthosteric PPI modulators used as a positive dataset in the SVM approach. To guarantee structural diversity of the compounds, the 52 small molecules present in 2P2I database were clustered on the basis of Tanimoto similarity criterion of 0.8 with the OptiSim algorithm implemented in the Tripos package leading to 40 non-redundant molecules that were used as the positive set in the SVM protocol.

Figure S2: Representation of the chemical diversity of the positive dataset and decoy. (**A**) A set of 2D BCUT descriptors were generated for the positive dataset and decoy. The first three principal components defined by these BCUT descriptors that illustrate the diversity of the compounds have been selected and used to represent the chemical space of the training dataset. Compounds are clustered based on their chemical properties calculated with the BCUT descriptors. Clusters of molecules are shown as red and black spheres for the positive dataset and decoy respectively. The average number of molecules in each cluster is 1.03 and 1.05 for the positive dataset and decoy respectively. (**B**) Projection of the same training dataset according to the first three principal components calculated with the 11 Dragon descriptors selected for the SVM models.

Figure S3: Flowchart of the external validation bioassay selection from Pubchem Bioassay. PPI related bioassays for which both a primary and confirmatory screenings were available have been selected through an advanced query. Manual inspection of the 17 series of bioassays revealed that only five corresponded to protein-protein disruption. Cell-based bioassays were discarded to avoid inactive molecules due to cell penetration and only take into account the potentiality of the small molecules to act as PPI modulators.

Figure S4: Distributions of MW, number of rings and number of hydrogen bond acceptors for the positive training dataset as well as the active and selected molecules from the 2 external validation datasets. The y axes represent the percentage of molecules in each plot.

Figure S5: 2D PCA plot of training dataset and external validation dataset. The training set is shown as red and grey spheres for the positive and decoy compounds respectively. Active compounds from AID1496 (top) and AID1531 (bottom) are shown as triangles and divided into true positives (light red) and false negatives (black). The insert in the bottom panel shows the entire 2D chemical space for the active compounds of AID1531.

Figure S6: Distributions of MW, number of rings and number of hydrogen bond acceptors for ChemDivPPi and its subset Eccentric and 3 'Ro5' compliant libraries, Maybridge, Otava and IBScreen. [a] Percentage of molecules selected by the SVM model. [b] Percentage of molecules with MW higher than 400g/mol. [c] Percentage of molecules with number of rings greater than 4. [d] Percentage of molecules with more than 4 hydrogen bond acceptors. The y axes represent the percentage of molecules in each plot.

Figure S1

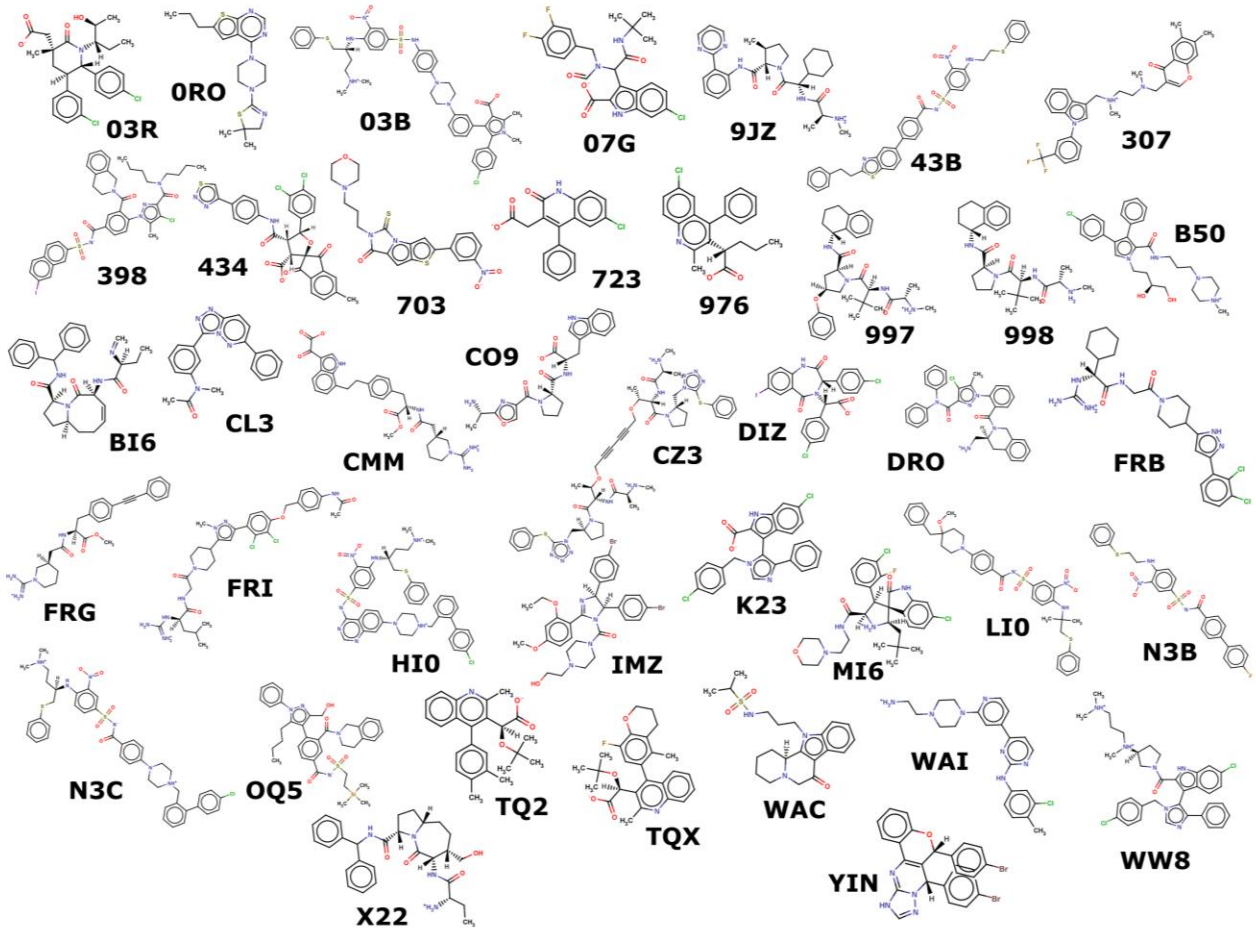

03R
0RO
03B
07G
9JZ
43B
307
398
434
703
723
976
997
998
B50
BI6
CL3
CMM
CO9
CZ3
DIZ
DRO
FRB
FRG
FRI
HI0
IMZ
K23
MI6
LI0
N3B
N3C
OQ5
TQ2
TQX
WAC
WAI
WW8
X22
YIN

Figure S2

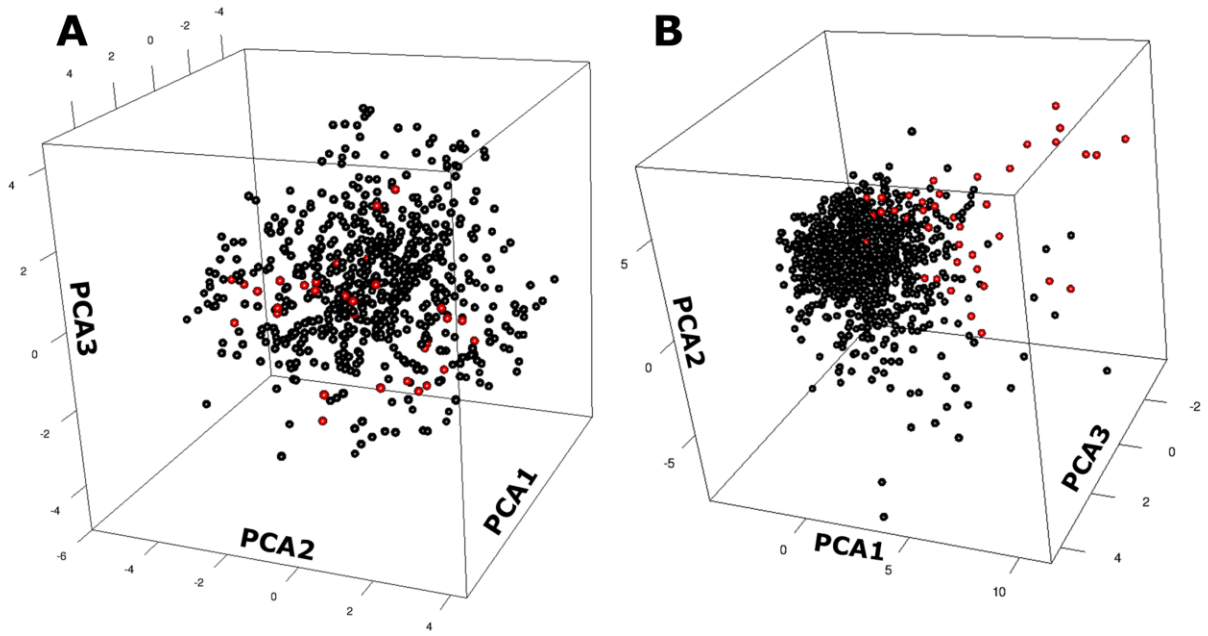


Figure S3

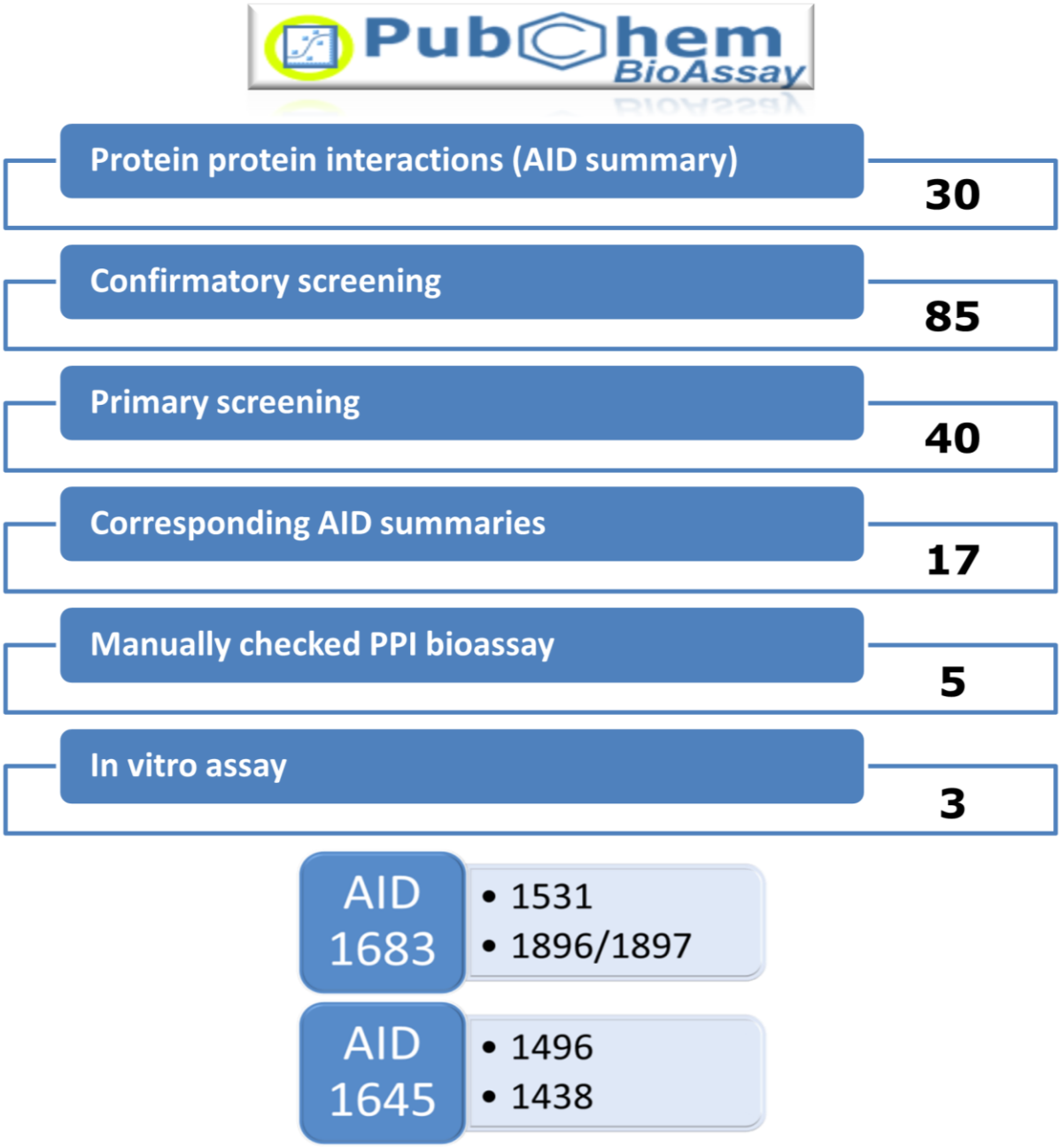

PubChem BioAssay
Protein protein interactions (AID summary)
30
Confirmatory screening
85
Primary screening
40
Corresponding AID summaries
17
Manually checked PPI bioassay
5
In vitro assay
3
AID 1683
• 1531
• 1896/1897
AID 1645
• 1496
• 1438

Figure S4

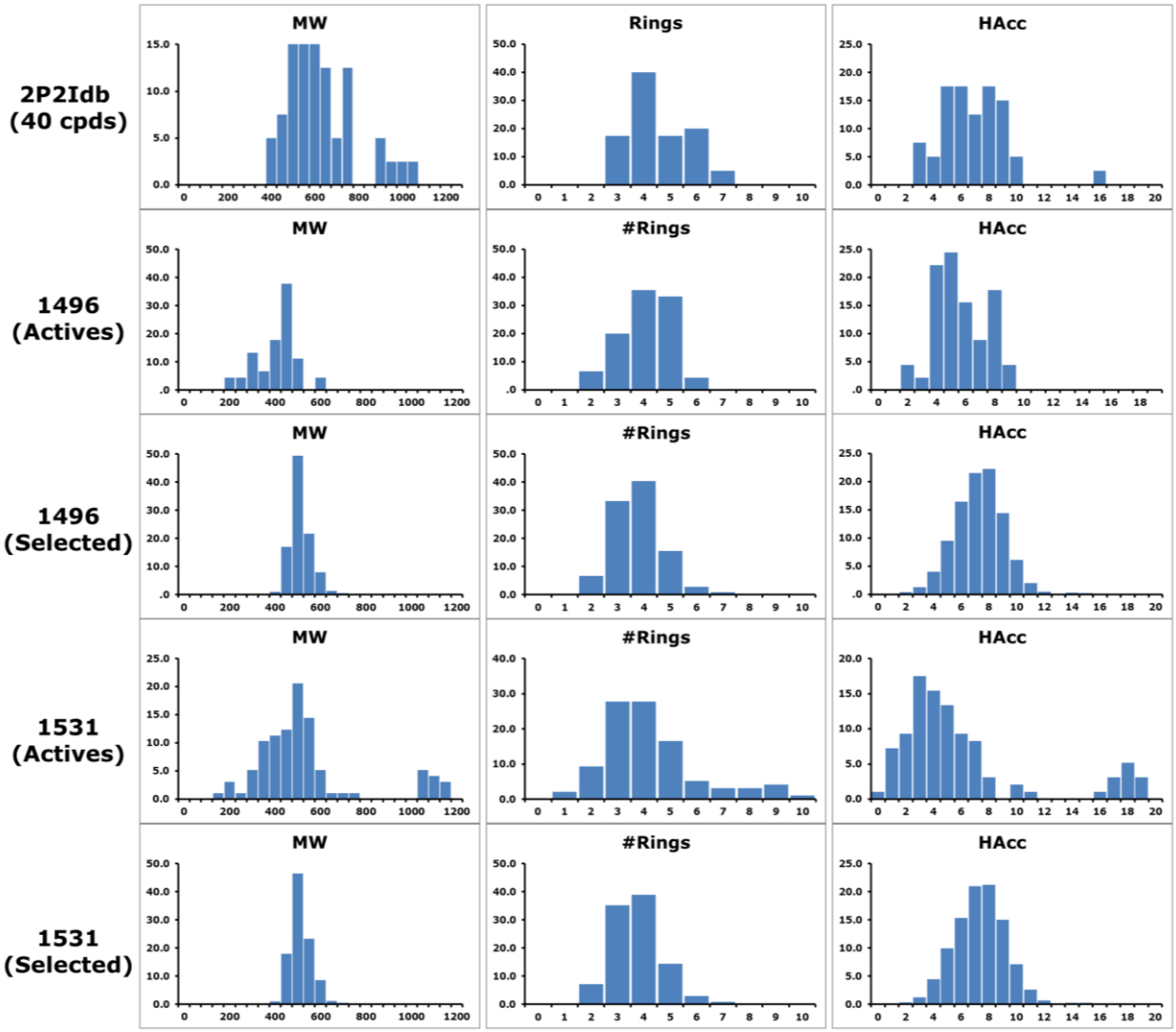

2P2Idb (40 cpds)
1496 (Actives)
1496 (Selected)
1531 (Actives)
1531 (Selected)
MW
Rings
#Rings
HAcc

Figure S5

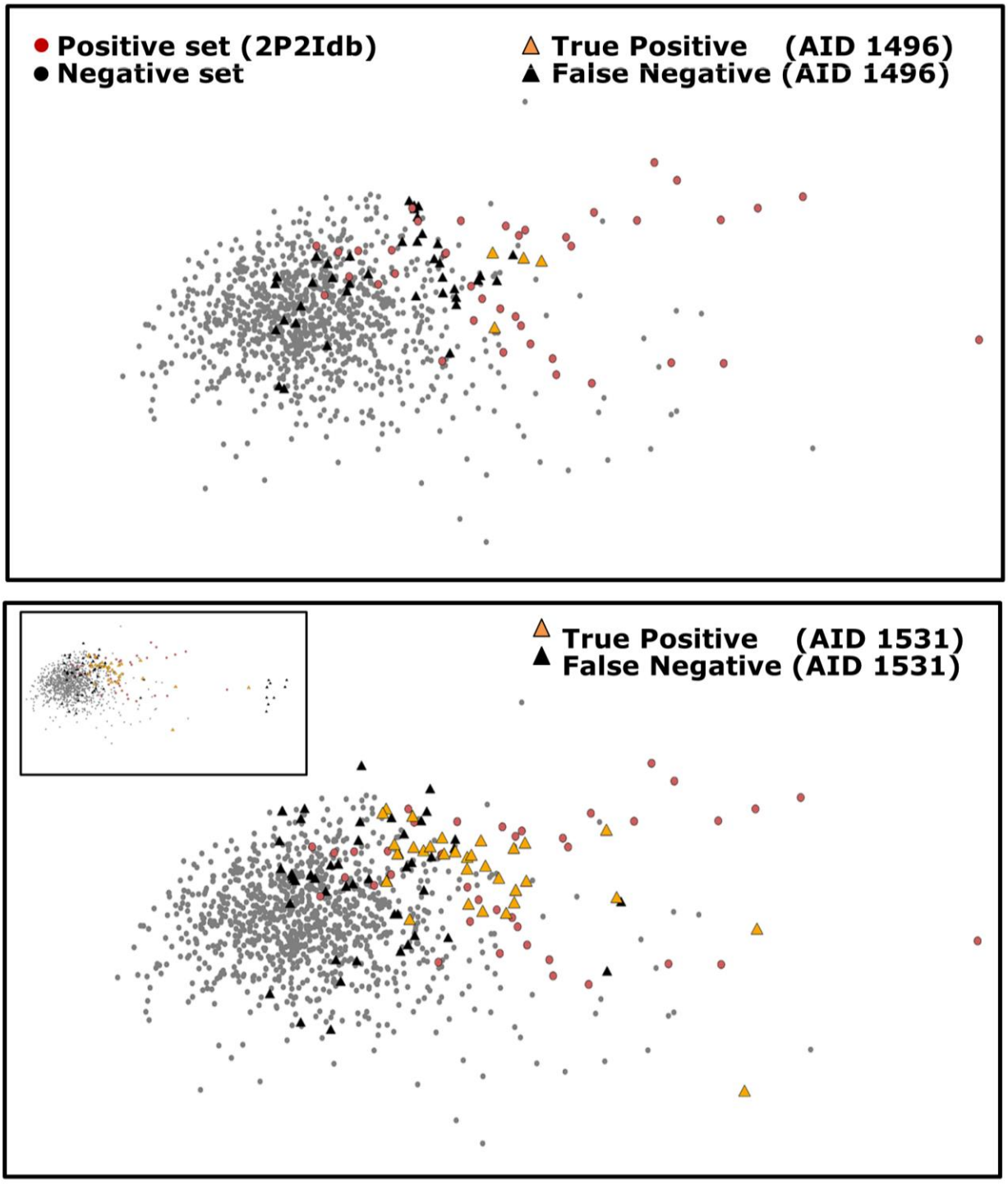
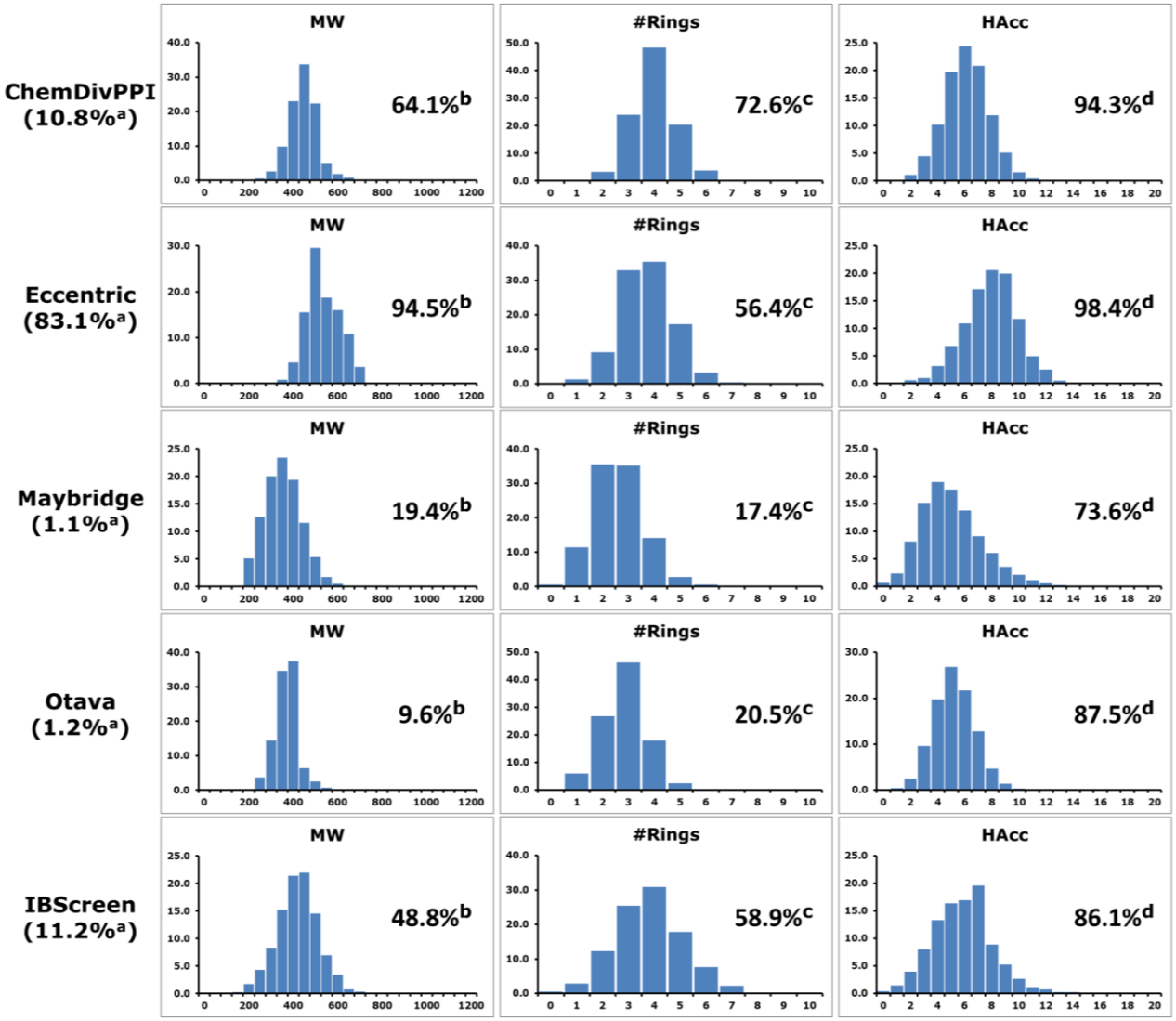

Positive set (2P2Idb)
Negative set
True Positive (AID 1496)
False Negative (AID 1496)
True Positive (AID 1531)
False Negative (AID 1531)

Figure S6


ChemDivPPI (10.8%[a])
Eccentric (83.1%[a])
Maybridge (1.1%[a])
Otava (1.2%[a])
IBScreen (11.2%[a])
MW
#Rings
HAcc
64.1%[b]
72.6%[c]
94.3%[d]
94.5%[b]
56.4%[c]
98.4%[d]
19.4%[b]
17.4%[c]
73.6%[d]
9.6%[b]
20.5%[c]
87.5%[d]
48.8%[b]
58.9%[c]
86.1%[d]