MedChemComm

RSC Publishing

CONCISE ARTICLE

2P2Ichem: focused chemical libraries dedicated to orthosteric modulation of protein–protein interactions†

Cite this: DOI: 10.1039/c3md00018d

Véronique Hamon, Jean Michel Brunel, Sébastien Combes, Marie Jeanne Basse, Philippe Roche* and Xavier Morelli*

We have recently developed a $2P2I_{HUNTER}$ learning machine tool for filtering potential orthosteric PPI modulators. In the present research article we applied our algorithm to 8.3 million compounds representing the main chemical providers commercially available to design a PPI-focused library, $2P2I_{REF}$, composed of 143 218 small molecules. Compounds corresponding to medicinally important privileged structures identified as core structures in numerous therapeutics were prioritized in a medicinal oriented version of this database ($2P2I_{PRIV}$ – 51 476 compounds). A diverse chemical library was generated using 2D molecular fingerprints ($2P2I_{DIV}$ – 8217 diverse compounds). The carbon bond saturation index (Fsp3) was used as a final filter (Fsp3 $\geq$ 0.4) to escape from flatland, another hurdle in the long path to the gold mine. We analyzed the resulting chemical space of this final library of 1683 compounds, $2P2I_{3D}$ and discuss the chemical moieties proposed to the community. This library will now be tested to evaluate its ability to enhance hit rates in general screening campaigns and is open to academic and private companies for collaborative prospects.

Received 15th January 2013
Accepted 28th February 2013

DOI: 10.1039/c3md00018d

www.rsc.org/medchemcomm

Introduction

Cells encode up to a half million protein–protein interactions (PPIs).[1,2] However, among these PPIs, only a small percentage is hypothesized to be disease relevant and to be amenable to disruption *via* small molecule drugs (*i.e.* to represent druggable targets).[3,4] This number represents untapped target opportunities and recently excited researchers who fight for 'high-hanging fruits'.[5–9] However, target identification and druggability assessment is a delicate process, as proteins operate within highly interconnected interactome networks and do not function in isolation. To help researchers understand this intricate network, the macromolecules and interactions within the web can be modeled as graphs of nodes and edges, respectively. Targeting protein nodes that possess a higher number of functional connections within the network, so-called hubs, affects the entire network and pinpoints these proteins as attractive targets for antibiotics, fungicides, pesticides and anticancer compound development.[10] In the PPI drug discovery field, hub-targeting processes and multi-target approaches are of great interest for network damage. In contrast, edge-specific perturbations confer distinct functional consequences compared with node removal. The disruption of information between functional modules that target bridges instead of hubs also represents an attractive concept for human-specific diseases. The ability to modulate these 'edgetics' in a particular signaling pathway, transcription factor-related diseases, metabolic dysfunction, or epigenetically related diseases has been recently demonstrated to be a key concept for bringing innovative drug discovery to patients.[11]

Regardless of whether hubs or bridges are the target, an improvement in success rates of the development of PPI modulators is essential for addressing the ever-increasing expectations of pharmaceutical research to discover new therapeutic targets derived from intensified genomics and proteomics programs.[12] The success of experimental screening techniques used to identify innovative molecules depends equally on the specificity of the input, such as the quality of target selection, and also the chemical libraries used to find the 'hidden gem' or active compound. These chemical libraries are clearly limited in their use for discovering modulators of PPIs (PPIMs), even considering the impressive progress in the field of medicinal chemistry regarding these undruggable targets.

A number of recent studies have confirmed a poor correlation between the chemical spaces of existing screening libraries and known PPIMs. Therefore, the improvement of compound selection dedicated to PPI chemical space represents an exciting challenge for the chemoinformatics community. In a seminal paper published in 2004, Pagliaro *et al.* highlighted the fact that the widely accepted Lipinski's standard chemoinformatics filters were not valuable for high quality enrichment factors towards PPI inhibition and urged the abolishment of

CRCM, CNRS UMR7258, Laboratory of Integrative Structural and Chemical Biology (ISCB), INSERM, U1068, Institut Paoli-Calmettes, Aix-Marseille Université, UM105, F-13009, Marseille, France. E-mail: philippe.roche@imm.cnrs.fr; xavier.morelli@inserm.fr; Fax: +33 (0)491 164 540; Tel: +33 (0)86 97 73 31

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3md00018d

preliminary filters when searching for small organic compounds as tools to study the effects of small compound inhibition.[13] In 2007, Neugebauer *et al.* developed a learning machine approach to decipher a corresponding chemical space using 25 compounds and a decision tree with 3 descriptors: molecular shape, presence of ester function and three-dimensional structure of the molecule.[14] Later, Higueruelo *et al.* proposed a larger analysis of 104 molecules that disrupt 17 PPIs retrieved from 40 papers assembled in the TIMBAL database.[15] Finally, Reynes *et al.* recently offered an original tool, the hit-profiler, to the scientific community.[16,17] They selected 66 diverse PPIMs and compared them with 557 traditional drugs through a decision tree involving a molecular shape parameter and multiplicity bond index. All these efforts permitted a clearer delineation of the PPIM chemical space and comparison with existing drugs. A consensus of these studies highlighted the observation that PPIMs tend to have a larger molecular size, a more hydrophobic nature, and structures with more rigid, aromatic scaffolds in combination with charged or polar functionalities. With these data, it is becoming increasingly clear that the chemical space of the PPIMs does not fully overlap with existing drugs. We proposed to further progress in this process to assemble a structural database with all of the successfully discovered orthosteric PPI inhibitors for which the three-dimensional structures of the PPI interfaces in both the free form and bound to the inhibitor form are known.[18,19] A general analysis of the molecular descriptors of these orthosteric inhibitors led us to propose a 'Rule-of-Four' to describe the chemical space covered by these compounds.[7] Finally, we compared recent metrics related to size, activity and polarity, such as ligand efficiency (LE), binding efficiency index (BEI) and surface efficiency index (SEI), for known drugs that target this specific ensemble and demonstrated that orthosteric PPIMs should not be rejected because of their low probability for development as drugs because they fall in a reasonable efficiency space.[20,21] We then developed our learning machine tool, $2P2I_{HUNTER}$, which is a tool for filtering orthosteric PPI modulators *via* a dedicated support vector machine.[22] The filtering protocol was validated using both external datasets from the PubChem bioassay and results from in-house screening campaigns. To characterize the applicability of our protocol, the $2P2I_{HUNTER}$ algorithm was applied to the main chemical libraries commercially available to the pharmaceutical industries and academic researchers, representing a total of more than 8.3 million compounds. We believe that the resulting chemical space identified here will provide the scientific community with concrete support in the search for PPI inhibitors during high-throughput screening (HTS) campaigns. Therefore, the $2P2I_{3D}$ database is proposed as the first academic chemical library dedicated to the HTS of PPIs.

Results

Building a reference library dedicated to PPI

We developed the 2P2I database, a hand-curated database dedicated to the structure of protein–protein complexes with known small molecule orthosteric inhibitors.[18] Detailed analysis and characterization of the chemical space of the 39 orthosteric PPI modulators (present in the version 0.95 of the 2P2I database) prompted us to propose a 'Rule-of-Four' to define the generic profile of PPI modulators, which is partially consistent with Lipinski's well-known 'Rule-of-Five'.[7] Using a combination of learning approaches with a set of 10 molecular descriptors, we developed the SVM-based $2P2I_{HUNTER}$ filtering tool to design focused chemical libraries dedicated to the inhibition of protein–protein complexes.[7] This tool has been validated with external PPI bioassays from PubChem. The next step was to use this tool to construct PPI libraries to be experimentally validated on various selected PPI targets. For this, we collected a set of 25 commercial chemical libraries from the large vendor subset of the ZINC database, which is a free database of commercially available compounds for virtual screening.[23] The complete list of chemical libraries composed of more than 8.3 million compounds is shown in Table 1, along with the number of compounds and web address. Collections of compounds corresponding with the version available in January 2012 were either downloaded directly from the provider's website or requested. Among these libraries, 3 corresponded with PPI targeted libraries and 2 with natural products libraries (Table 1). The size of the individual libraries varied from 363 compounds for the Enamine PPI library to more than 1.6 million compounds for the Uorsy screening compounds. These commercial libraries were filtered by the providers with Lipinski-like rules ($M_w \leq 500$, $A\log P \leq 5$, hydrogen bond acceptors ≤ 10 and hydrogen bond donors ≤ 5), as indicated by the high percentage of 'Ro5' compliant compounds contained within these libraries (data not shown). Each of these 25 chemical libraries was submitted to the version 0.95 of the $2P2I_{HUNTER}$ filtering tool, which was implemented in an R environment (http://www.cran.r-project.org). The percentage of compounds selected by the SVM model varied from 0.18 to 7.85, which represented a 40-fold rate (Table 2 and Fig. 1). This result indicates that the selected chemical libraries share different chemical spaces and validates the concept of low enrichment of PPI-like inhibitors. The highest percentage of selected molecules was observed with the IBScreen natural compounds collection (7.9%). The ZINC natural products library also leads to an above average percentage of selected compounds (4.3%). Natural compounds usually exhibit higher molecular weights than traditional drugs. The main objective of this study was to design PPI-focused chemical libraries, not to compare the different commercially available chemical libraries; therefore, for further treatments and analyses, individual sets of molecules selected by the combination of the SVM-model and 'Ro4' amounting to 233 727 compounds were pooled. Any duplicates were removed, leading to 143 218 unique compounds corresponding with our reference PPI library, $2P2I_{REF}$ (Fig. S1†). The first assessment of the library diversity was obtained by applying the Optisim procedure[24] implemented in the Sybyl package (Tripos, http://www.tripos.com). Tuning the Tanimoto similarity threshold at 0.8 selected a diverse set of 22 845 compounds (Fig. S1†).

To further characterize $2P2I_{REF}$ and estimate its structural diversity, we analyzed its population of distinct, two-ring

Table 1 List of the 25 chemical collections of compounds used to build the *in silico* PPI targeted libraries discussed in this work. For each chemical library, the provider, name of the collection, number of compounds, percentage of 'Ro5' compounds, and website of the provider are given. This initial set of chemical libraries contains 3 PPI targeted libraries and 2 natural product collections

Library	Collection	# Compounds	% Ro5	Website
ASDI	Global collection	16 656	99.3	http://www.frontierssi.com/
ASDI	HTS	65 119	95.7	http://www.frontierssi.com/
Asinex	Merged libraries	436 012	96.8	http://www.asinex.com/
AsisChem	Screening libraries	234 509	94.7	http://www.asischem.com/
ChemBridge	EXPRESS-Pick collection	503 803	98.9	http://www.chembridge.com
ChemDiv	Discovery chemistry collection	668 052	94.2	http://www.chemdiv.com
ChemDiv	**PPI**	123 000	95.9	http://www.chemdiv.com
DrugBank	FDA approved drugs	1412	88.8	http://www.epa.gov/ncct/dsstox/
Enamine	**PPI**	363	98.6	http://www.enamine.net
IBScreen	Screening collection	441 574	95.9	http://www.ibscreen.com/
IBScreen	Natural products	47 229	93.3	http://www.ibscreen.com/
Innovapharm	Synthetic organic compounds	316 334	93.8	http://www.innovapharm.com.ua/
LaboTest	OnStock	106 743	94.8	http://www.labotest.com/
LifeChemicals	**PPI**	31 143	97.8	http://www.lifechemicals.com/
MayBridge	Screening collection	55 717	97.1	http://www.maybridge.com/
MayBridge	HitFinder	14 400	98.7	http://www.maybridge.com/
NCI	Screening compounds	377 581	98.5	http://www.dtp.nci.nih.gov/
Otava	Tangible compounds library	487 428	99.4	http://www.otavachemicals.com/
Pharmeks	Main	259 523	91.6	http://www.pharmeks.com/
Princeton Biomolecular	Express stock	794 160	95.5	http://www.princetonbio.com/
Specs	Screening compounds	203 434	92.9	http://www.specs.net/
TimTech	ActiMol collection (HTS)	97 721	96.7	http://www.timtec.net/
Uorsy	Screening compounds	1 643 662	99.3	http://www.ukrorgsynth.com/
Vitas	HTS compounds	1 101 503	94.6	http://www.vitasmlab.com/
ZINC	Natural products	225 118	93.4	http://www.zinc.docking.org/

Table 2 Percentage of selected compounds from the 25 chemical libraries from big vendors after applying the SVM-based $2P2I_{HUNTER}$ tool (a) or a combination of $2P2I_{HUNTER}$ and 'Ro4' (b) (see Materials and methods for more details)

Library	% Selected $2P2I_{HUNTER}$[a]	% Selected $2P2I_{HUNTER}$ + Ro4[b]
IBScreen natural products	7.9	6.5
ChemDiv PPI	6.9	6.0
Pharmeks	6.1	5.5
Innovapharm	5.9	5.3
ChemDiv	5.9	5.3
FDA approved	5.7	3.7
Enamine PPI	5.5	3.6
Asinex	5.4	4.6
Specs	4.7	4.3
Asischem	4.4	4.0
ZINC natural products	4.3	3.5
Princetonbio	4.2	3.7
Vitas	4.1	3.8
LifeChemicals PPI	3.7	3.3
Asdi HTS	3.1	2.9
IBScreen screening collection	3.1	2.7
LaboTest	2.9	2.6
Uorsy	2.0	1.4
Timtec	1.9	1.7
NCI	1.4	0.7
ChemBridge	1.3	1.1
MayBridge screening collection	0.7	0.6
Otava	0.6	0.6
MayBridge Hit Finder	0.5	0.3
Asdi global collection	0.2	0.1

systems and the number of molecules associated with each of these systems. For this purpose, we used the ScaffoldTreeGenerator, which is the Java-based script at the origin of the interactive Scaffold Hunter program (http://www.scaffoldhunter.sourceforge.net) that builds a hierarchical classification of cyclic scaffolds.[25,26] Each structure was iteratively simplified into its parent chemical scaffolds by successively removing one ring entity at a time according to the prioritization rules. From this analysis, all two-cycle ancestor scaffolds were retrieved, and the corresponding molecule population that they were related to was determined for each scaffold. We found 3973 different two-ring scaffolds that were represented at least once in the $2P2I_{REF}$. These 3973 scaffolds corresponded with 140 534 compounds (98.2%) in the $2P2I_{REF}$; the remaining molecules could not be subdivided into two-ring scaffolds. The number of molecules in the $2P2I_{REF}$ containing each scaffold is shown in a bubble plot representation (Fig. 2). The top 50% of the molecules in the $2P2I_{REF}$ are represented by only 40 scaffolds (1% of the scaffolds), as shown in Fig. S2.†

As can be observed on the TreeMap histogram, there is a well-balanced distribution of structures between both classes of the fused rings and phenyl-substituted cycles. These basic frameworks are common to many different scaffolds, from indoles to benzopyrans and aryl piperidines to phenyl triazolidines. It is not surprising that the phenyl moiety is ubiquitous in compounds of potent pharmaceutical interest because aromatic systems have long been considered crucial in molecular recognition.

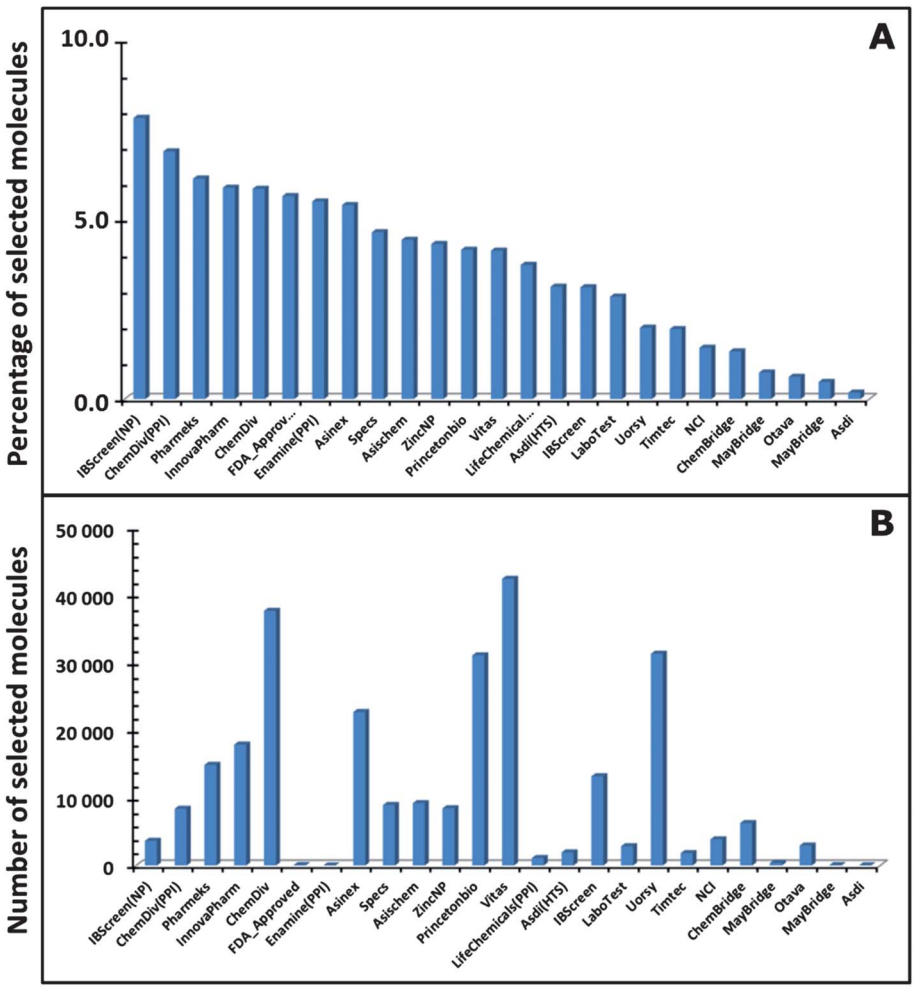


Fig. 1 Performance of the PPI filtration tool on 25 screening libraries. Each chemical library was filtered with a combination of the SVM-based $2P2I_{HUNTER}$ tool and the 'Ro4' rule. (A) Percentage of compounds selected. (B) Actual number of compounds selected.

Building a privileged PPI library ($2P2I_{PRIV}$)

Our objective was to produce a chemical library of reasonable size that could be further validated through the experimental screening of various selected PPI targets. To decrease the size of the library to a manageable size and increase the chances of including putative biologically active and drug-like compounds, we decided to discard all compounds without privileged scaffolds from the reference library. Privileged structures are molecular scaffolds that are frequently found in molecules with good pharmacological properties and capable of interacting with a large set of targets. For example, benzodiazepine derivatives are considered privileged because of their ability to structurally mimic beta turn substructures. These privileged scaffolds are highly represented in the drugs used in the pharmaceutical industry and either obtained from natural product derivatives or through *de novo* synthesis. We relied on a recently reported comprehensive survey to define a set of 54 privileged scaffolds.[27] A query search for each substructure representing a privileged scaffold was applied to the 143 218 model-filtered, non-redundant compounds from the reference library. Thirty-five of the privileged scaffolds were detected in at least one molecule in $2P2I_{REF}$ (Table 3). Because more than one privileged structure per molecule may be detected, some compounds were present in more than one of the 35 subgroups. Therefore, the next step was to pool the molecules from the individual groups and remove the duplicates. This led to a collection of 51 476 compounds that corresponded to the $2P2I_{PRIV}$ PPI-focused library (Fig. S1†). The most populated privileged structures correspond to attractive compounds for drug discovery (Fig. 3).

The indole scaffold (7889 molecules) represents one of the most important structural subunits for the discovery of new potent drug candidates.[28,29] The observation that many alkaloids contain the indole nucleus brought about massive research into indole chemistry. This gave rise to a vast number of biologically active products with a wide range of therapeutic uses, such as anti-inflammatory agents, phosphodiesterase inhibitors or cannabinoid receptor agonists and HMG-CoA reductase inhibitors. The quinoline scaffold (8570 molecules) is characterized by the presence of a double-ring structure containing a benzene ring fused to pyridine at two adjacent carbon atoms. These compounds have been widely used to synthesize molecules with medical benefits, including antimalarial,

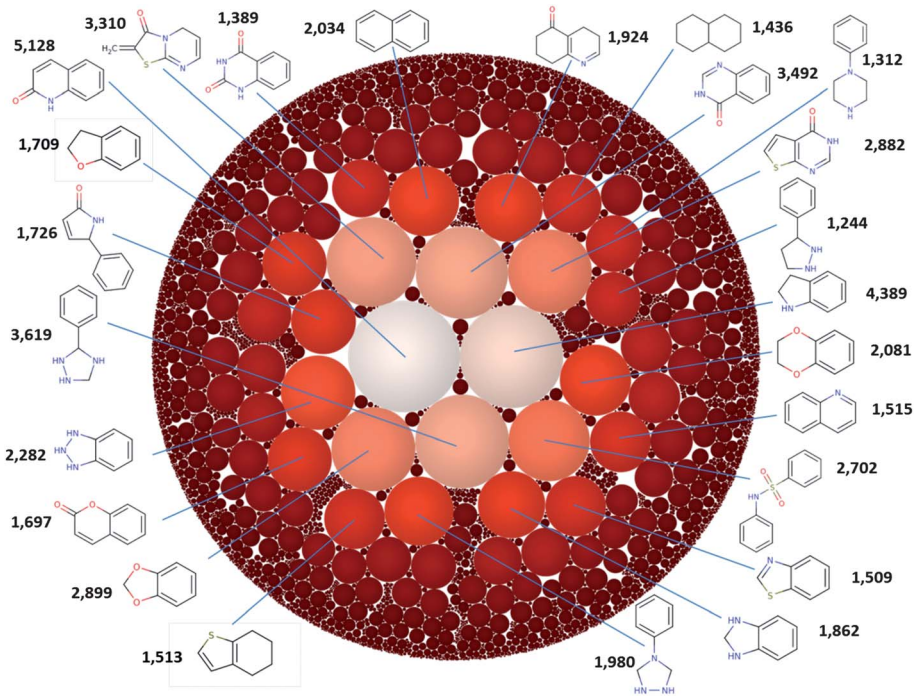


Fig. 2 Structural diversity of the $2P2I_{REF}$ chemical library. The Scaffold Hunter program (http://www.scaffoldhunter.sourceforge.net) was used to search for two-ring systems within $2P2I_{REF}$ compounds. Each structure was iteratively simplified into its parent chemical scaffolds by successively removing one ring entity at a time according to prioritization rules. From this analysis, all two-cycle ancestor scaffolds were retrieved and the corresponding molecule population they were related to was determined for each one. We found 3973 diverse two-ring scaffolds present in $2P2I_{REF}$ and the distribution of their population is shown here as a bubble plot representation. 2D structures of the most populated scaffolds are highlighted. This figure was generated with the hierarchical grouping visualization program TreeMap from Macrofocus (http://www.treemap.com/).

antimicrobial and anticancer activities.[30] This broad spectrum of biological and biochemical activities has been further facilitated by the synthetic versatility of quinolone, which can be easily substituted, allowing for the synthesis of numerous potent biologically active compounds, such as those that inhibit tyrosine kinases, the proteasome, tubulin polymerization and DNA repair in the context of cancer drug development and refinement.[31,32] Quinazolines[33] (5948 molecules) and quinazolinones[34,35] (5592 molecules) occupy a distinct and unique place in the medical field. This heterocyclic moiety has great medicinal and biological significance. A wide array of quinazoline drugs possess a variety of medicinal properties and act as analgesics, antibacterial and anticancer agents. Substituted 2-aryl-benzothiazepines (304 molecules) have recently emerged as important pharmacophores in a number of diagnostic and therapeutic settings.[36] Therefore, attractive features of drug candidates based on this scaffold include their synthetic accessibility. Examples of 2-aryl-benzothiazepines that are endowed with diagnostic/therapeutic activity include the 2-(4-aminophenyl)benzothiazole series, which has a remarkable activity profile for both the potential non-invasive diagnosis of Alzheimer's disease and as antitumor agents.

Building chemically diverse PPI libraries: $2P2I_{DIV}$ and $2P2I_{3D}$

A chemically diverse PPI library was then designed from the $2P2I_{PRIV}$ library using a three-step procedure (Fig. S1†). First, the chemical diversity of each ensemble of molecules corresponding with one privileged structure was assessed. The 35 diverse sets were then pooled, and the duplicates were removed. Finally, the chemical diversity of the privileged library was assessed using a Tanimoto coefficient of 0.8 to build the $2P2I_{DIV}$ PPI-focused library composed of 8217 compounds (Fig. S1†).

An increasingly accepted measure of the complexity of molecules is carbon bond saturation (Fsp3), where Fsp3 is the number of sp3-hybridized carbons divided by the total carbon count. It has been shown that Fsp3 correlates with improved solubility and leads to the faster transition of a compound from discovery to drug.[37,38] In addition, because of the physicochemical properties of the protein–protein interface, such as the presence of 2 or 3 non-aligned binding pockets, PPI modulators tend to be more three-dimensional than other drug molecules.[39] In a last filtering step, only those compounds with an Fsp3 greater than 0.4 were retained. The final $2P2I_{3D}$ PPI-focused library (Fig. S1†) is composed of 1683 diverse compounds with privileged structure and improved 3D-shape. As expected, compounds in $2P2I_{3D}$ share the same chemical space as known orthosteric PPI modulators used as a positive training dataset in the development of the SVM-based $2P2I_{HUNTER}$ tool (Fig. S3†). The $2P2I_{3D}$ chemical library has been further characterized using standard molecular descriptors (Fig. 4).

In the $2P2I_{3D}$ library, 51% of the compounds follow the widely used 'Rule-of-Five'.[40] A very similar percentage of PPI modulators that are present in the 2P2I database follow the

Table 3 List of privileged structures found in at least one molecule in the reference library $2P2I_{REF}$. A set of 54 privileged scaffolds defined in a recent review[27] was selected to perform a query search using smart codes in the reference PPI library composed of 143 218 compounds. The 2D structure and occurrence of scaffolds identified in at least one molecule are presented in this table. The number of diverse compounds in each of the 35 sets was estimated using Optisim implemented in TRIPOS[24]

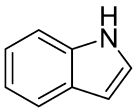
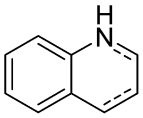
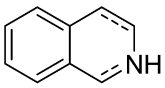
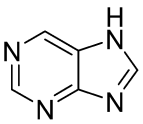
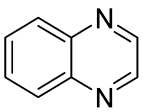
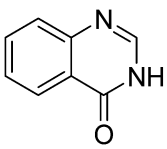
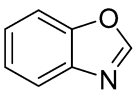
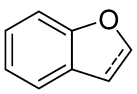
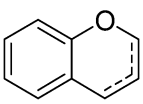
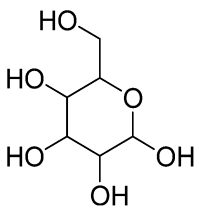
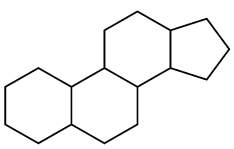
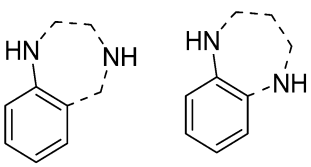
#	Privileged structures	Code smarts (MarvinSketch)	Structure	Total cpds	Diverse cpds
1	Indole	N1C=CC2=CC=CC=C12		7889	1265
2	Quinoline	[#6]=,:1-[#6]=,:[#7]C2=CC=CC=C2 [#6]=,:1		8570	666
3	Isoquinoline	[#6]1=,:[#6]C2=CC=CC=C2[#6]=,:[#7]1		374	95
4	Purine	N1C=NC2=NC=NC=C12		1025	129
5	Quinoxaline	C1=CC=C2N=CC=NC2=C1		1047	148
6	Quinazolinone	O=C1NC=NC2=CC=CC=C12		5592	443
7	Benzoxazole	O1C=NC2=CC=CC=C12		255	90
8	Benzofurane	[#6]1=,:[#6]C2=CC=CC=C2O1		1158	250
9	Chromene	[#6]1=,:[#6]C2=CC=CC=C2OC1		3305	387
10	Carbohydrate	OCC1OC(O)C(O)C(O)C1O		169	36
11	Steroid	C1CC2CCC3C(CCC4CCCCC34)C2C1		236	57
12	Benzodiazepine	[#6]~1~[#6]~[#7]~C2=C(~[#6]~[#7]~1) C=CC=C2, [#6]~1~[#6]~[#7] ~C2=C(~[#7]~[#6]~1)C=CC=C2		352	38

Table 3 (*Contd.*)

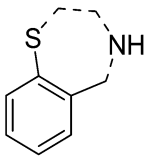
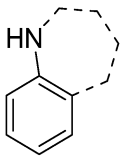
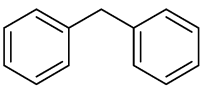
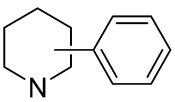
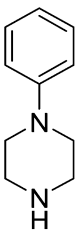
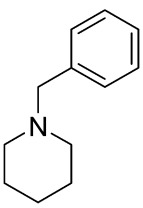
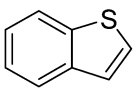
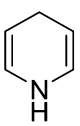
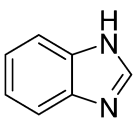
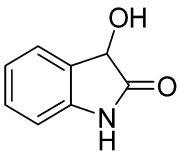
#	Privileged structures	Code smarts (MarvinSketch)	Structure	Total cpds	Diverse cpds
13	Benzothiazepine	[#6]~1~[#6]~[#16]~C2=C(~[#6]~[#7]~1)C=CC=C2		304	22
14	Benzazepine	[#6]~1~[#6]~[#6]~C2=C(~[#7]~[#6]~1)C=CC=C2		41	17
15	Diphenylmethane	C(C1=CC=CC=C1)C1=CC=CC=C1		3584	1541
16	Arylpiperidine	[$(C1CNCC(C1)C1=CC=CC=C1),$(C1CCC(NC1)C1=CC=CC=C1),$(C1CCN(CC1)C1=CC=CC=C1),$(C1CCC(NC1)C1=CC=CC=C1),$(C1CNCC(C1)C1=CC=CC=C1),$(C1CC(CCN1)C1=CC=CC=C1)]		1385	234
17	*N*-Arylpiperazine	C1CN(CCN1)C1=CC=CC=C1		3764	884
18	Benzylpiperidine	C(N1CCCCC1)C1=CC=CC=C1		1493	539
19	Benzothiophene	S1C=CC2=CC=CC=C12		375	83
20	Dihydropyridine	C1C=CNC=C1		3716	186
21	Benzimidazole	N1C=NC2=CC=CC=C12		2777	846
22	3-Substituted-3-hydroxy-2-oxindole	OC1C(=O)NC2=CC=CC=C12		25	9

Table 3 (*Contd.*)

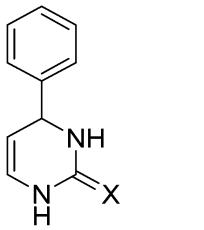
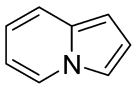
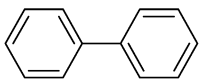
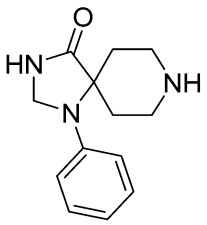
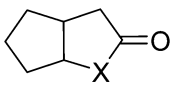
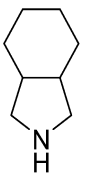
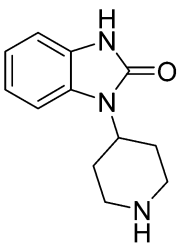
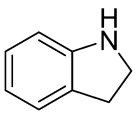
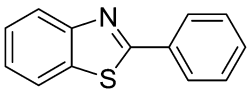
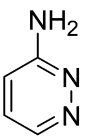
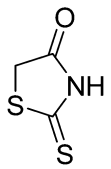
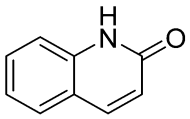
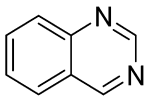
#	Privileged structures	Code smarts (MarvinSketch)	Structure	Total cpds	Diverse cpds
23	Aryldihydropyrimidone	O=C1NC=CC(N1)C1=CC=CC=C1, S=C1NC=CC(N1)C1=CC=CC=C1	 X = O or S	511	73
24	Indolizine	C1=CN2C=CC=CC2=C1		223	25
25	Biphenyl	C1=CC=C(C=C1)C1=CC=CC=C1		2009	882
26	Triazaspirodecanone	O=C1NCN(C2=CC=CC=C2)C11CCNCC1		4	1
27	5,5-*trans*-Lactam/lactone	O=C1CC2CCCC2N1, O=C1CC2CCCC2O1	 X = N orO	5	2
28	Hexahydroisoindole	C1NCC2CCCCC12		326	78
29	Benzimidazolone	O=C1NC2=CC=CC=C2N1C1CCNCC1		9	6
30	Indoline	C1CC2=CC=CC=C2N1		2132	262
31	2-Arylbenzothiazole	S1C2=CC=CC=C2N=C1C1=CC=CC=C1		88	37
32	Aminopyridazine	[#7]C1=NN=CC=C1		429	68

Table 3 (*Contd.*)

#	Privileged structures	Code smarts (MarvinSketch)	Structure	Total cpds	Diverse cpds
33	Rhodamine	O=C1CSC(=S)N1		3372	180
34	Pyranoquinolone	O=C1NC2=CC=CC=C2C=C1		6243	123
35	Quinazoline	C1=CC=C2N=CN=CC2=C1		5948	566

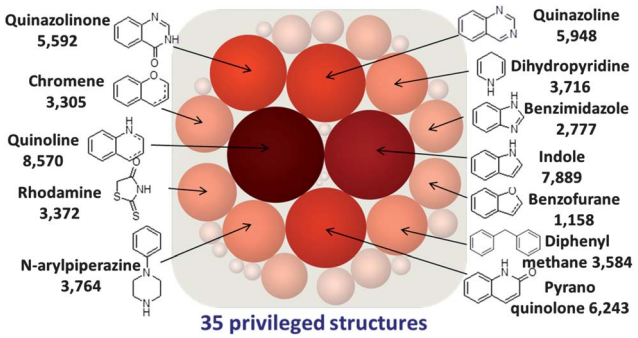


Fig. 3 Distribution in the TreeMap bubble-shaped representation of the 35 privileged structures inferred from all compounds in $2P2I_{PRIV}$. The 12 most populated scaffolds are identified on both sides and data for all scaffolds are presented in Table 3.

'Rule-of-Five'.[18,22] PPI modulators tend to be larger than other drug molecules.[7,15–17,22] In the $2P2I_{3D}$ PPI library, a shift towards higher values was also observed (average $M_w = 521.4 \pm 69.3$ g mol^{-1}). However, 53.4% of the compounds in $2P2I_{3D}$ have a M_w less than 500 g mol^{-1}, ranging from 405.7 to 971.4 g mol^{-1}. As expected, the hydrophobicity, expressed by *Alog P*, showed a significant right shift compared with oral drugs (average *Alog* $P = 4.0 \pm 2.4$) and wide distribution (ranging from −2.5 to 15.1). The protein–protein interfaces with known inhibitors are more hydrophobic than general PPIs, with fewer charged residues and more non-polar atoms.[18] It has been reported that the PPI compounds currently in development are larger and more lipophilic than non-PPI compounds, the hydrophobicity of PPI modulators could be seen as a drawback in the development of oral drugs because of reduced ADME properties and we may have to be careful in the future with the increase of these values in hit to lead development.[41,42] The number of rings is a key molecular descriptor included in the 'Rule-of-Four'. Compounds in $2P2I_{3D}$ contain between 1 and 9 rings, with an average value of 4.4 ± 1.0. Most compounds possess at least 4 rings (86% of the library). This number is correlated with the three-dimensionality of PPI compounds. Compared with enzymes, protein–protein interfaces do not reveal a deep pocket to accommodate a small molecule ligand; however, they usually contain a few transient pockets that can bind to different parts of a small molecule modulator. As a consequence, protein–protein modulators tend to adopt star, L, or T three-dimensional shapes rather than being linear.[39] PPIMs are less polar than other drugs, which is illustrated by the distribution profiles of hydrogen bond donors (HDon) and acceptors (HAcc) and the topological polar surface area (TPSA) (Fig. 4) (average HDon $= 1.9 \pm 1.3$, ranging from 0 to 9; average HAcc $= 6.9 \pm 1.9$, ranging from 1 to 15; and average TPSA $= 102.1 \pm 32.5$ Å^2, ranging from 15.4 to 298.2). Compounds with a TPSA of less than 140 Å^2 and fewer than 10 rotatable bonds generally show good oral bioavailability according to Veber.[43] The TPSA is less than 140 Å^2 for 88.9% of the compounds in $2P2I_{3D}$. The number of rotatable bonds (RBN) is a good indicator of the structural flexibility of a molecule. In $2P2I_{3D}$, 71.4% of the molecules possess less than 10 rotatable bonds (average RBN $= 10.0 \pm 2.9$, ranging from 4 to 26). The average value is significantly higher than other drugs, most likely because of the necessary structural adaptations of the orthosteric inhibitors to the few, non-aligned interfacial pockets. Finally, the analysis of the distribution of chirality pinpoints an average of 1.822 chiral centers per molecule and a median of 1 chiral center. This number should be compared to the positive set, which identifies an average value of 2.03 chiral centers per molecule and a median of 2 chiral centers per compound. These conservative numbers (*i.e.* with low chiral complexity) will thus likely increase the chances to find reasonable hits.[44]

The percentage of molecules in the final PPI chemical library, $2P2I_{3D}$, selected from each of the 25 initial collections of compounds is illustrated in Fig. 5. Notably, the highest percentage corresponds with FDA approved small molecule drugs, which is indicative of a high potential in medicinal chemistry. The other highest percentages were obtained with natural products and PPI-targeted chemical libraries. The entire

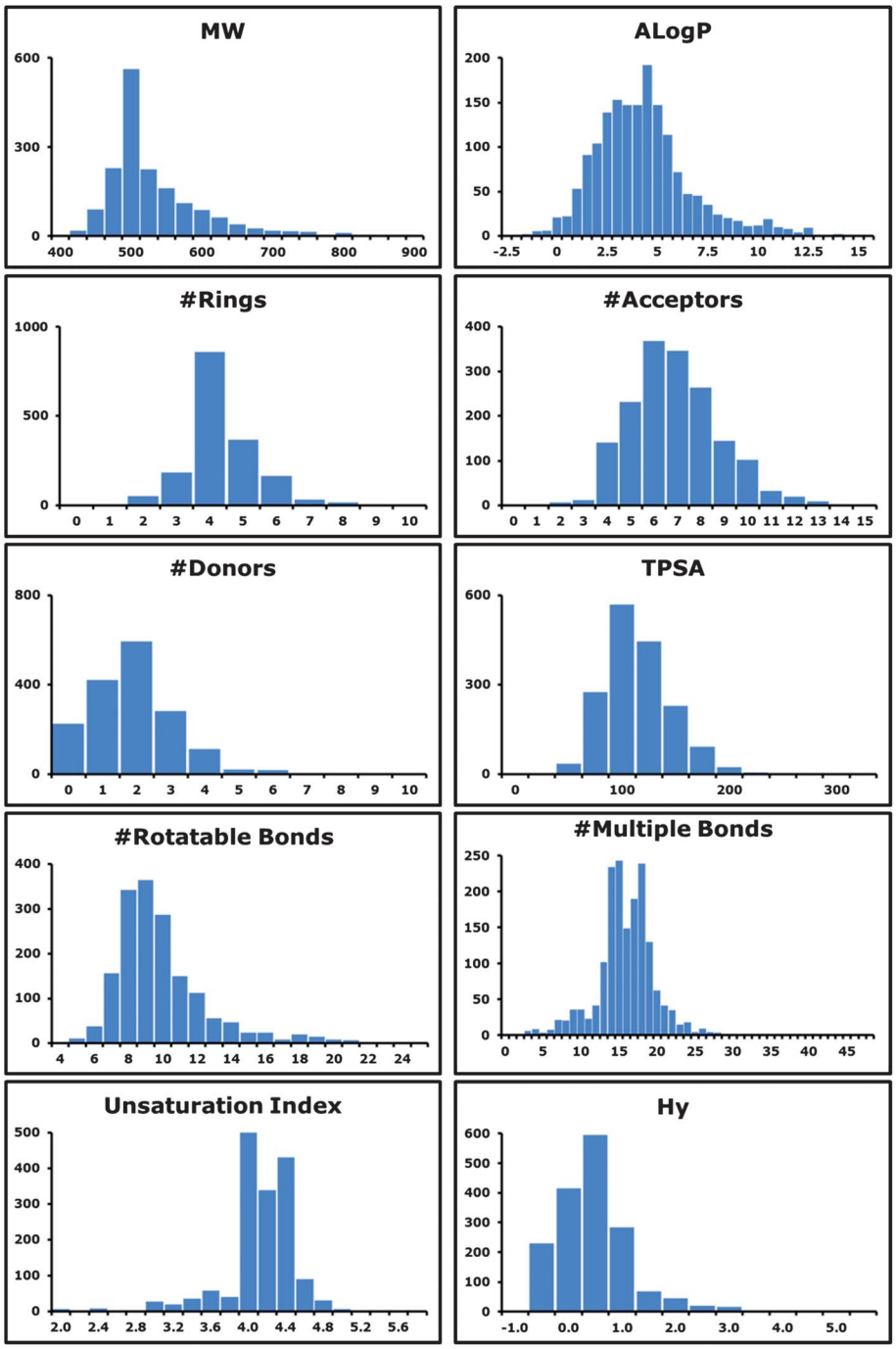


Fig. 4 Distribution of 10 key molecular descriptors for compounds in the final *in silico* diverse PPI library $2P2I_{3D}$. Molecular weight (MW), predicted octanol–water partition coefficient (*A*log *P*), number of rings (#Rings), hydrogen-bond acceptors (#Acceptors) and donors (#Donors), topological polar surface area (TPSA), number of rotatable bonds, number of multiple bonds, unsaturation index and Hydrophilic index (Hy).

set of compounds can be purchased from a limited number of providers (Fig. S4†). A total of 27.0%, 26.6% and 17.1% of the compounds can be obtained from ChemDiv, Uorsy, and Vitas respectively.

Conclusion

$2P2I_{HUNTER}$, a SVM-based algorithm for filtering potential orthosteric PPI modulators, was applied to 25 commercially available chemical libraries affording to a reference PPI-targeted library composed of 143 218 molecules ($2P2I_{REF}$). We then selected compounds containing medicinally important privileged scaffolds to conceive a PPI-focused library potentially relevant in medicinal chemistry ($2P2I_{PRIV}$, 51 476 compounds). A diverse subset of this database was built using 2D molecular fingerprints and Tanimoto coefficients ($2P2I_{DIV}$, 8217 diverse compounds). Analysis of the most representative privileged scaffolds revealed some patterns widely used in medicinal

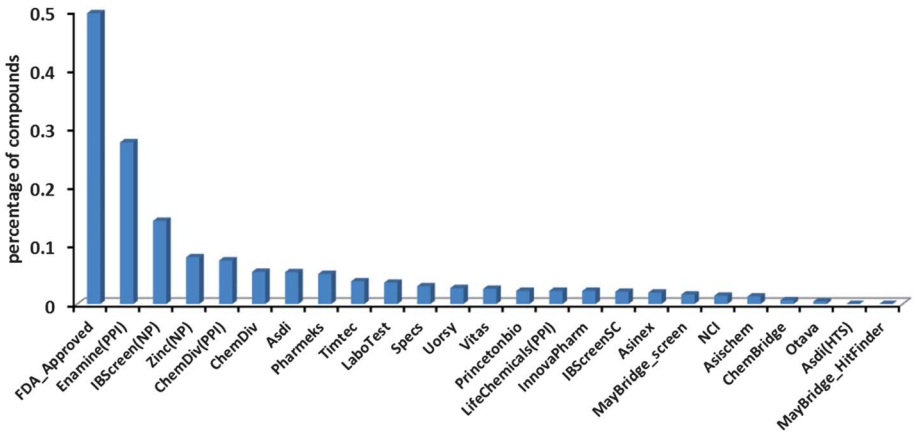


Fig. 5 Percentage of compounds in the final diverse privileged $2P2I_{3D}$ PPI library selected from each of the initial 25 chemical libraries.

chemistry such as indoles, quinolones, quinazolines, quinazolinones, and substituted 2-aryl-benzothiazepines. However, one should notice that these privileged scaffolds are mainly made of two rings and are therefore just one part of the compounds (average number of rings 4), therefore giving more space to originality and specificity. The final $2P2I_{3D}$ chemical library was built by removing flat compounds and selecting compounds with higher three-dimensional shape using the carbon bond saturation index (Fsp3) as a filter. As expected, the 1683 compounds in $2P2I_{3D}$ exhibit a different profile than standard molecule drugs or screening compounds for most molecular properties such as MW, number of rings, hydrophobicity, hydrogen bond donors and acceptors. However, these compounds possess interesting drug-like properties since a large number of them follow the oral bioavailability rules of Lipinski[40] or Veber.[43] The potential of the $2P2I_{3D}$ PPI-targeted library will now be assessed experimentally on selected PPI targets, and the hit rates compared to those obtained with standard screening libraries.

Materials and methods

List of providers

A set of 19 providers was selected from the ZINC database big vendor subset (http://www.zinc8.docking.org/vendor0/index_fsb.shtml and Table 1). FDA approved drugs from DrugBank and ZINC natural products were added to the collection. Compounds corresponding to early 2012 catalogues were collected directly from the vendor websites when possible or from the ZINC database.[23] A total of 25 libraries were finally selected including screening compounds, natural products and PPI targeted molecules representing a total of 8 129 319 compounds. The chemical libraries are listed in Table 1. Libraries dedicated to protein–protein interactions are indicated with (PPI) whereas natural products are indicated with (NP). Freely accessible SDF files of chemical databases were collected directly from commercial provider websites listed in Table 1 with the size of each database.

Preparation of compounds

All compounds collected from vendors and the ZINC database were incorporated into our in-house database, before extracting a list of unique structures. Unique structures were defined after being standardized, which consists of removal of any salts or lower molecular weight fragments, neutralisation of protonatable groups, and standardization of the structural representation of tautomeric moieties and charged moieties which cannot be neutralized. The list of unique structures therefore serves as a central reference table for all compounds included in the database, and not only those that are commercially available.

ChemAxon (http://www.chemaxon.com) tools were used to clean and homogenize libraries in view of applying the prediction model. Structures were prepared according to a classical protocol that consists in several steps as follows: firstly, the Checker module detects errors in valence, coordination, aromaticity and covalently bound counter ions and cleans them when possible. Compounds are then processed by the Standardizer module that runs successive transformations: the largest fragment is kept (salts discarded), molecules are dearomatized, aromatized and neutralized, and finally structures are cleaned in 2D. The last step is to determine major species at physiological pH 7.4 using the Cxcalc module.

'Rule-of-Four' definition

The 'Rule-of-Four' keeps a molecule if it obeys three of the following properties: $M_w \geq 400$ Da, *A*log $P \geq 4$, number of rings ≥ 4 and number of hydrogen bond acceptors ≥ 4.[7]

Molecular descriptors and filters

In order to submit the prepared libraries to our prediction tool $2P2I_{HUNTER}$, the 10 properties used for its machine learning process are calculated by Dragon version 6 (http://www.talete.mi.it) on the charged forms of compounds prepared as described above. These descriptors are: molecular weight; 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 *A*log *P*. The 11th descriptor corresponded to the sum of hydrogen bond donors

and acceptors. Molecules that generated at least one error in descriptors computation were discarded.

Molecular descriptors were calculated with DRAGON 6 and the filtering protocol combining Ro4 and $2P2I_{HUNTER}$ was applied to each individual library. The 233 727 molecules selected by the filtering tool were pooled together using the Chemaxon Instant JChem package (http://www.chemaxon.com/instantjchem/) and duplicates were removed by the overlap analysis function in a self-comparison mode that identifies multiple items of compounds in the whole dataset. A diverse set was built from the 143 218 remaining compounds using the OptiSim algorithm in Tripos and a Tanimoto coefficient of 0.8 leading to 22 845 diverse compounds.

Two-ring scaffold characterization

The java-based program ScaffoldTreeGenerator was used to build the hierarchical scaffold tree, a classification based on the ring systems contained in the molecules.[25] After reducing a molecule to its framework by cutting all terminal chains, it dissects it through an iterative removal of rings according to a set of prioritizing rules to decide which part is to be eliminated at each step. For each molecule, we used in-house scripts to retrieve the SMILES code of its 2-ring parent and infer the range of scaffold types and the population of compounds they represent.

Privileged structures search

We based our privileged structure search on a recent review providing a list of such scaffolds.[27] The 143 218 model-filtered non-redundant compounds from $2P2I_{REF}$ were imported in the Instant JChem environment and a substructure query search was carried out for each of the 54 identified privileged structures. Thirty-five of them were detected and the results are presented in Table 3. TreeMap (http://www.treemap.com/) is used to represent their bubble-shaped distribution (Fig. 3). All sets of molecules related to each one of these 35 scaffolds were exported into individual files for further treatment.

Acknowledgements

The authors would like to thank Bernard Chetrit for IT support.

References

1 M. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. An, M. Lappe and C. Wiuf, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 6959–6964.
2 K. Venkatesan, J. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K. Goh, M. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. Sahalie, S. Cevik, C. Simon, A. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. Cusick, F. Roth, D. Hill, J. Tavernier, E. Wanker, A. Barabási and M. Vidal, *Nat. Methods*, 2009, **6**, 83–90.
3 P. J. Hajduk, J. R. Huth and C. Tse, *Drug Discovery Today*, 2005, **10**, 1675–1682.
4 P. Buchwald, *IUBMB Life*, 2010, **62**, 724–731.
5 J. A. Wells and C. L. McClendon, *Nature*, 2007, **450**, 1001–1009.
6 M. E. Bunnage, *Nat. Chem. Biol.*, 2011, **7**, 335–339.
7 X. Morelli, R. Bourgeas and P. Roche, *Curr. Opin. Chem. Biol.*, 2011, **15**, 475–481.
8 A. Mullard, *Nat. Rev. Drug Discovery*, 2012, **11**, 173–175.
9 D. C. Fry, *Curr. Pharm. Des.*, 2012, **18**, 4679–4684.
10 A. F. Fliri, W. T. Loging and R. A. Volkmann, *J. Med. Chem.*, 2009, **52**, 8038–8046.
11 Q. Zhong, N. Simonis, Q. R. Li, B. Charloteaux, F. Heuze, N. Klitgord, S. Tam, H. Yu, K. Venkatesan, D. Mou, V. Swearingen, M. A. Yildirim, H. Yan, A. Dricot, D. Szeto, C. Lin, T. Hao, C. Fan, S. Milstein, D. Dupuy, R. Brasseur, D. E. Hill, M. E. Cusick and M. Vidal, *Mol. Syst. Biol.*, 2009, **5**, 321.
12 K. H. Bleicher, H. J. Böhm, K. Müller and A. I. Alanine, *Nat. Rev. Drug Discovery*, 2003, **2**, 369–378.
13 L. Pagliaro, J. Felding, K. Audouze, S. J. Nielsen, R. B. Terry, C. Krog-Jensen and S. Butcher, *Curr. Opin. Chem. Biol.*, 2004, **8**, 442–449.
14 A. Neugebauer, R. W. Hartmann and C. D. Klein, *J. Med. Chem.*, 2007, **50**, 4665–4668.
15 A. P. Higueruelo, A. Schreyer, G. R. J. Bickerton, W. R. Pitt, C. R. Groom and T. L. Blundell, *Chem. Biol. Drug Des.*, 2009, **74**, 457–467.
16 C. Reynes, H. Host, A. C. Camproux, G. Laconde, F. Leroux, A. Mazars, B. Deprez, R. Fahraeus, B. O. Villoutreix and O. Sperandio, *PLoS Comput. Biol.*, 2010, **6**, e1000695.
17 O. Sperandio, C. Reynès, A. Camproux and B. Villoutreix, *Drug Discovery Today*, 2010, **15**, 220–229.
18 R. Bourgeas, M.-J. Basse, X. Morelli and P. Roche, *PLoS One*, 2010, **5**, e9598.
19 M. J. Basse, S. Betzi, R. Bourgeas, S. Bouzidi, B. Chetrit, V. Hamon, X. Morelli and P. Roche, *Nucleic Acids Res.*, 2013, **41**, D824–D827.
20 C. Abad-Zapatero and J. T. Metz, *Drug Discovery Today*, 2005, **10**, 464–469.
21 C. Abad-Zapatero, O. Perišić, J. Wass, A. P. Bento, J. Overington, B. Al-Lazikani and M. E. Johnson, *Drug Discovery Today*, 2010, **15**, 804–811.
22 V. Hamon, R. Bourgeas, P. Ducrot, I. Theret, L. Xuereb, M. J. Basse, J. M. Brunel, S. Combes, X. Morelli and P. Roche, submitted for publication.
23 J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad and R. G. Coleman, *J. Chem. Inf. Model.*, 2012, **52**, 1757–1768.
24 R. D. Clark, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1181–1188.
25 A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch and H. Waldmann, *J. Chem. Inf. Model.*, 2007, **47**, 47–58.
26 A. Schuffenhauer, N. Brown, P. Ertl, J. L. Jenkins, P. Selzer and J. Hamon, *J. Chem. Inf. Model.*, 2007, **47**, 325–336.
27 M. E. Welsch, S. A. Snyder and B. R. Stockwell, *Curr. Opin. Chem. Biol.*, 2010, **14**, 347–361.
28 N. A. Ali, B. A. Dar, V. Pradhan and M. Farooqui, *Mini-Rev. Med. Chem.*, 2012, **22**, 98–119.

29 C. R. Prakash and S. Raja, *Mini-Rev. Med. Chem.*, 2012, **12**, 98–119.
30 V. R. Solomon and H. Lee, *Curr. Med. Chem.*, 2011, **18**, 1488–1508.
31 Beena and D. S. Rawat, *Med. Res. Rev.*, 2012, DOI: 10.1002/med.21262.
32 V. Facchinetti, C. R. Gomes, M. V. de Souza and T. R. Vasconcelos, *Mini-Rev. Med. Chem.*, 2012, **12**, 866–874.
33 G. Marzaro, A. Guiotto and A. Chilin, *Expert Opin. Ther. Pat.*, 2012, **22**, 223–252.
34 P. C. Sharma, G. Kaur, R. Pahwa, A. Sharma and H. Rajak, *Curr. Med. Chem.*, 2011, **18**, 4786–4812.
35 V. Kumar, M. Mishra, S. K. Rajput, S. Bajpai and R. K. Singh, *Parasitol. Res.*, 2012, **111**, 1851–1855.
36 J. B. Bariwal, K. D. Upadhyay, A. T. Manvar, J. C. Trivedi, J. S. Singh, K. S. Jain and A. K. Shah, *Eur. J. Med. Chem.*, 2008, **43**, 2279–2290.
37 F. Lovering, J. Bikker and C. Humblet, *J. Med. Chem.*, 2009, **52**, 6752–6756.
38 Y. Yang, O. Engkvist, A. Llinàs and H. Chen, *J. Med. Chem.*, 2012, **55**, 3667–3677.
39 D. Fry, *Small-Molecule Inhibitors of Protein–Protein Interactions*, Springer, Nutley, New Jersey, 2011.
40 C. Lipinski, F. Lombardo, B. Dominy and P. Feeney, *Adv. Drug Delivery Rev.*, 2001, **46**, 3–26.
41 P. D. Leeson and S. A. St-Gallay, *Nat. Rev. Drug Discovery*, 2011, **10**, 749–765.
42 J. Zuegg and M. A. Cooper, *Curr. Top. Med. Chem.*, 2012, **12**, 1500–1513.
43 D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, *J. Med. Chem.*, 2002, **45**, 2615–2623.
44 M. M. Hann, A. R. Leach and G. Harper, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 856–864.