Focused chemical libraries – design and enrichment: an example of protein–protein interaction chemical space

One of the many obstacles in the development of new drugs lies in the limited number of therapeutic targets and in the quality of screening collections of compounds. In this review, we present general strategies for building target-focused chemical libraries with a particular emphasis on protein–protein interactions (PPIs). We describe the chemical spaces spanned by nine commercially available PPI-focused libraries and compare them to our 2P2I3D academic library, dedicated to orthosteric PPI modulators. We show that although PPI-focused libraries have been designed using different strategies, they share common subspaces. PPI inhibitors are larger and more hydrophobic than standard drugs; however, an effort has been made to improve the drug-likeness of focused chemical libraries dedicated to this challenging class of targets.

Why is there a trend toward focused libraries?
The pharmaceutical industry is at a crossroads. This revolution is taking place at several levels: active policies regulating pharmaceutical spending, diminishing returns in R&D, health authorities establishing new requirements for drug approvals and ‘new approaches to science’ becoming more and more complex. This deep transformation has led pharmaceutical companies to reposition their strategies, including slowing expenditures on R&D. As a direct consequence of these multiple changes, the number of drugs that receive approval every year is a barometer of how well the pharmaceutical industry is facing these new challenges. 27 new drugs were approved in 2013, ten fewer than the year before; the unexpected increase since 2010 was not maintained. More disturbingly, in the last 5 years, only 11 molecules were issued annually, on average, by the ‘big pharmaceutical companies’ [1]. Thus, the need to rationalize R&D at every stage of the pipeline is clearly apparent.

To face these changes, two challenging sources of improvement are currently being considered in the early stages of drug discovery: on the one hand, the selection of new relevant targets, and on the other hand, the quality of the compound collections used in screening campaigns. In other words, the success of the experimental screening techniques used to identify innovative molecules depends equally on the specificity of the input, such as the quality of target selection, and on the chemical libraries used to find the ’hidden gems’ or active compounds that act on the selected targets.

The definition of the ‘ideal’ screening library has evolved over the years. In the early years of development, the pharmaceutical industry was guided by the concept of ‘one fits all’, in other words, the aim was to design very large and highly diverse screening collections that could be tested against all protein targets in exhaustive high-throughput screening (HTS). In an effort to reduce attrition rates in drug discovery, the quality of these libraries was improved by applying simple filters related to the concept of drug-likeness. Medicinal chemical knowledge based on the analysis of the physicochemical and absorption, distribution, metabolism, excretion and toxicity (ADMET) properties of known small-molecule drugs has led to the general acceptance of some simple rules of thumb filters, such as Lipinski’s ‘rule of five’ (Ro5) [2], Veber’s rule [3], Oprea’s rule [4] or the Pfizer
**Key terms**

**Drug-likeness**: Different rules of thumb have been proposed to assess the drug-likeness of chemical compounds. These rules are usually applied to remove, from screening collections, compounds that are unlikely to be developed as drugs due to their poor absorption, distribution, metabolism, excretion and toxicity properties. They are derived from the properties of known drugs or from bioavailability measurements of putative drugs.

**Descriptors**: Molecular descriptors are experimentally measured or calculated parameters that reflect the physicochemical and structural properties of small molecules. Numerous descriptors of various level of complexity are available today. They have been classified into 1D, 2D or 3D categories according to the dimensionality of the structure that is needed to calculate them.

**Quantitative estimate of drug-likeness**: To circumvent the use of simple cut-offs, quantitative estimate of drug-likeness, reflecting the distribution of eight physiologically relevant molecular descriptors, has been proposed. Quantitative estimate of drug-likeness provides a score between 0 and 1, which allows compounds to be ranked. It has been used to quantify the ‘chemical beauty’ of drugs.

**Target-focused chemical library**: Correspond to collections of small molecule compounds that have been selected for their propensity to interact with a family of related targets. Targets within the family can be rather similar (such as kinases) or more diverse (such as protein–protein interactions). They are used to find primary hit compounds in high-throughput screening campaigns.

3/75 toxicity rule [5], which are used to remove undesired compounds from screening collections [6]. These rules rely on the application of simple cut-offs for some standard molecular descriptors such as molecular weight (MW), octanol-water partition coefficient (LogP) or topological polar surface area (TPSA). More quantitative estimates of drug-likeness have been developed to avoid the use of binary cut-offs [7–11].

Although highly effective, screening strategies based on massive and highly diverse collections of compounds involve huge costs and major resources, and the ability to perform such studies is therefore limited to only a handful of screening facilities worldwide. In addition, this global approach has delivered lower hit rates than expected, especially for emerging classes of targets. Depending on the assay, hit rates typically range between 0.1 and 5% [12]. Focusing chemical libraries toward a particular family of targets represents an effective alternative to improve the hit discovery process [13,14].

Historically, target-focused chemical libraries have been designed mainly to address G-protein-coupled receptors (GPCRs), kinases and nuclear receptors. More recently, as knowledge has grown regarding cellular protein interaction networks and their role in numerous cell disorders, protein–protein interactions (PPIs) have emerged as a new class of promising therapeutic targets [15–17]. The increasing amount of structural data concerning both PPI targets and their small-molecule inhibitors has resulted in the development of dedicated structural databases [18–22]. On the one hand, analysis of the properties of protein–protein interfaces, especially hotspots and interfacial pockets, has resulted in the development of scoring functions and dedicated servers to assess the druggability of PPIs [23–30]. On the other hand, characterization of the profiles of PPI disruptors through the detailed analysis of their physicochemical properties has allowed the development of PPI-focused libraries [15,18,22,31–37].

**General methods for the construction of focused chemical libraries**

Designing target-focused chemical libraries requires 3D structural knowledge for various members of the target family and/or a set of known active small molecules to drive compound selection. This structural knowledge is accessible through the constant development of structural-biology validation methods. Today, there are more than 90,000 protein structures in the protein data bank (PDB) [38–41], including approximately 1000 3D structures of drugs from the Drugbank database bound to their respective targets [42]. Meanwhile, more and more active compounds are continuously being discovered and characterized through HTS or drug-discovery programs.

Focused chemical libraries were first developed for GPCRs, protein kinases, proteases, nuclear receptors or polymerases, which correspond to the most studied therapeutic targets that have led to marketed drugs. Hundreds or even thousands of small-molecule modulators have been experimentally tested, characterized and optimized for these targets. Furthermore, data can often be cross-correlated inside one target family because key features of the substrate binding mode or mode of action of these enzymes are generally conserved across the entire family. Serine/threonine protein kinases are perfect examples of these properties. Kinases are among the most represented enzymes in the PDB and the serine/threonine kinase subfamily represents the third most abundant family with nearly 1500 PDB entries. Lessons learned from one tyrosine kinase can often be extended to the entire family, or at least to the closest members. In the case of such large target families, the design of focused libraries can be driven by a ‘divide and conquer’ strategy based on the types of inhibitors and their inhibitory mechanisms [43–45].

Focused chemical libraries are generally designed by selecting compounds from larger and more diverse collections of compounds using chemoinformatics tools based on ligand knowledge or target knowledge or,
ideally, a combination of both. The rationale for these approaches is briefly discussed below (more exhaustive descriptions of these methods can be found in the literature [46–55]).

**Ligand-based techniques**

The most popular approaches to the design of focused target-specific libraries are based on the properties of known active compounds. Ligand-based techniques have a number of inherent advantages; for instance, protein structures are not needed and only a set of reference active compounds is required. Ligand-based methods can be divided into descriptor- or pharmacophore-based techniques and shape recognition-based techniques.

**Descriptor- or pharmacophore-based techniques**

Descriptor-based techniques rely on the use of various parameters that describe the constitutional, electronic, geometrical, topological and molecular properties of small molecules. The basic assumption of descriptor-based techniques is that molecules with similar descriptors share similar properties and, therefore, similar activities towards a given target family. There is a wide range of possible descriptors with different levels of complexity (1D, 2D and 3D) and pharmacophores can be considered as advanced 3D descriptors. Creating a focused library requires the selection of relevant descriptors and/or pharmacophores that can best segregate potential ligands of the targeted family from among all other small molecules. Descriptor selection is most frequently accomplished through computer-assisted data mining using support vector machines, decision trees or neural networks [56,57].

Ligand-based selection is generally performed by applying similarity-search methods based on a known active chemical scaffold to a large and diverse database. This process is often referred as ‘scaffold hopping’, and relies on rapid and powerful database scanning for comparisons and alignments of molecular fingerprints to identify similar chemical entities. 2D and 3D ligand pharmacophore alignments can also be used to identify similarity subsets in the database.

**Shape-recognition-based techniques**

Given the assumption that shape complementarity is a prerequisite for the binding of a ligand to its target, it is reasonable to suppose that molecules with similar shapes could be active against similar targets. Shape-based virtual screening has therefore become a popular method of identifying small-molecule compounds that resemble a reference structure known to be active against a protein target of interest. Even if shape-recognition algorithms were not originally developed to address the design of family-focused libraries, they are well-suited to this task because they offer rapid computation and efficient computer-resource consumption. Based on known structures of some ligands for the target family, one can perform rapid 3D-shape searches in large chemical databases and retrieve compounds that share similar shapes with known ligands. Shape-recognition-based techniques can be divided into two subcategories: superposition- and nonsuperposition-based techniques. One of the most widely used algorithms for superposing and comparing compounds is Rapid Overlay of Chemical Structures [58–61]. This program was designed to perform large-scale 3D database searches using a superposition method that finds similar but nonintuitive compounds. The two most popular nonsuperposition-based algorithms are EShape3D and Ultrafast Shape Recognition (USR) [62], which rely on eigenvectors and moments, respectively, as descriptors. In addition to their high-speed efficiency, the advantage of these methods is that the descriptors that are used are independent of molecular orientation and position. USR is an extremely fast algorithm (it has been shown to be able to compare a query compound to a database of 690 million conformers in seconds to minutes), which has prompted the addition of pharmacophoric descriptors and topological information to the USR algorithm to increase the pertinence of the selected compounds [63–65].

Shape-similarity techniques are powerful because of their speed, but, more importantly, they often result in the identification of new inhibitors with innovative chemical scaffolds that can be used as novel leads in drug discovery.

**Target-based techniques**

3D structural data of protein targets have been used for many years to bias library design because of the progress in docking methods that are able to correctly predict the correct binding poses and energies of interactions. Docking is a successful approach for hit identification in structure-based virtual screening [66,67]. A high-throughput docking strategy is feasible for family-focused design if the targeted family contains a conserved active site. Gozalbes et al. [68] have developed a systematic high-throughput docking strategy to design a kinase-focused library. These authors docked 123 structurally diverse kinase ligands into three representative kinase 3D crystal structures, which led to the determination of score thresholds for each kinase that a query compound must meet to be assessed as a kinase-like inhibitor. Proof of concept was first established through virtual screening of two collections of compounds: a collection of more than 2500 drugs and drug-like compounds as a negative control, and a kinase-targeted library of 1440 compounds. The
strategy was then experimentally validated by testing 60 compounds from the kinase-targeted library on 41 kinases from five different families. The 60 compounds were split into those that passed all thresholds and those that did not (30 compounds in each group). The overall hit enrichment was 6.7-fold higher in the first group, thus validating the approach for the generation of kinase-targeted libraries and the identification of scaffolds with high kinase inhibitory potential [68].

However, the application of docking to the design of focused libraries suffers two main limitations: it requires knowledge of the 3D structure or a high quality model of the target; and it incurs a high computational cost. Unlike ligand-based techniques, which are able to screen up to millions of compounds per second, docking is limited to only a few compounds per second. However, recent developments in graphics processing unit (GPU)-based docking algorithms have dramatically improved the computation efficiency and scalability, thereby permitting the use of the docking approach on a larger scale [69,70].

Ballester et al. have combined ligand- and target-based techniques by using both USR and docking on two versions of an antibacterial target [71]. Using three known ligands as a query for the USR search in the 9 million-compound ZINC database, the authors selected 4379 compounds of similar shape and then processed them in four various hierarchical screenings using docking protocols. The 148 compounds with the highest docking scores were screened in vitro against the two dehydroquinases targets. The protocols identified 100 new inhibitors with calculated $K_i$ values ranging from 4 to 250 $\mu$M. Most importantly, more than 50 new active molecular scaffolds were discovered, illustrating the benefits that a wide application of prospectively validated in silico screening tools is likely to offer in antibacterial hit identification [71].

An example: PPI-focused libraries

Importance of PPIs as therapeutic targets

The need for innovation in drug discovery challenges pharmaceutical research to look beyond traditional targets in order to place pharmaceutical R&D back on track. Among the possible approaches to addressing this necessary selection of new relevant targets, PPI modulation is attracting considerable interest as knowledge increases regarding cellular protein interaction networks and their role in numerous cell disorders [72]. The human interactome, which is estimated to consist of between 130,000 and 650,000 PPIs [73,74], represents a large reservoir of potential new therapeutic drug targets. This perspective has propelled a new surge of research on this class of targets, which is often regarded as being poorly druggable because of the 3D structural and biophysical complexity of the interfacial features of transiently formed complexes.

Despite these difficulties, PPIs are becoming more accepted and popular targets because of the increasing number of successes in the development of PPI inhibitors [75–99]. A recent survey of approximately 50 screening platforms from academia, small biotech companies and pharmaceutical industries has revealed that PPIs are screened on almost all platforms [100]. However, the low hit rates that are generally observed in HTS campaigns against PPI targets have guided a general effort in both academia and industry toward the design of dedicated libraries to cover the chemical space of PPI inhibitors.

Academic efforts to design chemical libraries

The earliest efforts to develop small-molecule compounds capable of disrupting PPIs were based on the mimicry of secondary structural elements of the interacting protein partner, such as beta-turns, alpha helices, beta strands or polyproline helices [101–103]. These peptidomimetics, designed to reproduce the orientations of key amino-acid side chains, constitute an excellent source of compounds for the construction of PPI-focused libraries. However, this class of compound is restricted to protein–peptide interactions and cannot be used for globular PPIs. In the last decade, a large number of nonpeptidomimetic small-molecule compounds capable of disrupting protein–protein complexes have been identified. Several studies have been undertaken to characterize the chemical properties of these known PPI inhibitors by comparing them to other small-molecule compounds [15,18,22,31,32,34,37,111–113]. On average, PPI modulators are relatively hydrophobic, rigid, large (high MW), non-planar and non-linear compounds that often contain multiple aromatic residues that differentiate them from standard drugs. These characteristic properties have been used to design PPI-focused libraries by extracting putative PPI inhibitors from a large screening collection of compounds using supervised learning methods, such as decision trees [31,34,112] or support vector machines [37]. We have developed a medicinally oriented diverse PPI-focused library by applying our 2P2I3DHUNTER SVM algorithm to 8.3 million compounds from the main commercially available chemical providers [36]. This library of 1664 compounds has been plated and is currently being tested against several structurally diverse protein–protein targets, including PDZ domains, bromodomains and protein–peptide interactions. The properties of the 2P2I3DHUNTER academic library are discussed below in comparison with available commercial libraries.

An alternative approach has been developed by Fry et al. [35]. Instead of using information regard-
ing the PPI inhibitors, these authors used structural information concerning protein–protein complexes that contain alpha-helical binding epitopes to guide the design of novel PPI scaffolds. The resultant PPI-targeting libraries were evaluated against a panel of PPI targets (MDM2, MDMX, BCL2, BCLXL, MCL1, BIR2, BIR3, JNK1 and NUR77) and revealed hits in the low–medium micromolar range corresponding to ligand-efficiency values in the range 0.15–0.24, which are reasonable for PPI inhibitors [18,120].

Commercially available libraries
Several commercial PPI-focused libraries containing from 378 to 125,418 compounds that were designed using machine learning approaches, rules of thumb such as Ro4, scaffold hopping or dedicated synthesis are available from several providers (Table 1). Here, we present the collections of small-molecule compounds specifically dedicated to PPI targets that can be easily purchased.

Asinex
Asinex (Moscow, Russia) has used dedicated synthesis to design active compounds against PPI targets that lack potential ADMET and solubility liabilities. Asinex used an algorithm based on combining highly hydrophilic 3D-like scaffold cores that are enriched in hydrogen-bond acceptors and donors with PPI-specific lipophilic peripheries to build a collection of 11,177 compounds.

ChemDiv
The design of the ChemDiv (CA, USA) PPI libraries was primarily inspired by the concept of ‘escaping from flatland’ [121]. The global library of PPI modulators consists of 125,418 compounds subdivided into 11 subsets corresponding to specific PPI targets (e.g., MDM2, PDZ and CD16A), chemical families of scaffolds (cyclic Ugi-based compounds, spiro compounds and eccentric compounds) or shapes of recognition elements (beta-turns, helices, beta-sheets, strands and loops). The Eccentric subset contains 6684 nonreactive compounds that correspond to original scaffolds that lie outside the field of heteroaromatic compounds [122]. This subset has been shown to cover a chemical space similar to the one described by our $2P2I_{HUNTER}$ SVM model [36].

Life Chemicals
Life Chemicals (ON, Canada) offers three PPI-focused libraries. The similarity PPI library (23,532 compounds) was prepared by comparing the Life Chemicals stock collection to compounds in the Timbal database [18,21], using a Tanimoto similarity cut-off of 85%. The ‘Ro4’ PPI library (4364 compounds) was obtained by applying the Ro4 filter to the screening collection of compounds [15]. A decision-tree approach adopted from the literature with two descriptors (RDF070m and Ui) was used to design the machine learning set, which is composed of 869 compounds [31,32]. In the

<table>
<thead>
<tr>
<th>Provider</th>
<th>Library</th>
<th>Source</th>
<th>Method</th>
<th>Number of compounds</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asinex (Moscow, Russia)</td>
<td>PPI</td>
<td>Commercial</td>
<td>Dedicated synthesis</td>
<td>11,177</td>
<td>[114]</td>
</tr>
<tr>
<td>ChemDiv (CA, USA)</td>
<td>PPI</td>
<td>Commercial</td>
<td>PPI-biased chemistry</td>
<td>125,418</td>
<td>[115]</td>
</tr>
<tr>
<td>ChemDiv Eccentric</td>
<td>Commercial</td>
<td>PPI-biased chemistry</td>
<td>6684</td>
<td>[115]</td>
<td></td>
</tr>
<tr>
<td>Life Chemicals (ON, Canada)</td>
<td>Similarity</td>
<td>Commercial</td>
<td>Similarity search</td>
<td>23,532</td>
<td>[116]</td>
</tr>
<tr>
<td>Life Chemicals Machine learning</td>
<td>Commercial</td>
<td>Decision tree</td>
<td>869</td>
<td>[116]</td>
<td></td>
</tr>
<tr>
<td>Life Chemicals Ro4</td>
<td>Commercial</td>
<td>Ro4</td>
<td>4364</td>
<td>[116]</td>
<td></td>
</tr>
<tr>
<td>Otava (ON, Canada)</td>
<td>Tree™</td>
<td>Commercial</td>
<td>Decision tree</td>
<td>1332</td>
<td>[117]</td>
</tr>
<tr>
<td>Otava</td>
<td>Analog™</td>
<td>Commercial</td>
<td>Similarity search</td>
<td>1027</td>
<td>[117]</td>
</tr>
<tr>
<td>Otava BRD4</td>
<td>Commercial</td>
<td>BRD4 focused</td>
<td>373</td>
<td>[118]</td>
<td></td>
</tr>
<tr>
<td>iSCB/CRCM (Marseille, France)</td>
<td>2P2I3D</td>
<td>Academic</td>
<td>Machine learning and Ro4</td>
<td>1664</td>
<td>[119]</td>
</tr>
</tbody>
</table>

For each library, the provider, name of the collection, source, general methods to prepare the library, number of compounds and link to the website of the provider are given.
PPI: Protein–protein interaction; Ro4: Rule of four.
original papers, it was demonstrated that RDF070m partially correlates with MW at MW <400, and, therefore, compounds with MW ≥400 were discarded from this library.

**Otava**

The PPI-focused libraries designed by Otava (ON, Canada) comprise compounds that were selected on the basis of recently published PPI inhibitor models and scaffold hopping. The ‘Tree™ library’ (1332 compounds) was designed using a decision-tree algorithm based on several molecular-shape and functional-group descriptors [112]. The ‘Analogs™ library’ (1027 compounds) was prepared using a similarity search (ECFP4 fingerprints and a Tanimoto cut-off of 40%) with known active PPI inhibitors from the TIMBAL database [18,21]. These two libraries were filtered using Lipinski rules while considering the higher MW and LogP values of known PPI inhibitors. The MW constraints were set between 300 and 700 g. mol⁻¹, and the LogP cut-off was between 1 and 6. In addition, compounds with reactive groups as well as promiscuous inhibitors, were removed. The ‘receptor-based BRD4 inhibitors focused library’ was designed by docking each compound into the binding site of the BRD4 x-ray crystal structure (PDB ID: 4J0S).

**PPI chemical-space coverage**

**Standard drug-like molecular descriptors**

Standard molecular descriptors that are commonly used to assess the drug-likeness of small-molecule compounds were calculated for the nine commercial libraries and for our academic library 2P2I 3D (Figure 1 & Table 2). PPI inhibitors are known to present high MW compounds. However, the median MW value is below 500 g. mol⁻¹ for all chemical libraries, and 90% of the compounds exhibit MWs that range from 300 to 600 g. mol⁻¹, which reflects the desire to design libraries with medicinally driven potential. It is important to note that although the high MW of PPI inhibitors has long discouraged medicinal chemists, some recent successes of orally available drugs, such as navitoclax from Abbott Laboratories (IL, USA; in Phase 2 despite a MW of 974.6 g. mol⁻¹), have demonstrated that high MW is not necessarily an obstacle to the development of drugs against PPI targets [124–127]. The vast majority of PPI inhibitors in the focused libraries possess fewer than five hydrogen-bond donors (a median of approximately one) and fewer than ten hydrogen-bond acceptors (a median of approximately six). Most compounds in PPI-focused libraries are compliant with Lipinski’s ‘Ro5’ for these two descriptors. The low number of hydrogen-bond donors is related to the hydrophobic nature of protein–protein interfaces with few hydrophilic residues. PPI inhibitors are considered to be highly hydrophobic compounds as reflected in the relatively high LogP values. For all libraries, the median MLogP value is below 4.15, as represented by the cut-off value in Lipinski’s ‘Ro5’. Although PPI inhibitors are hydrophobic compounds, the currently available PPI-focused libraries were prepared such that most compounds with high LogP values were discarded. Although PPI inhibitors are generally described as anti-Ro5 compounds, analysis of the available PPI libraries demonstrates that a large number of compounds are compliant with Lipinski’s ‘Ro5’ with values ranging from 81.2 to 98.9%, with an average of 91.7 ± 6.2% (Figure 1, left panels). Moreover, it should be noticed that there is currently a general tendency to incorporate compounds that exhibit larger MWs and higher hydrophobicities than those of orally available drugs into commercial screening collections [128].

Veber et al. have shown that a polar surface area of less than or equal to 140 Å² and a number of rotatable bonds of fewer than or equal to ten are correlated with oral bioavailability in rats [3]. The median values for these two criteria in the PPI-focused libraries are below the cut-offs defined by Veber (Figure 1 & Table 2). As a consequence, the percentage of compounds that follow Veber’s rule is relatively high, ranging from 64.4% for the Eccentric library from Chemdiv, to 100% for the machine learning set from Life Chemicals with an average of 93.4 ± 10.5% (Table 3 & Figure 2).

According to the definitions of Lipinski et al. [2], Veber et al. [3], Oprea [4], Walters and Murcko [129], and Rishston [130], the Eccentric library contains less drug-like compounds than the other PPI-focused libraries (Table 3 & Figure 2). However, even for this library, which contains new and unusual ring scaffolds [122], the majority of the compounds satisfy the various drug-likeness conditions.

It has been shown that PPI inhibitors contain more rings than standard drugs [15,18,32], which could be correlated with the high propensity of aromatic amino acids to act as hot-spots at protein–protein interfaces [133]. Many small-molecule inhibitors have therefore been designed to mimic interfacial interactions, including pi/pi, CH/pi and hydrophobic packing interactions. Most compounds in the available PPI libraries contain between three and five rings (Figure 1 & Table 2).

PPI inhibitors are more three-dimensional and possess more asymmetric centers than compounds developed for other targets, such as enzymes and receptors [132]. PPI-focused chemical libraries have therefore been designed to ‘escape from flatland’ [121], as indicated by the fact that the Fsp3 ratio (i.e., the number of sp3 hybridized carbons divided by the total carbon count) is greater than 0.3 for most compounds in the
Figure 1. Physicochemical profile of compounds from protein–protein interaction libraries. Properties of the compounds from the various protein–protein interaction-focused libraries are compared using box plots for eight standard molecular descriptors. The thin bars represent the complete range distribution for a given property and boxes correspond to 90% of the compounds (from 5 to 95% of the distribution). Median values are shown through a color change (from dark to light orange), whereas average values are indicated as dark brown circles. Cut-off values used in standard drug-likeness rules are highlighted using dotted lines.

†Carbon bond saturation index corresponding to the number of sp³ hybridized carbons to the total number of carbon atoms.

MW: Molecular weight; Ro4: Rule of four; TPSA: Topological polar surface area.
PPI-focused libraries (Figure 1 & Table 2). Compounds from the three Otava libraries exhibit lower Fsp3 values, while ‘Ro4 library’ from Life Chemicals and our academic 2P2I 3D library, which has been filtered to remove most compounds with low Fsp3 values, contain more three-dimensional compounds.

On average, although the PPI-focused libraries were prepared using different approaches, they share some common properties, as indicated by the analysis of their standard molecular descriptors (Figure 1 & Table 2). To better compare the chemical space spanned by the various libraries, we used principal component analysis.

### Key terms

**Delimited reference chemical subspace**: Represent an easy way of visualizing and comparing chemical subspaces. They are used to create a visual contour representative of the densest portion of the chemical space spanned by a set of compounds in a reduced-dimension space.

**Chemical spaces**: The chemical space of a set of compounds can be seen as the region sampled by these molecules in the universe of all chemical compounds. It is highly dependent on the set of descriptors that are used to describe the compounds. They can be easily visualized using dimensionality reduction methods such as principal component analysis.

### Table 2. Average values and standard deviations of nine physicochemical parameters for each protein–protein interaction-focused library.

<table>
<thead>
<tr>
<th>Provider</th>
<th>Library</th>
<th>MW (g.mol⁻¹)</th>
<th>H-bond acceptors</th>
<th>H-bond donors</th>
<th>MLogP</th>
<th>Rotatable bonds</th>
<th>Rings</th>
<th>Multiple TPSA (Å²)</th>
<th>Fsp3†</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChemDiv (Moscow, Russia)</td>
<td>PPI</td>
<td>460.9 ± 35.4</td>
<td>6.0 ± 1.4</td>
<td>1.1 ± 1.1</td>
<td>2.5 ± 1.9</td>
<td>6.6 ± 1.9</td>
<td>4.4 ± 0.7</td>
<td>18.4 ± 3.6</td>
<td>71.3 ± 19.4</td>
</tr>
<tr>
<td>ChemDiv (CA, USA)</td>
<td>PPI</td>
<td>425.6 ± 63.1</td>
<td>6.2 ± 2.7</td>
<td>1.1 ± 1.1</td>
<td>2.8 ± 1.5</td>
<td>5.5 ± 2.2</td>
<td>4.0 ± 0.9</td>
<td>17.1 ± 4.6</td>
<td>76.1 ± 26.9</td>
</tr>
<tr>
<td>ChemDiv (ON, Canada)</td>
<td>PPI</td>
<td>414.4 ± 74.9</td>
<td>7.9 ± 2.0</td>
<td>1.1 ± 1.1</td>
<td>2.6 ± 1.6</td>
<td>6.8 ± 3.1</td>
<td>3.7 ± 0.0</td>
<td>18.5 ± 6.5</td>
<td>124.6 ± 38.5</td>
</tr>
<tr>
<td>Life Chemicals (ON, Canada)</td>
<td>Similarity</td>
<td>391.6 ± 59.0</td>
<td>6.3 ± 1.6</td>
<td>1.1 ± 1.1</td>
<td>2.3 ± 1.4</td>
<td>6.2 ± 1.8</td>
<td>36.6 ± 0.8</td>
<td>16.1 ± 4.2</td>
<td>95.7 ± 29.6</td>
</tr>
<tr>
<td>Life Chemicals (ON, Canada)</td>
<td>Machine</td>
<td>383.4 ± 14.9</td>
<td>6.2 ± 1.1</td>
<td>1.0 ± 0.9</td>
<td>3.2 ± 1.0</td>
<td>4.9 ± 1.3</td>
<td>38.8 ± 0.6</td>
<td>20.1 ± 2.4</td>
<td>79.0 ± 12.7</td>
</tr>
<tr>
<td>Life Chemicals (ON, Canada)</td>
<td>Learning</td>
<td>465.1 ± 44.1</td>
<td>6.8 ± 1.3</td>
<td>1.0 ± 0.7</td>
<td>3.0 ± 1.5</td>
<td>6.4 ± 1.9</td>
<td>4.3 ± 0.6</td>
<td>16.3 ± 3.0</td>
<td>100.9 ± 26.6</td>
</tr>
<tr>
<td>Life Chemicals (ON, Canada)</td>
<td>Ro4</td>
<td>330.3 ± 44.1</td>
<td>5.6 ± 1.3</td>
<td>1.0 ± 0.9</td>
<td>2.7 ± 1.3</td>
<td>5.7 ± 2.1</td>
<td>31.1 ± 0.8</td>
<td>16.6 ± 4.2</td>
<td>91.1 ± 29.2</td>
</tr>
<tr>
<td>Life Chemicals (ON, Canada)</td>
<td>Tree™</td>
<td>390.3 ± 46.9</td>
<td>6.1 ± 1.4</td>
<td>1.0 ± 0.9</td>
<td>2.6 ± 1.4</td>
<td>5.5 ± 2.1</td>
<td>35.7 ± 0.7</td>
<td>17.5 ± 2.9</td>
<td>210.5 ± 27.5</td>
</tr>
<tr>
<td>Life Chemicals (ON, Canada)</td>
<td>Analogy™</td>
<td>382.7 ± 46.9</td>
<td>6.1 ± 1.4</td>
<td>1.0 ± 1.0</td>
<td>2.6 ± 1.1</td>
<td>5.5 ± 2.1</td>
<td>35.7 ± 0.7</td>
<td>17.5 ± 2.9</td>
<td>210.5 ± 27.5</td>
</tr>
<tr>
<td>Life Chemicals (ON, Canada)</td>
<td>Machine</td>
<td>383.4 ± 14.9</td>
<td>6.2 ± 1.1</td>
<td>1.0 ± 0.9</td>
<td>3.2 ± 1.0</td>
<td>4.9 ± 1.3</td>
<td>38.8 ± 0.6</td>
<td>20.1 ± 2.4</td>
<td>79.0 ± 12.7</td>
</tr>
<tr>
<td>Life Chemicals (ON, Canada)</td>
<td>Machine</td>
<td>498.9 ± 46.2</td>
<td>7.4 ± 1.7</td>
<td>1.0 ± 1.0</td>
<td>2.7 ± 1.6</td>
<td>8.3 ± 2.0</td>
<td>45.9 ± 0.9</td>
<td>22.0 ± 4.4</td>
<td>103.4 ± 30.0</td>
</tr>
<tr>
<td>Life Chemicals (ON, Canada)</td>
<td>Machine</td>
<td>518.5 ± 63.8</td>
<td>7.0 ± 1.9</td>
<td>2.0 ± 1.2</td>
<td>17.2 ± 2.6</td>
<td>96.2 ± 27</td>
<td>4.4 ± 1.0</td>
<td>16.3 ± 3.8</td>
<td>102.8 ± 30.6</td>
</tr>
<tr>
<td>iSCB/CRCM (Marseille, France)</td>
<td>2P2I 3D</td>
<td>518.5 ± 63.8</td>
<td>7.0 ± 1.9</td>
<td>2.0 ± 1.2</td>
<td>17.2 ± 2.6</td>
<td>96.2 ± 27</td>
<td>4.4 ± 1.0</td>
<td>16.3 ± 3.8</td>
<td>102.8 ± 30.6</td>
</tr>
</tbody>
</table>

The various molecular descriptors were calculated with Dragon 6 [123].

†Carbon bond saturation index corresponding to the number of sp3 hybridized carbons to the total number of carbon atoms.

<table>
<thead>
<tr>
<th>Provider</th>
<th>Library</th>
<th>MW (g.mol⁻¹)</th>
<th>H-bond acceptors</th>
<th>H-bond donors</th>
<th>MLogP</th>
<th>Rotatable bonds</th>
<th>Rings</th>
<th>Multiple TPSA (Å²)</th>
<th>Fsp3†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future Med. Chem. 2014 6(11)</td>
<td>Review</td>
<td>Zhang, Betzi, Morelli &amp; Roche</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1298

**Future Med. Chem.** (2014) 6(11)
A general overview of the various libraries confirms that all PPI libraries cover a similar subspace of the chemical space, as could be expected for target-specific libraries (cyan contour in Figure 4, middle top panel). There is an obvious tendency to incorporate mainly drug-like compounds in commercial libraries, as evidenced by the chemical space they occupy, which primarily lies in the drug-like subspace. This is

<table>
<thead>
<tr>
<th>Provider</th>
<th>Library</th>
<th>Lipinski (%)</th>
<th>Veber (%)</th>
<th>Oprea (%)</th>
<th>Walters (%)</th>
<th>Rishton (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asinex (Moscow, Russia)</td>
<td>PPI</td>
<td>93.5</td>
<td>97.8</td>
<td>86.8</td>
<td>95.0</td>
<td>96.9</td>
</tr>
<tr>
<td>ChemDiv (CA, USA)</td>
<td>PPI</td>
<td>94.1</td>
<td>97.8</td>
<td>89.7</td>
<td>96.5</td>
<td>96.8</td>
</tr>
<tr>
<td>ChemDiv</td>
<td>Eccentric</td>
<td>81.2</td>
<td>64.4</td>
<td>63.0</td>
<td>73.9</td>
<td>74.4</td>
</tr>
<tr>
<td>Life Chemicals (ON, Canada)</td>
<td>Similarity</td>
<td>98.2</td>
<td>99.6</td>
<td>94.2</td>
<td>99.0</td>
<td>98.4</td>
</tr>
<tr>
<td>Life Chemicals</td>
<td>Machine learning</td>
<td>95.6</td>
<td>100.0</td>
<td>97.6</td>
<td>99.6</td>
<td>99.7</td>
</tr>
<tr>
<td>Life Chemicals</td>
<td>Ro4</td>
<td>92.1</td>
<td>98.5</td>
<td>85.1</td>
<td>94.6</td>
<td>95.1</td>
</tr>
<tr>
<td>Otava (ON, Canada)</td>
<td>Tree™</td>
<td>98.9</td>
<td>99.5</td>
<td>97.1</td>
<td>97.2</td>
<td>97.4</td>
</tr>
<tr>
<td>Otava</td>
<td>Analogs™</td>
<td>96.4</td>
<td>98.6</td>
<td>95.0</td>
<td>99.2</td>
<td>99.0</td>
</tr>
<tr>
<td>Otava</td>
<td>BRD4</td>
<td>84.9</td>
<td>91.7</td>
<td>73.7</td>
<td>84.5</td>
<td>87.6</td>
</tr>
<tr>
<td>iSCB/CRCM (Marseille, France)</td>
<td>2P2I3D</td>
<td>82.3</td>
<td>85.7</td>
<td>67.9</td>
<td>78.5</td>
<td>83.5</td>
</tr>
</tbody>
</table>

Minimum                      | 81.2 | 64.4 | 63.0 | 73.9 | 74.4         |
Maximum                      | 98.9 | 100.0 | 97.6 | 99.6 | 99.7         |
Average ± SD                 | 91.7 ± 6.2 | 93.4 ± 10.5 | 85.0 ± 11.9 | 91.8 ± 8.9 | 92.9 ± 7.9 |

For each library, the percentage of compounds that comply with standard drug-like rules is indicated. The different scores were computed using Dragon 6 [123]. PPI: Protein–protein interaction; Ro4: Rule of four; SD: Standard deviation.
particularly true for the machine learning and similarity sets from Life Chemicals, Otava BRD4 and, to a slightly lesser extent, ChemDiv. This observation can be correlated with high percentages of low MW compounds in these libraries (Figure 1). Concerning the academic library, although the majority of the compounds lie within the drug-like subspace, a significant number of compounds lie outside this chemical space extending in the nondrug-like and pharmaceutical subspaces. The DRCS analysis using MOE2D descriptors clearly indicates that most compounds from the various PPI libraries occupy a specific chemical space that lies between drug-like and nondrug-like regions. This observation reflects both the different properties of PPI inhibitors compared with standard drugs (higher MW and higher LogP) and the desire, especially by commercial providers, to remove potential ADMET and solubility liabilities by applying drug-like filters in the construction of screening libraries.

**Future perspective**

In this work, we reviewed the growing interest in the design of focused libraries, with a particular emphasis on PPI-targeted libraries. We determined that the vast majority of libraries focused on PPI that are currently available through commercial providers were developed using pharmacological or physicochemical filters, in addition to machine-learning methods using the general profile of PPI inhibitors as a guide. These filters or algorithms were applied to extract potential PPI inhibitors from the large screening collections of the major providers, which had previously been filtered through other drug-likeness filters, such as the well-known Lipinski’s ‘Ro5’. Thus, the resulting compounds from these libraries are a selection of existing...
Focused chemical libraries – design & enrichment: an example of protein–protein interaction chemical space

Review

Figure 4. Projection of protein–protein interaction-focused libraries on delimited reference chemical subspace contours. The various PPI libraries were projected in the MOE 2D delimited reference chemical space as described in Colliandre et al. [139] using the open-source software screening assistant [141,142]. Compounds are shown by black dots. The Prestwick library, which contains 1200 small-molecule drugs, is shown as a reference. Three contours representing drug-like (Lipinski+, in blue), non-drug-like (Lipinski- in green) and pharmaceuticals (in red) that were defined in the original paper from a set of more than 6.5 million compounds are shown. The contour representing the PPI subspace defined in this review from all available libraries is shown with a cyan contour (top middle panel). The cumulated variance for the first two principal components is 44% (54% for the first three components).

PC1: First principal component; PC2: Second principal component; PPI: Protein–protein interaction.

compounds (a selected subpopulation) more than a de novo synthesis of compounds according to rules identified through the analysis of successfully validated PPI modulators. This important bias is clearly an obstacle to innovation and certainly the Achilles’ heel limiting the expected success of addressing this challenging class of targets. The discovery of innovative compounds with original and dedicated scaffolds will certainly occur once the field is sufficiently mature to yield compounds that have not been proposed in the current libraries.

The available interactome is vast compared with the number of PPI targets that can presently be modulated using small molecules. We are, therefore, far from covering the entire chemical space of druggable PPI targets. As knowledge increases, specifically structural information regarding both the target and the inhibitors, it will become possible to better characterize the
general profile of inhibitors for each family of PPI targets, which will lead to the development of more specific and selective compounds. As a result, it will be possible to design more efficient PPI-focused libraries dedicated to each class of PPI targets.

In the meantime, an obvious track for the design of new original scaffolds is related to the three-dimensionality of the compounds. There is already a general tendency toward attempting to escape from ‘flatland’ by removing the flat and hydrophobic compounds that pollute commercial libraries and generate false positives or nonspecific hits. Compounds with a higher number of asymmetric carbons (natural products or those obtained through organic synthesis) constitute another possible track that should be investigated in greater detail in the future. These innovations (e.g., three-dimensionality, number of asymmetric carbons or new original scaffolds) represent a way to avoid the clear trend toward the development of compounds of higher MW, increased hydrophobicity and greater number of aromatic rings. The main challenge in the future most likely lies in the hands of organic and medicinal chemists in the field who are striving to design novel compounds dedicated

---

### Executive Summary

#### Why is there a trend toward focused libraries?
- The number of new molecular entities approved annually has been in constant decline over the last 50 years. A possible track to overcome this crisis in the pharmaceutical industry involves the expansion of the number of druggable targets and the proper design of the collection of compounds.
- The quality of the collection of compounds used in high or medium throughput screening campaigns is essential to increase hit rates, lower attrition rates and reduce the cost of drug development. It is commonly admitted nowadays that chemical libraries need to be diverse. For the design of medicinally oriented libraries, absorption, distribution, metabolism, excretion and toxicity filters are generally applied to remove undesired compounds.
- The design of small or medium size diverse focused-libraries dedicated to a particular class of targets is regarded as an efficient way to enhance the quality of small molecule compound collections when prior knowledge regarding reference compounds or the target is available. Methods generally used to focus chemical libraries involve ligand-based and structure-based approaches.

#### Protein–protein interaction-focused libraries
- Despite their reputation for being poorly druggable targets, protein–protein interactions (PPI) are emerging as a new class of innovative targets. They are involved in numerous biological pathways and cell disorders, such as cancer pathogenesis, which makes them an attractive class of therapeutic targets.
- Historically, the design of focused libraries dedicated to PPI involved modified peptides and peptidomimetics that reproduce secondary structural elements at the protein interface. PPI are considered as difficult targets for the design of small-molecule inhibitors through drug discovery because of the properties of the binding interface. A major breakthrough came with the characterization of interfacial hotspot residues accounting for most of the binding energy between the two partners. Since then, thousands of small-molecule inhibitors have been developed for more than 60 targets. Several dedicated databases have been developed to gather knowledge on both PPI druggable targets and their modulators.
- A general profile of a PPI inhibitor has been defined based on the increasing number of success stories. Compared to standard drugs, PPI modulators are larger, more hydrophobic and more three-dimensional, with a greater number of rings and higher aromaticity.
- Several commercially available PPI-focused libraries have been designed using machine learning approaches, rules of thumb such as the ‘rule of four’, scaffold hopping or dedicated synthesis. We have built 2P2I3D, a small academic diverse PPI-focused library, by applying a supervised vector machine algorithm on 8.3 million compounds representing the major chemical providers commercially available. Compounds that did not contain privileged scaffolds identified as core structures in numerous therapeutics and compounds with low three-dimensionality were excluded from the final library composed of 1664 small molecules.

#### PPI chemical-space coverage
- Delimited reference chemical space analyses revealed that, although PPI-focused libraries presented in this review have been prepared using different strategies, most compounds share a common subspace that lies between ‘drug-like’ and ‘nondrug-like’ regions. The various libraries exhibit a relatively high percentage of drug likeness, showing that undesirable compounds have usually been removed during the preparation of the chemical libraries.

#### Future perspective
- Compounds in most PPI-focused libraries available today were extracted from large screening collections. A better characterization of the structure and dynamics of protein interfaces, through success stories in drug development, will result in the design of more potent and more selective modulators. In return, the general profile of PPI inhibitors will be better characterized and it will be possible to design more efficient focused libraries dedicated to subclasses of PPI targets. The next generation of small molecules dedicated to PPI targets should involve the design of new original scaffolds specifically addressing the specific chemical space of this difficult class of target.
to this difficult class of targets, which most certainly represents a bright future for the pharmaceutical industry.

Acknowledgements
We would like to thank P Bonnet, S Bourg and V Le Guilloux from the group of Structural Bioinformatics and Chemoinformatics (Institut de Chimie Organique et Analytique [ICOA], UMR 7311) for helpful discussions regarding delimited reference chemical subspaces tools.

References
Papers of special note have been highlighted as:
• of interest; •• of considerable interest


• Describes the advancement in drug discovery efforts against protein–protein interactions, ‘the unmined biology gold reserve’.


• Describes the first hand-curated database dedicated to small molecule protein–protein interaction inhibitors and provides general molecular properties of the modulators.


23 Kozakov D, Hall DR, Chuang GY et al. Structural conservation of druggable hot spots in protein–protein interactions of interest; •• of considerable interest
Authors used a variety of small probe molecules in combination with conformational sampling of the target to define and predict druggable sites at the protein interface. Sugaya N, Furuya T Jr. PLAS: an integrative system for assessing the druggability of protein–protein interactions. BMC Bioinformatics 12, 50 (2011).


Rose PW, Bi C, Buhlm BF et al. The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res. 41(Database issue), D475–D482 (2013).

DrugBank database. www.drugbank.ca


focused chemical libraries – design & enrichment: an example of protein–protein interaction chemical space


70 Wu J, Hong B, Takeda T, Guo JT. High performance transcription factor-DNA docking with GPU computing. *Proteome Sci.* 10(Suppl. 1), S17 (2012).


92 Silvian L, Enyedy I, Kumaravel G. Inhibitors of protein–protein interactions: new methodologies to tackle


• Represents the first attempt at characterizing the chemical space spanned by protein–protein interaction inhibitors using principal component analyses.


• In this pioneer review article, the authors analysed the 3D structures of six discontinuous protein–protein interfaces, for which small molecule inhibitors had been developed, to characterize the properties of druggable targets and the associated small molecule inhibitors.

114 ASINEX’s Protein–Protein Interactions (PPI) Library. [www.asinex.com/PPILibrary.html](http://www.asinex.com/PPILibrary.html)


118 Bromodomain Containing Protein 4 Brd4-Focused Library. [www.otavachemicals.com/targets/bromodomain-containing-protein-4-brd4-focused-library](http://www.otavachemicals.com/targets/bromodomain-containing-protein-4-brd4-focused-library)

119 2P2I Hunter. [http://2p2idb.cnrs-mrs.fr/2p2i_hunter.html](http://2p2idb.cnrs-mrs.fr/2p2i_hunter.html)


• The authors demonstrate that three-dimensionality and the presence of chiral centers both correlate with success as compounds progress from discovery, through clinical testing, to drugs.


123 Dragon 6. [www.talete.mi.it](http://www.talete.mi.it)


125 Ackler S, Mitten MJ, Foster K et al. The Bcl-2 inhibitor ABT-263 enhances the response of multiple chemotherapeutic
Focused chemical libraries – design & enrichment: an example of protein–protein interaction chemical space

Review


- Describes the method to define delimited reference chemical subspace using principal component analyses and its application to comparing chemical libraries.
