Rationalizing the chemical space of protein–protein interaction inhibitors

Olivier Sperandio[1,2], Christelle H. Reynès[2], Anne-Claude Camproux[2] and Bruno O. Villoutreix[1,2]

[1] CDithem Platform/IGM, 27 rue Juliette Dodu, 75010 Paris, France
[2] Inserm UMR-S973/MTi, University Paris Diderot, 35 Rue Hélène Brion, 75205 Paris Cedex 13, France

Protein–protein interactions (PPIs) are one of the next major classes of therapeutic targets, although they are too intricate to tackle with standard approaches. This is due, in part, to the inadequacy of today's chemical libraries. However, the emergence of a growing number of experimentally validated inhibitors of PPIs (i-PPIs) allows drug designers to use chemoinformatics and machine learning technologies to unravel the nature of the chemical space covered by the reported compounds. Key characteristics of i-PPIs can then be revealed and highlight the importance of specific shapes and/or aromatic bonds, enabling the design of i-PPI-enriched focused libraries and, therefore, of cost-effective screening strategies.

Protein–protein interactions (PPIs) are associated with the most crucial biological processes in life, and their malfunction can be tracked to numerous disease states. Among the estimated 650,000 interactions that regulate human life [1], a sizable number should be druggable [2–7], as suggested by the growing number of PPI systems successfully targeted by drug-like compounds and the recent progress of two compounds to clinical trials [8]. Although a vast array of high-throughput, fragment-based and *in vitro* and *in silico* screening technologies have been developed over the past 15 years [9], identifying PPI modulators is still challenging [3,5–7,10,11] for multiple reasons. Unlike G-protein-coupled receptor (GPCRs) or enzymes, PPIs have not evolved to bind small molecules, so the druggability of the system studied has to be probed in a case-by-case scenario. The nature of the interfaces (i.e. flat, large, with or without preformed cavities and with the presence of water molecules) and their diversity (obligate or non-obligate, homo- or hetero-multimers) make them a highly versatile class of targets. Furthermore, the assessment of their druggability might guarantee the identification of good binders but cannot guarantee the identification of a good drug [12].

Corresponding author:. Sperandio, O. (olivier.sperandio@univ-paris-diderot.fr)

Deriving information about i-PPIs from interface properties

In the light of previous studies of protein–protein interface properties and of some known inhibitors, a series of observations can be made in an attempt to shed light on the nature of the PPI inhibitor (i-PPI) chemical space while providing hints about the conceptual and technical hurdles ahead. First, the seminal study of Clackson and Wells [13] defined 'hot-spot regions', thereby clarifying some key misconceptions about the likelihood of finding drug-like i-PPIs – namely, the size of the interface is not crucial and the energetic of the complex formation does not disqualify small molecular weight molecules. Hot spots are responsible for most of the binding free energy, and the surface area of these 'high-affinity regions' is much smaller than the entire interface and, therefore, compatible with the overall dimension of a drug-like compound. Second, the interface region, in general, consists of a core and a rim region; the amino acid composition of the rim region is similar to the one of the surface of the rest of the protein, and the core region is enriched in aromatic residues [10]. Several studies have underlined the preponderance of tyrosine, tryptophan, phenylalanine and methionine at the protein–protein interface [14–16]. These investigations have provided drug designers with crude but important indications about the chemistry that is probably required to inhibit PPIs (i.e some aromatic and/or hydrophobic functions).

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2009.11.007

Along this line of reasoning but this time from the small molecule side, aromatic structures were often noticed on i-PPIs, although a possible correlation between the physicochemical nature of the interface region and of the i-PPI chemical space has never been truly investigated [3,6,17].

ADME/Tox characteristic of i-PPIs

From an energetic standpoint (and still from the ligand side), the ligand efficiency coefficient used to analyze protein–ligand interactions during the course of drug discovery programs has been revisited for i-PPIs and estimated to be 0.24 kcal/mol per heavy atom [5]. This is lower than most kinase inhibitors (0.3–0.4 kcal/mol) but on the same order of magnitude of many protease inhibitors (0.25–0.35 kcal/mol). Consequently, an i-PPI with a K_d of 10 nM is expected to have a molecular weight (MW) of approximately 645 Da when regular orally available drugs, most of the time, are below 500 Da. The well-known Lipinski's rule of five (RO5) is, thus, violated and the situation worsens as one investigates natural i-PPIs often characterized by MW above 1000 Da [6], while other submicromolar i-PPI compounds tend to be in the 500–900 Da range [5]. However, the recent clinical trial of the orally bioavailable ABT-263 (MW of 975 Da) [18] suggests that the RO5 and related absorption, distribution, metabolism, excretion and toxicity (ADME/Tox) concepts might need to be tailored to i-PPIs, either in terms of allowing MW violations or in terms of gaining greater knowledge about pharmacokinetic/pharmacodynamic events authorizing larger compounds to reach the desired targets without severe reactions with antitargets.

Deriving i-PPI properties from interface plasticity

Although a significant plasticity can be observed on large protein–protein interfaces (>1500 Å^2), this plasticity is less pronounced than for the rest of the protein surface [19]. In terms of drug development, the design of constrained or more rigid i-PPIs might be beneficial to the affinity because a more flexible ligand could incur a larger energetic (entropic) penalty. The same is true if the inhibitor can bind the apo-conformation of the target without any protein-induced fit. Nevertheless, it should not be assumed that the binding site for an inhibitor can be systematically observed from static structures of either the apo-protein or the protein–protein complexes [5]. Flexibility, however, is known to be a major hurdle in most rational drug design programs [4,20–26]. This plasticity is expected in many PPIs, as illustrated by more than 30 diverse inhibitors of the p53/MDM2 interaction. This suggests that it could be beneficial to treat the problem at a more global level to avoid being trapped in the chemistry of the ligands or in the structure of the receptor. As such, it might be possible to define some essential physicochemical properties that would need to be present in any i-PPI. In the case of p53/MDM2, the existence of a pharmacophore known as the 'thumb–index–middle' finger designed to mimic the α-helix motif of p53 that binds the N-terminus of MDM2 [27] demonstrates the efficacy of combining global properties such as shape and polarity to address the wide chemical variety of i-PPIs in this system.

Rationalizing the i-PPI chemical space to build focused libraries

Although the i-PPI chemical space is still mostly an unknown territory, analysis of known i-PPIs via statistical means and with a medicinal chemistry perspective could open new ways to assist the rational design of i-PPIs. Indeed, if one assumes that a crucial step toward the design of i-PPIs is to identify hits and that high-throughput screening using conventional compound collections will usually lead to very low hit rate, a crucial move is to devise a tool that improves the quality of the starting library. A possible solution here is to follow the focused library concept [28] used for regular targets (e.g. enzymes and GPCRs) but to shape it to the singularity of i-PPIs. As in regular targets, it should be possible to minimize the biomolecular or *in silico* screening that would be required to successfully target PPIs by designing a focused library enriched in i-PPIs. But this supposes to apply empirical or statistically derived rules that would be soft enough to tackle generic PPIs and, thus, be target independent. This exercise would realign the chemical space window of compound collections with the chemical requirements of i-PPIs and enhance the hit rate, allowing investigators to focus on hit optimizations rather than on screening yet another collection. In addition, and keeping in mind the benefit of focused collections, the overall approach tailored to i-PPIs should not only reduce waste by eliminating *a priori* compounds that are unlikely to impede protein–protein complex formation but also lead to enhanced potency and/or specificity of the binders. To this end, it is interesting to learn from the past developments of i-PPIs and attempt to gather precious information about their chemical space specificities. The primary design of i-PPIs was based on a variety of chemical scaffolds, but some key properties were rapidly identified and initiated the quest for specific chemical motifs that could translate peptide-like i-PPIs into drug-like i-PPIs. These first i-PPIs, historically, have been heavier than regular drugs, more hydrophobic and more rigid with multiple rings in succession, and often contain aromatic functions [2,3,12,29–32] (Fig. 1).

More recently, two studies focused on the newly reported i-PPIs to determine their properties at a global level and in a target-independent manner. The chemical space of 19 published i-PPIs initially not filtered for ADME/Tox properties and targeting four PPI complexes was studied using MOE (http://www.chemcomp.com/) descriptors (namely topological surface area, $S \log P$ and MW) in a principal component analysis (PCA) [30]. The study used a 3D projection of the PCA with the Chemical Diversity database, the MayBridge database and the Asinex database and showed that approximately 50% of the 19 i-PPIs were covered by the diversity space of the three commercial databases. The application of ADME filters (Lipinski's RO5) reduced the size of the Chemical Diversity database from 119,475 to 85,248 compounds and the list of 19 i-PPIs to 8 compounds with only three PPI systems (Bak-BH3/Bcl-xL, p53/MDM2 and NGF/p75) well covered by the chemical space of the filtered Chemical Diversity database. This study essentially highlighted the fact that today's commercial libraries are not fully adequate for targeting PPI for therapeutic purposes.

More recently, Neugebauer *et al.* [33] used machine learning strategies to profile i-PPIs in an attempt to define a specific physicochemical pattern. They used a set of 25 true i-PPIs and a set of 1137 non-i-PPI molecules and built a decision tree to select three descriptors able to discriminate true i-PPIs from non-i-PPIs. The most relevant descriptor, SHP2, used at the top of the decision tree was alone responsible for a tenfold enrichment of the learning dataset (the others were another shape descriptor and the number

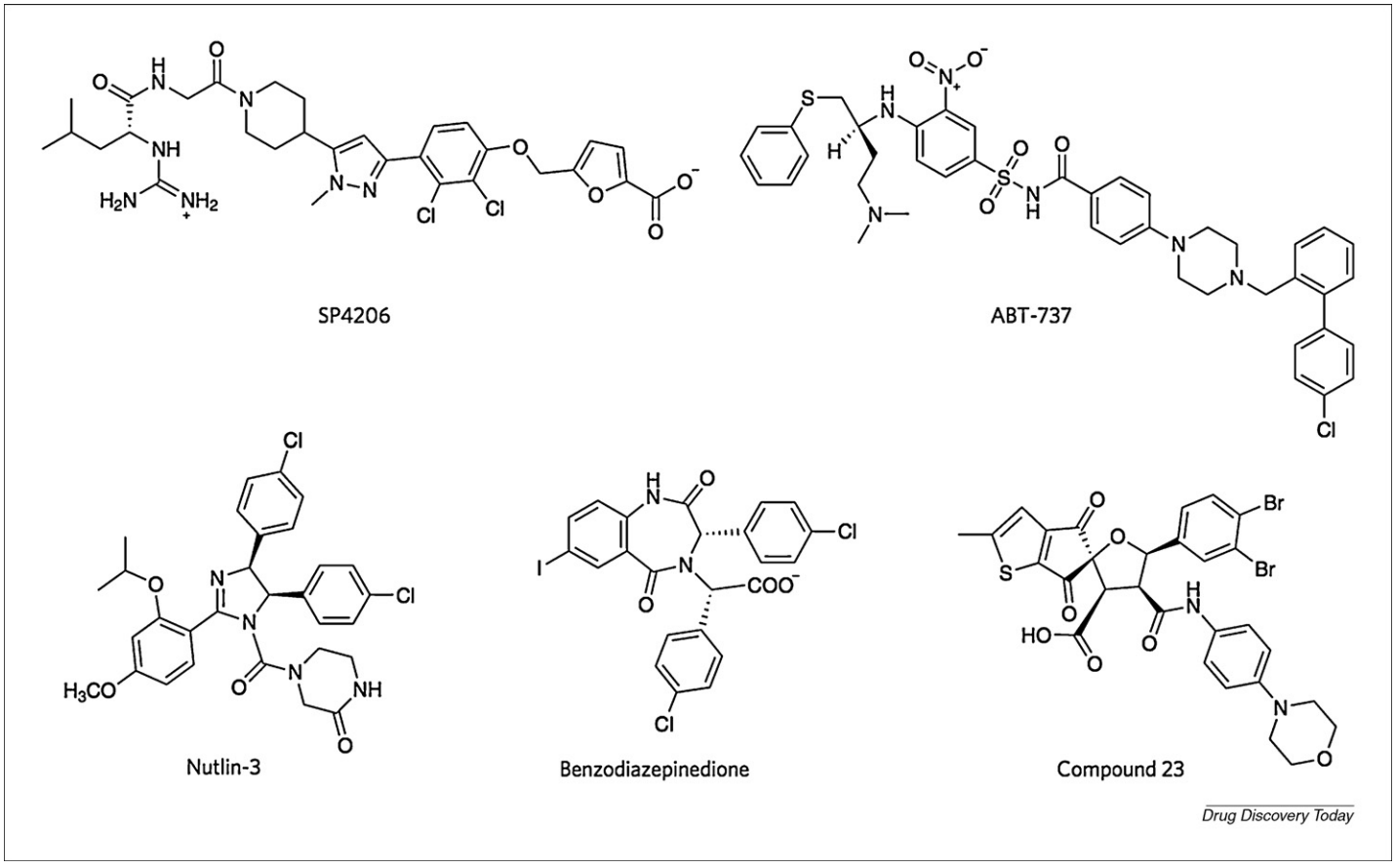


FIGURE 1

Examples of five i-PPIs with a $K_i < 1$ μM, a MW above 500 Da and at least three aromatic rings.

of ester function). SHP2 refers to a molecular shape descriptor introduced by Randic [34,35] and further suggests the importance of a shape factor when designing i-PPIs.

Revisiting i-PPIs: case study in a diverse collection test set

The following analysis aims to revisit the chemical space of i-PPIs to define rules that could enable cost-effective hit screening with improved prospects for clinical success.

Descriptive analysis

To further describe and quantify the physicochemical properties of i-PPIs versus regular drugs in a target-independent manner and to determine rules for designing i-PPI-enriched chemical libraries, we have gathered two subsets of compounds: 145 experimentally validated i-PPIs and the 4857 existing drugs taken from the small subset of DrugBank (Sept 2008). Both subsets were ADME/Tox filtered ($100 < MW < 900$; $0 <$ H-bond donors < 8; $0 <$ H-bond acceptors < 12; $-5 < A \log P < 6$; $0 <$ nb Rotatable bonds < 20; $0 < TPSA < 160$; plus one allowed rule violation) and processed through a chemical-fingerprint-based clustering approach (pairwise Tanimoto index < 0.8) to ensure reasonable chemical diversity. This whole process obtained 66 diverse drug-like i-PPIs and 557 'regular' drugs. A series of 1666 molecular descriptors was then calculated using the program E-Dragon to obtain as much information as possible to characterize the two subsets and attempt to distinguish them (http://www.talete.mi.it/). The distribution of some of these key descriptors is represented in Fig. 2, along with their box plot and the associated *P*-values (Student's *t*-test) of the descriptors for the two working subsets (66 i-PPIs and 557 drugs). We also reported the *P*-values of the *t*-test of the same descriptors when normalized by the MW to compare the same properties of the two subsets at equivalent MW.

As seen from the distributions, the MW of i-PPIs is significantly heavier than those of regular drugs ($mean_{i\text{-}PPI} = 421$ Da; $mean_{Drugs} = 341$ Da; *P*-value = 4.923E−09). The same is true for the octanol/water partition coefficient ($A \log P$), which confirms that i-PPIs tend to be more hydrophobic than regular drugs ($mean_{i\text{-}PPI} = 3.58$; $mean_{Drugs} = 2.61$; *P*-value = 5.706E−06). The median value (appropriate parameter when dealing with discrete descriptors) for H-Bond acceptors shows that half of the i-PPIs have at least six H-Bond acceptors as opposed to drugs that have only five (*P*-value = 0.00483). Interestingly, the median value for the number of H-Bond donors is the same for the two subsets with one H-Bond donor for at least half of them in each case. By contrast, the median value for the number of rotatable bonds is one point higher for the i-PPIs (nb Rotatable bonds = 6) as opposed to drugs (nb Rotatable bonds = 5). This might seem to violate 'the rigidity factor' that was previously thought to be crucial for i-PPIs, but this difference is only slightly significant (*P*-value ~ 0.05, see the results below after MW normalization). The average value of the topological polar surface area (TPSA) is also higher for i-PPIs than for regular drugs ($mean_{i\text{-}PPI} = 89$; $mean_{Drugs} = 71$, *P*-value = 2.659E−05). This is in apparent contradiction with the $A \log P$ results, which express a

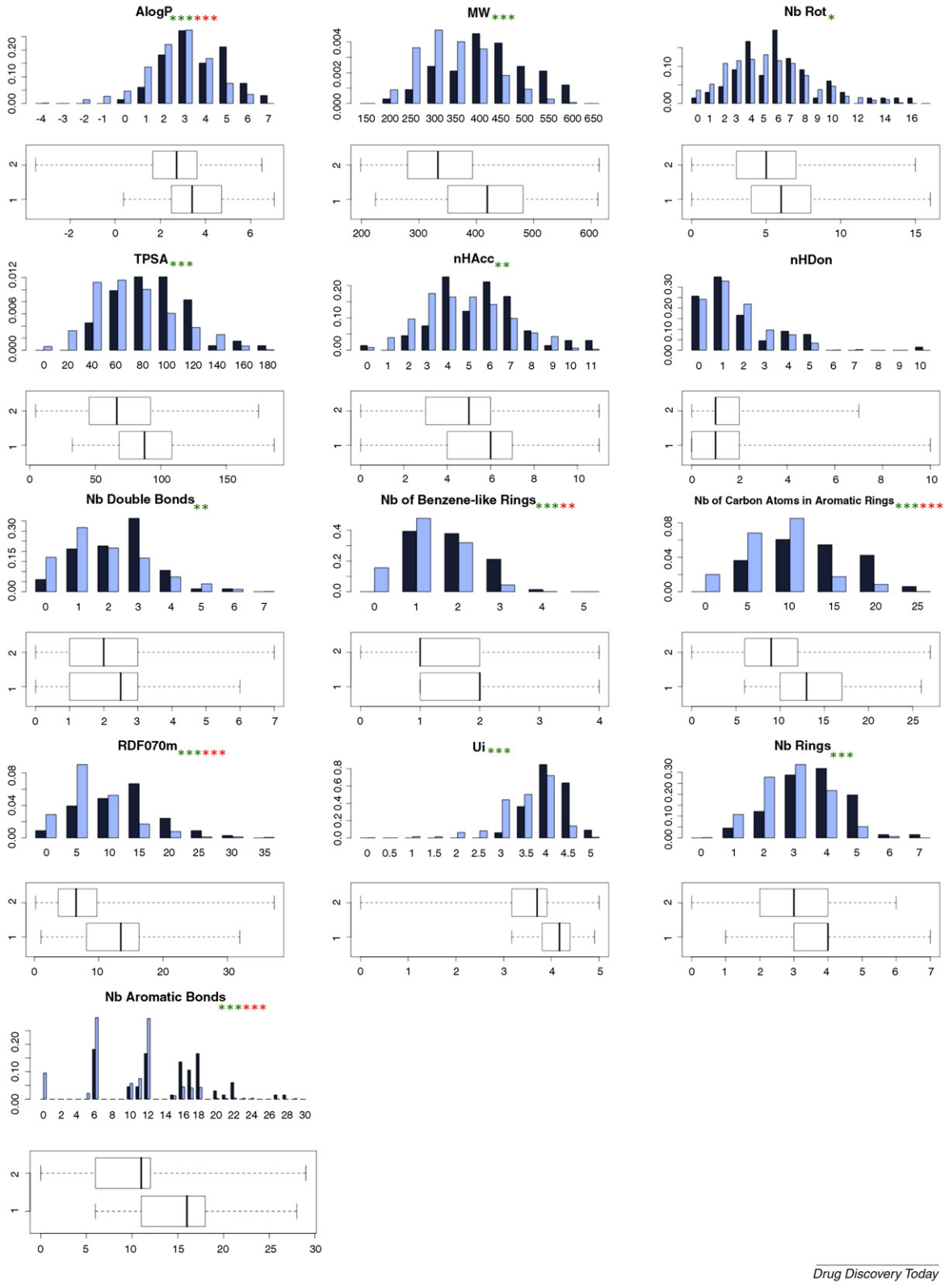


FIGURE 2

Distribution of several molecular descriptors calculated in the two subsets (66 i-PPIs [1 in box plot and dark in histogram] and 557 drugs [2 in box plot and cyan in histogram]). Descriptors are flagged with different degrees of discrimination based on their Student's *t*-test *P*-value, from not significant (no star) via modest, $0.01 < P\text{-value} < 0.05$ (*) and moderate, $0.001 < P\text{-value} < 0.01$ (**) to very significant, $P\text{-value} < 0.001$ (***). The same flags are associated with the same descriptors when normalized by the molecular weight but with red stars: *, ** and ***. The box plots indicate the minimum value, the first quartile, the median, the third quartile and the maximum value.

higher hydrophobic character for i-PPIs, but it has be to kept in mind that on average, i-PPI compounds are bigger, so they are more likely to have extra polar atoms. This might contribute more to the polar surface area, as opposed to log *P*, which is a macroscopic property that takes into account apolar and polar contributions to the overall lipophilicity of the molecules. It is, therefore, interesting to consider the *P*-value of the same descriptors when normalized by the MW to evaluate the difference of those properties at equivalent MW. In this case, one can see that the *A* log *P* difference is still significant ($P\text{-value}_{A\ \log P/\mathrm{MW}} = 0.0397$) but that TPSA is not ($P\text{-value}_{\mathrm{TPSA/MW}} = 0.52$) and nor is the number of rotatable bonds ($P\text{-value}_{\mathrm{Nb\ Rot/MW}} = 0.24$), thereby confirming previous observations about the generic chemical nature of i-PPIs. Among the other descriptors that separate the two subsets significantly (*P*-values < 0.05), we examined the median values of a series of interesting chemical descriptors. Here, more than half of the i-PPIs and the 'regular drugs' have, respectively, 17 and 12 multiple bonds (descriptor that contains the number of aromatic bonds), 2 and 1 benzene-like rings, and 13 and 9 carbon atoms in aromatic rings. All these descriptors remain significantly discriminative when normalized by the MW. This gives an interesting overview of an ideal i-PPI compound when combined with our previous observations for *A* log *P*, MW, TPSA and number of H-Bond acceptors and donors.

Proposition of decision trees

Decision trees are multivariate machine learning tools that are used to classify data items by posing consecutive questions (most of the times, 'yes or no' questions). The questions form a hierarchical structure encoded as a tree [36]. When the questions are

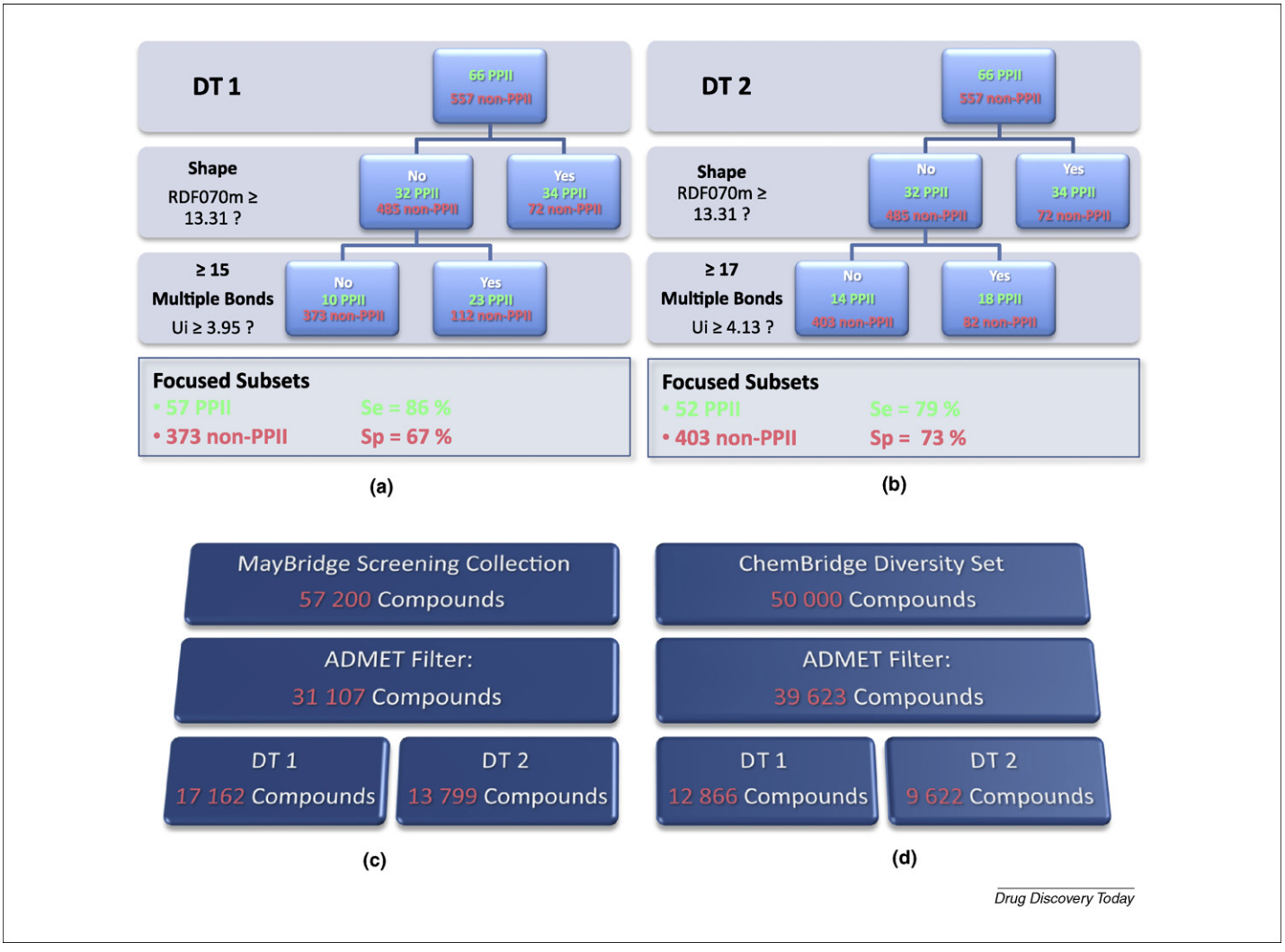


FIGURE 3

Definition and size reduction effect of the decision trees. **(a,b)** Application of the two decision trees on the two subsets (66 i-PPIs and 557 drugs). The two descriptors RDF070m (shape) and Ui (Nb of multiple bonds) are successively used to separate the subsets. One can see the correctly identified i-PPIs and regular drugs, for example, 57 i-PPIs out of 66 and 373 drugs out of 557 for DT 1. The two trees have complementary performances to separate the subsets. DT 1 has a higher sensitivity than DT 2 ($Se_{DT1} = 84.9\%$ versus $Se_{DT2} = 75.8\%$) (i.e. a better ability to identify true i-PPIs), whereas DT 2 has a higher specificity ($Sp_{DT1} = 70.2\%$ versus $Sp_{DT2} = 76.5$) (i.e. a better capacity to discard non-i-PPIs). For example, this means for DT 1 that 84.9% of the true i-PPIs have either a RDF070m > 13.31 or a UI > 3.95 that is at least 15 multiple bonds (aromatic or double bonds). **(c,d)** From the 57,200 molecules initially present in the MayBridge Screening collection, 31,107 molecules passed the soft ADME/Tox filter. Subsequently, 17,162 molecules passed the first tree (DT 1) and 13,799 passed the second tree (DT 2). A similar evaluation carried out on the diversity set of the ChemBridge database that initially contained 50,000 compounds led to an intermediate library of 39,623 compounds satisfying the ADME/Tox filters and, ultimately, 12,866 compounds with the first tree and 9622 compounds with the second tree.

molecular descriptor-associated thresholds, this brings the net advantage to provide to the medicinal chemists a comprehensive description of the relevant physicochemical features of the compounds by reducing the number of descriptors needed. In fact, decision trees can offer a significant advantage over other methods, such as support vector machine or artificial neural networks, that result from the combination of all the descriptors at once and usually lack interpretability [36].

To go one step further than a descriptive analysis of i-PPIs, we proposed two decision trees (DT) that were constructed with the aim of getting the best possible sensitivity and specificity, respectively (i.e. to profile the most efficiently true i-PPIs on the one hand and to discard non-i-PPIs on the other hand). Notably, the best combination of descriptors unraveled the two same descriptors for the two proposed trees, RDF070m and Ui, although with different thresholds. These two descriptors clearly have the ability

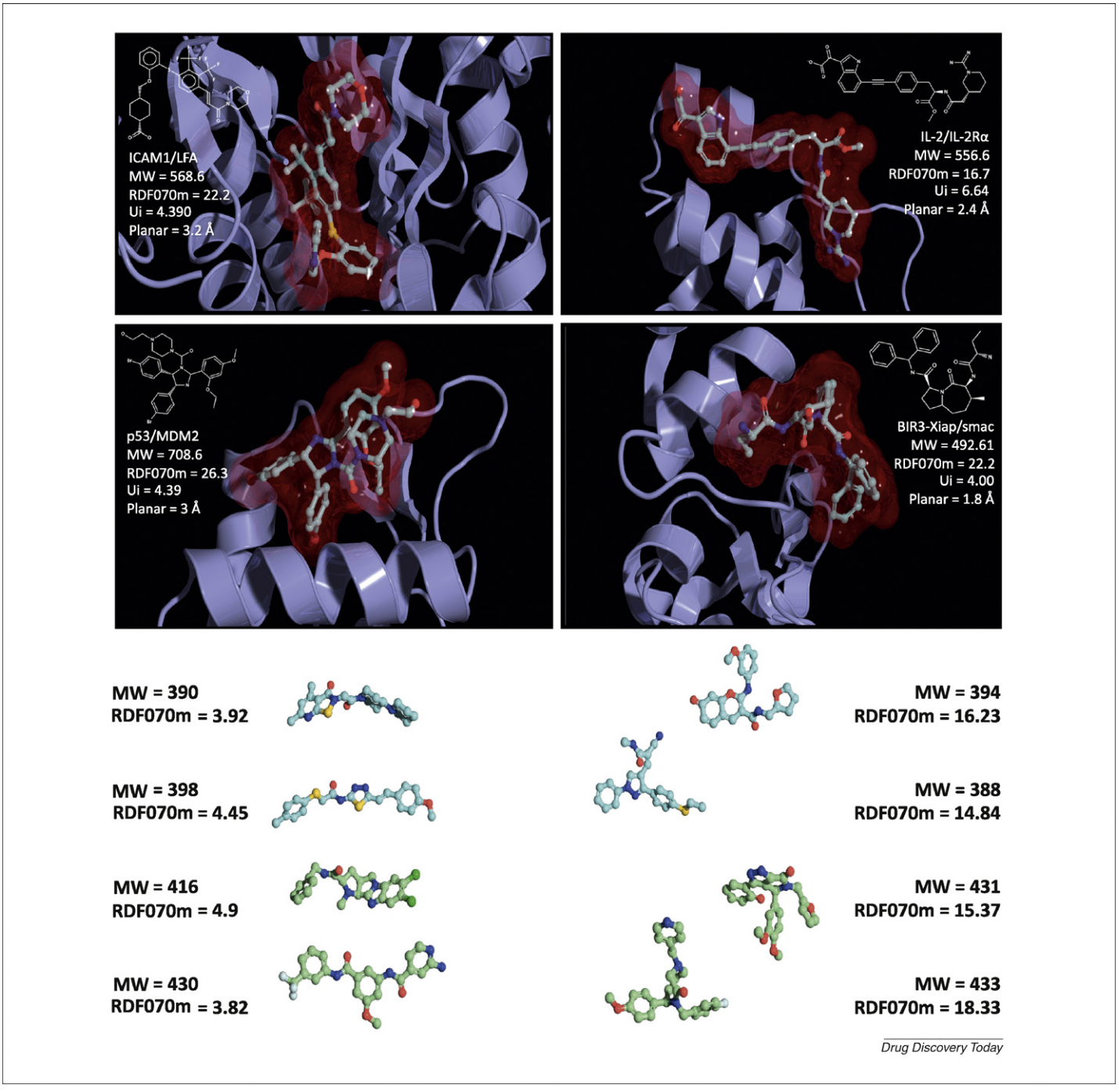


FIGURE 4

Importance of the RDF070m descriptor on molecular shape. Impact of the molecular shape on the value of the molecular descriptor RDF070m. In the upper half of the figure, four systems PPI with one of the partner and one synthetic inhibitor. The MW of the ligand, its RDF070m, its Ui for information and the planarity of the protein cavity are reported on the figure. In the lower half of the figure, four i-PPIs from the PubChem BioAssay AID1434 (cyan carbon) and four inactive compounds from the same assay (green carbon) are shown (modeled with the tool Omega2 http://www.eyesopen.com). One can see that I-shaped compounds have low RDF070m values and that star-, L- or T-shaped molecule have higher RDF070m values.

to distinguish the two subsets significantly, as suggested by their distributions and associated *P*-values ($P\text{-value}_{RDF070m}$ = 5.74E−08, $P\text{-value}_{Ui}$ = 3.256E−16) (Fig. 2). As can be seen in Fig. 2, RDF070m is capable of discriminating the two subsets significantly, even when normalized by the MW. This is not the case for Ui, but the two descriptors do not operate on the same populations, which is the main strength of decision trees as a multivariate method. RDF070m is a radial distribution function, or RDF (*r*), descriptor weighted by the atomic masses using a sphere radius *r* of 7 Å as the associated probability distribution function. Ui is the unsaturation index, directly linked to the number of multiple bonds, which contains double, triple and aromatic bonds. The two best decision trees are both based on RDF070m ($threshold_{RDF070m} > 13.31$) and Ui ($Threshold\text{-}DT1_{Ui} > 3.95$; $threshold\text{-}DT2_{Ui} > 4.13$) (Fig. 3a,b). Although the definition of the unsaturation index Ui is associated with float values, it relies on the sum of integers and, as such, can be traced back to a discrete number of multiple bonds. Thus, the Ui thresholds for DT 1 and DT 2 are 3.95 and 4.13, meaning, respectively, a minimal number of 15 and 17 aromatic or double bonds if one considers the number of triple bonds as negligible (with 0.1–0.6% of triple bonds on average on several databases).

We observed a poor correlation between RDF070m and Ui ($r^2_{RDF070m/Ui} = 0.34$), which confirms that they provide low redundancy and good complementarities in discriminating PPI from non-PPI inhibitors. Moreover, we noticed that RDF070m partially correlates with MW ($r^2_{RDF070m/MW} = 0.8$). However, this correlation can be observed only at lower MWs (MW < ~400), and it must be repeated that, as opposed to Ui, RDF070m remains significantly discriminative of true i-PPIs when normalized by the MW (*P*-value < 0.001, Fig. 2). This is important because RDF070m stands at the top of the trees. RDF descriptors are known as shape descriptors and are usually used as a multiple-value code calculated at different interatomic distance thresholds (here, we just have 7 Å) and can be associated with various atomic properties (like the atomic weight here, but it can be partial charges, polarizability, among others). These descriptors were successfully used to study active compounds on Vitamin D receptor [37], flavonoid compounds as inhibitors of aldose reductase [38]. Interestingly, these descriptors were also used to predict 3D structures from their infrared spectra in which specific substructures are, by definition, associated with a specific signal, such as the presence or absence of multiple bonds in a given region of the compounds [39,40]. Indeed, it seems that the distribution of the atomic masses and the shape of the molecules is one appropriate generic property of i-PPIs. Interestingly, RDF070m partially correlates with the shape descriptor identified by Neugebauer, SHP2 ($r^2_{RDF070m:SHP2} = 0.71$), in another subset of i-PPIs but was found to be more significant than SHP2 to discriminate true i-PPIs. Nevertheless, this confirms previous observations made on the importance of the molecular shape for i-PPIs. To empirically observe what type of impact RDF070m has on the shape of the compounds, we evaluated this value in four PPI complexes, one protein–protein complex and one synthetic inhibitor (ICAM1/LFA, IL-2/IL-2Rα, p53/MDM2 and Xiap-BIR3/smac) and in compounds taken from the PubChem Bioassay AID1434 (Fig. 4). It is clear in Fig. 4 that this descriptor has higher values when the molecules have more ramifications and are star-, L- or T-shaped. Conversely, I-shaped molecules have lower values. Finally, to further stress the prevalence of specific shapes observed within i-PPI structures, we noticed that several p53/MDM2 inhibitors satisfying the 'thumb–index–middle' finger pharmacophore cited above and that were not present in the two initial subsets (66 i-PPIs and 557 drugs) have high values for RDF070m.

The importance of the second yet unraveled descriptor, Ui, supports the more and more consensual concept of privileged i-PPI chemical properties such as aromatic rings, especially when considering the correlation coefficient obtained between Ui and the number of aromatic bonds ($r^2_{Ui/nAB} = 0.92$), the number of multiple bonds ($r^2_{Ui/nBM} = 0.95$) and, to a minor extent, the number of benzene-like rings ($r^2_{Ui/nBnz} = 0.75$). Interestingly, when one considers the proportion of multiple bonds in various databases and by looking at Fig. 2 and the distribution of multiple, aromatic and double bonds, some pertinent observations can be made. In an i-PPI compound, there are on average ~39.4% of multiple bonds, among which ~35.9% are aromatic bonds, 3.5% are double bonds and 0.58% are triple bonds. If a typical i-PPI has at least 15 multiple bonds, it has in fact ~13–14 aromatic bonds and ~1–2 double bonds. This obviously represents a rule of thumb more than a law.

Application to database size reduction

To evaluate the reduction effect of using the decision trees on commercial databases, the two models have been applied on the MayBridge Screening collection and the ChemBridge diversity set

BOX 1

Experimental assessment of the decision trees

Figure I. Application of the decision trees to three PubChem BioAssays: AID432, AID1434 and AID689. The three subsets have previously been filtered for ADME/Tox properties using the parameters cited in the text for the design of the two subsets (66 PPII and 557 drugs). The number of remaining drug-like compounds is specified for each tree, along with the number of remaining drug-like hits and the corresponding sensitivity (Se) and specificity (Sp).

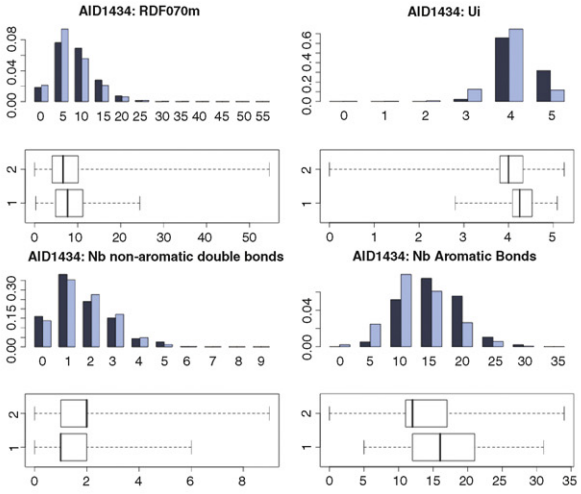
Figure II. The distribution of the two descriptors RDF070m and Ui applied to AID1434 shows that both descriptors are significantly discriminative toward true PPIIs ($P\text{-value}_{RDF070m}$ = 7.875E−06; $P\text{-value}_{Ui}$ = 2.2E−16) and that the number of aromatic bonds is by itself capable of significantly separating the two subsets ($P\text{-value}_{Nb\ Arom\ Bonds}$ = 2.2E−16), whereas the number of double bonds is not ($P\text{-value}_{Nb\ Doub\ Bonds}$ = 0.08507). Ultimately, it is important to consider in this largest dataset (117,533 molecules) that among the 894 drug-like hits identified by the assay and the 722 identified correctly by DT 1, 139 compounds were identified by the RDF070m (19.25%) descriptor and 583 by the Ui descriptor (81.75%). Similar observations could be made in the other screening assays (AID432 and AID689), such that the most important descriptor seems to be the unsaturation index and, therefore, the number of aromatic bonds. The results of the two DTs on the three PubChem Bioassays (Fig. 4b) suggest that both descriptors are significantly capable of separating true PPII from non-PPII and that Ui is by far the descriptor that identifies most of the true PPII (81.75%) as opposed to RDF070m (19.25%) confirming the prevalence of aromatic bonds in PPII.

Figure III. Scheme summarizing the application of the decision trees to a drug-like chemical library to create a focused library enriched in PPII using the two important descriptors relying on shape (RDF070m) and on π-electrons and aromatic bonds (Ui). Then follows the rest of the drug discovery pipeline with the screening of the focused library and the identification and optimization of hits to obtain drug candidates.

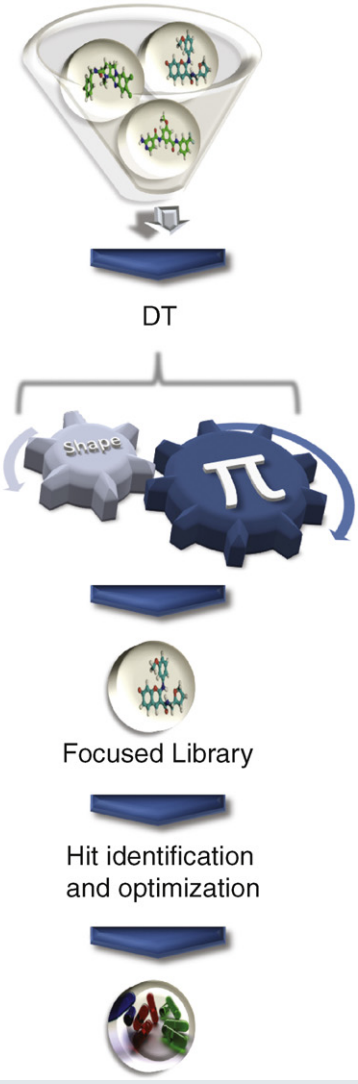
(i)

PubChem BioAssay	Total Nb of drug-like compounds	Total Nb of drug-like Hits	Decision Tree	Nb drug-like compounds	Nb remaining hits	Se (%)	Sp (%)
AID432 BFL-1/Bid	46,476	10	D.T. 1	26,429	8	80	43
			D.T. 2	21,400	7	70	54
AID1434 CBFb/CBFa	117,533	894	D.T. 1	69,635	722	81	41
			D.T. 2	56,265	621	70	52
AID689 EphA4/ephrin-A	37,114	38	D.T. 1	22,463	33	87	40
			D.T. 2	18,660	27	71	50

(ii)



(iii)



after ADME/Tox filtering, as mentioned above (Fig. 3c,d). In the case of the diversity set of the ChemBridge database, the use of the second tree represents a size reduction of more than 75% from the ADME/Tox version of the ChemBridge diversity set. This gives a concept of the financial cost saving that could be accomplished by screening the focused library as opposed to the initial library.

Application to i-PPI enrichment

Finally, to evaluate the pertinence of the decision trees in experimental screening results, they were applied to three PubChem Bioassay test sets: AID432, AID1434 and AID689 (Box 1a), which are experimental screening assays of three different PPIs (BFL-1/Bid, CBFb/CBFa and EphA4/ephrin-A). It is clear that the combination of RDF070m and Ui descriptors is capable of discriminating the true i-PPIs from the non-i-PPIs (Box 1b) relatively well and is, on average, with the first tree capable of identifying from 81% to 87% of the true i-PPIs while removing approximately 40% of the non-i-PPIs and with the second tree capable of identifying more than 70% of the true i-PPIs while removing more than half of the non-i-PPIs. The use of such trees, the schematic representations of

which are reported in Box 1c, can be useful to design focus chemical libraries in a PPI-independent manner. In this case, however, it is understandable that the level of reduction that can be obtained is not as good as the one observed in other systems such as GPCRs, ion channels and kinases. In the future, it could be valuable to design PPI-specific focused chemical libraries, but additional knowledge and investigations are needed to build relevant mathematical models.

Clues for the design of i-PPI-enriched focused libraries

The descriptors of the decision trees presented herein, RDF070m and Ui, are easy to evaluate because relying on simple chemical properties such as pairwise interatomic distances (RDF070m) and counts of multiple bonds (Ui). RDF070m is clearly related to molecular shape, and high values favor more ramified structures. Ui is a direct evaluation of the number of multiple bonds. An ideal i-PPI compound would need a crucial number of multiple bonds (15–17) and/or possess a molecular shape favoring a diversified spatial distribution of the R-groups around a central scaffold. Thus, we suggest that the decision trees presented herein could be of precious help in the design of focused libraries enriched in i-PPIs starting either from any commercial compound libraries or from a newly designed combined and/or virtual chemical library to bias their chemical space toward the specificities of i-PPIs. The trees could also be used in combination with the other physicochemical properties described here to assist hit-to-lead programs and favor specific chemical moieties.

Concluding remarks

The importance of PPIs in life and disease states and their potential as a credible and huge pool of putative therapeutic targets will necessitate changes in the way libraries and hits are designed. These macromolecular systems, although highly challenging, will become the object of major research investments. It is, therefore, of primary importance to ease the discovery of new chemical entities modulating PPIs. In achieving these aims, *in silico* tools and chemoinformatics technologies will have a major role. A proper understanding of the i-PPI chemical space is still ahead, but new ideas and concepts have started to emerge in this matter. The shape, hydrophobicity and aromaticity of i-PPIs seem to be a proper way to characterize them as a whole, but more effort should be put into extending our comprehension of their specificity. In addition, a more detailed characterization of i-PPIs based on the 3D topological nature of PPIs, depending on whether one deals with α-helices interacting with groove, inter-protein beta-sheets or a loop-based interaction, and so on will be required. Such dichotomy in the process of characterizing i-PPIs would probably help to design even more focalized chemical libraries and, therefore, alleviate the discovery of novel drugs in a cost-effective manner while also assisting chemical biology projects.

References

1 Stumpf, M.P. *et al.* (2008) Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. U. S. A.* 105, 6959–6964
2 Berg, T. (2008) Small-molecule inhibitors of protein–protein interactions. *Curr. Opin. Drug Discov. Dev.* 11, 666–674
3 Fry, D.C. (2008) Drug-like inhibitors of protein–protein interactions: a structural examination of effective protein mimicry. *Curr. Protein Pept. Sci.* 9, 240–247
4 Villoutreix, B.O. *et al.* (2008) In silico–*in vitro* screening of protein–protein interactions: towards the next generation of therapeutics. *Curr. Pharm. Biotechnol.* 9, 103–122
5 Wells, J.A. and McClendon, C.L. (2007) Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* 450, 1001–1009
6 Whitty, A. and Kumaravel, G. (2006) Between a rock and a hard place? *Nat. Chem. Biol.* 2, 112–118
7 Fuller, J.C. *et al.* (2009) Predicting druggable binding sites at the protein–protein interface. *Drug Discov. Today* 14, 155–161
8 Arkin, M.R. and Whitty, A. (2009) The road less traveled: modulating signal transduction enzymes by inhibiting their protein–protein interactions. *Curr. Opin. Chem. Biol.* 13, 284–290
9 Shoichet, B.K. (2004) Virtual screening of chemical libraries. *Nature* 432, 862–865
10 Chene, P. (2006) Drugs targeting protein–protein interactions. *ChemMedChem* 1, 400–411
11 Eyrisch, S. and Helms, V. (2007) Transient pockets on protein surfaces involved in protein–protein interaction. *J. Med. Chem.* 50, 3457–3464
12 Arkin, M. (2005) Protein–protein interactions and cancer: small molecules going in for the kill. *Curr. Opin. Chem. Biol.* 9, 317–324
13 Clackson, T. and Wells, J.A. (1995) A hot spot of binding energy in a hormone–receptor interface. *Science* 267, 383–386
14 Lo Conte, L. *et al.* (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* 285, 2177–2198
15 Ma, B. and Nussinov, R. (2007) Trp/Met/Phe hot spots in protein–protein interactions: potential targets in drug design. *Curr. Top. Med. Chem.* 7, 999–1005
16 Reichmann, D. *et al.* (2007) The molecular architecture of protein–protein binding sites. *Curr. Opin. Struct. Biol.* 17, 67–76
17 Arkin, M.R. and Wells, J.A. (2004) Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nat. Rev. Drug Discov.* 3, 301–317
18 Park, C.M. *et al.* (2008) Discovery of an orally bioavailable small molecule inhibitor of prosurvival B-cell lymphoma 2 proteins. *J. Med. Chem.* 51, 6902–6915
19 Cole, C. and Warwicker, J. (2002) Side-chain conformational entropy at protein–protein interfaces. *Protein Sci.* 11, 2860–2870
20 Clark, D.E. (2008) What has virtual screening ever done for drug discovery? *Expert Opin. Drug Discov.* 3, 841–851
21 Davies, J.W. *et al.* (2006) Streamlining lead discovery by aligning *in silico* and high-throughput screening. *Curr. Opin. Chem. Biol.* 10, 343–351
22 Keskin, O. *et al.* (2008) Characterization and prediction of protein interfaces to infer protein–protein interaction networks. *Curr. Pharm. Biotechnol.* 9, 67–76
23 Kitchen, D.B. *et al.* (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* 3, 935–949
24 Rochet, J.C. (2007) Novel therapeutic strategies for the treatment of protein-misfolding diseases. *Expert Rev. Mol. Med.* 9, 1–34
25 Stockwell, B.R. (2004) Exploring biology with small organic molecules. *Nature* 432, 846–854
26 Thanos, C.D. *et al.* (2006) Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc. Natl. Acad. Sci. U. S. A.* 103, 15422–15427
27 Domling, A. (2008) Small molecular weight protein–protein interaction antagonists: an insurmountable challenge? *Curr. Opin. Chem. Biol.* 12, 281–291
28 Orry, A.J. *et al.* (2006) Structure-based development of target-specific compound libraries. *Drug Discov. Today* 11, 261–266
29 Keskin, O. *et al.* (2008) Principles of protein–protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.* 108, 1225–1244
30 Pagliaro, L. *et al.* (2004) Emerging classes of protein–protein interaction inhibitors and new tools for their development. *Curr. Opin. Chem. Biol.* 8, 442–449
31 Yin, H. and Hamilton, A.D. (2005) Strategies for targeting protein–protein interactions with synthetic agents. *Angew. Chem. Int. Ed. Engl.* 44, 4130–4163
32 Fry, D.C. (2006) Protein–protein interactions as targets for small molecule drug discovery. *Biopolymers* 84, 535–552
33 Neugebauer, A. *et al.* (2007) Prediction of protein–protein interaction inhibitors by chemoinformatics and machine learning methods. *J. Med. Chem.* 50, 4665–4668
34 Randic, M. (1995) Molecular profiles novel geometry-dependent molecular descriptors. *New J. Chem.* 19, 781–791

35 Randic, M. (1995) Molecular shape profiles. *J. Chem. Inf. Comput. Sci.* 35, 373–382
36 Kingsford, C. and Salzberg, S.L. (2008) What are decision trees? *Nat. Biotechnol.* 26, 1011–1013
37 Gonzalez, M.P. *et al.* (2006) *In silico* studies using radial distribution function approach for predicting affinity of 1 alpha,25-dihydroxyvitamin D(3) analogues for vitamin D receptor. *Steroids* 71, 510–527
38 Fernandez, M. (2005) Quantitative structure–activity relationship to predict differential inhibition of aldose reductase by flavonoid compounds. *Bioorg. Med. Chem.* 13, 3269–3277
39 Hemmer, M. (2000) Prediction of three-dimensional molecular structures using information from infrared spectra. *Anal. Chim. Acta* 420, 145–154
40 Hemmer, M. (1999) Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectrosc.* 19, 151–164