Analysis of solvent-exposed and buried co-crystallized ligands: a case study to support the design of novel protein–protein interaction inhibitors

Daniela Trisciuzzi[1], **Orazio Nicolotti**[1], **Maria A. Miteva**[2,3] **and Bruno O. Villoutreix**[2,3]

[1] Dipartimento di Farmacia – Scienze del Farmaco, Università degli Studi di Bari 'Aldo Moro', Via E. Orabona, 4, Bari I-70126, Italy
[2] Université Paris Diderot, Sorbonne Paris Cité, Molécules Thérapeutiques In Silico, INSERM UMR-S 973, Paris, France
[3] INSERM, U973, Paris, France

Molecular descriptors have been used to characterize and predict the functions of small molecules, including inhibitors of protein–protein interactions (iPPIs). Such molecules are valuable to investigate disease pathways and as starting points for drug discovery endeavors. iPPIs tend to bind at the surface of macromolecules and the design of such compounds remains challenging. Here, we report on our investigation of a pool of interpretable molecular descriptors for solvent-exposed and buried co-crystallized ligands. Several descriptors were found to be significantly different between the two classes and were further exploited using machine-learning approaches. This work could open new perspectives for the rational design of focused libraries enriched in new types of small drug-like molecules that could be used to prevent PPIs.

Introduction

Over the past few decades, interpretable molecular descriptors have been extensively used to investigate small molecules in an attempt to define or predict many different types of properties, including drug-likeness, to visualize the chemical space, or to study quantitative structure–toxicity relationships [1–8]. For instance, after analyzing oral drugs and clinical candidates, Lipinski and colleagues outlined that, in general, poor absorption or permeation is more likely when the molecular weight (MW) is >500 Da, the calculated log P is >5, and there are more than five H-bond donors (nHDon) or more than ten H-bond acceptors (nHAcc) [9]. Thus, the rule of 5 (Ro5) describes molecular properties important for the pharmacokinetics of a drug in the human body, including absorption, distribution, metabolism, excretion, and indirectly toxicity (ADME-Tox). However, the analysis by Lipinski *et al*. did not include natural products and antibiotics. Other descriptors frequently considered when evaluating bioavailability include flexibility: drug-like compounds should, in general, have fewer than ten rotatable bonds (RBN) and a polar surface area (PSA) < 140 Å^2 [10]. However, there are approved drugs, including small chemical compounds, small peptides, and natural products, that are orally available and/or cell permeable despite violating the Ro5. These compounds are in regions of the chemical space often defined to the extended Ro5 (eRo5). In the case of the eRo5, examples of cut-off values for several molecular descriptors are: MW 500–700 Da, cLogP 0–7.5, nHDon ≤ 5, nHAcc ≤ 10, PSA ≤ 200 Å^2, and RBN ≤ 20. There are also drugs beyond the Ro5 (bRo5) with calculated properties within the following ranges: MW >700 Da, cLogP <0 or >7.5, nHDon >5, nHAcc >10, PSA >200 Å^2, or RBN >20 [11,12]. Related studies carried out on macrocycles highlighted molecules that could still be administrated orally with a MW <1000 Da, cLogP <10, PSA <250, and nHDon <5 [13,14].

Other investigations were carried out to relate possible ranges of descriptors to specific targets or mechanisms of action or even to set a preliminary filter where specific types of molecule would be found. The latter case is seen, for instance, when studying bitter molecules, because these compounds normally show MW ≤ 700 and the AlogP is in between -3 and 7 [15]. Investigations of physicochemical properties of molecules acting on different target

Corresponding author: Villoutreix, B.O. (bruno.villoutreix@inserm.fr)

1359-6446/© 2018 Elsevier Ltd. All rights reserved.
https://doi.org/10.1016/j.drudis.2018.11.013

classes, such as G-protein-coupled receptors (GPCRs), nuclear receptors, and kinases, have been performed [16]. For example, a recently reported database of protein kinase inhibitors indicated the following ranges of values: MW between 309-617, cLogP between 1.4 and 6.7, nHDon between 0 and 4, nHAcc between 3 and 11, PSA between 54 and 140, and RBN between 1 and 11 [17]. Overington and co-workers examined a large set of bioactive molecules retrieved from the ChEMBL database unveiling molecular properties able to discriminate allosteric from nonallosteric compounds [18]. Along this line, Wanga and co-workers provided an 'allosteric-like' filter (i.e., MW ≤600; 3 ≤cLogP ≤7; RBN ≤6; 2 ≤total number of rings ≤5, with maximum two rings in the largest ring system) for the identification of putative allosteric modulators. Such a filter can also be exploited for the generation of focused libraries for screening campaigns or as a guide for drug design and optimization of allosteric hits [19]. Several other studies were conducted to investigate the physicochemical property ranges of compounds acting as iPPIs [20–26]. For instance, if we take ~1500 iPPIs from various databases, remove outliers and look at value ranges for several descriptors, iPPIs tend to have a MW between 200 and 900, cLogP from −1 to 9.5, nHAcc between 2 and 12, nHDon between 0 and 6, PSA between 20 and 185, and RBN between 1 and 15. Most direct iPPIs often bind in more solvent-exposed pockets comprising three to five small subpockets [27] and, more recently, it was suggested that pockets on RNA could be substantially similar to iPPI-binding pockets [28]. Building on these observations, it is possible that small molecules that would have an iPPI profile could also target RNA molecules. This could be of interest given that RNAs, similar to PPIs, are implicated in many human diseases.

Here, we compare co-crystallized ligands that tend to be solvent exposed at the surface of a protein receptor (referred here as 'solvent-exposed ligands' or 'solvent-friendly binders', comprising molecules having, for instance, one fragment more buried into a receptor cavity while the remaining fragments are essentially solvent-exposed) with co-crystallized ligands that are more buried inside the targets (named here 'buried', including, for instance, ligands that have only two methyl groups solvent exposed while the remaining groups are buried in the receptor) using interpretable computed physicochemical properties and molecular descriptors. We are particularly interested in solvent-friendly binders for their numerous potential applications in biology, including exploration of the human interactome. To the best of our knowledge, this is the first study devoted to this topic and it could be valuable for the rational design of molecules preventing PPIs (i.e., such molecules could replace monoclonal antibodies that inhibit PPIs) or interfering with RNA targets. Moreover, the wealth of information provided herein could also help to set cut-off descriptor values when designing specific types of ligands or to prepare focused libraries enriched in molecules that are more likely to remain at the surface of a target. Orthosteric iPPIs tend to be solvent exposed and could be used to gain insights about solvent-friendly binders, but there are not many small molecules co-crystallized at the surface of a protein–protein interface (~670 iPPIs in the 2P2I database, version June 2018) [23]. By contrast, there are thousands of ligands co-crystallized with proteins that could be used to gain novel knowledge and to assist the design of the aforementioned solvent-friendly binders.

To carry out our investigation, we generated two data sets that should allow us to study the molecular profiles of solvent-exposed and buried co-crystallized ligands. We first analyzed high-quality 3D experimental structures and then computed the solvent-accessible surface area (SASA) for each compound [29]. We calculated the SASA values of cognate ligands within and without their protein partners. From these computations, we designated two sets of compounds categorized as either solvent-exposed or buried according to the percentage of the fraction of the co-crystallized ligand accessible to the solvent. Then, interpretable molecular descriptors were collected and a random forest model was constructed to estimate the relative importance of the different descriptors with regard to the solvent-exposed class.

Case study

Generation of the initial data set

All analyses were conducted using the PDBbind v.2017 database, a curated collection of high-quality 3D crystallographic data of biomolecular complexes retrieved from the Protein Data Bank (PDB) [30] that are annotated with experimental binding affinity information. In total, 14 761 protein–ligand complexes were initially downloaded from the PDBbind server (www.pdbbind.org.cn/) [31]. The protonation states used at PDBbind is as follows for protein receptors: Asp and Glu were considered negatively charged, whereas Lys and Arg residues were considered positively charged. For small ligands, carboxylic, sulfonic, and phosphoric acid groups were considered negatively charged, whereas aliphatic amine, guanidine, and amidine groups were assumed to be positively charged [31]. Ligands covalently bound to the receptors or peptide-like ligands (e.g., having amino acid residues) were excluded from the analysis. Furthermore, to remove very large compounds, we applied a soft filtering protocol on specific molecular descriptors. The minimum and maximum value ranges tolerated to select the initial set of small molecules were as follows: 250 <MW <900, −5 <AlogP <6, 0 <nHDon <8, 0 <nHAcc <12; 0 <RBN <20, 0 <TPSA <160. In doing so, 7424 compounds were selected, hereafter referred to as 'PDBbind-focus-DB'.

Solvent accessible surface area calculation

The SASA represents the area of surface traced by a center of a hypothetical solvent sphere that rolls over the van der Waals surface of the molecule [32]. The SASA value of each crystallographic ligand of the PDBbind-focus-DB was computed in the presence and absence of the protein partner, using the FreeSASA program [29]. For the SASA calculation, default settings were used (i.e., the spherical solvent probe has a radius of 1.4 Å). All the hydrogen atoms and HETATM molecules were included in the calculation, whereas all water molecules and metal ions were removed. A fundamental step in the SASA calculation is the assignment of a certain class and radius to each atom. For the computation of the solvent-exposed area of each protein residue, the default FreeSASA library provided by Tsai *et al*. [33] using the *ProtOr* radii for the recognition of the 20 standard amino acids was used. By contrast, the van der Waals radii taken from [34] were applied for the ligand atoms. The fraction of ligand exposed to the solvent of each complex in the PDBbind-focus-DB was calculated by the percentage of the relative SASA of the ligand (named here %rSASA) calculated using Eq. (1):

$$\%\text{rSASA} = \frac{PL - SASA}{L - SASA} \times 100 \quad (1)$$

where the PL-SASA and L-SASA values represent the SASA values of the crystallographic ligand computed in the presence and absence of the protein partner, respectively. The distribution of the PDBbind-focus-DB in function of the %rSASA is depicted in Fig. 1.

The ligands were then classified into two main categories based on their corresponding %rSASA values as follows: (i) if the % rSASA ≤ mean value −1 standard deviation (SD), the ligands were classified as buried (or mainly buried) into the protein bound partner; or (ii) if the %rSASA ≥ mean value +1 SD, the ligands were classified as solvent exposed (or more exposed).

The third class of compounds (mean −1 SD <% rSASA < mean +1 SD) was not considered further here. At the end of this process, 1121 crystallographic ligands were categorized as mainly buried and 1140 crystallographic ligands were labeled as essentially solvent exposed (the SMILES for the ligands together with the PDB codes and the ligand categories are provided in the supplemental information online). Four examples with different levels of exposition to the solvent as identified by our approach are shown in Fig. 2 [35–38].

Molecular descriptors

For each ligand belonging to the solvent-exposed and buried classes, 42 molecular descriptors were computed (Table 1). We computed 35 descriptors that had been previously selected for their interpretability [40]. These molecular descriptors encode intuitively important chemical–physical properties, such as size and shape, polarity, polarizability, hydrogen bonding, lipophilicity, flexibility, and rigidity. Seven additional interpretable descriptors were also considered. Thus, 35 descriptors [39] were calculated using the free E-Dragon server [40] and seven additional descriptors were computed using the FAF-Drugs web server [41,42] and the DataWarrior package [43]. Among these descriptors, three encode the notion of molecular complexity in a different manner than those computed by E-Dragon, namely the number of stereocenters, Fsp^3 [44], the complexity, and the shape index [43].

To analyze the nonbonded interactions occurring between the protein receptors and ligands, several other descriptors were computed using the BINding ANAlyzer *BINANA) program, a python-implemented algorithm available at http://rocce-vm0.ucsd.edu/data/sw/hosted/binana/ [45]. The default setting parameters were used, including the hydrophobic, salt bridge, and HB cut-offs. This means here that a HB is identified if the HB donor was within 4.0 Å of the HB acceptor, and the angle θ formed between the donor, acceptor, and hydrogen atom was no greater than 40°. The pi-interaction, the cation-pi, and T-stacking (or edge-face) were computed with the default settings.

Analysis of the solvent-exposed and buried molecules

The main receptor families present in our test case

First, a clustering analysis of the most representative protein families was performed on all the proteins belonging to the two different groups. According to the Enzyme Commission classification system, a numerical code (i.e., EC number) expressing the reaction that the enzyme catalyzes was assigned. Six main protein categories [oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5), and ligases (EC 6)] and a class of miscellaneous proteins were identified. All the six enzyme classifications for both the solvent-exposed and buried classes are depicted in Fig. S1 in the Supplemental information online. Interestingly, no crystallized ligands of the miscellaneous buried class were involved in iPPIs, whereas ∼20% of the compounds of the miscellaneous solvent-exposed class could be flagged as iPPIs. This information was obtained by comparing our protein–

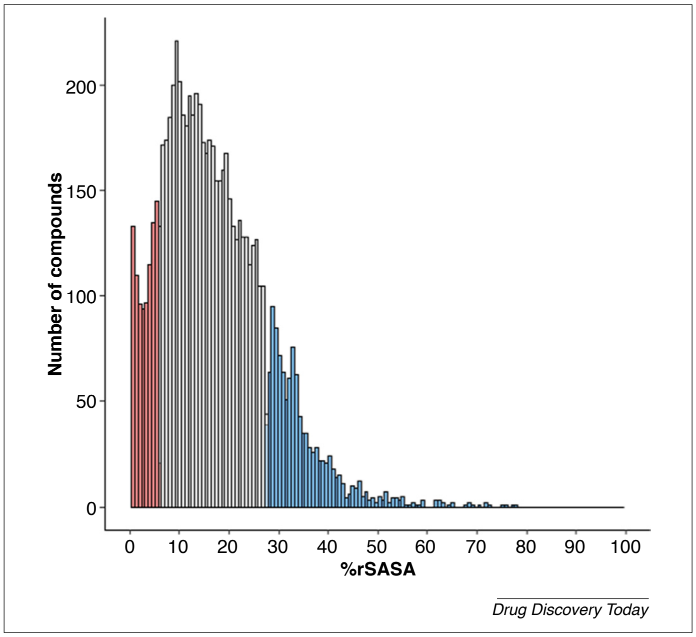
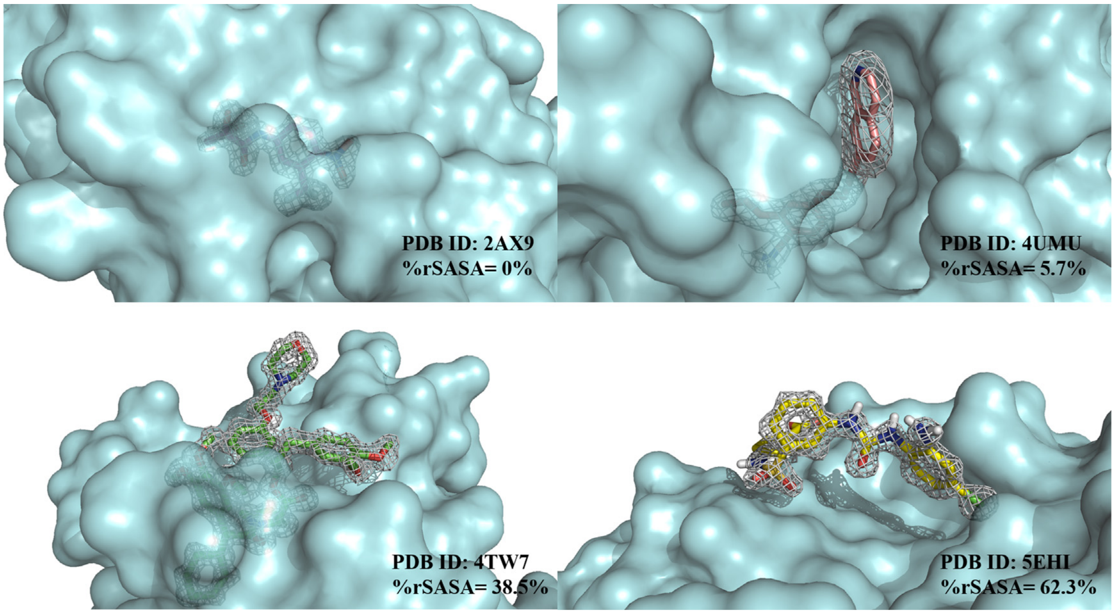


FIGURE 1

Distribution of the PDBbind-focus-DB as a function of the percentage of relative solvent-accessible surface area (%rSASA). Thus, a ligand can be completely buried (%rSASA = 0) or more exposed (e.g., %rSASA = 60%). The buried and solvent-exposed classes are depicted in red and blue, respectively. The remaining molecules are depicted in white.



Drug Discovery Today

FIGURE 2

Four ligands with different percentage of relative solvent accessible surface area (%rSASA) values. All the target proteins [Protein Data Bank (PDB) codes: 2AX9 (androgen receptor) [35]; 4UMU (maternal embryonic leucine zipper kinase) [36]; 4TW7 (FK506-binding protein 51) [37]; and 5EHI (NS5 methyltransferase) [38]] are shown as cyan-colored surfaces, whereas the ligands are rendered as sticks (magenta and pink indicate the buried category and green and yellow indicate the exposed category). The 2Fo-Fc electron density map contoured at the 1.2 sigma level is shown. As far as the structure of 5EHI is concerned, the unit cell shows that a crystallographic neighbor protein chain covers part of the ligand, leaving it 40% solvent exposed. However, the recombinant methyltransferase is a monomer in solution, suggesting that the cognate ligand could indeed be essentially solvent exposed when measuring its activity *in vitro*.

ligand complexes with the structural files available at the 2P2I$_{db}$ v2.0 website [23]. Of importance, many solvent-exposed ligands are not known iPPIs.

Solvent-exposed and buried ligands: diversity

We then investigated the chemical diversity of the solvent-exposed and buried ligands using fingerprint-based clustering analysis and 2D-Rubber Band Scaling (2D-RBS). After removing duplicates and some compounds for which descriptors could not be computed, 1017 solvent-exposed and 967 buried ligands were investigated. In total, 663 (i.e., 68.2%) and 710 (i.e., 68.5%) clusters (highest similarity falls below 0.8 Tanimoto to define the clusters) were obtained for the buried and solvent-exposed class, respectively. In this respect, a Tanimoto similarity equal to 0.8 was employed by using the FragFp DataWarrior fingerprint, a binary fingerprint similar to the MDL keys that relies on a dictionary of 512 predefined structure fragments. This notion of diversity can be visualized using the 2D-RBS approach computed with DataWarrior (Fig. S2 in the Supplemental information online). Overall, the compounds were relatively diverse in both classes.

Solvent-exposed and buried data sets: analysis of the molecular descriptors

The descriptor mean values were computed for the entire data sets and are reported in Fig. S3 in the Supplemental information online. The mean values of several physicochemical descriptors were significantly higher for the ligands belonging to the solvent-exposed class compared with those of the buried class (Table 1). Conceptually, our observations appear reasonable because, for instance, it is intuitive that larger ligands should belong to the solvent-exposed class because such ligands probably need to be larger to bind with good affinity to the receptors, a situation that is often observed in the case of iPPIs [22,23,25–27]. Also, we note similar trends for descriptors such as MW, logP, RBN or nHAcc when comparing the profiles of our solvent-exposed co-crystallized compounds with those reported for iPPIs [26]. Along this line of reasoning, it is interesting that a decision tree (DT) that made use of a shape descriptor and counting for the presence of multiple bonds was developed in 2010 to discriminate potential iPPIs from non-iPPI compounds [26]. Using this approach, ~40% of the buried class are predicted to be potential iPPIs, whereas ~70% of the solvent-exposed class are flagged as potential iPPIs, suggesting that our solvent-exposed co-crystallized ligands could be of interest to modulate PPIs.

We also investigated using several physicochemical and topological descriptors whether molecular complexity discriminated the two ligand classes [46,47]. Molecular complexity is intimately related to several major aspects of drug development, comprising target selectivity, synthetic accessibility, and potential success in preclinical and clinical phases [44,48,49], including compound safety [50]. To this end, we analyzed the number of chiral atoms, Fsp^3, MW, and the DataWarrior complexity index, which gives a measure of complexity of the entire molecule and the shape index [43]. Fig. S4 in the Supplemental information online shows the complexity distribution of the two data sets based on these five complexity metrics. Here, only two descriptors discriminated

TABLE 1

Molecular descriptors

Categories	Descriptor code	Description	Web server	*P* values[a]
Size- and shape-related descriptors	MW	Molecular weight	E-Dragon	1.51E-49
	nAT	Number of atoms	E-Dragon	5.93E-55
	nC	Number of carbon atoms	E-Dragon	3.08E-39
	nSK	Number of non-H atoms	E-Dragon	4.84E-56
Polarity	TPSA(NO)	Topological polar surface area using N and O	E-Dragon	4.52E-28
	TPSA(Tot)	Topological polar surface area using N, O, S, and P	E-Dragon	4.78E-26
	Hy	Hydrophilic factor	E-Dragon	4.13E-26
Polarizability	Sp	Sum of atomic polarizabilities (scaled on C atoms)	E-Dragon	1.32E-53
	AMR	Ghose-Crippen molar refractivity	E-Dragon	1.01E-50
	Total Charges	Formal total charge	FAF-Drugs	3.55E-05
Hydrogen bond capability	nHDon	Number of donor atoms for hydrogen bonds (N and O)	E-Dragon	2.40E-11
	nHAcc	Number of acceptor atoms for hydrogen Bonds (N, O, and F)	E-Dragon	6.21E-20
	nN	Number of nitrogen atoms	E-Dragon	3.40E-29
	nO	Number of oxygen atoms	E-Dragon	1.09E-09
	nROH	Number of aliphatic hydroxy groups	E-Dragon	3.33E-05
	nArOH	Number of aromatic hydroxy groups	E-Dragon	1.61E-05
Lipophilicity	ALOGP	Ghose-Crippen octanol-water partition coefficient	E-Dragon	5.12E-04
	XlogP3[b]	Logarithm of n-octanol-water partition coefficient	FAF-Drugs	2.0E-01
Flexibility, rigidity	nBT	Number of bonds	E-Dragon	7.78E-54
	nCIC	Number of rings	E-Dragon	1.77E-14
	RBN	Number of rotatable bonds	E-Dragon	8.66E-39
	RBF	Rotatable bond fraction	E-Dragon	1.06E-10
Complexity	Fsp^3	Number of sp3 hybridized carbons/total carbon count	FAF-Drugs	4.31E-03
	Stereocenters	Number of chiral centers	FAF-Drugs	8.03E-03
	Molecular complexity index	Use the number of unique connected subgraphs	DataWarrior	6.43E-12
	Shape index	Compounds more linear or more spherical	DataWarrior	5.43E-07
Constitutional and functional descriptors	Sv	Sum of atomic van der Waals volumes (scaled on C atom)	E-Dragon	5.93E-54
	Se	Sum of atomic Sanderson electronegativities (scaled on C atom)	E-Dragon	2.97E-56
	Mv	Mean atomic van der Waals volume (scaled on C atom)	E-Dragon	1.14E-06
	Me	Mean atomic Sanderson electronegativity (scaled on C atom)	E-Dragon	7.15E-06
	nBO	Number of non-H bonds	E-Dragon	1.68E-51
	nBM	Number of multiple bonds	E-Dragon	2.65E-14
	ARR	Aromatic ratio	E-Dragon	1.08E-03
	nDB	Number of double bonds	E-Dragon	3.31E-15
	nAB	Number of aromatic bonds	E-Dragon	2.46E-07
	nX	Number of halogens	E-Dragon	9.48E-07
	nBnz	Number of benzene-like rings	E-Dragon	5.17E-09
	nCar	Number of aromatic carbon atoms (sp2)	E-Dragon	3.07E-08
	n_amid[c]	Number of amides	E-Dragon	7.93E-37
	Ui	Unsaturation index	E-Dragon	2.65E-14
	LAI	Lipinski alert index (drug-like index)	E-Dragon	3.11E-03
	Lipinski_Violation	Number of Ro5 violations	FAF-Drugs	1.48E-07

[a] Wilcox's t-test *P* values highlight statistically significant differences between the two groups of co-crystallized ligands.

[b] XlogP3 is benchmarked on a set of 1300 molecules with experimental logP values from US National Cancer Institute.

[c] Total number of amides is given by the sum of the number of aliphatic and aromatic primary, secondary, and tertiary amides.

solvent-exposed from buried ligands: MW and DataWarrior molecular complexity. The solvent-exposed ligands tend to be slightly less linear (and more 3D) than buried molecules, a property that is also observed with iPPIs [51].

Chemical space visualization of solvent-exposed and buried ligands

To investigate further our solvent-exposed and buried data sets, we used principal component analysis (PCA). In our analysis, PC1 represented ~48% of the total variance, PC2 ~25%, and PC3 ~13%. The first two axes of the PCA (~73% of the variance) are shown in Fig. 3. The first axis is mainly characterized by MW and aspects of polarizability, whereas the second axis essentially represents the hydrophilicity and lipophilicity. This analysis shows that solvent-exposed and buried molecules tend to occupy different regions of the chemical space.

Bioactivity and ligand–protein interaction analysis

The bioactivity distribution of the solvent-exposed and buried classes was computed to compare the experimental binding affini-

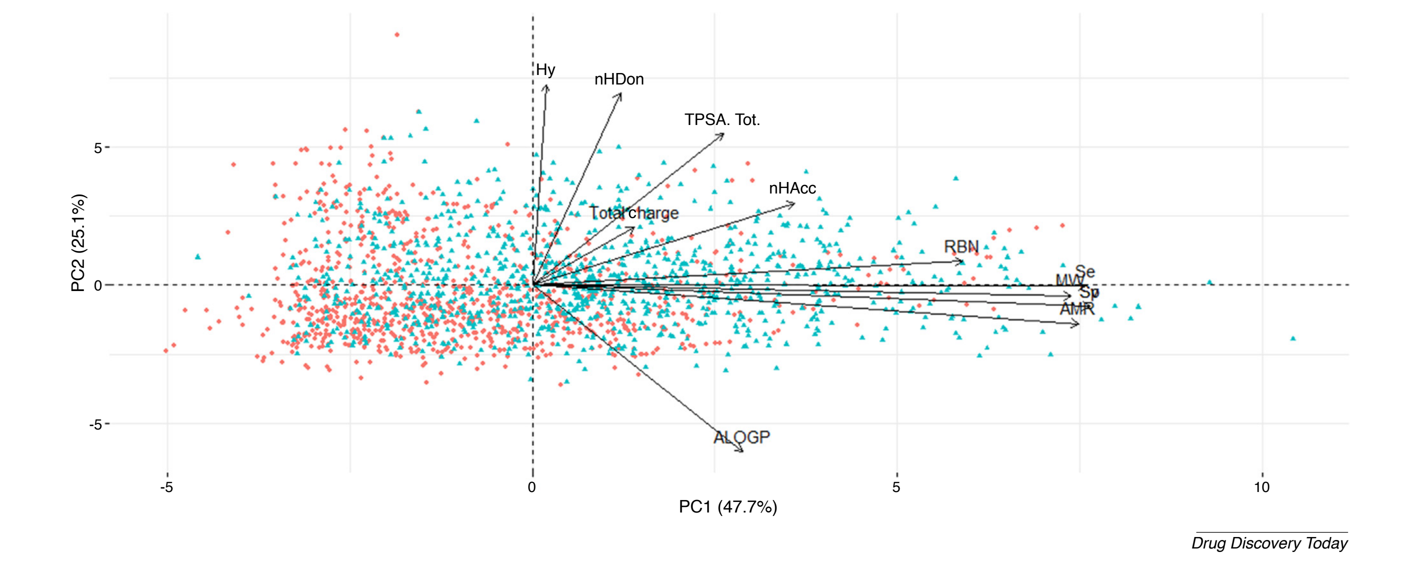


FIGURE 3

Principal component analysis (PCA) for the compound data sets. The buried ligands are in red and the solvent-exposed molecules are in blue. The molecular descriptors used were: MW, TPSA(Tot), AMR, total charges, nHDon, nHAcc, ALOGP, Se, Sv, Sp, Hy, and RBN. The first (PC1) and second (PC2) components explained ~48% and ~25% of the variance, respectively. All the data were centered and scaled to unit variance. The circle of correlation indicates that the first axis is characterized by the size of the compounds (e.g., MW), and an aspect of polarizability (e.g., Sp and AMR), whereas the second axis is represented by the hydrophilicity and lipophilicity of the compounds. Please see the main text for definitions of abbreviations.

ty data. We computed pChEMBL (or pIC_{50}), a parameter that is used, for example, in the ChEMBL database [52]. More specifically, this approach enabled us to compare different types of affinity value using: $pIC_{50} = -\log$ (molar X), where X represents a value of bioactivity expressed as K_i, K_d, and IC_{50}. For instance, a pIC_{50} value of ~8.5 corresponds to an affinity of almost 4 nM. The distribution of bioactivity values for the two classes of compound is shown in Fig. 4. Overall, as expected, the buried ligands show higher affinity for the targets compared with the solvent-exposed ones. Given that solvent-exposed ligands are generally larger than buried ones, we used various metrics to measure the binding energy per atom, such as ligand efficiency (LE, a measure of the activity normalized by the number of non-H atoms) [53,54], lipophilic LE (LLE; calculated from the pIC50 of the compound and its cLogP) [53,55,56] or LE lipophilic price (LELP = logP/LE). Specifically, LELP is negative when logP is negative, and the higher the absolute value of LELP, the less drug-like the lead compound [43,57]. The accepted lower limit of LE is ~0.3 kcal/mol during early-stage drug discovery [53,55,58], whereas acceptable LLE values tend to be >3 for lead compounds and >5 for clinical candidates [56]. For an acceptable lead, in general, LELP fits the following values: $-10 < \text{LELP} < 10$, although it has been suggested that the closer LELP is to zero in the positive range, the better; overall, the desirable range for LELP is between 0 and 7.5 [57]. As shown in Fig. S5 in the supplemental information online, we observed that the LE mean value for solvent-exposed compounds (LE $\text{mean}_{\text{solvent-exposed}} = 0.39$) was lower than that for the buried class (LE $\text{mean}_{\text{buried}} = 0.47$), whereas LLE and LELP mean values tended to be similar (LLE mean $_{\text{solvent-exposed}} = 6.23$ and LLE $\text{mean}_{\text{buried}} = 6.19$, whereas LELP $\text{mean}_{\text{solven-texposed}} = 5.48$ and LELP mean $_{\text{buried}} = 4.81$). In this regard, the larger size of the solvent-exposed compounds explains in part the lower LE mean value.

For a comparison with our solvent-exposed and buried data sets, it is known that PPI inhibitors tend to be more solvent exposed than traditional active site inhibitors of regular targets and they are larger, enabling them to reach the same level of potency with higher log P values [25,59–62]. Using a LE filter of 0.30 kcal/mol per heavy atom and an LLE >5.00, it was found that only 14.5% and 4.5% of iPPIs pass the LE and LLE filters, respectively [62]. Here, for the solvent-exposed class, 82.9% of the molecules passed the LE filter and 69.7% passed the LLE filter.

To gain additional insights into the key molecular interactions occurring in the two classes, we used the 3D structures of the protein–ligand complexes available to compute the various nonbonded interactions between the small molecules and the receptors with the BINANA package (Fig. S6 in the Supplemental information online). This showed that nonbonded interactions are significantly different between the two classes apart from π–π, cation–π, and salt bridges. In a buried site, the ligand can contact almost the entire surface of the receptor, which explains in part why hydrophobic interactions are more frequent in highly efficient ligands [63], as also confirmed by our analysis (i.e., more hydrophobic contacts in the buried class and molecules that tend to have higher affinity values compared with exposed ligands).

Is it possible to discriminate the two classes of molecule based on computed descriptors?

To investigate whether it is possible to discriminate the two classes of molecule based on computer descriptors, we first built simple DTs with all the descriptors. We used the DT classifier algorithm implemented in Scikit-learn [64] (http://scikit-learn.org) and tuned several hyperparameters. The train_test_split method was used to split the data sets into a training set (70%) and a test set

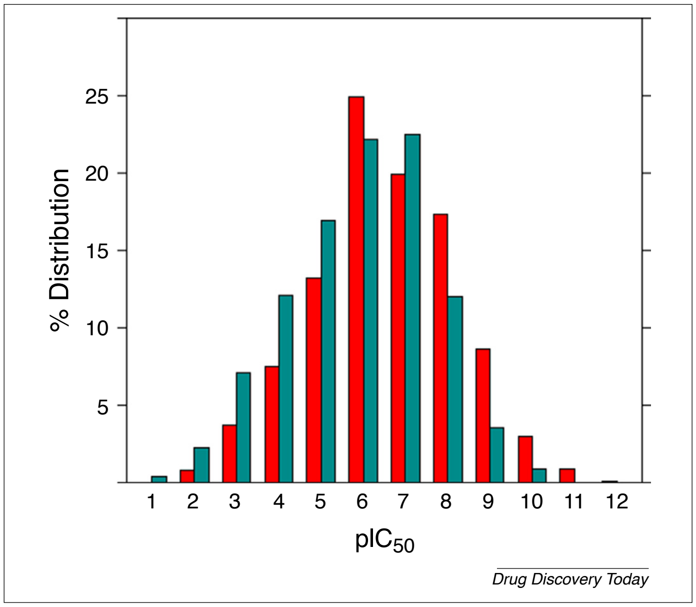


FIGURE 4

Bioactivity distribution of the exposed (cyan) and buried (red) classes expressed using the pIC$_{50}$ notation.

(30%) with the built-in stratification option turned on to have the same proportion of class labels in the training and test subsets as the input data set. Several hyperparameters (gini/entropy information gain, max_depth, min_samples_split, and min_samples_leaf) were investigated by plotting receiver operating characteristic area under the curve (ROC AUC) for the training and test sets to define values where the tree perfectly predicted the training data but failed to generalize the findings on the test data. The final DT (e.g., max_depth = 5) showed an accuracy of 76% on the training set and 68% on the validation set. This initial step suggested that it could be possible to build a model to filter out, for instance, molecules that are likely to be buried.

We then explored molecular descriptors that could be important for discriminating the two classes of compound. We estimated the relative importance of the different molecular descriptors with regard to the solvent-exposed class. We had already observed in the PCA plot that the two classes of compound are located in different regions. To identify descriptors that best explain the relationship between the structure and property of our molecules, we constructed a random forest (RF) model [65]. We used the RF

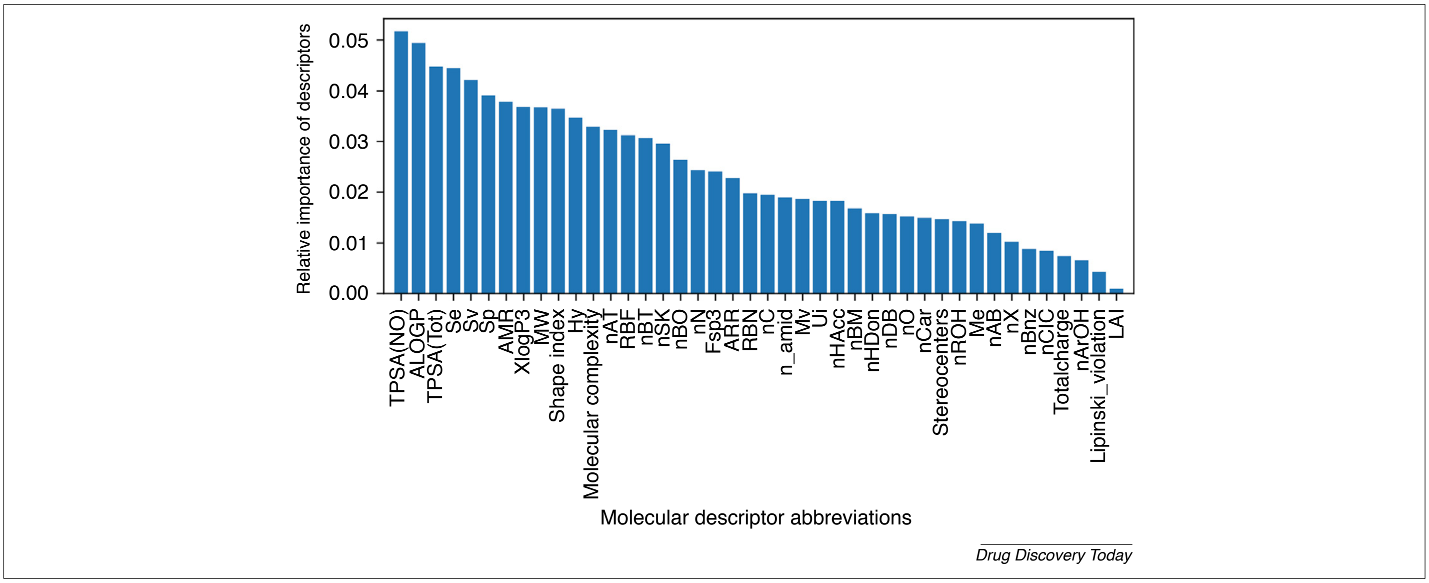


FIGURE 5

Relative molecular descriptor importance inferred from the random forest model. The bar plot underlines the relative importance of the descriptors inferred from the Random Forest classifier model trained to discriminate between solvent-exposed and buried molecules. The importance of the descriptors is sorted from highest [TPSA(NO)] to lowest (LAI). The *x*-axis reports the abbreviations for the molecular descriptors as explained in Table 1 (in the main text) and the y-axis the relative importance of the descriptors.

classifier algorithm implemented in Scikit-learn [66] (http://scikit-learn.org) to create an ensemble of 500 trees and randomly selected subsets of descriptors following the approach of Raschka *et al*. [66]. The feature importance ultimately helped to analyze the molecules by assigning scores to descriptors based on usefulness in building this model. Figure 5 shows a bar plot of the descriptor importance values, normalized to sum up to 1. We then took these top five–ten descriptors and redeveloped RF models. Hyperparameters (n_estimators, max_depth, max_features, min_samples_leaf, and gini/entropy) were assessed using a grid search algorithm with fivefold cross-validation. A final model was built and had an accuracy of 86% for the training set and 74% for the test set (sensitivity = 71%, specificity = 76%, ROC AUC = 0.81). Y-randomization was employed to ensure that there was no chance correlation between the selected descriptors and the buried/solvent-exposed classes. After random permutation of the class labels, models were rebuilt and found to perform significantly worse when applied to an external validation set not used during the training.

Based on these preliminary results, our case study indicates that it is possible to develop predictive models to select molecules that prefer to bind at the surface of a receptor. The following stepwise approach could be used for the rational design of collections enriched in iPPIs: (i) take as input a generic compound collection; (ii) compute molecular descriptors; (iii) run a first filtering step developed to filter out non-inhibitors of PPIs, such as the DT reported in Ref. [26] trained on known inhibitors of PPIs and on putative non-inhibitors; and (iv) run a second filtering step using a statistical model built on our co-crystallized solvent-exposed and buried data sets (see Supplemental information online). For projects aiming at the perturbation of a disease pathway, superficial binders could be valuable and in this case the use of a buried-exposed filter alone could be valuable.

We further searched for the presence of different types of chemical fragment among the two classes of molecule using the molBLOCKS tool [67], a package that breaks down molecules into fragments according to a predefined set of chemical rules (i.e., we applied the RECAP rules to cut bonds and generate fragments [68]). However, we did not identify specific fragments that could help design buried or solvent-exposed compounds.

Concluding remarks

Molecular descriptors and data set analyses have been used in many areas of drug discovery and chemical biology, from the preparation of a chemical library enriched in more bioavailable compounds to the preparation of focused collections dedicated to the modulation of PPIs. In our case study, we were interested in determining properties that highlight molecules that prefer to bind in deep binding pockets (buried compounds) from molecules that remain more solvent exposed at the surface of a target. A combination of a few interpretable molecular descriptors could be used to partially discriminate the two classes. Based on our preliminary investigations, it should be possible to develop machine-learning models to generate focused collections enriched in either type of compound. Such collections could be of interest for various types of experimental and *in silico* screening studies (e.g., screening entire PPI pathways), especially when the 3D structure of the targets is not known. We believe that generating collections enriched in compounds that would favor binding at the surface of a macromolecule could have direct application in the development of chemical probes that modulate PPIs and possibly RNA molecules.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.drudis.2018.11.013.

References

1 Benet, L.Z. *et al.* (2016) BDDCS, the Rule of 5 and druggability. *Adv. Drug Deliv. Rev.* 101, 89–98

2 Hann, M.M. and Oprea, T.I. (2004) Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* 8, 255–263

3 Price, D.A. *et al.* (2009) Physicochemical drug properties associated with in vivo toxicological outcomes: a review. *Expert Opin. Drug Metab. Toxicol.* 5, 921–931

4 Gleeson, M.P. (2008) Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* 51, 817–834

5 Walters, W.P. (2012) Going further than Lipinski's rule in drug design. *Expert Opin. Drug Discov.* 7, 99–107

6 Xue, L. and Bajorath, J. (2000) Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screen.* 3, 363–372

7 Nicolotti, O. *et al.* (2002) Multiobjective optimization in quantitative structure–activity relationships: deriving accurate and interpretable QSARs. *J. Med. Chem* 45, 5069–5080

8 Gissi, A. *et al.* (2014) An alternative QSAR-based approach for predicting the bioconcentration factor for regulatory purposes. *ALTEX* 31, 23–36

9 Lipinski, C.A. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46, 3–26

10 Veber, D.F. *et al.* (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623

11 Doak, B.C. *et al.* (2014) Oral druggable space beyond the rule of 5, insights from drugs and clinical candidates. *Chem. Biol.* 21, 1115–1142

12 Matsson, P. *et al.* (2016) Cell permeability beyond the rule of 5. *Adv. Drug Deliv. Rev.* 101, 42–61

13 Giordanetto, F. and Kihlberg, J. (2014) Macrocyclic drugs and clinical candidates: what can medicinal chemists learn from their properties? *J. Med. Chem.* 57, 278–295

14 Whitty, A. *et al.* (2016) Quantifying the chameleonic properties of macrocycles and other high-molecular-weight drugs. *Drug Discov. Today* 21, 712–717

15 Dagan-Wiener, A. *et al.* (2017) Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. *Sci. Rep.* 7, 12074

16 Morphy, R. and Rankovic, Z. (2006) The physicochemical challenges of designing multiple ligands. *J. Med. Chem.* 49, 4961–4970

17 Carles, F. *et al.* (2018) PKIDB: a curated, annotated and updated database of protein kinase inhibitors in clinical trials. *Molecules* 23, E908

18 van Westen, G.J.P. *et al.* (2014) Chemical, target, and bioactive properties of allosteric modulation. *PLoS Comput. Biol.* 10, e1003559

19 Wang, Q. *et al.* (2012) Toward understanding the molecular basis for chemical allosteric modulator design. *J. Mol. Graph. Model.* 38, 324–333

20 Lagorce, D. *et al.* (2017) Computational analysis of calculated physicochemical and ADMET properties of protein–protein interaction inhibitors. *Sci. Rep.* 7, 46277

21 Labbé, C.M. *et al.* (2016) iPPI-DB: an online database of modulators of protein–protein interactions. *Nucleic Acids Res.* 44, D542–D547

22 Higueruelo, A.P. *et al.* (2013) TIMBAL v2, update of a database holding small molecules modulating protein–protein interactions. *Database* 2013 bat039

23 Basse, M.-J. *et al.* (2016) 2P2Idb v2, update of a structural database dedicated to orthosteric modulation of protein–protein interactions. *Database* 2016 baw007

24 Villoutreix, B.O. *et al.* (2012) A leap into the chemical space of protein–protein interaction inhibitors. *Curr. Pharm. Des.* 18, 4648–4667

25 Wells, J.A. and McClendon, C.L. (2007) Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* 450, 1001–1009

26 Sperandio, O. *et al.* (2010) Rationalizing the chemical space of protein–protein interaction inhibitors. *Drug Discov. Today* 15, 220–229

27 Fuller, J.C. *et al.* (2009) Predicting druggable binding sites at the protein–protein interface. *Drug Discov. Today* 14, 155–161

28 Warner, K.D. *et al.* (2018) Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug Discov.* 17, 547–558

29 Mitternacht, S. (2016) FreeSASA: an open source C library for solvent accessible surface area calculations. *F1000Research* 5, 189

30 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242

31 Liu, Z. *et al.* (2017) Forging the basis for developing protein–ligand interaction scoring functions. *ACC Chem. Res.* 50, 302–309

32 Lee, B. and Richards, F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55, 379–400

33 Tsai, J. *et al.* (1999) The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* 290, 253–266

34 Mantina, M. *et al.* (2009) Consistent van der Waals radii for the whole main group. *J. Phys. Chem. A* 113, 5806–5812

35 Bohl, C.E. *et al.* (2005) Structural basis for accommodation of nonsteroidal ligands in the androgen receptor. *J. Biol. Chem.* 280, 37747–37754

36 Johnson, C.N. *et al.* (2015) Structure-based design of type II inhibitors applied to maternal embryonic leucine zipper kinase. *ACS Med. Chem. Lett.* 6, 31–36

37 Gaali, S. *et al.* (2015) Selective inhibitors of the FK506–binding protein 51 by induced fit. *Nat. Chem. Biol.* 11, 33–37

38 Benmansour, F. *et al.* (2017) Discovery of novel dengue virus NS5 methyltransferase non-nucleoside inhibitors by fragment-based drug design. *Eur. J. Med. Chem.* 125, 865–880

39 Larsson, J. *et al.* (2007) ChemGPS-NP: tuned for navigation in biologically relevant chemical space. *J. Nat. Prod.* 70, 789–794

40 Tetko, I.V. *et al.* (2005) Virtual computational chemistry laboratory: design and description. *J. Comput. Aided Mol. Des.* 19, 453–463

41 Lagorce, D. *et al.* (2015) FAF-Drugs3, a web server for compound property calculation and chemical library design. *Nucleic Acids Res.* 43, W200–W207

42 Lagorce, D. *et al.* (2017) FAF-Drugs4, free ADME-tox filtering computations for chemical biology and early stages drug discovery. *Bioinformatics* 33, 3658–3660

43 Sander, T. *et al.* (2015) DataWarrior: an open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model.* 55, 460–473

44 Lovering, F. *et al.* (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* 52, 6752–6756

45 Durrant, J.D. and McCammon, J.A. (2011) BINANA: a novel algorithm for ligand-binding characterization. *J. Mol. Graph. Model.* 29, 888–893

46 Sheridan, R.P. *et al.* (2014) Modeling a crowdsourced definition of molecular complexity. *J. Chem. Inf. Model.* 54, 1604–1616

47 Selzer, P. *et al.* (2005) Complex molecules: do they add value? *Curr. Opin. Chem. Biol.* 9, 310–316

48 Lovering, F. (2013) Escape from Flatland 2, complexity and promiscuity. *MedChemComm* 4, 515–519

49 Clemons, P.A. *et al.* (2010) Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci. U. S. A.* 107, 18787–18792

50 González-Medina, M. *et al.* (2016) Chemoinformatic expedition of the chemical space of fungal products. *Future Med. Chem.* 8, 1399–1412

51 Fry, D. *et al.* (2013) Design of libraries targeting protein–protein interfaces. *ChemMedChem* 8, 726–732

52 Bento, A.P. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090

53 Hopkins, A.L. *et al.* (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* 9, 430–431

54 Cavalluzzi, M.M. *et al.* (2017) Ligand efficiency metrics in drug discovery: the pros and cons from a practical perspective. *Expert Opin. Drug Discov.* 12, 1087–1104

55 Schultes, S. *et al.* (2010) Ligand efficiency as a guide in fragment hit selection and optimization. *Drug Discov. Today Technol.* 7, e157–e162

56 Leeson, P.D. and Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* 6, 881–890

57 Keserü, G.M. and Makara, G.M. (2009) The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov.* 8, 203–212

58 Hajduk, P.J. (2006) Fragment-based drug design: how big is too big? *J. Med. Chem.* 49, 6972–6976

59 Higueruelo, A.P. *et al.* (2009) Atomic interactions and profile of small molecules disrupting protein–protein interfaces: the TIMBAL database. *Chem. Biol. Drug Des.* 74, 457–467

60 Labbé, C.M. *et al.* (2013) iPPI-DB: a manually curated and interactive database of small non-peptide inhibitors of protein–protein interactions. *Drug Discov. Today* 18, 958–968

61 Morelli, X. *et al.* (2011) Chemical and structural lessons from recent successes in protein–protein interaction inhibition (2P2I). *Curr. Opin. Chem. Biol.* 15, 475–481

62 Laraia, L. *et al.* (2015) Overcoming chemical, biological, and computational challenges in the development of inhibitors targeting protein–protein interactions. *Chem. Biol.* 22, 689–703

63 de Freitas, R.F. and Schapira, M. (2017) A systematic analysis of atomic protein–ligand interactions in the PDB. *Med Chem Comm* 8, 1970–1981

64 Pedregosa, F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830

65 Teixeira, A.L. *et al.* (2013) Random forests for feature selection in QSPR models — an application for predicting standard enthalpy of formation of hydrocarbons. *J. Cheminf.* 5, 9

66 Raschka, S. *et al.* (2018) Automated inference of chemical discriminants of biological activity. *Methods Mol. Biol.* 1762, 307–338

67 Ghersi, D. and Singh, M. (2014) molBLOCKS: decomposing small molecule sets and uncovering enriched fragments. *Bioinformatics* 30, 2081–2083

68 Lewell, X.Q. *et al.* (1998) RECAP — retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 38, 511–522