Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties

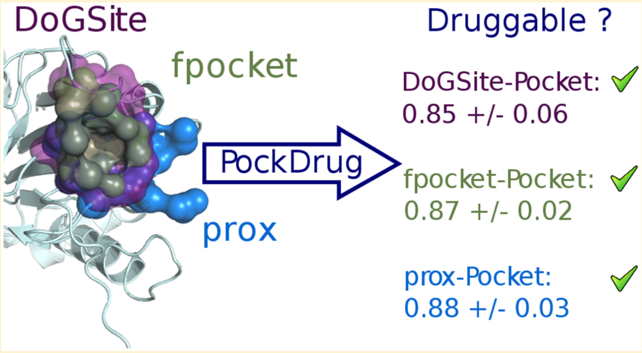
Alexandre Borrel,[†,‡,§] Leslie Regad,[†,‡] Henri Xhaard,[§] Michel Petitjean,[†,‡] and Anne-Claude Camproux*[,†,‡]

[†]INSERM, UMRS-973, MTi, Paris, France

[‡]University Paris Diderot, Sorbonne Paris Cité, UMRS-973, MTi, Paris, France

[§]University of Helsinki, Division of Pharmaceutical Chemistry, Faculty of Pharmacy, Helsinki, Finland

Supporting Information



ABSTRACT: Predicting protein druggability is a key interest in the target identification phase of drug discovery. Here, we assess the pocket estimation methods' influence on druggability predictions by comparing statistical models constructed from pockets estimated using different pocket estimation methods: a proximity of either 4 or 5.5 Å to a cocrystallized ligand or DoGSite and fpocket estimation methods. We developed PockDrug, a robust pocket druggability model that copes with uncertainties in pocket boundaries. It is based on a linear discriminant analysis from a pool of 52 descriptors combined with a selection of the most stable and efficient models using different pocket estimation methods. PockDrug retains the best combinations of three pocket properties which impact druggability: geometry, hydrophobicity, and aromaticity. It results in an average accuracy of 87.9% ± 4.7% using a test set and exhibits higher accuracy (~5−10%) than previous studies that used an identical apo set. In conclusion, this study confirms the influence of pocket estimation on pocket druggability prediction and proposes PockDrug as a new model that overcomes pocket estimation variability.

INTRODUCTION

Proteins—or their binding pockets—exhibit unequal abilities to be modulated with noncovalent high-affinity drug-like small molecules, an ability that is commonly referred to as druggability, as first defined by Hopkins and Groom in 2002.[1] In a drug-discovery setting, drug-like and target druggability properties are conducive to oral bioavailability and clinical progression, as underlined by recent druggability studies on oxazolidinones compounds conducted by Xue et al.[2] or on cofactor-independent phosphoglycerate mutase targets conducted by Crowther et al.[3] Druggability is challenging to assess experimentally, and only two seminal studies have been published to date. First, the Hadjuk et al. study[4] compiled NMR fragment screening data for 28 binding sites on 23 different proteins with ~1500 to ~12 000 fragment-like/lead-like compounds, resulting in 1 to 65 hits, (i.e., hit rates of 0.01 to 0.78). Second, the Cheng et al. study[5] compiled biochemical screening data for 34 prokaryotic and eukaryotic drug protein targets with libraries ranging from 50 000 to 1 million compounds and extracted and internally compared a data set of 37 275 common compounds for 22 targets, (i.e., hit rates varying from 0.06 to 3.85). Additional evidence can be found for proteins found undruggable in HTS screens, which makes druggability highly relevant to compound design.[3] Differences in target druggability can also be observed in the reverted HTS approach, where drug-like molecules are screened against, e.g., a collection of recombinant proteins (see Table 2 in Brown et al. study[6]). Druggability, or the lack thereof, is therefore a major factor in the ~60% failure rate estimated for drug discovery projects at the lead identification or optimization phase; another factor being the ability of the protein target to be disease-modifying.[6] Consequently, the computational evaluation of target druggability prior to the investment of resources has become crucial for the clinical progression of compounds.[6,8−10]

Many computational approaches have been developed to predict target druggability before extensive time and money are invested and to reduce the high failure rate. These computational prediction methods involve pocket estimation and characterization using different physicochemical and geometrical descriptors,[4,10−15] followed by statistical model optimization. Identifying the atoms that form the binding pocket is therefore a key issue in druggability predictions (for a review of pocket estimation methods, see Pérot et al.[16]). Two types of binding pocket estimation methods are used: (i) the

Received: October 6, 2014
Published: April 2, 2015

DOI: 10.1021/ci5006004

identification of all amino acids (or atoms) based on information on the position of a ligand to assert the pocket boundaries (referred to as "observed pocket estimation" in this manuscript) and (ii) predictive methods that are fully independent of the ligand position and automatically predict the amino acids (or atoms) that form the surface of potential binding cavities (referred to as "predicted pocket estimation" in this manuscript). The predicted pocket estimation approaches can be divided into three classes: energy-, geometry-, and evolution-based approaches.[16] Geometry-based pocket detection appears to be faster and more robust with respect to structural variations or missing atoms in the input structures compared with energy-based algorithms; therefore, geometry-based algorithms yield satisfactory pocket predictions in terms of the tested binding sites identified.[12] However, if the ligand overlap in the predicted pockets is satisfactory, there is no consensus pocket estimation method, as the predicted pocket boundaries are difficult to determine and depend on certain arbitrary selections in the pocket estimations.[17] Thus, certain estimation methods are important for the estimation of some particular types of pockets; for example, LIGSITE[18] is efficient in detecting small-molecule binding pockets, and DoGSite can be used to predict some coagulation factor Xa- or catalytic antibody-binding pockets.[19] The estimation of the same binding site by different approaches may result in different estimated pockets and boundaries that may therefore influence the values of the computed pocket descriptors. Accordingly, in addition to the statistical model used, the choice of pocket estimation method influences the computational prediction of druggability. Presently, none of the proposed druggability models can perform with different estimation methods. Indeed, each druggability model proposed in the literature is attached to one particular pocket estimation method. Recently, four druggability prediction models[10−12,14] were compared by Desaphy et al.[11] using the same test set extracted from the Krasowski[14] binding site set. These druggability models were respectively based on two observed estimation methods (DrugPred[14] and Desaphy's model)[11] and on two predicted estimation methods (fpocket score[12] and SiteMap[10]). Desaphy et al.[11] concluded that the recently proposed models DrugPred[14] and Desaphy's model,[11] which were trained and applied using observed pockets, exhibit the best druggability predictions (accuracies of 89%) compared with the accuracies of 76% and 65% obtained with fpocket score[12] and SiteMap,[10] respectively, using predicted pockets. The recently proposed DoGSiteScorer[15] model was trained and applied using predicted pockets[19] for a set of holo and apo proteins, and the model yielded similar druggability prediction results compared with those obtained in previous studies using predicted pockets.[10,12,15] However, computational druggability models trained on observed pockets are only directly applicable to binding pockets, whereas those trained on predicted pockets exhibit inferior performance but can be extended to predict the druggability of pockets in their apo forms. In this context, enhancing the performance of druggability prediction models by developing models that only minimally depend on the choice of pocket estimation method for both holo and apo pockets is overstatement for drug discovery. In this study, a druggability model, referred to as PockDrug, that can efficiently be used with several pocket estimation methods was developed. The first step in the statistical protocol to develop PockDrug included the optimization of linear discriminant analysis (LDA) models corresponding to three different descriptor combinations (from a pool of 52 geometrical and physicochemical descriptors) by cross-validation using one holo training set. In the second step, the seven best LDA models were selected by testing their performances when applied to an independent test set, which was estimated using four pocket estimation methods, based on a binding ligand proximity of 4 or 5.5 Å or detected by DoGSite[19] or fpocket.[20] The prediction of the seven best LDA models was then averaged into one model which we call PockDrug. PockDrug was then compared with the druggability results reported in the literature on an identical test set and validated using one independent apo set.

MATERIALS AND METHODS

Protein Data Sets. *Two Proteins Data Sets Were Used.* First, the largest currently freely available data set, i.e., the "NonRedundant dataset of Druggable and Less Druggable binding sites" (NRDLD), constructed by Krasowski et al.,[14] was used. The NRDLD set contains 113 nonredundant complexed proteins sharing a pairwise sequence identity of less than 60%. The protein set includes a large diversity of enzymes, such as oxidoreductases, ligases, and hydrolases. The definition of druggability used here was the ability of a protein to bind a drug-like molecule, as defined by Krasowski and co-workers;[14] thus, 71 binding sites were classified as druggable (i.e., binding site that noncovalently binds small drug-like ligands, which are orally available and do not require administration as prodrugs), and 42 were classified as less druggable. This terminology is preferred to "non-druggable" because nondruggability is only inferred from a limited set of tested molecules. In agreement with the definition of "less druggable" proposed by Krasowski,[14] a protein is less druggable if it does not meet any of the following requirements: (1) At least one ligand is orally available as judged by Lipinski's "rule-of-five." (2) The ligands must have a clogP greater than −2;. (3) The ligand efficiency of at least one of the ligands that fulfill criteria 1 and 2 is at least 0.3 kcal mol^{-1}/heavy atom.

The second data set (later referred to as Apo139), which includes 139 apo proteins, was collected from the "Druggable Cavity Directory" (DCD) database (http://fpocket.sourceforge.net/DCD), with at least one corresponding available holo form in the DCD database. It includes 139 proteins in the apo form, all of which have an equivalent holo form; based on the classification proposed by Schmidtke and Barril,[12] 132 of these were classified as druggable, and seven, classified as nondruggable by Schmidtke and Barril,[12] were annoted as less druggable in this analysis. Overall, the deformation between holo and apo proteins is limited; the mean root-mean-square deviation (RMSD) for all α carbon atoms is 0.89 Å ± 1.08 Å.

Pocket Estimation. Different observed and predicted pocket estimation methods were applied to estimate pockets.

Observed Pocket Estimation Based on Ligand Proximity. The observed pocket estimation, referred to as prox, corresponds to the extraction of protein atoms localized within a fixed distance threshold from a binding ligand. In the literature, various distance thresholds ranging from 4 to 6 Å have been used for pocket identification, particularly in pocket comparison studies.[21−24] Distance thresholds translate into the inclusion of various protein-to-ligand molecular interactions; e.g., from 2.2 to 3.5 Å includes hydrogen bond interactions, and a longer threshold starting from 4.0 Å also includes aromatic interactions and hydrophobic contacts. In this study, a range of distance thresholds was examined, i.e., from 3 to 7 Å, using druggability prediction models that were specifically built for observed pocket estimation based thresholds of 4 Å (prox4) and 5.5 Å (prox5.5).

Predicted Pocket Estimation Methods for Apo or Holo Proteins: fpocket and DoGSite. The predicted pocket estimation methods considered here are based on two automated geometry-based methods that investigate all of the cavities of a protein independently of any ligand proximity information.

The first predicted pocket estimation method considered was fpocket,[6] recently used as a pocket estimation method in the Tang and Altman[25] and Dance[26] studies. This method is based on the decomposition of a 3D protein into Voronoi polyhedrals. fpocket extracts all of the pockets from the protein surface using spheres of varying diameters. Its advantages include ease of use, source code availability, calculation speed (i.e., it requires only a few seconds to estimate all protein pockets), and satisfactory performance in terms of overlaying known binding sites with the predicted sites.[12]

The other predicted pocket estimation method considered was DoGSite,[19] recently used as a pocket estimation method in the Masini et al.[27] and Sivakumar and Niranjali Devaraj[28] studies. This pocket estimation method is based on a grid that spans the area surrounding the protein with a difference of Gaussian (DoG) filter, which is used to identify the protein surface positions suitable for accommodating ligand atoms. DoGSite can be used to detect all of the pockets on the protein surface and splits the detected pockets into subpockets.

In this study, the predicted pocket of interest from the entire set of candidate cavities detected was selected as the one that best overlaps the ligand binding site of interest in the holo form or the corresponding holo form of an apo protein. This selection was performed after superimposition of the apo and corresponding holo proteins using the TMalign[29] software. Seven holo proteins corresponding to 11 apo proteins included in Apo139 are included in the NRDLD set.

Different Estimated Pocket Data Sets. The NRDLD set was estimated using four different observed and predicted pocket estimation methods, namely prox4, prox5.5, fpocket, and DoGSite, which resulted in different estimated pocket sets, referred to as the prox4-NRDLD, prox5.5-NRDLD, fpocket-NRDLD, and DoGSite-NRDLD sets, respectively. The NRDLD set was then split into one training set and one test set based on an identical division used by Krasowski et al.[14] and Desaphy et al.[11] The 76 binding sites of the training set were estimated by four estimation methods, and the results are referred to as prox4-training, prox5.5-training, fpocket-training, and DoGSite-training. The fpocket-training set includes 74 estimated binding pockets instead of the 76 from the training set of Krasowski et al. because two pockets could not be estimated using fpocket (from proteins 1RNT and 1QS4). The NRDLD test set was estimated by four estimation methods, which resulted in exactly 37 estimated pockets (as in the Krasowski et al.[14] and Desaphy et al.[11] pocket test set), referred to as prox4-test, prox5.5-test, fpocket-test, and DoGSite-test sets.

Only the two predicted pocket estimation methods, i.e., fpocket and DoGSite, can be applied to estimate Apo139, and this yielded two 139-estimated-pocket validation sets, referred to as fpocket-Apo139 and DoGSite-Apo139, respectively.

Pocket Comparison. The overlap between two estimated pockets was quantified using two scores:

- The Score of Overlap (SO) indicates the overlap between the two pocket estimates, i.e., pocket1 and pocket2, as follows:

$$\mathrm{SO} = \frac{100 \times \mathrm{Ncommon}}{\mathrm{Nestim1pocket} + \mathrm{Nestim2pocket} - \mathrm{Ncommon}}$$

where Npocket1 and Npocket2 are the number of atoms in pocket1 and pocket2, respectively, and Ncommon is the number of atoms common to pocket1 and pocket2. SO yields values between 0 and 100%. An SO value of 100% indicates maximum overlap between the pair of estimated pockets used.

- Relative overlap (RO) was defined by Schmidtke and Barril[12] and indicates the overlap in terms of the exposed atoms between the two pockets for the same binding site:

$$\mathrm{RO} = 100 \times \frac{\mathrm{SApocket1} \cap \mathrm{SApocket2}}{\mathrm{SApocket1}}$$

where SApocket1 and SApocket2 are the solvent-accessible areas of pocket1 and pocket2, respectively, computed using NACCESS.[30] An RO value closer to 100% indicates all exposed areas in pocket1 are included in pocket2.

Pocket Characterization. We computed a set of 52 descriptors composed of 36 physicochemical and 16 geometrical descriptors to characterize each estimated pocket.

Geometrical Descriptors. We used 11 volume and shape descriptors computed using RADI[31] based on the convex hull, which is the smallest convex envelope that includes all of the atoms in the pocket. These 11 geometric descriptors are only based on the Cartesian coordinates of the atoms that compose the convex hull (Figure 1). In this study, the pocket volume, which is referred to as VOLUME_HULL, is strongly correlated with the DoGSite volume descriptor,[19] as quantified by their correlation coefficient of 0.98 (p value $<10^{-12}$) obtained using DoGSite-NRDLD. Three principal moments of inertia and the numbers of pocket atoms and residues are also considered as descriptors. These descriptors are listed in Table 1.

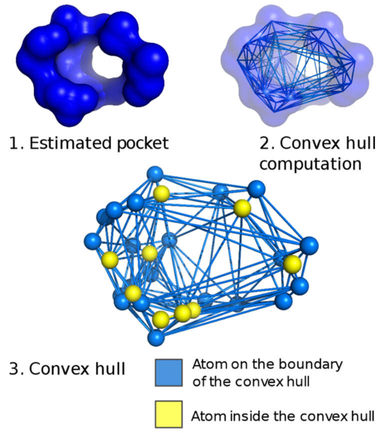

1. Estimated pocket
2. Convex hull computation
3. Convex hull
Atom on the boundary of the convex hull
Atom inside the convex hull

Figure 1. Representation of a pocket convex hull computed on phosphoenolpyruvate binding site on pyruvate phosphate dikinase (1KC7) when estimated by prox4. Atoms in blue represented atoms on the convex hull and those in yellow correspond to atoms included within the convex hull. The ratio between atoms on the convex hull and atoms within the convex hull was used to compute pocket shape descriptors detailed in Table 1.

Physicochemical Descriptors. Thirty-six physicochemical descriptors, detailed in Table 1, were used to describe the main characteristics, such as the pocket polarity, hydrophobicity,

Table 1. A List of the 52 Descriptors Used to Characterize the Pockets and Construct the Druggability LDA Models[a]

	name of descriptors	description	references
		hydrophobicity descriptors	
1	p_hydrophobic_residues	proportion of hydrophobic residues in pocket (C, G, A, T, V, L, I, M, F, W, Y, H, K)	
2	p_Hyd_atom	frequency of Hyd atoms in pocket	Milletti et al.[32]
3	hydrophobicity_pocket	hydrophobicity pocket estimated with solvent accessibility computed using NACCESS[30] software.	Burgoyne et al.[33]
4	hydrophobicity_kyte	hydrophobicity based properties of residues	Kyte et al.[35]
		aromatic descriptors	
5	p_aromatic_residues	frequency of aromatic residues in pocket (F, Y, H, W)	
6	p_Car_atom	frequency of Car atoms in pocket	Milletti et al.[32]
		polarity descriptors	
7	p_polar_residues	frequency of polar residues in pocket (C, D, E, H, K, N, Q, R, S, T, W, Y)	
8	polarity_pocket	polarity pocket estimated with sum of oxygen and nitrogen divide by sum of nitrogen, oxygen and carbon	Eyrish and Helms[34]
		physicochemical descriptors	
9	p_charged_residues	frequency of charged residues in pocket (D, E, R, K, H)	
10	p_negative_residues	frequency of negative residues in pocket (D, E)	
11	p_positive_residues	frequency of positive residues in pocket (H, K, R)	
12	charge	global charge in pocket based of charged residues (D, E, R, K, H) and ion included in pocket.	
13	p_aliphatic_residues	frequency of positive residues in pocket (I, L, V)	
14	p_Nlys_atom	frequency of Nlys atoms in pocket	
15	p_Ntrp_atom	frequency of Ntrp atoms in pocket	
16	p_Ocoo_atom	frequency of Ocoo atoms in pocket	Milletti et al.[32]
17	p_Cgln_atom	frequency of Cgln atoms in pocket	Milletti et al.[32]
18	p_Ccoo_atom	frequency of Ccoo atoms in pocket	Milletti et al.[32]
19	p_Carg_atom	frequency of Carg atoms in pocket	Milletti et al.[32]
20	p_S_atom	frequency of S atoms in pocket	Milletti et al.[32]
21	p_Otyr_atom	frequency of Otyr atoms in pocket	Milletti et al.[32]
22	p_Ooh_atom	frequency of Ooh atoms in pocket	Milletti et al.[32]
23	p_O_atom	frequency of O atoms in pocket	Milletti et al.[32]
24	p_N_atom	frequency of N atoms in pocket	Milletti et al.[32]
25	p_ND1_atom	frequency of ND1 atoms in pocket	Milletti et al.[32]
26	p_NE2_atom	frequency of NE2 atoms in pocket	Milletti et al.[32]
27	p_pro_residues	frequency of proline in pocket (P)	
28	p_small_residues	frequency of small residues in pocket (C, V, T, G, A, S, D, N, P)	
29	p_tiny_residues	frequency of tiny residues in pocket (A, C, G, S)	
30	p_main_chain_atom	frequency of main chain atom in pocket	
31	p_side_chain_atom	frequency of side chain atom in pocket	
32	p_C_atom	frequency of C atoms in pocket	Milletti et al.[32]
33	p_sulfur_atom	frequency of sulfur in pocket	
34	p_carbon_atom	frequency of carbon in pocket	
35	p_oxygen_atom	frequency of oxygen in pocket	
36	p_nitrogen_atom	frequency of nitrogen in pocket	
		volume descriptors	
37	RADIUS_HULL	radius of the smallest enclosing sphere	Petitjean[31,43]
38	SURFACE_HULL	surface of convex hull	
39	DIAMETER_HULL	longest distance in the convex hull	Petitjean[31,43]
40	VOLUME_HULL	volume of convex hull	RADI software[31]
41	FACE	number of faces in convex hull	RADI software[31]
42	SMALLEST_SIZE	distance separating the two closest slabs enclosing the hull	RADI software[31]
43	RADIUS_CYLINDER	radius of the smallest height cylinder enclosing the hull	Petitjean[31,43]
44	%_ATOM_CONVEX	frequency of pocket atoms in convex hull boundary	RADI software[31]
45	C_ATOM	number of atoms in pocket	
46	C_RESIDUE	number of residues in pocket	

Table 1. continued

	name of descriptors	description	references
		shape descriptors	
47	PSI	PSI: Pocket Sphericity Index is the ratio of the radius of the largest sphere inscribed in the hull to radius of the smallest enclosing sphere. Closer PSI is to 1, more spherical the pocket is. A small PSI value indicates that the pocket hull is flat.	Petitjean
48	PCI	PCI: Pocket Convexity Index is the ratio of the mean of the squared distances of the atoms to their hull to the squared radius of the largest inscribed sphere. When PCI = 0, all pocket atoms lie on the convex hull boundary. A PCI value close to 1 indicates a large proportion of pocket atoms interior to the pocket and far from the hull, so that the convex shape approximation is very bad.	Petitjean
49	CONVEX_SHAPE_COEFFICIENT	shape coefficient	Petitjean[31,43]
50	INERTIA_1	largest eigenvalue of inertia matrix	RADI software[31]
51	INERTIA_2	second largest eigenvalue of inertia matrix	RADI software[31]
52	INERTIA_3	smallest eigenvalue of inertia matrix	RADI software[31]

[a]The set is divided into six classes: hydrophobic descriptors, aromatic descriptors, polar descriptors, other physicochemical descriptors, volume descriptors, and shape descriptors. The descriptors PSI and PCI were developed using the framework herein. These descriptors are intended to supply pocket shape information.

atom and residue compositions, aromaticity, and solvent accessibility.[32−35] Some pocket properties were differently quantified; for example, the hydrophobicity information was calculated using the solvent exposition calculation[33,34] obtained using the NACCESS software[30] or the residue hydrophobicity score.[35] This large pool of descriptors, which exhibits some redundancy, is useful at this stage for selecting the descriptors that are most involved in the druggability pocket.

Statistical Protocol. *Supervised Learning Method: LDA.* The supervised learning method used in this analysis was LDA, which can determine which variable discriminates between two or more classes and derive a classification model that predicts the group membership of new observations, based on Fisher's linear discriminant methods.[36] LDA enables the identification of optimal linear descriptor combinations to discriminate druggable from less druggable pockets. The advantage of LDA models is that they provide direct information regarding which descriptors are important in pocket druggability. Moreover, LDA provides an easy model for interpreting a probability of druggability from 0 to 1 that corresponds to the probability that a pocket is druggable; a pocket with a score greater than 0.5 is considered druggable.

Quality Criteria. The model quality was estimated using two statistical criteria: Matthew's correlation coefficient (MCC) and accuracy. MCC is an overall rate that is pertinent for druggability optimization because it considers the imbalance ratio between druggable and less druggable pockets in the data set. The ability to predict druggable pockets is assessed based on the model's sensitivity and ability to predict less druggable pockets through specificity:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{MCC} = \frac{\text{TP}\cdot\text{TN} - \text{FP}\cdot\text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

True positive (TP) refers to the number of druggable pockets that are correctly classified, and false negative (FN) is the number of druggable pockets that are incorrectly classified. True negative (TN) is the number of less druggable pockets that are correctly classified, and false positive (FP) is the number of less druggable pockets that are incorrectly classified.

l-Fold Cross-Validation. l-fold cross-validation was used during PockDrug construction. The first step of the cross-validation procedure consisted of randomly splitting the set into one training subset and one validation subset without changing the ratio between the druggability classes. The second step consisted of training the model using the training subset and to assess its performance using the validation subset. Steps 1 and 2 were repeated l times with random splits. The performances obtained with l training and l validation subsets were averaged and presented with the corresponding standard deviations

Construction of Druggability Prediction Models. PockDrug was constructed based on three steps using the NRDLD set, as illustrated in Figure 2.

Step 1: Selection of p LDA models using the NRDLD training set

LDA models were trained using the NRDLD training set, and the p best LDA models were selected using the MCC performance by cross-validation (l = 10). The main goal of this step was to select a set of p parsimonious LDA models (combining a minimal number of descriptors) with good MCC performances by cross-validation. The number of descriptors was increased in a stepwise manner among the 52 pocket descriptors used to characterize the estimated pockets (see detailed protocol in Supporting Information, Figure S3). This LDA model training step was applied using four estimated training sets, but only the p best models trained using the fpocket-training set were kept as the most efficient for the subsequent steps in the protocol.

Step 2: Selection of the k best LDA models that are stable and efficient regardless of the choice of pocket estimation method using one independent NRDLD test set

To select the most efficient models, regardless of the choice of pocket estimation method, the p LDA models selected in step 1 were reduced to the k LDA models that exhibited the best performance using the NRDLD test set (not used for training) with different pocket estimation methods. We selected

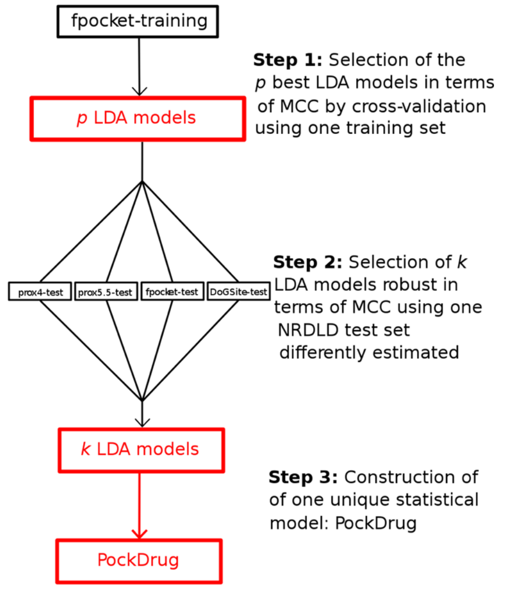

fpocket-training
Step 1: Selection of the p best LDA models in terms of MCC by cross-validation using one training set
p LDA models
prox4-test
prox5.5-test
fpocket-test
DoGSite-test
Step 2: Selection of k LDA models robust in terms of MCC using one NRDLD test set differently estimated
k LDA models
Step 3: Construction of of one unique statistical model: PockDrug
PockDrug

Figure 2. Representation of the three steps to construct PockDrug: (Step 1) Selection of the *p* best LDA models in terms of MCC by cross-validation using NRDLD training set (fpocket-train) with a minimal number of descriptors. (Step 2) Selection of *k* ($k < p$) LDA models among *p* models selected in step 1, robust in terms of MCC using NRDLD test set, estimated using four different estimation methods. (Step 3) Construction of the unique statistical druggability model PockDrug, which provides one druggability probability, computed as the mean of the *k* probabilities of the *k* LDA models selected in step 2. The estimated NRDLD subsets are indicated in black and different selected LDA models steps in red.

a low number (*k*) of the "best" LDA models that provided the best and similar results to the MCC results using the four estimated NRDLD test sets (i.e., prox4-test, prox5.5-test, fpocket-test, and DoGSite-test sets).

Step 3: PockDrug: average probability of *k* best models

The *k* best LDA models were then combined into PockDrug. PockDrug is a statistical model that takes an estimated pocket as input and provides as output one druggability probability, which is computed as the mean of the *k* probabilities of the *k* models, and an associated standard deviation that indicates the druggability probability confidence on the *k* models.

PockDrug Validation Using One apo Pocket Set. PockDrug was then validated using the apo pocket set Apo139 estimated using two different predicted pocket estimation methods (i.e., fpocket-Apo139 and DoGSite-Apo139) to assess the robustness of PockDrug based on the choice of apo pocket estimation method.

Druggability Model Comparison. The PockDrug druggability prediction performances were then compared in terms of the MCC and accuracy on both holo and apo pockets with the druggability results reported in the literature.

The efficiencies of PockDrug for holo pockets were compared using the NRDLD test set. The PockDrug comparisons concerning observed pockets were made against the DrugPred[14] and Desaphy et al.'s[11] model trained using the NRDLD set. Desaphy et al.'s model[11] used Volsite descriptors; it is based on geometry/energy but uses ligand information to identify the binding pocket boundaries. DrugPred uses the shape of a pseudoligand formed by the docked poses of 995 molecules to detect the boundaries of the unliganded binding cavities.[14]

The PockDrug comparisons concerning predicted pockets were made against three druggability prediction models: SiteMap,[10] fpocket druggability score,[12] and DoGSiteScorer[15] (available on http://dogsite.zbh.uni-hamburg.de/). These models are based on predicted pocket estimation methods that do not rely on any ligand information to identify the boundaries of the binding pocket.[12]

The PockDrug comparisons concerning apo pockets were made against the fpocket druggability score and DoGSiteScorer obtained using Apo139.

RESULTS AND DISCUSSION

Pocket Estimation and Characterization. *Pocket Estimations.* The set of 113 NRDLD binding sites were estimated using different observed and predicted pocket estimation methods.

Observed pocket estimation was performed based on proximity to the ligand, i.e., using the prox method, with distance thresholds ranging from 3 to 7 Å. The choice of the threshold to estimate observed pockets was recently shown to have a strong influence on the pocket descriptors, e.g., as shown by Krotzky and Rickmeyer[37] using the ATP binding site. We found that the overlap (SO) between the observed pockets obtained with different distance thresholds and predicted pockets (estimated by the fpocket and DoGSite methods) increased as the threshold was increased to 5.5 Å and then decreased progressively using the NRDLD set, as shown in Figure S1. A distance threshold of 5.5 Å for prox maximizes the SO between the observed pockets and corresponding predicted pockets. We then decided to consider two different distance thresholds to identify the observed pockets using the prox method: one stringent threshold of 4 Å, which is referred to as prox4, as used by Krasowski,[14] and a longer threshold of 5.5 Å, which is referred to as prox5.5. Prox4 enables the extraction of a well-defined pocket limited to short interactions with a ligand (as hydrogen bonds or ionic interactions), whereas prox5.5 enables the identification of all significant contact points and a more complete environment of the binding site.

Thus, the NRDLD set was estimated using four estimation methods: prox4, prox5.5, fpocket, and DoGSite. fpocket-NRDLD exhibits less overlap with prox5.5-NRDLD (SO = 36% ± 12%) than DoGSite-NRDLD (SO = 49% ± 18%). These relatively weak average SO values between these four NRDLD estimated pocket sets, which exhibited a minimal value of 29% ± 13% between prox4 and DoGSite, illustrates the difficulties in finding a common definition for pocket boundaries.[38] However, the four NRDLD estimated pocket sets exhibited a good mean RO (relative to the number of common exposed atoms between the pocket estimations) between the observed and predicted pockets from 76% ± 23% between prox4 and DoGSite to 85% ± 20% between prox5.5 and fpocket. These high RO values show that the predicted pockets, as estimated by fpocket or DoGSite, include the correct exposed atoms in the pocket that interact with the ligand, as estimated by prox4 or prox5.5.

Pocket Characterization. These four estimated NRDLD pocket sets were characterized using the set of 52 pocket descriptors detailed in Table 1. The mean and standard deviation of the 52 descriptors set are computed and compared using four pocket sets, see Table S3. Most of the geometrical descriptors and some physicochemical ones are highly dependent on the estimation methods used. For instance, the mean count of atoms (C_ATOM) is highly dependent on the

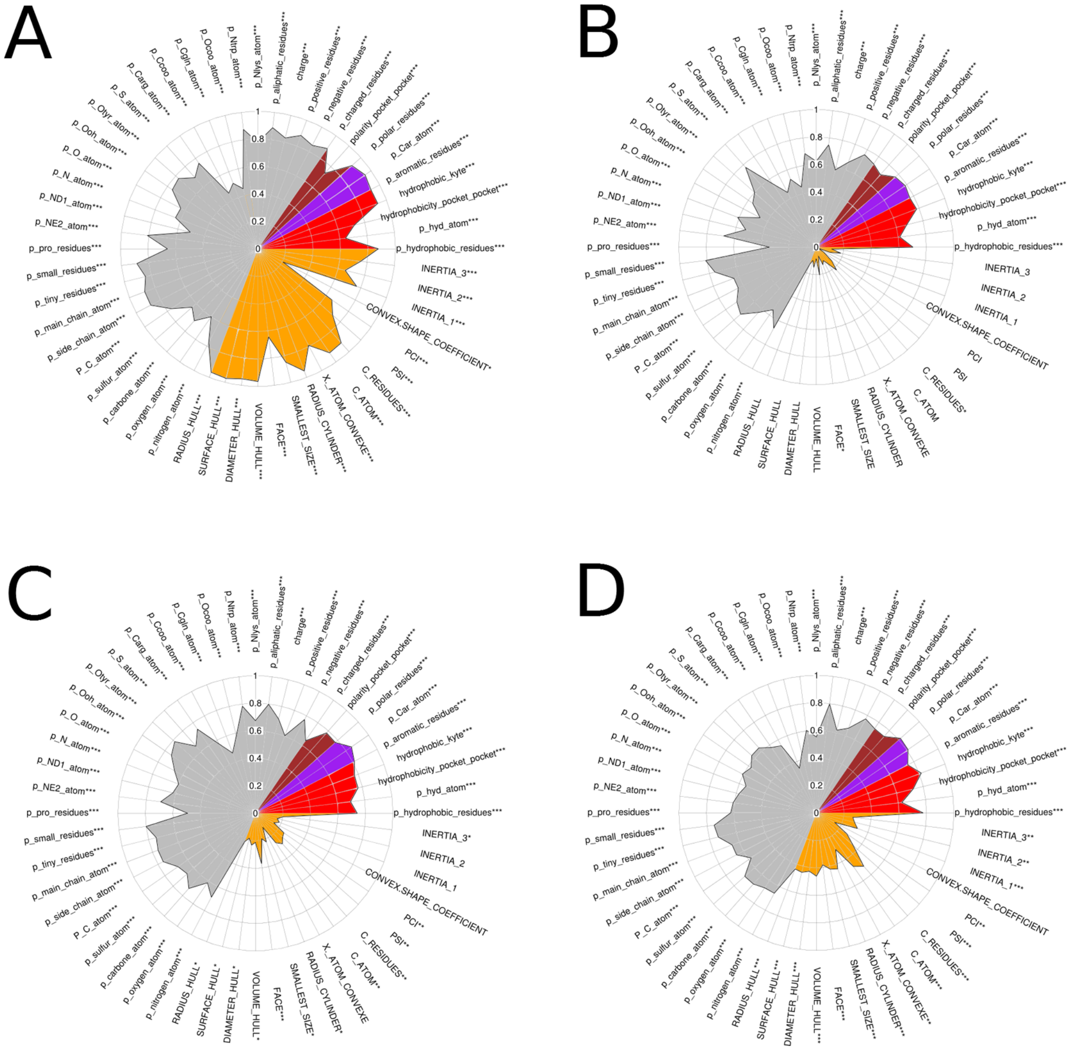


Figure 3. Representation on radar plot of linear correlation coefficients for 52 descriptor set between NRDLD set estimated by different methods on radar plots: (A) between prox4 and prox5.5, (B) between prox5.5 and fpocket, (C) between prox5.5 and DoGSite, (D) between fpocket and DoGSite estimations. The descriptors are colored based on their type, see Table 1: red for hydrophobic descriptors, brown for polar descriptors, purple for aromatic descriptors, yellow for geometrical descriptors, and gray for other physicochemical descriptors. Statistical significance corresponding to p value <0.05, <0.01, and <0.001 are indicated respectively by *, **, and ***.

pocket estimation method considered: prox4-NRDLD exhibited a significantly lower mean C_ATOM, namely 37 ± 15, than that obtained with prox5.5-NRDLD, namely 96 ± 32 (p value < 2.2×10^{-16}). fpocket estimates smaller pockets on average, as quantified by the mean C_ATOM; the value of 88 ± 57 obtained for fpocket-NRDLD was significantly smaller than the value of 116 ± 63 obtained for DoGSite-NRDLD (p value < 5.8×10^{-4}) in agreement with fpocket-NRDLD weaker overlap with prox5.5-NRDLD than DoGSite-NRDLD. Some others physicochemical descriptors are not significantly dependent on pocket estimation methods, such as polarity and aromaticity descriptors. We observe that pocket hydrophobicity quantification is more or less dependent on the pocket estimation methods, according to the hydrophobicity descriptor considered.

The comparison of the 52 pocket descriptors using the observed NRDLD pockets estimated using prox4 and prox5.5 revealed linear correlation coefficients with a high mean value of 0.75 ± 0.15, as shown in Figure 3A. Forty-one percent of the descriptors presented a high correlation of more than 0.8, corresponding primarily to some geometrical or physicochemical descriptors based on residue information, whereas atomistic information, which depends more on the exact boundaries of the pockets, are less correlated on average (0.70 ± 0.14). This is consistent with the fact that similar but larger pockets were estimated by increasing the distance threshold, resulting in a proportional pocket volume and similar shapes but the inclusion of supplementary atoms in the estimated pockets.

The comparison of the 52 pocket descriptors between the observed and predicted pockets revealed relatively weak mean correlation coefficients: 0.49 ± 0.27, 0.55 ± 0.26, and 0.57 ± 0.18 between prox5.5 and fpocket, between prox5.5 and DoGSite, and between DoGSite and fpocket, respectively. The corresponding 52 correlation coefficients are illustrated in Figure 3. Using the 16 pocket geometrical descriptors, we noted a relatively weak correlation between the four estimated pocket sets (less than 0.4). This weak correlation of geometrical descriptors is consistent with the influence of the pocket estimation methods on the pocket geometry, shape, and boundaries, as described by Gao et al.[17] Conversely, using the 36 physicochemical descriptors, we found that the hydrophobic, aromatic, and polarity environment information was well

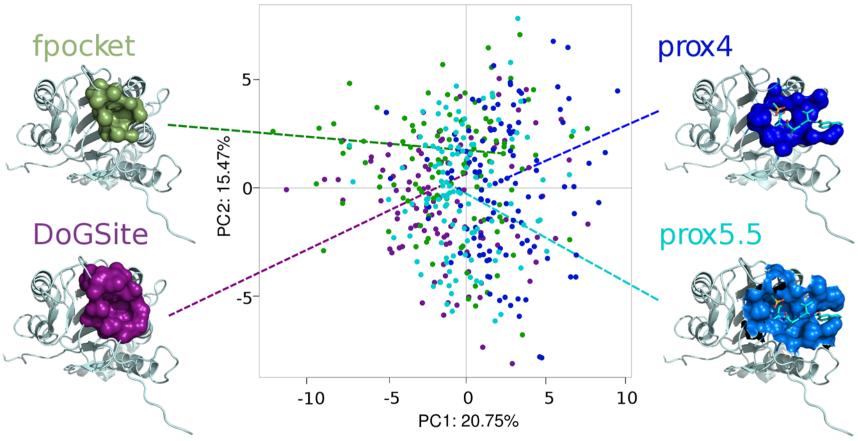


Figure 4. Representation of four estimated NRDLD pocket sets: prox4-NRDLD, prox5.5-NRDLD, fpocket-NRDLD, and DoGSite-NRDLD on first PCA plane computed using the 52 descriptor set. The first PCA plane explains more than 35% of the variability: the first principal component PC1 explains 20.75%, and the second principal component PC2 explains 15.47%. Different estimated pockets are colored based on the estimation method used: blue for prox4-NRDLD, light blue for prox5.5-NRDLD, green for fpocket-NRDLD, and purple for DoGSite-NRDLD. Interleukin-1beta binding site (1BMQ) estimated by the four pocket estimation methods is illustrated.

correlated between the four estimated pocket sets. For example, the hydrophobic_kyte, p_aromatic_residues, and p_polar_residues descriptors always exhibited correlation values greater than 0.7.

Finally, geometric pocket descriptors are confirmed to be highly dependent on the pocket estimation methods, specifically between observed and predicted pockets, while some physicochemical descriptors, such as polar and hydrophobic, appear to be stable and robust with respect to pocket estimations.

A principal component analysis (PCA) was performed based on the 52 descriptors using the four NRDLD estimated pocket sets. Figure 4 presents the projections of the four pocket sets on the first PCA plane, which explains more than 35% of the descriptor variability. The 52 descriptors sampled different parts of the first PCA plane (Figure 5), and certain geometrical or physicochemical descriptors were highly correlated between them. The main explanation for the first principal component (PC) is geometrical descriptors. For example, large pockets associated with high VOLUME_HULL or C_ATOM values were located on the left portion of the first PCA axis. The main explanation for the second PC is physicochemical descriptors; hydrophobic and aromatic descriptors were inversely correlated with the polar descriptors. Globally, the four pocket sets appeared close in the first PCA plane (Figure 4), which is consistent with high RO values between different pocket estimations. Notably, the prox4-NRDLD pockets were localized to the right of the first PCA plane, corresponding to rather small pockets compared with the three other estimated pocket sets (as quantified by the smaller mean C_ATOM value). In addition, the four pocket estimations for one binding site vary and may not be positioned close on the first PCA plane, which is consistent with the relatively weak overlap between the four estimated pockets sets according to the SO. This observation is exemplified by the interleukin-1beta inhibitor binding site (1BMQ) shown in Figure 4. The pocket druggability information (druggable versus less druggable pockets) was then visualized on the first PCA plane (Figure 6). The druggable pockets were primarily located in the same portion of the PCA space, regardless of the pocket estimation method used, illustrating that druggability information may overcome the inaccuracy of the estimation methods. Consistent with Krasowski et al.'s PCA results,[14] the druggable pockets primarily corresponded to large, hydrophobic and weakly polar pockets on the first PCA plane, which demonstrates that the 52 pocket descriptors may potentially discriminate druggable and less druggable pockets. However, this discrimination must be improved using a statistical learning method, such as LDA models, to select the optimal combination of the descriptors involved in druggability.

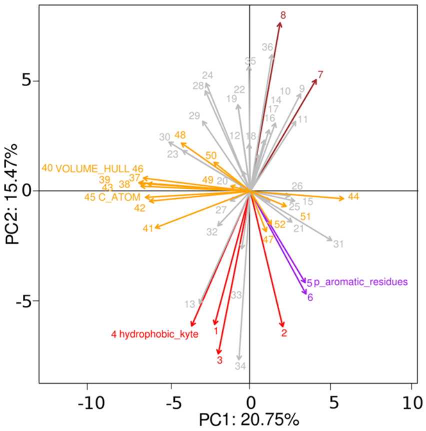


Figure 5. Projection of the 52 descriptor set on the first PCA plane, explaining more than 35% of the variability. The descriptors are colored based on their type: red for hydrophobic descriptors such as hydrophobic_kyte, brown for polar descriptors, purple for aromatic descriptors such as p_aromatic_residues, yellow for geometrical descriptors such as VOLUME_HULL and C_ATOM, and gray for other physicochemical descriptors. The descriptors are indicated by their number, listed in Table 1.

PockDrug Construction and Performances. *PockDrug Model Construction.* In the first step of the PockDrug model construction, p = 184 LDA models including three descriptors were selected using the fpocket-training set. These 184 LDA models had quite similar results on training and validation subsets obtained by cross-validation; see the detailed protocol in the Supporting Information, Table S1. The second step consisted of selecting the most robust LDA models in terms of MCC and accuracy performances using the NRDLD test set

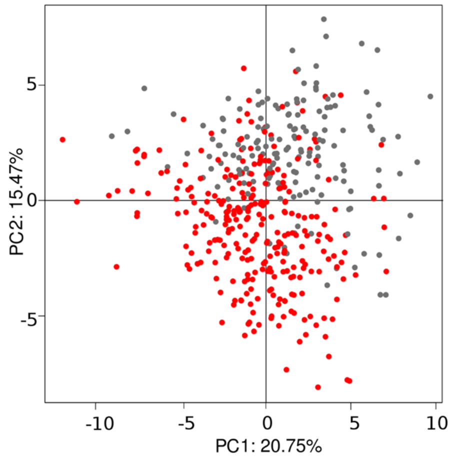


Figure 6. Projections of the NRDLD pockets estimated by four different estimation methods (prox4-NRDLD, prox5.5-NRDLD, fpocket-NRDLD, and DoGSite-NRDLD) in the first PCA plane, explaining more than 35% of variability. Estimated pockets are colored based on their druggability classes: the druggable pockets are in red, and the less druggable pockets are in gray.

estimated by four different methods, as shown in Figure 2. This resulted in the k = 7 best LDA models that provide similar druggability probabilities, as confirmed by a mean standard deviation of only 0.031 for the seven probabilities obtained by pocket. These seven models discriminate druggable from less druggable pockets with mean probabilities of 0.87 ± 0.15 for the druggable pockets and 0.18 ± 0.15 for the less druggable pockets using four estimated test sets, which corresponds to a high mean confidence probability. These seven best LDA models were combined into PockDrug.

PockDrug Performances Using Differently Estimated NRDLD Test Set. To ensure that the difference between the druggability results were due to the quality of the druggability models and not to the pocket estimation methods upon which the druggability models are based, these comparisons were made using the same pocket estimation method when available. For example, comparisons of PockDrug to the fpocket score were performed using pocket sets (apo or holo) estimated by fpocket and to DoGSiteScorer using pocket sets estimated by DoGSite.

The PockDrug performances in terms of accuracy, sensibility, specificity, and MCC using different estimated NRDLD test sets are summarized in Table 2. The mean and associated standard deviation performances obtained on the seven best LDA models used in PockDrug are indicated in Table S2. Globally, PockDrug exhibited good performances with the four estimated NRDLD test sets: accuracy greater than 83% and MCC greater than 0.650. The PockDrug performances with the prox4-test and prox5.5-test sets exhibited accuracies of 94.8% and 83.8% and MCC values of 0.885 and 0.655, respectively, as shown in Table 3. This finding illustrates the influence of pocket estimation methods on computational druggability predictions. Comparisons with the Desaphy et al.[11] and DrugPred[14] models with performances are summarized in Table 3 and were made using the same test set, differently estimated. Thus, ensuring that the differences between the model performances are due to the models and not to the quality of the pocket estimation methods used is difficult. For example, PockDrug outperforms (more than 0.10) the Desaphy et al.[11] and DrugPred[14] models in terms of the MCC results when using the prox4-test set but exhibits inferior performance when using the prox5.5-test set. For prox4, the estimation method kept only the atoms in the protein that were in close interaction with the ligand, and the estimated pockets precisely surrounded the ligand. This indicates that this estimation method is susceptible to the inclusion of druggability information without being affected by more distance atoms, which can explain the good PockDrug performance. The application of PockDrug to the prox5.5-test set resulted in inferior performance. The prox5.5 method extracts fuzzier pockets, is thus susceptible to the inclusion of significant points of binding to different types of ligand, and exhibits more overlap with the predicted pockets. Thus, the PockDrug performance on the prox5.5-test set resembles the PockDrug performance obtained with the predicted test sets (i.e., fpocket-test and DoGSite-test sets). When applied to the predicted fpocket-test and DoGSite-test sets, PockDrug exhibited accuracy greater than 86.5% and an MCC value higher than 0.71. It outperformed the recently proposed druggability models[12,15] fpocket score and DoGSiteScorer when applied on the NRDLD test set estimated by identical pocket estimation methods. PockDrug increased the accuracy by at least 10% and the MCC by at least 0.20. These results demonstrate that PockDrug performs better than the literature concerning predicted pocket and robustly compared with different pocket estimation methods.

Table 2. PockDrug Performances Using prox4-test, prox5.5-test, fpocket-test, and DoGSite-test Pocket Sets

druggability model	PockDrug			
estimated pocket test sets	prox4-test	prox5.5-test	fpocket-test	DoGSite-test
accuracy	94.6	83.8	86.5	86.5
sensitivity	95.7	95.7	95.7	100
specificity	92.9	64.3	71.4	64.3
MCC	0.885	0.655	0.712	0.727

Table 3. Performances of the Druggability Models in the Literature on NRDLD Test Set, Estimated by Different Pocket Estimation Methods

druggability models	Volsite[11]	DrugPred[14]	SiteMap[10]	fpocket[12]	DoGSite[15]
estimated pocket test sets	NRDLD test set estimations corresponding to associated druggability model.[10,11,14]			fpocket-test	DoGSite-test
accuracy	89	89	65	76	76
sensitivity				74	70
specificity				79	86
MCC	0.77	0.77	0.24	0.51	0.54

PockDrug Extrapolation to apo Pockets. The robustness of PockDrug with respect to different pocket estimation methods is promising for extrapolation to apo pockets. PockDrug can be directly applied to the prediction of the druggability of apo pockets using pocket prediction estimation methods. It yielded reasonable MCC values of 0.450 and 0.515 for fpocket-Apo139 and DoGSite-Apo139, respectively, due to the imbalance between the number of druggable and less druggable pockets in the apo set. However, PockDrug yielded high accuracies of 91.4% and 93.5%, respectively, due to its high sensitivity (Table 4). The PockDrug performances on apo pockets were then compared with the performances of recently proposed

Table 4. PockDrug, fpocket Score, and DoGSiteScorer Performances Using fpocket-Apo139 and DoGSite-Apo139 Sets

druggability models	PockDrug		fpocket score[12]	DoGSiteScorer[15]
estimated apo pocket sets	fpocket-Apo139	DoGSite-Apo139	fpocket-Apo139	DoGSite-Apo139
accuracy	91.4	93.5	47.5	79.1
sensitivity	92.4	94.7	44.7	78.8
specificity	71.4	71.4	100	85.7
MCC	0.450	0.515	0.198	0.328

druggability prediction models using identical estimation methods: fpocket druggability score[12] using fpocket-Apo139 and DoGSiteScorer[15] using DoGSite-Apo139. PockDrug outperformed these two recently proposed models; the MCC increased by at least 0.10, and the accuracy increased by at least 13% (Table 4). These results demonstrate that PockDrug efficiently predicts the druggability of an apo pocket, and its performance is stable regardless of the choice of predicted pocket estimation method.

PockDrug Characteristics. PockDrug is based on a consensus of seven LDA models, each of which includes only three descriptors. PockDrug involves nine different descriptors distributed in a balanced way on three key characteristics: hydrophobic information corresponding to only one descriptor, geometrical information corresponding to six descriptors, and two other physicochemical descriptors corresponding to aromatic and hydroxyl group information, as shown in Figure 7A. These seven models combined hydrophobic information with geometrical information, i.e., volume information complemented by aromatic or hydroxyl group characteristics (Figure 7B). Five LDA models involve similar descriptor combinations with identical hydrophobic and aromatic descriptors (named hydrophobic_kyte and p_aromatic_residues) but different geometrical descriptors as the third descriptor. The six geometrical descriptors exhibited high correlations on fpocket-NRDLD, i.e., greater than 0.82, as shown in Figure S2. Two other LDA models combine geometrical and hydrophobic information with hydroxyl group information rather than aromatic residue information. This atomistic information, i.e., named p_Otyr_atom, characterizes a hydrogen bond donor group: a hydroxyl group in the tyrosine side chain (Otyr), which plays a key role in binding drug-like molecules. This group is linked to aromatic residue information (p_aromatic_residues) in the manner it directly influences the pocket aromaticity.

This finding demonstrates that the addition of the effects of aromatic or hydroxyl groups on geometrical and hydrophobic information results in similar druggability prediction performance. This result can be explained by the contribution of aromatic or hydroxyl groups with pocket-ligand interactions (H-bond, $\pi-\pi$ stacking, and aromatic interactions), which favor interactions between drug-like molecules and pockets. Their effects were statistically demonstrated by comparing seven LDA models based on three descriptors with LDA models based on only two descriptors in terms of their performance when applied to the fpocket-test set (Table 5). The mean model accuracy increased from 83.3% ± 1.0%, which was obtained with the combination of only hydrophobicity and geometrical information, to 86.5% ± 0.0% after the addition of aromaticity

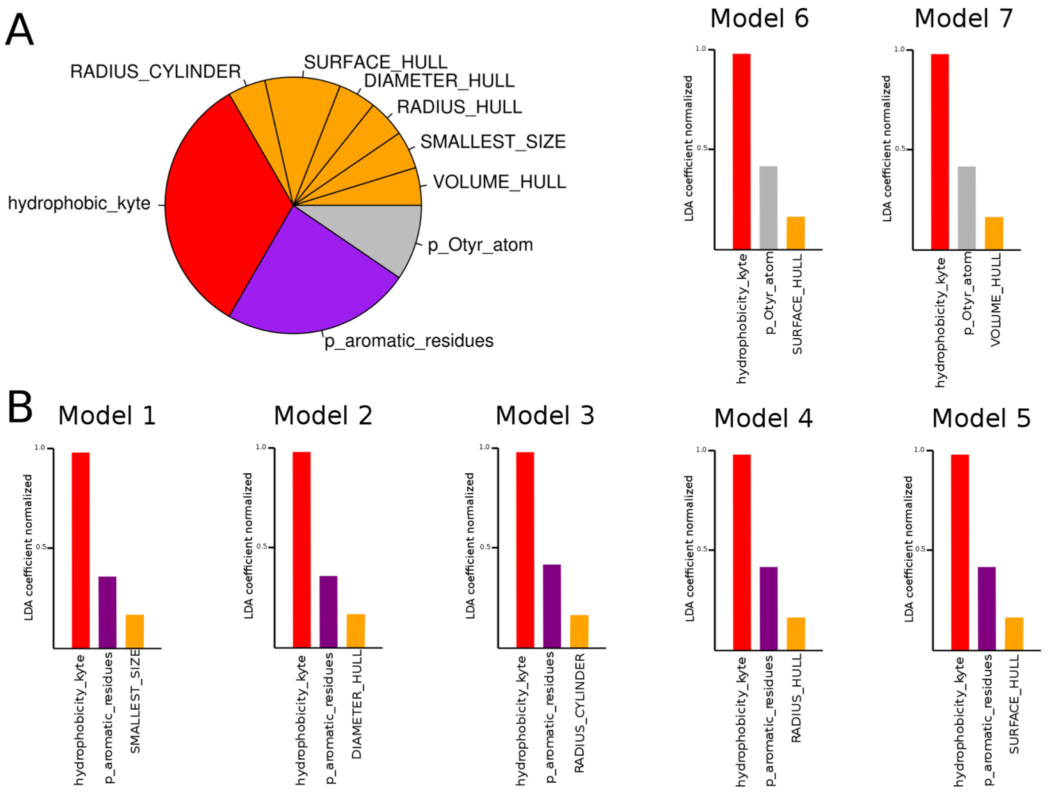
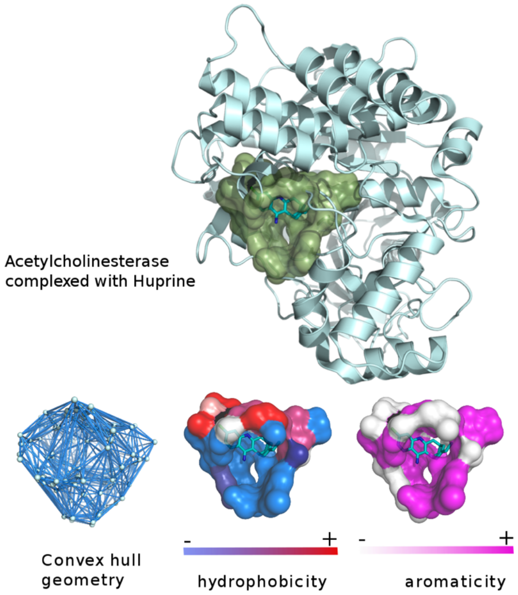


Figure 7. (A) Pie chart showing the distribution of the nine descriptors included in PockDrug. (B) Representation of the descriptors combination corresponding to the k = seven best LDA models of PockDrug. Each descriptor is represented by its normalized LDA coefficient. Descriptors associated with hydrophobicity are in red, with aromaticity in purple, geometric descriptors in yellow, and others in gray.

Table 5. Performances of LDA Models Using fpocket-test Based on a Combination of Two Descriptors Amongst Those Included in PockDrug: Hydrophobicity, Geometric, and Physicochemical with Aromatic

descriptors	hydrophobicity + geometric	hydrophobicity + physicochemical with aromaticity	geometric + physicochemical with aromaticity
estimated pocket test sets		fpocket-test	
accuracy	83.3 ± 1.0	74.8 ± 9.2	63.3 ± 1.3
sensitivity	93.5 ± 2.2	94.2 ± 4.1	100 ± 0.0
specificity	66.7 ± 5.3	42.9 ± 31.9	3.0 ± 3.5
MCC	0.643 ± 0.021	0.393 ± 0.282	0.089 ± 0.105

or hydroxyl group descriptors, and a 20% greater specificity was obtained, demonstrating the importance of adding this information for druggability assessments. Figure 8 illustrates the pocket descriptors of the huprine binding site on acetylcholinesterase (1E66) and the druggability probability associated with different descriptors used to predict pocket druggability.



Druggability model probability.
- for LDA models using only geometric and hydrophobicity information: 0.72+/- 0.03
- for LDA models using geometric, hydrophobicity and physicochemical with aromaticity information: 0.92 +/- 0.02

Figure 8. Representation of the huprine binding site on acetylcholinesterase (1E66), estimated by fpocket. Three pocket descriptor types are represented: the geometry with pocket convex hull, the hydrophobicity, and aromaticity. The average druggability probabilities were computed from LDA models using either only two descriptors (geometric and hydrophobicity), 0.72 ± 0.02, or using three descriptors (geometric, hydrophobicity, and physicochemical with aromaticity), 0.92 ± 0.03.

The PockDrug descriptor analysis showed that the druggable pockets tend to be larger and more hydrophobic and to promote the presence of aromatic or hydroxyl group, as shown in Figure 9. PockDrug improves the druggability prediction through the exclusion of less druggable large pockets with high hydrophobicity from the druggable pocket class by considering whether the environment includes many aromatic residues or atoms. High hydrophobicity and a specific pocket shape influence all recent pocket druggability models.[4,10−15,39] Nisius et al.[40] noted that polar and ionic interactions, in addition to nonpolar interactions, significantly contribute to drug−target interactions. On the basis of physicochemical information, the pocket environment has been described as enriched in aromatic residues, such as Phe, His, Trp, and Tyr residues,[41] but this study provides the first demonstration of the contribution of aromatic residues or hydroxyl groups to druggability.

The originality of PockDrug is that it was constructed to overcome the limits and inaccuracies of pocket estimations and be efficient with different pocket estimation methods. Volume descriptors are included in PockDrug, as in various druggability models,[4,10−15] despite their dependence on the pocket estimation method, which can be explained by the essential contribution of volume to druggability. The volume descriptors were systematically combined with hydrophobicity and aromatic (or Otyr composition) descriptors, which are less dependent on the estimation uncertainties (see Table S3 and Figure 3). These two hydrophobicity and aromatic physicochemical descriptors are highly autocorrelated (mean of 0.81 ± 0.06) between the four NRDLD estimated pocket sets, illustrating their robustness with respect to the estimation uncertainties. The P_Otyr_atom was less robust with respect to the pocket estimation uncertainties but still exhibited a relatively high mean correlation coefficient (0.67 ± 0.05 in average); thus, this descriptor was selected for combination with hydrophobicity and volume information.

PockDrug selects seven coherent descriptor combinations from the descriptors that exhibit both the properties of being connected to druggability and robustness with respect to the pocket estimation uncertainties to efficiently predict the druggability of different estimated pockets.

■ CONCLUSION

In conclusion, we propose an efficient druggability prediction model named PockDrug that successfully selects the physicochemical descriptors involved in druggability and is able to overcome pocket estimations and boundaries. This model can thus be applied for the analysis of both holo and apo pockets estimated using different methods. The influence of the pocket detection method on druggability prediction has been poorly studied to date, and this study illustrates the influence of the estimation method used on the druggability prediction. Interestingly, the findings demonstrate that a more precise pocket estimation, which is well-defined by a short interaction to the ligand (4 Å is considered in this study), allows the druggability prediction model to achieve its best performance.

The statistical models (in R and R script) associated with the use of PockDrug for pockets in which pocket descriptors are computed are available in the Supporting Information. A Web site is currently in development to facilitate the implementation of PockDrug, which is now available for any pocket estimation methods at http://pockdrug.rpbs.univ-paris-diderot.fr/.

The diversity of the protein set used to train PockDrug, which includes different enzymes or transmembrane proteins, suggests that PockDrug is applicable for a large set of proteins if the pockets are correctly estimated. PockDrug should be applicable to membrane proteins, but one must account for the membrane. Limitations could be application to interfaces, such as protein−protein interactions, in which no pocket is apparent. In terms of field applications, PockDrug should be useful (i) for

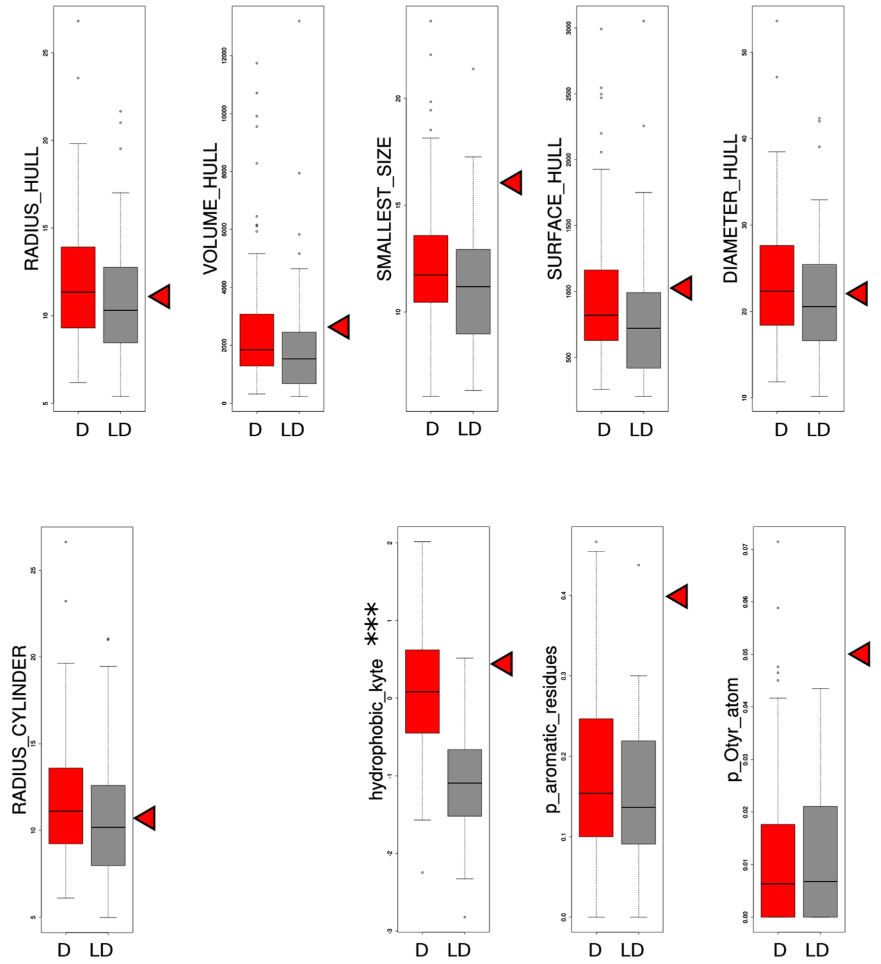


Figure 9. Boxplot distribution of nine descriptors included in PockDrug for druggable pockets (in red) and less druggable pockets (in gray) using fpocket-NRDLD. The Student's *t* test between druggable and less druggable associated with *** correspond to *p* values < 0.001. The descriptor values of the acetylcholinesterase binding site (1E66) estimated by fpocket (illustrated in Figure 8) are indicated on the right of the boxplot by a red triangle: 0.045 for the p_Otyr_atom, 0.4 for the p_aromatic_residues, −0.39 for the hydrophobic_kyte, 16.19 for the SMALLEST_SIZE, 1025.0 for the SURFACE_HULL, 22.3 for the DIAMETER_HULL, 11.44 for the RADIUS_HULL, 10.7 for the RADIUS_CYLINDER, and 2688.5 for the VOLUME_HULL.

druggability assessment to evaluate whether a compound development project is worth starting after an X-ray structure is available (for example, in academic settings), (ii) for ranking/prioritizing the pockets of an orphan protein to choose one available for compound development in structure-based drug design, (iii) in the above case, finding a druggable secondary pocket would be the most common case, and (iv) for identifying a target in a disease-modifying pathway to compare druggability predictions and select which protein is interesting to study (less likely to be in academic settings).

In the future, the proposed model could be used for the in-depth study of a large pocket descriptor profile, in agreement with pocket-ligand profile correspondence and the dual classification proposed by Pérot al.,[42] to not only predict the pocket druggability but also more precisely characterize the profile of ligands that would susceptibly bind the pocket profile.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Zip archive "PockDrug.zip" containing a R script "PockDrug.R," R environment including statistical models combined tutorial and pocket input files example. Excel files containing the 52 descriptors computed on different estimated pocket sets (used to construct and validate PockDrug): NRDLD-training, prox4-test, prox5.5-test, DoGSite-test, fpocket-test, fpocket-Apo139, and DoGSite-Apo139. Figure S1 shows values of the mean Score of Overlap (SO) (with the associated standard deviations) between the observed and predicted NRDLD pockets related to the distance threshold from 3 to7 Å used by the prox pocket estimation methods. Figure S2 illustrates the correlation matrix between the 52 descriptors using fpocket-NRDLD. The PockDrug construction protocol is detailed concerning step 1: this step is used to minimize the number of descriptors combined in the best LDA models. Figure S3 shows the first step of the statistical protocol used to construct PockDrug, and Table S1 shows the performances of LDA models on the fpocket-training set using a combination of *n* descriptors. Table S2 summarizes the mean performances with standard deviations of the seven LDA models included in PockDrug using the NRDLD test estimated by prox4, prox5.5, fpocket, and DoGSite and Apo139 sets estimated by fpocket and DoGSite. Table S3 summarizes the means and standard deviations of the 52 pocket descriptors computed on estimated pocket sets: prox4-NRDLD, prox5.5-NRDLD, fpocket-NRDLD, DoGSite-NRDLD, and *p* values associated with their comparisons. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: anne-claude.camproux@univ-paris-diderot.fr.

Present Address

UMRS-973, MTi, Université Paris Diderot, 35 Rue Hélène Brion, 75205 Paris Cedex 13, case courier 7113.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to Hiba Abi Hussein, Pr. Taboureau, and Dr. Flatters for helpful discussions.

■ ABBREVIATIONS

LDA, Linear Discriminant Analysis; NRDLD, Non Redundant data set of Druggable and Less Druggable binding sites; DCD, Druggable Cavity Directory; MCC, Matthew's Coefficient Correlation; PCA, Principal Components Analysis; RO, Relative Overlap; SO, Score of Overlap; PC, Principal Component; HTS, High-throughput screening; 3D, three-dimensional; RMSD, root-mean-square deviation

■ REFERENCES

(1) Hopkins, A. A. L.; Groom, C. R. C. The Druggable Genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727−730.

(2) Xue, T.; Ding, S.; Guo, B.; Zhou, Y.; Sun, P.; Wang, H.; Chu, W.; Gong, G.; Wang, Y.; Chen, X.; Yang, Y. Design, Synthesis, and Structure − Activity and Structure − Pharmacokinetic Relationship Studies of Novel [6,6,5] Tricyclic Fused Oxazolidinones Leading to the Discovery of a Potent, Selective, and Orally Bioavailable FXa Inhibitor. *J. Med. Chem.* **2014**, *57*, 7770−7791.

(3) Crowther, G. J.; Booker, M. L.; He, M.; Li, T.; Raverdy, S.; Novelli, J. F.; He, P.; Dale, N. R. G.; Fife, A. M.; Barker, R. H.; Kramer, M. L.; Van Voorhis, W. C.; Carlow, C. K. S.; Wang, M.-W. Cofactor-Independent Phosphoglycerate Mutase from Nematodes Has Limited Druggability, as Revealed by Two High-Throughput Screens. *PLoS Negl. Trop. Dis.* **2014**, *8*, e2628.

(4) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. *J. Med. Chem.* **2005**, *48*, 2518−2525.

(5) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-Based Maximal Affinity Model Predicts Small-Molecule Druggability. *Nat. Biotechnol.* **2007**, *25*, 71−75.

(6) Brown, D.; Superti-Furga, G. Rediscovering the Sweet Spot in Drug Discovery. *Drug Discovery Today* **2003**, *8*, 1067−1077.

(7) Cheng, A. C. Predicting Selectivity and Druggability in Drug Discovery. *Annu. Rep. Comput. Chem.* **2008**, *4*, 23−37.

(8) Fauman, E. B.; Rai, B. K.; Huang, E. S. Structure-Based Druggability Assessment−Identifying Suitable Targets for Small Molecule Therapeutics. *Curr. Opin. Chem. Biol.* **2011**, *15*, 463−468.

(9) Lipinski, C. a. Filtering in Drug Discovery. *Annu. Reports Comp Chem.* **2005**, 155−168.

(10) Halgren, T. a. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377−389.

(11) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287−2299.

(12) Schmidtke, P.; Barril, X. Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *J. Med. Chem.* **2010**, *53*, 5858−5867.

(13) Perola, E.; Herman, L.; Weiss, J. Development of a Rule-Based Method for the Assessment of Protein Druggability. *J. Chem. Inf. Model.* **2012**, *52*, 1027−1038.

(14) Krasowski, A.; Muthas, D.; Sarkar, A.; Schmitt, S.; Brenk, R. DrugPred: A Structure-Based Approach to Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *J. Chem. Inf. Model.* **2011**, *51*, 2829−2842.

(15) Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining Global and Local Measures for Structure-Based Druggability Predictions. *J. Chem. Inf. Model.* **2012**, *52*, 360−372.

(16) Pérot, S.; Sperandio, O.; Miteva, M. a; Camproux, A.-C.; Villoutreix, B. O. Druggable Pockets and Binding Site Centric Chemical Space: A Paradigm Shift in Drug Discovery. *Drug Discovery Today* **2010**, *15*, 656−667.

(17) Gao, M.; Skolnick, J. A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLoS Comput. Biol.* **2013**, *9*, e1003302.

(18) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins. *J. Mol. Graphics Modell.* **1997**, *15* (359−363), 389.

(19) Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the Topology of Active Sites: On the Prediction of Pockets and Subpockets. *J. Chem. Inf. Model.* **2010**, *50*, 2041−2052.

(20) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10*, 168.

(21) Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A Simple and Fuzzy Method to Align and Compare Druggable Ligand-Binding Sites. *Proteins* **2008**, *71*, 1755−1778.

(22) Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−135.

(23) Yeturu, K.; Chandra, N. PocketMatch: A New Algorithm to Compare Binding Sites in Protein Structures. *BMC Bioinformatics* **2008**, *9*, 543.

(24) Feldman, H. J.; Labute, P. Pocket Similarity: Are Alpha Carbons Enough? *J. Chem. Inf. Model.* **2010**, *50*, 1466−1475.

(25) Tang, G. W.; Altman, R. B. Knowledge-Based Fragment Binding Prediction. *PLoS Comput. Biol.* **2014**, *10*, e1003589.

(26) Dance, I. A Molecular Pathway for the Egress of Ammonia Produced by Nitrogenase. *Sci. Rep.* **2013**, *3*, 3237.

(27) Masini, T.; Kroezen, B. S.; Hirsch, A. K. H. Druggability of Non-Mevalonate Pathway Enzymes. *Drug Discovery Today* **2013**, *18*, 1256−1262.

(28) Sivakumar, S.; Niranjali Devaraj, S. Tertiary Structure Prediction and Identification of Druggable Pocket in the Cancer Biomarker - Osteopontin-C. *J. Diabetes Metab. Disord.* **2014**, *13*, 13.

(29) Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.

(30) Hubbard, S. J.; Thornton, J. *NACCESS*, version 2.1.1, 1992.

(31) Petitjean, M. *RADI*, version 4.0, 2014. http://petitjeanmichel.free.fr/itoweb.petitjean.freeware.html.

(32) Milletti, F.; Vulpetti, A. Predicting Polypharmacology by Binding Site Similarity: From Kinases to the Protein Universe. *J. Chem. Inf. Model.* **2010**, *50*, 1418−1431.

(33) Burgoyne, N. J.; Jackson, R. M. Predicting Protein Interaction Sites: Binding Hot-Spots in Protein-Protein and Protein-Ligand Interfaces. *Bioinformatics* **2006**, *22*, 1335−1342.

(34) Eyrisch, S.; Helms, V. Transient Pockets on Protein Surfaces Involved in Protein-Protein Interaction. *J. Med. Chem.* **2007**, *50*, 3457−3464.

(35) Kyte, J.; Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157*, 105−132.

(36) Fisher, R. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* **1936**, *7*, 179−188.

(37) Krotzky, T.; Rickmeyer, T. Extraction of Protein Binding Pockets in Close Neighborhood of Bound Ligands Makes Compar-

isons Simple due to Inherent Shape Similarity. *J. Chem. Inf. Model.* **2014**, *54*, 3229−3237.

(38) Gao, M.; Skolnick, J. APoc: Large-Scale Identification of Similar Protein Pockets. *Bioinformatics* **2013**, *29*, 597−604.

(39) Cuchillo, R.; Pinto-gil, K.; Michel, J. A Collective Variable for the Rapid Exploration of Protein Druggability. *J. Chem. Theory Comput.* **2015**, *11*, 1292−1307.

(40) Nisius, B.; Sha, F.; Gohlke, H. Structure-Based Computational Analysis of Protein Binding Sites for Function and Druggability Prediction. *J. Biotechnol.* **2011**, *159*, 123−134.

(41) Villar, H. O.; Kauvar, L. M. Amino Acid Preferences at Protein Binding Sites. *FEBS Lett.* **1994**, *349*, 125−130.

(42) Pérot, S.; Regad, L.; Reynès, C.; Spérandio, O.; Miteva, M. a.; Villoutreix, B. O.; Camproux, A.-C. Insights into an Original Pocket-Ligand Pair Classification: A Promising Tool for Ligand Profile Prediction. *PLoS One* **2013**, *8*, e63730.

(43) Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331−337.