RESEARCH ARTICLE

Optimal Ligand Descriptor for Pocket Recognition Based on the Beta-Shape

Jae-Kwan Kim[1‡], Chung-In Won[1‡], Jehyun Cha[2], Kichun Lee[3], Deok-Soo Kim[1,2]*

1 Voronoi Diagram Research Center, Hanyang University, Seoul, Korea, **2** School of Mechanical Engineering, Hanyang University, Seoul, Korea, **3** Department of Industrial Engineering, Hanyang University, Seoul, Korea

‡ These authors contributed equally to this work.
* dskim@hanyang.ac.kr

OPEN ACCESS

Citation: Kim J-K, Won C-I, Cha J, Lee K, Kim D-S (2015) Optimal Ligand Descriptor for Pocket Recognition Based on the Beta-Shape. PLoS ONE 10(4): e0122787. doi:10.1371/journal.pone.0122787

Academic Editor: Paul Taylor, University of Edinburgh, UNITED KINGDOM

Received: November 10, 2014

Accepted: February 17, 2015

Published: April 2, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: J-KK, C-IW, JC, and D-SK were supported by the National Research Foundation grant funded by MSIP (No. 2012R1A2A1A05026395), Republic of Korea. KL was supported by the grant (201400000002667) funded by Small and Medium Business Administration (SMBA), Republic of Korea. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Structure-based virtual screening is one of the most important and common computational methods for the identification of predicted hit at the beginning of drug discovery. Pocket recognition and definition is frequently a prerequisite of structure-based virtual screening, reducing the search space of the predicted protein-ligand complex. In this paper, we present an optimal ligand shape descriptor for a pocket recognition algorithm based on the beta-shape, which is a derivative structure of the Voronoi diagram of atoms. We investigate six candidates for a shape descriptor for a ligand using statistical analysis: the minimum enclosing sphere, three measures from the principal component analysis of atoms, the van der Waals volume, and the beta-shape volume. Among them, the van der Waals volume of a ligand is the optimal shape descriptor for pocket recognition and best tunes the pocket recognition algorithm based on the beta-shape for efficient virtual screening. The performance of the proposed algorithm is verified by a benchmark test.

Introduction

Drug discovery is a time consuming, costly process. One of the most critical processes in drug-discovery is identification of predicted hit where virtual screening as an *in silico* method screens a chemical library against a target protein [1–3]. For this purpose, the pharmacophore of a pocket can be used for virtual screening [4, 5]. Based on its effectiveness and the rapid accumulation of three-dimensional molecular structures, structure-based virtual screening is becoming more widespread. Over 100,000 experimentally determined biomolecular structures are cataloged in the Protein Data Bank (PDB) [6], and millions of rational biomolecular models are cataloged in the MODBASE [7], the SWISS-MODEL [8] and the PMDB [9]. Successful cases of structure-based virtual screening include Gleevec targeting a tyrosine kinase [10], Agenerase and Viracept for HIV protease [11]. Other successful cases are reviewed in [11–13].

A common approach in structure-based virtual screening is docking simulation which attempts to find the best binding of a ligand to a receptor by solving the energy minimization problem where the search space is exponential, making it hard to solve [14, 15]. In order to reduce computation, docking algorithms usually predict a potential binding site called a *pocket*, which is the concave region on the molecular boundary, to place an initial ligand for the energy minimization process [16–19].

There are three approaches in pocket recognition. The *grid-based approach* defines the lattice of the space occupied by a receptor, infers the relations among the grid points in the lattice to extract the exterior boundary of the molecule, and recognizes the depressed regions on the boundary [20–23]. A *sphere-coating approach* places a set of artificial spherical probes around the receptor and infers the relations among the probes for a pocket [24–26]. However, both approaches are rather heuristic and do not guarantee a quality solution in spite of heavy computational requirement. The *computational geometry approach* is based on the formal computational geometry theory of the proximity among atoms to recognize the receptor boundary and the shape of a pocket. The (weighted) alpha-shape based method [27, 28] and the beta-shape based method [29] belong to this category.

Most previous pocket recognition studies regarded the largest concave region on the receptor boundary as a pocket, ignoring the ligand characteristics. However, different ligands may bind to different sites on the boundary of an identical receptor. For example, c-Myc protein, which is overexpressed in the majority of human cancers, is known to have three independent binding sites corresponding to three different types of ligands: Ligands 10074-G5, 10074A4, and 10058-F4 [30] bind to 366–375, 375–385, and the 402–409 residues of c-Myc, respectively [31]. If the biggest pocket is only considered for virtual screening, drug candidates corresponding to the other two binding sites cannot be found. Hence, it is desirable to reflect the ligand characteristics during the pocket recognition process as its shape is the most important ligand characteristic. Reports for other cases are also available [32–34].

In this paper, we propose optimization of a ligand shape descriptor for pocket recognition based on the beta-shape so that the recognized pocket can be better used for virtual screening. We first present the formalization of our earlier pocket recognition algorithm [29] in the context of the beta-shape. We avoid the (weighted) alpha-shape due to the following reason. The alpha-shape was originally defined for points using the ordinary Voronoi diagram of points [35] and was used for reasoning the spatial properties of point clouds or molecular structures assuming that all atoms were of an identical size. However, poly-sized atomic model (i.e., different atom types had different radii) was more realistic for analyzing molecular structure. To reflect the size difference among different atom types, the weighted alpha-shape, which was based on the power diagram of the poly-sized atomic model, replaced the alpha-shape [36]. However, it turned out that the power diagram, and thus the weighted alpha-shape as well, was not based on the Euclidean distance but on the power distance which could be interpreted as the tangential distance from the boundary of spherical atoms. Due to this property, the topology structure of the weighted alpha-shape can be incorrect for reasoning the proximity between non-intersecting atoms and is not necessarily offset-invariant. The lack of offset-invariance causes the limitation of the weighted alpha-shape for many important applications of molecular structure.

Then, we present the optimal shape descriptor of a ligand for pocket recognition. This is based on an efficient algorithm to extract the molecular boundary using the beta-shape, a structure derived from the Voronoi diagram of the molecule [37]. Using the beta-shape and the optimized shape descriptor, effective pockets can be efficiently recognized and used for the docking algorithm called the BetaDock [38, 39]. The molecular graphics in this paper were created using BetaMol, a molecular modeling, visualization, and analysis program freely available from *http://voronoi.hanyang.ac.kr/software.htm* [40].

Approach

Pocket recognition using the beta-shape

For the proximity among the atoms on the molecular boundary, the concept of the beta-shape has been proposed [37]. Fig. 1(a) shows a two-dimensional molecule. Fig. 1(b) shows the Connolly surface (green curve) corresponding to the red circular probe where the radius is β. Suppose that the Connolly surface is straightened by substituting the straight edges for the circular arcs and the planar triangles for the spherical triangles where their vertices are the centers of the related atoms. The straightened object bounded by the planar facets is the *beta-shape* of the molecule. Fig. 1(c) shows the beta-shape of a molecule corresponding to the red circular probe in Fig. 1(b). The beta-shape concisely provides the precise proximity among the atoms on the molecular boundary with respect to the probe. Fig. 1(d), (e), and (f) show the van der Waals model of a protein (PDB id 1oq5), its Connolly surface for water molecule with 1.4Å radius, and the corresponding beta-shape. We note here that the beta-shape is efficiently computed from the *quasi triangulation* which is the dual structure of the *Voronoi diagram of atoms*. The details are reported in [37, 41–43] and readers are recommended to download the `BetaConcept` program from VDRC (http://voronoi.hanyang.ac.kr) to explore the properties of the beta-shape.

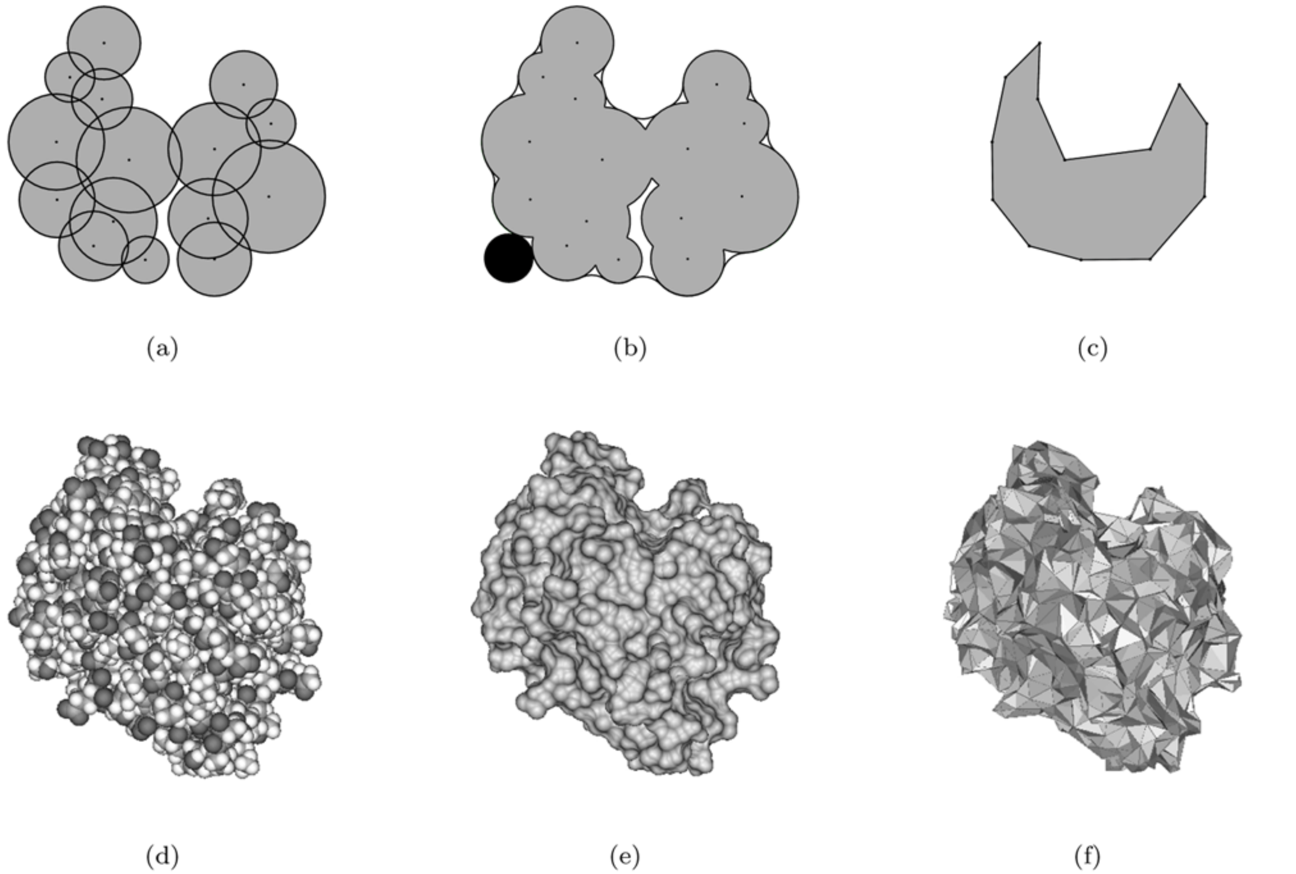


Fig 1. A schematic diagram of a molecule and its beta-shape. Figure drawn by using the `BetaConcept`[44] and `BetaMol` program freely available from VDRC. (a) A two-dimensional molecule, (b) A two-dimensional molecule and its Connolly surface corresponding to the red circular probe, and (c) the beta-shape corresponding to the probe, (d) the van der Waals model of a protein (PDB id 1oq5), (e) the Connolly surface for water molecule (with 1.4Å radius), and (f) the corresponding beta-shape.

doi:10.1371/journal.pone.0122787.g001

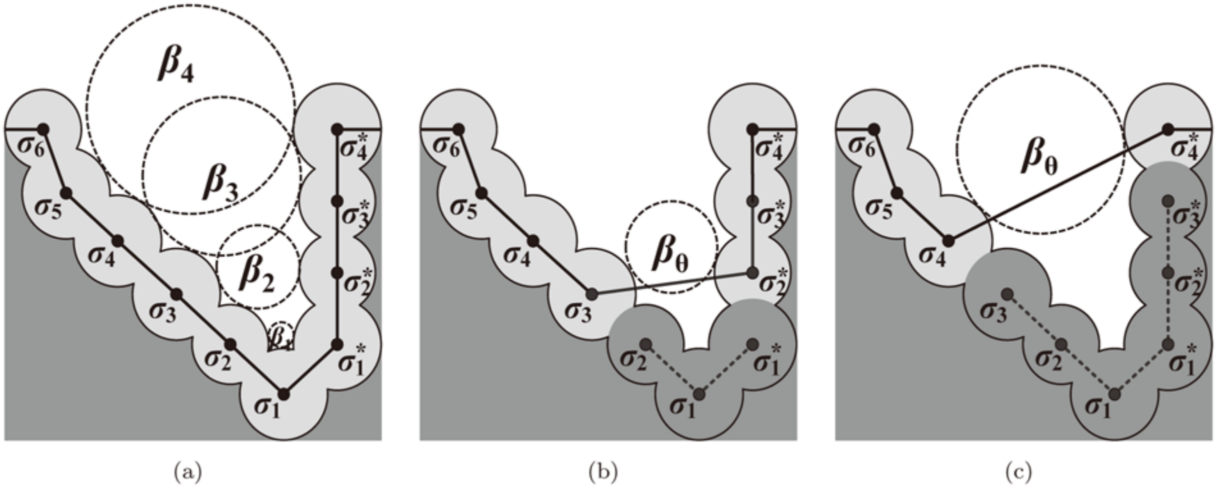


Fig 2. The idea of pocket recognition using the beta-shape. (a) Empty tangent balls defining the exposure intervals of each atom on the boundary. (b) The pocket $\{\sigma_1, \sigma_2, \sigma_1^*\}$ where $\beta_2 < \beta_\theta \leq \beta_3$. (c) The pocket $\{\sigma_1, \sigma_2, \sigma_3, \sigma_1^*, \sigma_2^*, \sigma_3^*\}$ where $\beta_3 < \beta_\theta \leq \beta_4$.

doi:10.1371/journal.pone.0122787.g002

Fig. 2 shows a two-dimensional schematic diagram showing the idea of pocket recognition using the beta-shape. Suppose that the figure depicts a subset of the beta-shape corresponding to the probe of water. Consider that the small circle σ or σ^* is an atom on the molecular boundary and the shaded region is the molecular interior. The atoms on the slanted wall in the left are numbered σ_1 through σ_6, and those on the vertical wall are numbered σ_1^* through σ_4^*. There are four dotted circles β_1, β_2, β_3 and β_4 in Fig. 2(a) where each is in contact with the boundary of the three atoms. For convenience, suppose that β_1, β_2, β_3 and β_4 also denote the radii of the corresponding circles where $0 \leq \beta_1 < \beta_2 < \beta_3 < \beta_4$. Let π be a spherical open probe with the radius β_π.

In Fig. 2(a), the smallest circle β_1 is in contact with σ_1, σ_2 and σ_1^*. Consider a probe π smaller than β_1 (i.e., $\beta_\pi \leq \beta_1$). Then, π can touch the boundary of all atoms implying that all atoms are exposed to π. However, if β_π is greater than β_1, π can no longer touch σ_1 and σ_1 is not exposed to π. Hence, σ_1 is exposed when $0 \leq \beta_\pi \leq \beta_1$, and the interval $[0,\beta_1]$ is called the *exposure interval* for σ_1. Consider β_2, which is in contact with the three atoms σ_2, σ_3 and σ_2^*. Then, σ_2 is similarly exposed when $0 \leq \beta \leq \beta_2$. The exposure interval of σ_3 is $[0,\beta_3]$. A similar observation holds for the other atoms. Therefore, each boundary atom is associated with an exposure interval.

Fig. 2(b) and (c) illustrate how to use the exposure interval in pocket recognition. Let β_θ be the threshold value to recognize a pocket. Suppose that $\beta_2 < \beta_\theta \leq \beta_3$. This implies that the atoms σ_1 and σ_2 (σ_1^* and σ_2^* as well) are not exposed to π when $\beta_\pi = \beta_\theta$. Then, the boundary of the beta-shape corresponding to $\pi = \beta_\theta$ is shown as the solid polyline in Fig. 2(b). Hence, the boundary no longer includes the three atoms σ_1, σ_2 and σ_1^* and the depressed, buried region consisting of σ_1, σ_2 and σ_1^* can be regarded as a pocket. Therefore, the atoms that constitute a pocket can be easily identified by checking the exposure interval of each atom. Fig. 2(c) shows a larger pocket. A lager β_θ tends to define a larger pocket and a smaller β_θ tends to define a smaller pocket. As different β_θ values define different pockets, it is important to find the optimal value of β_θ. The threshold β_θ is essential for the shape and size of the pockets. For details, see [45].

L-descriptor: descriptor of the ligand shape

Drug-like ligands ordinarily consist of 20 to 70 atoms [46] where each can have various conformations [47]. The conformation of a ligand instance affects the binding between the ligand and its receptor, and the primary factor of the binding is the ligand shape. Therefore, an appropriate consideration of the ligand shape is necessary. There are algorithms for computing the possible ligand conformations so that each conformation can be treated as a ligand instance in virtual screening [48]. The pocket recognition algorithm above uses the threshold β_θ whose optimal value for a given pair of ligands and receptors should be inferred to form the measure of the ligand shape. We call this measure the *L-descriptor*.

We examine six types of L-descriptor for a ligand: β_θ_mes, β_θ_PC1, β_θ_PC2, β_θ_PC3, β_θ_vdW and β_θ_beta. The β_θ_mes is the radius of the minimum enclosing sphere (mes), which is the smallest sphere that contains all the ligand atoms (Fig. 3(a)). The values of β_θ_PC1, β_θ_PC2 and β_θ_PC3 are obtained from the bounding box of a ligand that is computed by the principal component analysis (PCA) [49]. Let PC1 be the first principal component denoting the greatest variance of the data set. Similarly, let PC2 and PC3 be the second and the third principal components denoting the second and third greatest variance, respectively. Then, the length of each edge of the PCA-induced bounding-box is used as β_θ_PC1, β_θ_PC2, or β_θ_PC3. See Fig. 3(b) for examples of β_θ_PC1 and β_θ_PC2 in the plane. Two volume measures are also investigated. Let $Vol(vdW)$ be the volume of the vdW-model of a ligand. Consider a sphere whose volume is also $Vol(vdW)$. Then, the radius of the sphere is β_θ_vdW (Fig. 3(c)). For computation of $Vol(vdW)$, refer to [50]. Let $Vol(\beta)$ be the volume of the beta-shape corresponding to the spherical probe of a water molecule. Then, the radius of the sphere with the volume $Vol(\beta)$ is β_θ_beta (Fig. 3(d)). Fig. 4 shows the three-dimensional counterpart of the L-descriptors for three ligands found from protein complexes in PDB.

Methods

Definition of an optimal pocket

Consider a complex consisting of a receptor molecule M^R (the gray object in Fig. 5(a)) and its bound ligand molecule M^L (the green object the same figure) where both are defined by atom sets. Let ∂M^R be the boundary of the van der Waals model of M^R and $d(q,M^R)$ the minimum Euclidean distance between two points q and $x \in \partial M^R$. ∂M^L and $dist(q,M^L)$ are similarly defined. Let $IIF^\infty = \{q_1,q_2,q_3,\ldots\}$ be the surface (the blue curve in Fig. 5(b)) which is the locus of q_i where $dist(q_i,M^R) = dist(q_i,M^L)$. In other words, IIF^∞ is the mid-surface between M^R and M^L emanating to infinity. Let $IIF \subset IIF^\infty$ be the trimmed surface (the red curve in Fig. 5(d)) of IIF^∞ using the probe of a water molecule as a cutter (the red ball in Fig. 5(c)) [51]. Then, IIF is

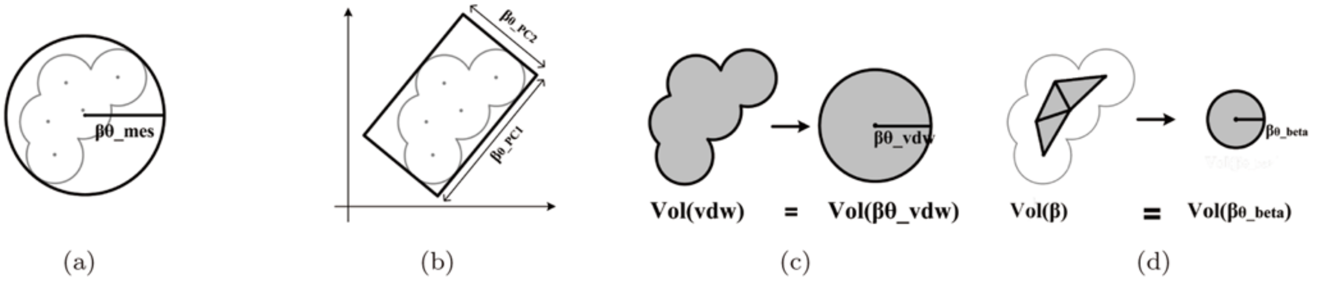


Fig 3. L-descriptor types in the plane. (a) The minimum enclosing sphere and β_θ_mes, (b) the bounding box by PCA, β_θ_PC1, and β_θ_PC2, (c) the van der Waals model of the ligand and β_θ_vdW, and (d) the beta-shape of the ligand and β_θ_beta.

doi:10.1371/journal.pone.0122787.g003

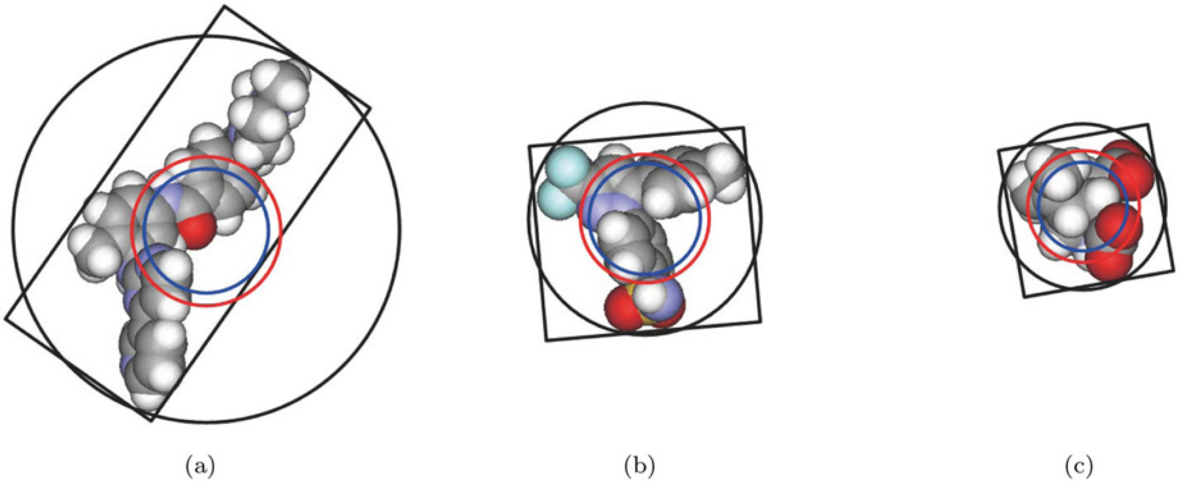


Fig 4. Some of the proposed L-descriptor types. The black circle denotes the minimum enclosing sphere; the red circle denotes the sphere whose volume is identical to the volume of the van der Waals model of the ligand; the blue circle denotes the sphere whose volume is identical to the volume of the beta-shape; the black rectangle denotes the bounding box of the PCA analysis. The PDB accession codes that contains the complex with the shown ligands are as follows: (a)1t46, (b)1oq5, and (c) 1tt1.

doi:10.1371/journal.pone.0122787.g004

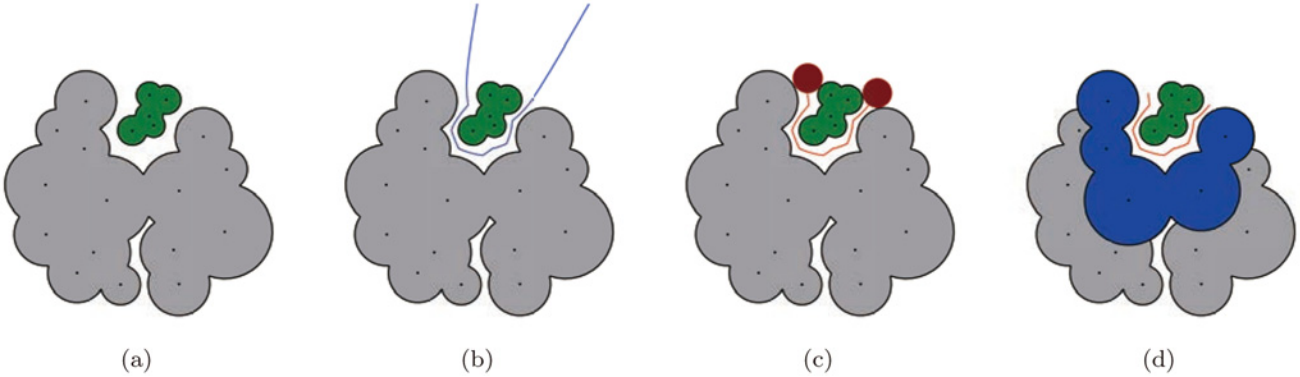


Fig 5. The interaction interface (IIF) of a two-dimensional molecule complex and the optimal pocket defined by *IIF*. The gray and green objects are a receptor molecule M^R and a ligand molecule M^L, respectively. (a) A two-dimensional molecule complex, (b) IIF^∞ shown as the blue curve, (c) *IIF* shown as the red curve trimmed by the red circle, and (d) the optimal pocket consisting of the five blue atoms and *IIF*.

doi:10.1371/journal.pone.0122787.g005

called the *interaction interface* between M^R and M^L. Let $\Pi \subset M^R$ be the set of receptor atoms (the blue five atoms in Fig. 5(d)) which defines *IIF*. Then, we call Π the *optimal pocket* in this paper. Π is called optimal in the sense that a complex consisting of a receptor and a ligand is crystalized, and its structure is solved in its entirety. For the details, see [52].

Evaluation of a recognized pocket

In a binary decision problem, a decision made by a classifier can be represented in a confusion matrix [53]. Recall that Π denotes the optimal pocket. Let $\Pi^c = B–\Pi$ where B is the set of atoms on the receptor boundary. In other words, Π^c is the boundary atoms except those in the optimal pocket. Let $\hat{\Pi}$ be the recognized pocket by the proposed algorithm. Then, $\hat{\Pi}^c = B - \hat{\Pi}$ is the boundary atoms except those in the recognized pocket.

Table 1. Confusion matrix for pocket evaluation.

	In recognized pocket ($\hat{\Pi}$)	Not in recognized pocket ($\hat{\Pi}^c$)
In optimal pocket (Π)	True Positive (T^+)	False Negative (F^-)
Not in optimal pocket (Π^c)	False Positive (F^+)	True Negative (T^-)

doi:10.1371/journal.pone.0122787.t001

We can now define the confusion matrix for pocket recognition as in Table 1. The atoms in $\Pi \cap \hat{\Pi}$ are called true positive (T^+); The atoms in $\Pi^c \cap \hat{\Pi}^c$ are called true negative (T^-); The atoms in $\Pi^c \cap \hat{\Pi}$ are called false positive (F^+); The atoms in $\Pi \cap \hat{\Pi}^c$ are called false negative (F^-). Hence, *true positive*(T^+) refers to the positive atoms correctly recognized as positive; *False positive*(F^+) refers to the negative atoms incorrectly recognized as positive; *True negative*(T^-) refers to the negative atoms correctly recognized as negative; *False negative*(F^-) refers to the positive atoms incorrectly recognized as negative.

Given the confusion matrix, various metrics can be defined for the evaluation of the quality of a recognized pocket. The true positive rate, *TPR*, is the proportion of the correct atoms in the recognized pocket (T^+) against the atoms in the optimal pocket (both T^+ and F^-). *TPR* is also referred to as the *recall rate R*, or the *sensitivity S*. The false positive rate, *FPR*, is the proportion of the incorrect atoms of the recognized pocket (F^+) against the atoms which do not belong to the optimal pocket (both T^- and F^+). The *specificity*, *SP*, is the proportion of the correct atoms not in the recognized pocket (T^-) against the atoms not in the optimal pocket (both T^- and F^+). The *precision*, *P*, is the proportion of the correct atoms in the recognized pocket (T^+) against the atoms in the recognized pocket (both T^+ and F^+). The *accuracy*, *AC*, is the proportion of correct atoms in the recognized pocket (both T^+ and T^-) against all atoms in the boundary *B*. In this paper, these are called the *primary metrics* from the confusion matrix and summarized in Table 2.

There are trade-offs among the primary metrics. A good recognized pocket should have high *TPR* and low *FPR* values. An overestimated, large pocket tends to have higher values for both *TPR* and *FPR* because there can be both many correctly identified atoms and many incorrectly identified atoms at the same time. An underestimated, small pocket tends to have a low *FPR* value (because the pocket size is small and thus there is a lower chance to have incorrect atoms) and a low *TPR* value (because the chance to have correct atoms is also lower). This trade-off is conveniently represented in the Receiver Operator Characteristic (ROC) graph which is useful for visualizing the performance of classifiers [54]. In the ROC-graph, the

Table 2. Primary metrics of the confusion matrix.

Primary metric	Equation
True Positive Rate (TPR)	$TPR = \frac{T^+}{T^+ + F^-} = \frac{n(\Pi \cap \hat{\Pi})}{n(\Pi)}$
False Positive Rate (FPR)	$FPR = \frac{F^+}{T^- + F^+} = \frac{n(\Pi^c \cap \hat{\Pi})}{n(\Pi^c)}$
Precision (P)	$P = \frac{T^+}{T^+ + F^+} = \frac{n(\Pi \cap \hat{\Pi})}{n(\hat{\Pi})}$
Specificity (SP)	$SP = \frac{T^-}{T^- + F^+} = \frac{n(\Pi^c \cap \hat{\Pi}^c)}{n(\Pi^c)} = 1 - FPR$
Accuracy (AC)	$AC = \frac{T^+ + T^-}{T^+ + F^+ + T^- + F^-} = \frac{n((\Pi \cap \hat{\Pi}) \cup (\Pi^c \cap \hat{\Pi}^c))}{n(B)}$
Sensitivity (S) = Recall (R)	$S = R = TPR$

doi:10.1371/journal.pone.0122787.t002

horizontal and vertical axes denote *FPR* and *TPR*, respectively. Hence, the coordinate ($FPR = 0$, $TPR = 1$) denotes the perfect pocket recognition. In the ROC-graph, the more upper-left a coordinate is, the better the performance. Given the operating points in the ROC-graph, a smooth ROC-curve can be computed with the assumption of binormal distribution. Then, the *area under the ROC-curve*, *AUC*, is a measure combining both *TPR* and *FPR* that is interpreted as the average sensitivity over all of the specificity range. In other words, *AUC* is the probability that a pocket recognizer will select a randomly chosen pocket atom higher than a randomly chosen atom not in a pocket.

It is usual that the number of atoms that do not belong to the optimal pocket significantly exceeds the number of atoms belonging to the optimal pocket. In other words, $n(\Pi^c) >> n(\Pi)$. Since $\Pi^c \cap \hat{\Pi} \subseteq \hat{\Pi}$ and $\hat{\Pi} \approx \Pi$, the numerator of FPR is usually significantly smaller than its denominator. Thus, even a large change in F^+ does not result in a significant change in the *FPR*. Hence, in pocket recognition, a ROC-graph tends be optimistic in that most recognized pockets and algorithms are likely to have low *FPR* regardless of the performance in reality.

The PR-graph denotes the coordinate system where the horizontal and vertical axes are the recall *R* and the precision *P*, respectively. Note that the precision *P* captures the size of the correctly recognized pocket because $\Pi \cap \hat{\Pi} \subseteq \Pi$ and $\Pi \approx \hat{\Pi}$. In the PR-graph, there is a trade-off between *R* and *P*. If all the atoms of an optimal pocket are perfectly predicted, $R = 1$, and if no atom of an optimal pocket is predicted at all, $R = 0$. If all the atoms of a recognized pocket are correct (i.e., there is no noise atoms in a recognized pocket), $P = 1$, and if all the atoms of a recognized pocket are noise atoms, $P = 0$. Hence, perfect pocket recognition occurs at the coordinates ($R = 1$, $P = 1$). Therefore, the more upper-right a coordinate is, the better the performance.

An overestimated, large pocket tends to have a high *R* (due to having many correct atoms) but a small *P* (because there are many noise atoms as well). On the other hand, an underestimated, small pocket tends to have a high *P* (because the size is small and it has lower chance to have noise atoms) but has a low *R* (because the chance to have correct atoms is lower).

Normalized Mutual Information [55], *NMI*, is a measure of information transmission which is based on Shannon's Entropy. Entropy measures are widely used in comparing true data with predicted data. Among those possible measures, entropy measures focus on the amount of the cross-section together with the match of total amount. Given a confusion matrix, the following four entropy values can be defined: the row entropy $H(x)$, the column entropy $H(y)$, and two conditional entropies $H(x|y)$ and $H(y|x)$

$$H(x) = -\sum_i p_i \log_2 p_i, \tag{1}$$

$$H(y) = -\sum_j p_j \log_2 p_j, \tag{2}$$

$$H(x|y) = \sum_j p_j \left[-\sum_i \frac{p_{ij}}{p_i} \log_2 \frac{p_{ij}}{p_j} \right], \tag{3}$$

$$H(y|x) = \sum_i p_i \left[-\sum_j \frac{p_{ij}}{p_j} \log_2 \frac{p_{ij}}{p_i} \right] \tag{4}$$

where p_i and p_j represent the empirical probabilities of the predicted and true examples, respectively, and p_{ij} is their joint probability. Then, NMI is defined as

$$NMI = \frac{H(x) - H(x|y)}{H(x)}. \tag{5}$$

The NMI contains more details of the confusion matrix which is not accounted for by other metrics [56]. The likelihood ratio test, LR, is a related metric that statistically compares the maximum likelihood of an unrestricted model with a restricted model [57] and is defined as

$$LR = 2 \sum_{i,j} Observed \log \left[\frac{Observed}{Expected} \right] \tag{6}$$

implying

$$\begin{aligned} LR \quad = \quad & 2\{N \log N + T^+ \log T^+ + F^- \log F^- + T^- \log T^- + F^+ \log F^+ \\ & - (T^+ + F^+) \log (T^+ + F^+) - (T^+ + F^-) \log (T^+ + F^-) \\ & - (T^- + F^+) \log (T^- + F^+) - (T^- + F^-) \log (T^- + F^-)\}. \end{aligned} \tag{7}$$

Table 3. Evaluation metrics.

	Secondary metric	Equation
ROC-based metrics	Balanced accuracy (BA)	$BA = \frac{S+SP}{2}$
	Geometric mean 2 (G2)	$G2 = \sqrt{S \times SP}$
	Euclidean distance (ED)	$ED = \sqrt{(S-1)^2 + (SP-1)^2}$
	Youden index (YI)	$YI = S+SP-1$
Precision-based metrics	F-measure (*f*)	$f = \frac{2 \times S \times P}{S+P}$
	Geometric mean 1 (G1)	$G1 = \sqrt{S \times P}$
	Predictive summary index (PSI)	$PSI = NPV+P-1$
	Negative Predictive Value (NPV)	$NPV = \frac{T^-}{T^- + F^-} = \frac{n(Op^c \cap Rp^c)}{n(Rp^c)}$
Ordinal association metrics	Gamma (γ)	$\gamma = \frac{T^+ \cdot T^- - F^+ \cdot F^-}{T^+ \cdot T^- + F^+ \cdot F^-}$
	Tau-b (τ_b)	$\tau_b = \frac{T^+ \cdot T^- - F^+ \cdot F^-}{\sqrt{(T^+ + F^-)(T^- + F^+)(T^+ + F^+)(T^- + F^-)}}$
	Tau-c (τ_c)	$\tau_c = \frac{4(T^+ \cdot T^- - F^+ \cdot F^-)}{(T^+ + T^- + F^+ + F^-)^2}$

doi:10.1371/journal.pone.0122787.t003

Both the *LR* and *NMI* are based on information entropy, which is loosely similar to the variance of the entries in the confusion table Table 2. Note also that the metric derived from the information entropy is independent of the ligand size.

In addition, we tested eleven more secondary metrics for the proposed six L-descriptors in Table 3: four based on ROC, four based on the precision, and three based on the ordinal association. The four metrics related to ROC graph are as follows: The balanced accuracy (*BA*) is defined as the numerical mean of S and *SP*[58]. The geometric mean 2 (*G*2) is the geometric mean of *S* and *SP*[59]. The Euclidean distance from an ideal classification (*ED*) is the combination of *S* and *SP* that measures the distance from an ideal classification in ROC space, where *S* and *SP* both equal one [56]. Youden index (*YI*) is the sum of the *S* and *SP* minus one and is a measure of goodness for diagnostic tests [60].

The four metrics related to PR graph are as follows: The F-measure (*f*) is a harmonic mean of *P* and *S* and was first used by Lewis and Gale for assessing text classification effectiveness and [61]. The geometric mean 1 (*G*1) is the geometric mean of *P* and *S*[59]. The predictive summary index (*PSI*) is the sum of *P* and *NPV* minus one and was developed as a measure of goodness for diagnostic tests [62]. The negative predictive value (*NPV*) is the proportion of the correct atoms out of the computed pockets (T^-) against the atoms out of the computed pocket (both T^- and F^-).

The ordinal association metrics have been used for the analysis of cross classifications with ordinal categories. The gamma (γ) is the estimated difference between the probability of concordance and the probability of discordance and has a range $1 \leq \gamma \leq 1$ [63]. The Kendall's τ_b makes an adjustment for ties when it measures the proportion of concordant and discordant pairs. The Kendall's τ_c is a variant of τ_b, which makes an adjustment for table size in addition to a correction for ties [64]. Both τ_b and τ_c has range $1 \leq \tau_b, \tau_c \leq 1$.

From the results of the ROC-graph and PR-graph, it is important to note the following: i) The *AUC* of ROC-curve can mislead because the curve cannot reflect the low sensitivity of smaller L-descriptor, and ii) the *AUC* of PR-curve can also mislead because the curve cannot reflect the low precision of larger L-descriptor. This phenomenon resides in the various secondary metrics based on the ROC-graph and PR-graph.

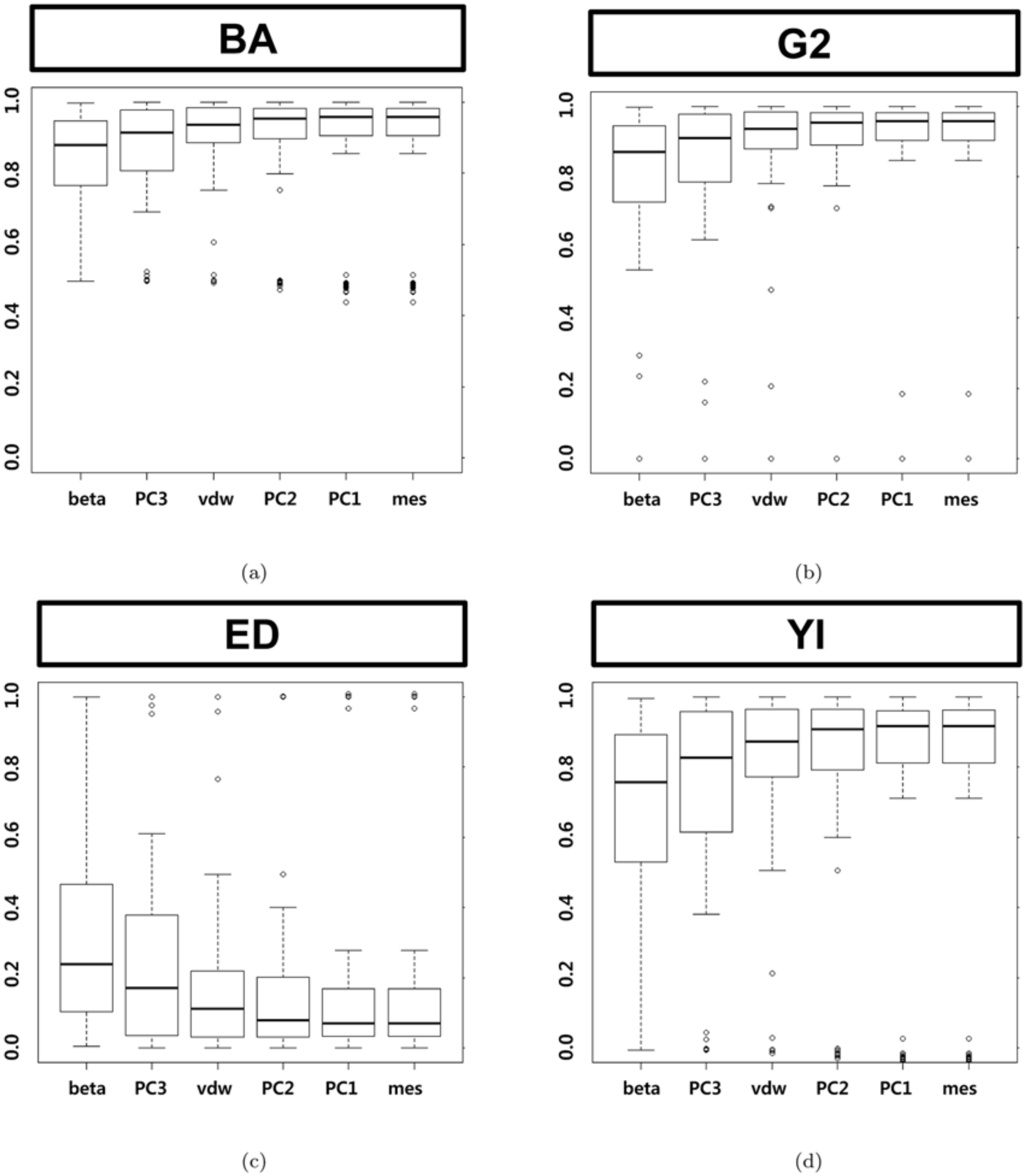
Fig. 6 shows the results of the ROC-based metrics which is based on sensitivity and specificity. Fig. 7 shows the results of metrics based on precision. These PR-based metrics mislead because the metrics cannot reflect the low precision of larger L-descriptor. Negative predictive value cannot discriminate among the L-descriptor types at all, because an optimal pocket has larger negative cases than positive cases. In all metrics, it turns out that the van der Waals volume consistently belongs to the group of L-descriptors showing better performance.

Results

Experimental materials and methods

The experiment was done using the Astex Diverse Set (ADS) consisting of 85 high resolution protein-ligand complexes containing drug-like compounds [65]. The optimal pocket Π of each receptor was computed from the bound complex, and the corresponding recognized pocket $\hat{\Pi}$ was computed from each receptor after the bound ligand was removed.

Consider an effective, optimal pocket related to a given ligand, and suppose that there is more than one depressed region on the receptor boundary that can be considered as a *pocket candidate*. Obviously, the larger the number of pockets used in the docking simulation, the better the solution quality, and the more time a computation takes. In this experiment, we assumed that the optimal pocket corresponds to one of the five biggest pocket candidates in



(a)

(b)

(c)

(d)

Fig 6. Box plots by ROC-based metrics of the six shape descriptors. (a) Balanced accuracy, (b) geometric mean 2, (c) Euclidean distance and (d) Youden index.

doi:10.1371/journal.pone.0122787.g006

terms of the number of atoms belonging to each pocket candidate. In fact, in most of the cases in our experiment, the optimal pocket belonged to one of the two biggest pocket candidates.

A ligand may have rotational bonds that can generate various conformations. In this experiment, we used two conformations for each ligand to check the effect of a ligand's conformation change: i) the native conformation found in the crystal structure and ii) the minimum energy conformation that was calculated by the MM2 method using ChemOffice software [66]. Fig. 8 shows two such examples.

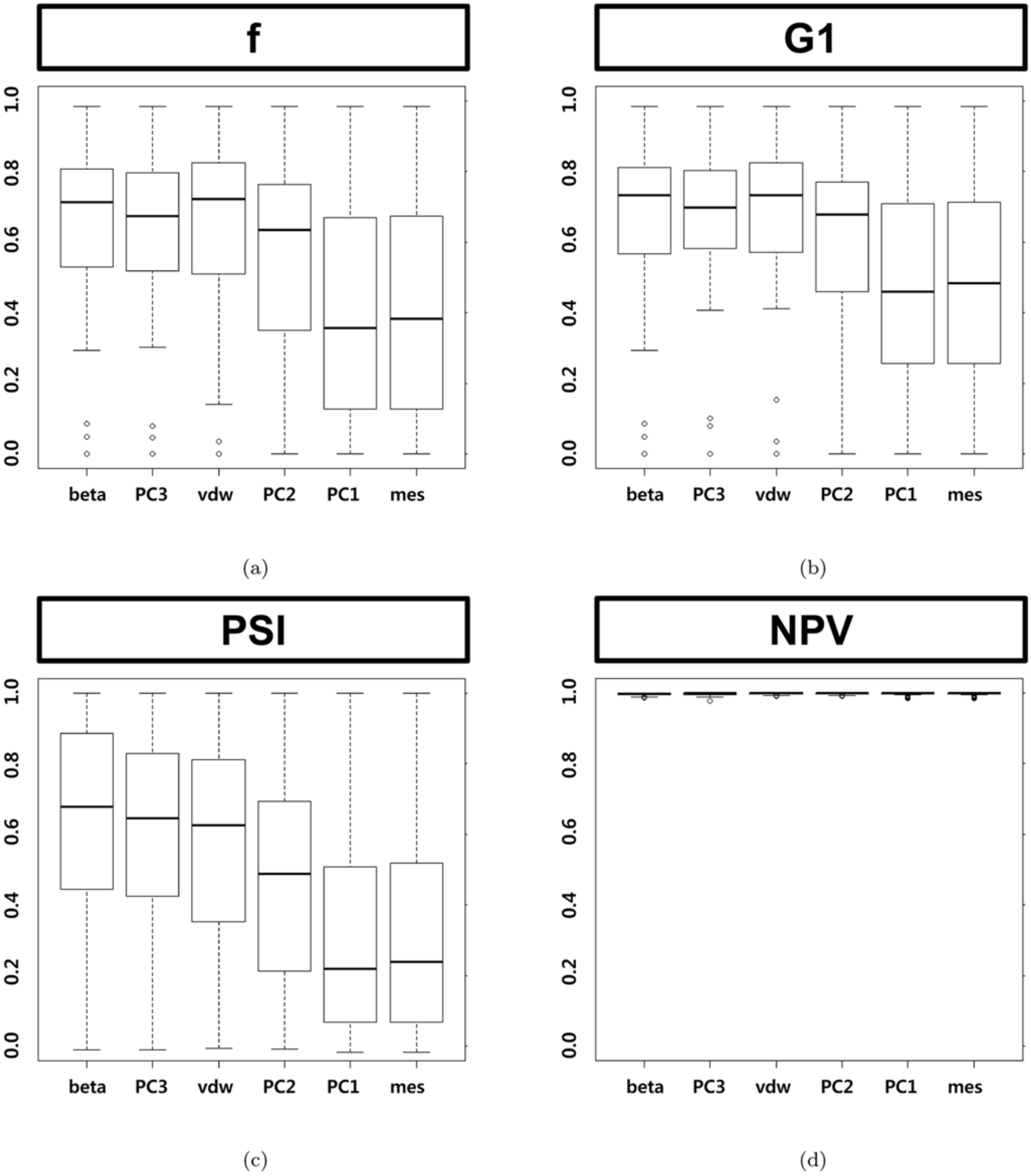


Fig 7. Box plots by Precision-based metrics of the six shape descriptors. (a) F-measure, (b) geometric mean 1 and (c) predictive summary index (d) negative predictive value.

doi:10.1371/journal.pone.0122787.g007

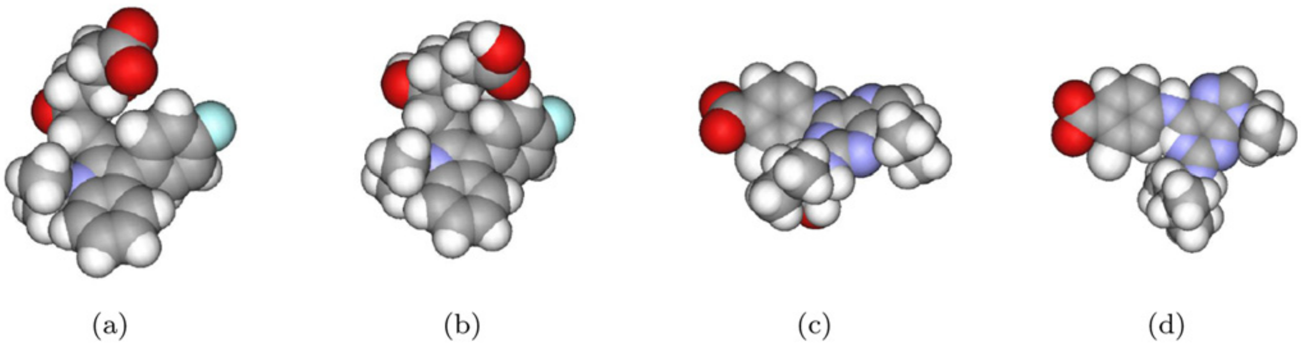


Fig 8. Two different conformations of two ligands: the native state and the minimum energy state. The minimized energy conformation is calculated by MM2 in ChemOffice software. (a) and (b) the native and the minimum energy conformations of 1hwi, respectively; (c) and (d) those of 1v0p.

doi:10.1371/journal.pone.0122787.g008

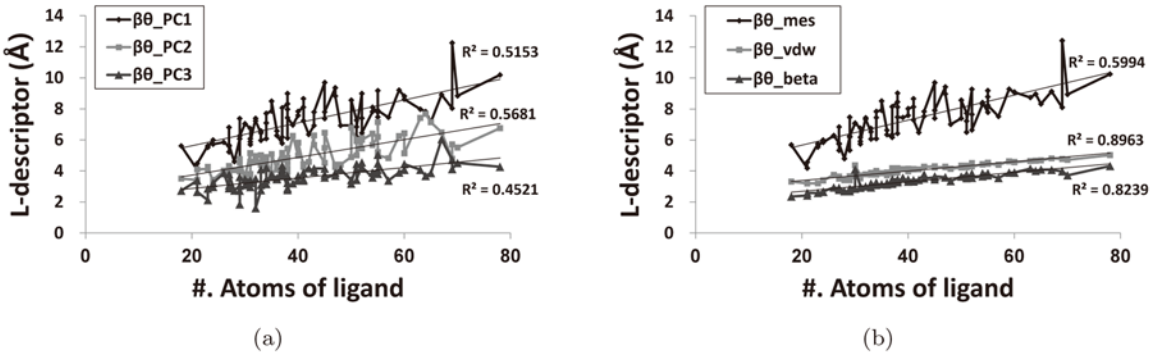


Fig 9. L-descriptor curves with respect to the ligand size. R^2 (the coefficient of determination) is a statistical measure of how close the data are to the fitted regression line. The p-values of the six linear regressions are all less than 10^{-11}.

doi:10.1371/journal.pone.0122787.g009

L-descriptors and ligand size

Fig. 9 shows the curves for the L-descriptors vs. the ligands ordered in their sizes. The six L-descriptors are divided into two graphs: Fig. 9(a) for the PC1, PC2, and PC3; Fig. 9(b) for the minimum enclosing sphere, the van der Waals volume, and the beta-shape volume. The L-descriptors tend to increase with respect to the ligand size, and their average values are in the following order (Within the parentheses are the averages):

$$\begin{aligned} &\beta_\theta\_beta(3.35) < \beta_\theta\_PC3(3.60) < \beta_\theta\_vdW(4.04) \\ &\quad < \beta_\theta\_PC2(4.96) < \beta_\theta\_PC1(7.21) < \beta_\theta\_mes(7.41). \end{aligned} \tag{8}$$

When $\beta_X < \beta_Y$ in Equation (8), we say that β_X is *smaller* than β_Y and β_Y is *bigger* than β_X.

Pocket evaluation

Fig. 10 compares the six L-descriptor types with four primary metrics; the sensitivity *S*, the precision *P*, the specificity *SP*, and the accuracy *AC*. The horizontal axis denotes the L-descriptors in the order given in Equation (8). The vertical axis denotes the metric values. Fig. 10(a) shows that a bigger L-descriptor tends to produce a higher sensitivity value than a smaller one. This implies that a bigger L-descriptor tends to produce a larger recognized pocket which has a

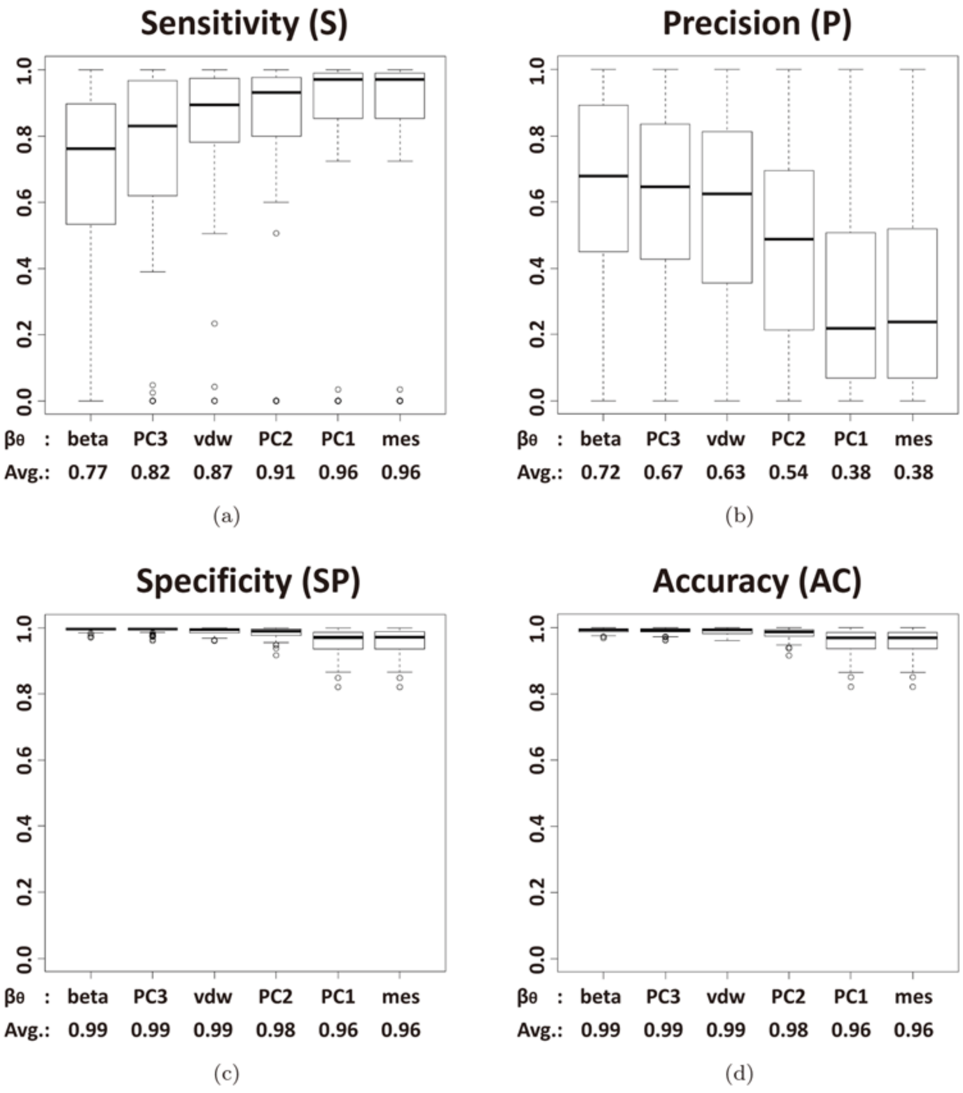


Fig 10. Box plots by primary metrics of the six types of L-descriptor. (a) Sensitivity, (b) precision, (c) specificity, and (d) accuracy.

doi:10.1371/journal.pone.0122787.g010

higher chance to have more correct atoms. On the other hand, Fig. 10(b) shows that a smaller L-descriptor tends to have a higher value of precision than a bigger one. This implies that a larger pocket has a higher chance to have incorrect atoms in a recognized pocket. This observation thus shows the trade-offs among the sensitivity and the precision. Fig. 10(c) and (d) shows that the specificity and the accuracy cannot properly discriminate the L-descriptor types.

Fig. 11 and Fig. 12 show the ROC-graphs and the PR-graphs of the six L-descriptor types, respectively, in the order as before. In the ROC-graphs in Fig. 11, the *FPR* tends to be small because there are many boundary atoms which do not belong to the optimal pocket. Note that the window of the horizontal-axis is given between 0 and 0.2. From these graphs, we observe that Fig. 11(c) and (d) shows the best distribution of the *FPR* and *TPR* values. Fig. 11(a) and (b) shows rather widely distributed *TPR* values and Fig. 11(e) and (f) shows rather widely

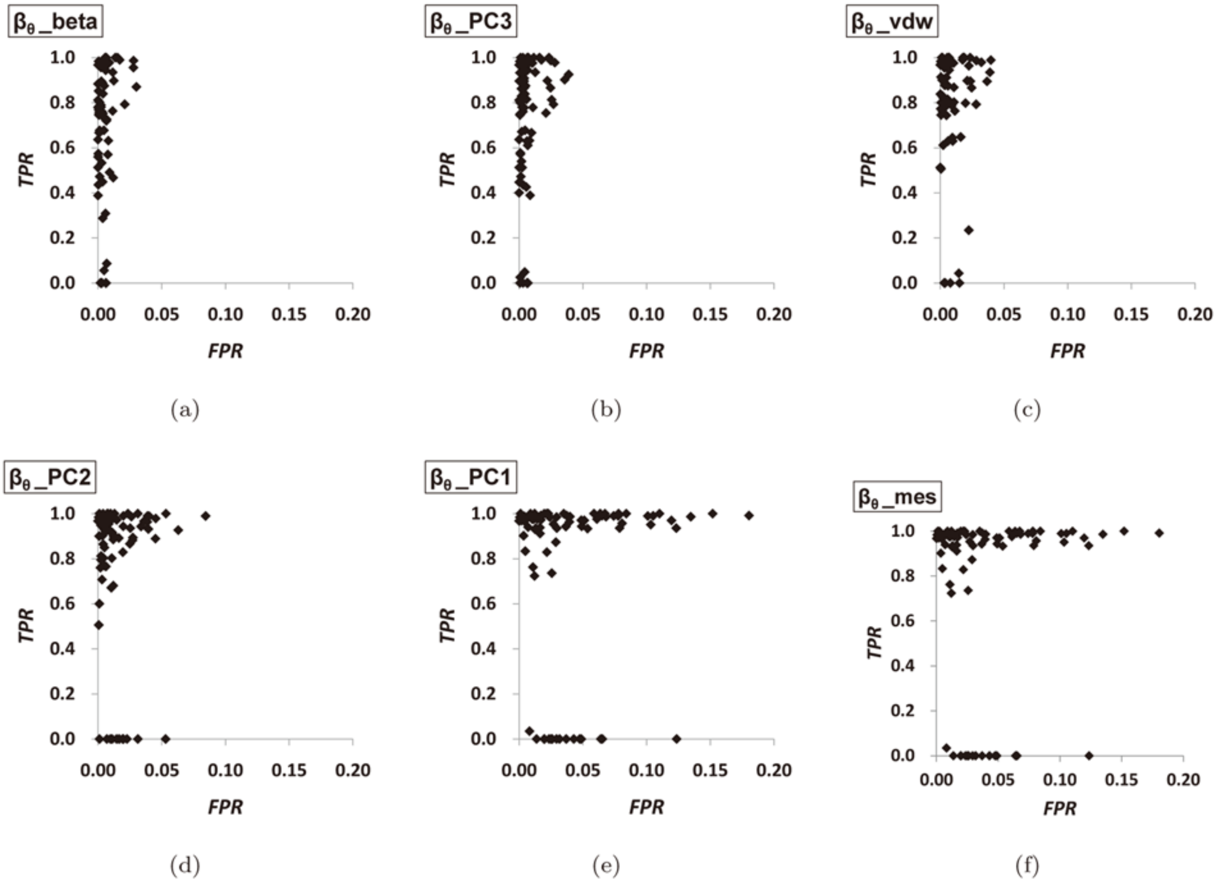


Fig 11. The ROC-graph of the L-descriptors. (a) the beta-shape volume, (b) the PC3, (c) the van der Waals volume, (d) the PC2, (e) the PC1, and (f) the minimum enclosing sphere.

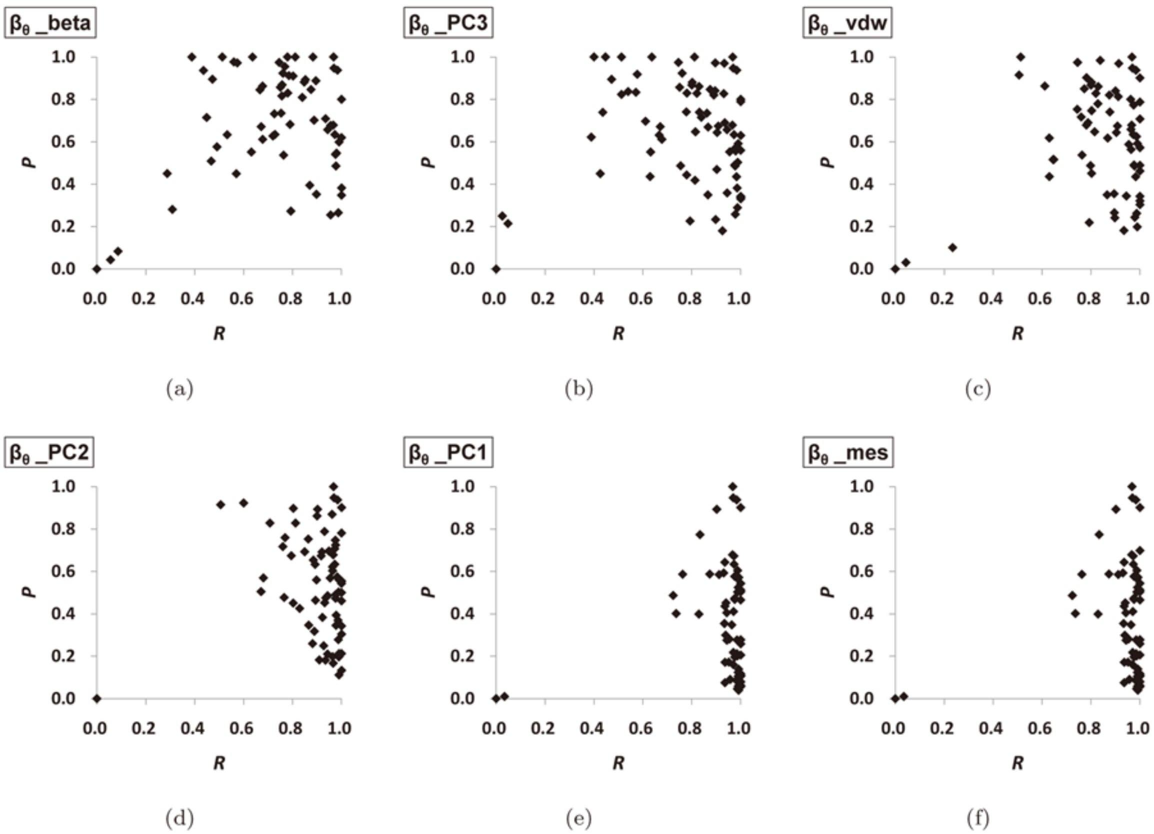
doi:10.1371/journal.pone.0122787.g011

distributed *FPR* values. Recall that the perfect match occurs at the point ($FPR = 0, TPR = 1$). In the PR-graphs in Fig. 12, we observe that Fig. 12(c) (β_θ_*vdW*) and (d) (β_θ_*PC*2) show the best distribution of the *R* and *P* values. Fig. 12(a) and (b) shows rather widely distributed *R* values and Fig. 12(e) and (f) shows that the *P* values are rather downward distributed. Recall that a perfect match occurs at the point ($R = 1, P = 1$).

Fig. 13(a) and (b) shows the normalized mutual information *NMI* and the likelihood ratio *LR*, respectively, and both suggest that β_θ_*vdw* and β_θ_*PC*2 are better than the others. The value of β_θ_*vdW* is again slightly better than β_θ_*PC*2. From a statistical view point, however, it is difficult to make a clear statement of their superiority. In this regard, we performed further statistical tests with additional eleven metrics and summarized the result in S1 Table of the supplementary material. The test clearly shows that the van der Waals volume of L-descriptors is consistently better measure than the others. For details, see the "Section 4. Secondary metrics tested" in the Supplementary material.

Optimal L-descriptor: the van der Waals volume

Fig. 14 shows some examples of recognized pockets using the six L-descriptor types from the two receptors (PDB accession codes: 1jd0 and 1s19) in the Astex Diverse Set. The *NMI* metric of each recognized pocket is shown in the figure. Fig. 14(a) shows 1jd0 (the carbonic anhydrase



(a) (b) (c)

(d) (e) (f)

Fig 12. The PR-graph of the L-descriptors. (a) the beta-shape volume, (b) the PC3, (c) the van der Waals volume, (d) the PC2, (e) the PC1, and (f) the minimum enclosing sphere.

doi:10.1371/journal.pone.0122787.g012

XII-acetazolamide complex), which has a small ligand consisting of 18 atoms. In this case, β_θ*_PC*3 and β_θ*_beta* are totally incorrect in that any atom of the optimal pocket is not contained within the recognized pocket. The value of β_θ*_PC*1 and β_θ*_mes* computes relatively large pockets compared to the size of the optimal pocket. Fig. 14(b) shows 1s19 (the vitamin D nuclear receptor-calcipotriol complex), which has a large ligand consisting of 70 atoms. In this case, β_θ*_PC*1 and β_θ*_mes* computes pockets that are too large compared to the size of the optimal pocket. In both cases, the β_θ*_vdw* and the β_θ*_PC*2 consistently predict good quality pockets.

Let l^{bound} and l^{opt} be the ligand conformations found in the crystal structure and in the minimum energy conformation, respectively. Let $\beta_{\theta_X}^{Y}$ be the value of l^{Y} for the L-descriptor type X of l^{opt}, where X is one of the six L-descriptor types and $Y \in \{bound,opt\}$. Fig. 15 shows the graphs for $\Delta L = \beta_{\theta_X}^{bound}$ - $\beta_{\theta_X}^{opt}$ for the Astex Diverse Set. Note that the graph of β_θ*_vdW* and β_θ*_beta* show less fluctuations compared to the other four; this implies that they are less sensitive to ligand conformation and less affected by the flexibility of the ligand. The fluctuation in the four graphs other than Fig. 15(a) and (c) implies that the corresponding L-descriptors are very sensitive to the ligand's flexibility. From the experiment, we conclude that β_θ*_vdW* is optimal in that it yields a consistently good performance regardless of ligand size and conformational change.

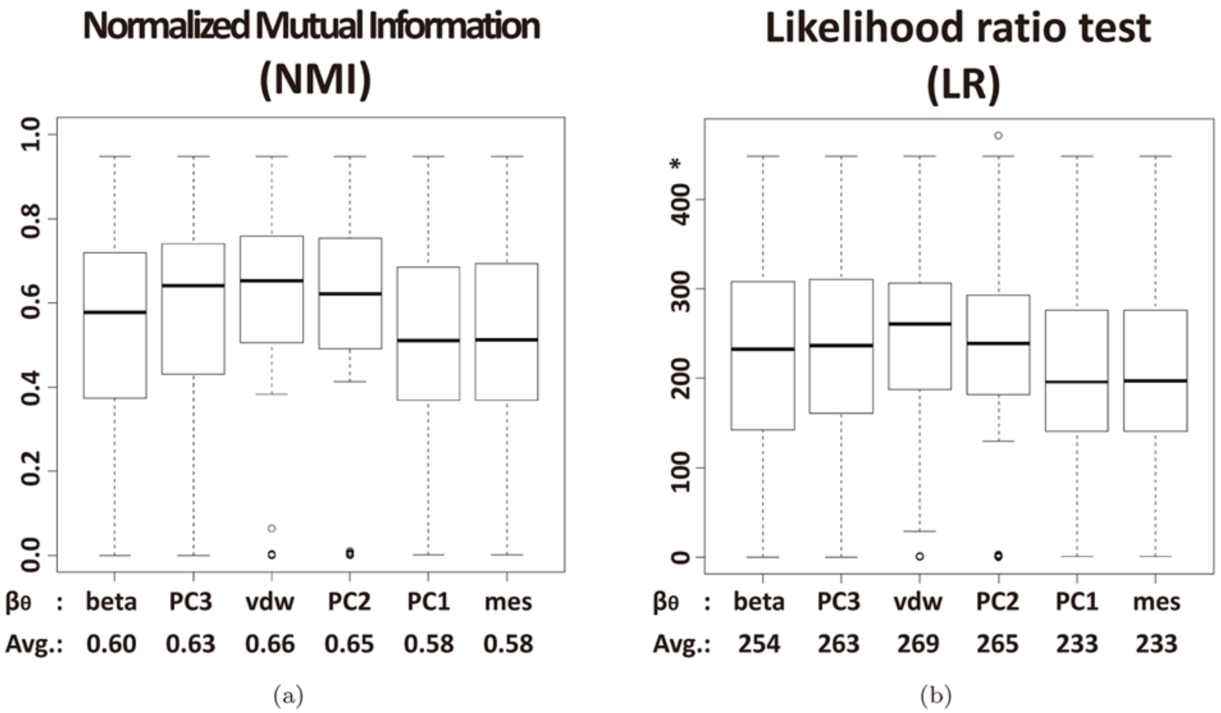


Fig 13. Box plots by entropy-based metrics of the six types of L-descriptor. (a) normalized mutual information and (b) Likelihood ratio. *Note that the y-axis scale of the *LR* plot is different from the *NMI* plot's.

doi:10.1371/journal.pone.0122787.g013

Benchmark

We benchmarked the proposed method against the STP (surface triplet propensities) algorithm [67] for recognizing the pockets of each protein in the Astex Diverse Set after removing the drug-like compounds. The STP algorithm assigns a score, called a patch score ranging between 0 to 100, to each and every atom of a protein. A higher value of the score implies that the atom has a higher probability to belong to a pocket. The STP algorithm selects those atoms whose scores are greater than a given threshold as the constituent of a predicted pocket. Thus, a higher patch score as a threshold selects fewer atoms than a lower one does. Be aware that the proposed method of this paper produces multiple components of boundary mesh where each can be a pocket candidate.

Fig. 16 shows the optimal pocket (Fig. 16(a)), the pocket computed by the proposed method (Fig. 16(b)), and the one by the STP method (Fig. 16(c) through (f)) for a protein (PDB Accession code: 1jd0). The bound compound is visualized as a set of blue sticks (for the reference purpose), the atoms belonging to pockets are visualized as colored balls, and the rest of the protein structure is visualized as gray line segments. The red balls in Fig. 16 (a) are the atoms of the optimal pocket; The green balls in Fig. 16 (b) are the atoms of the best matched component produced from the proposed algorithm; The yellow balls in Fig. 16 (c), (d), (e), and (f) are the atoms recognized by the STP method for the threshold values 80, 60, 40, and 20, respectively. Fig. 17 shows another example (PDB Accession code: 1s19). Experiments with other proteins show similar results.

The examples above show that the proposed method seems very powerful without any parameters and perhaps better than the STP method. This claim is asserted by the following

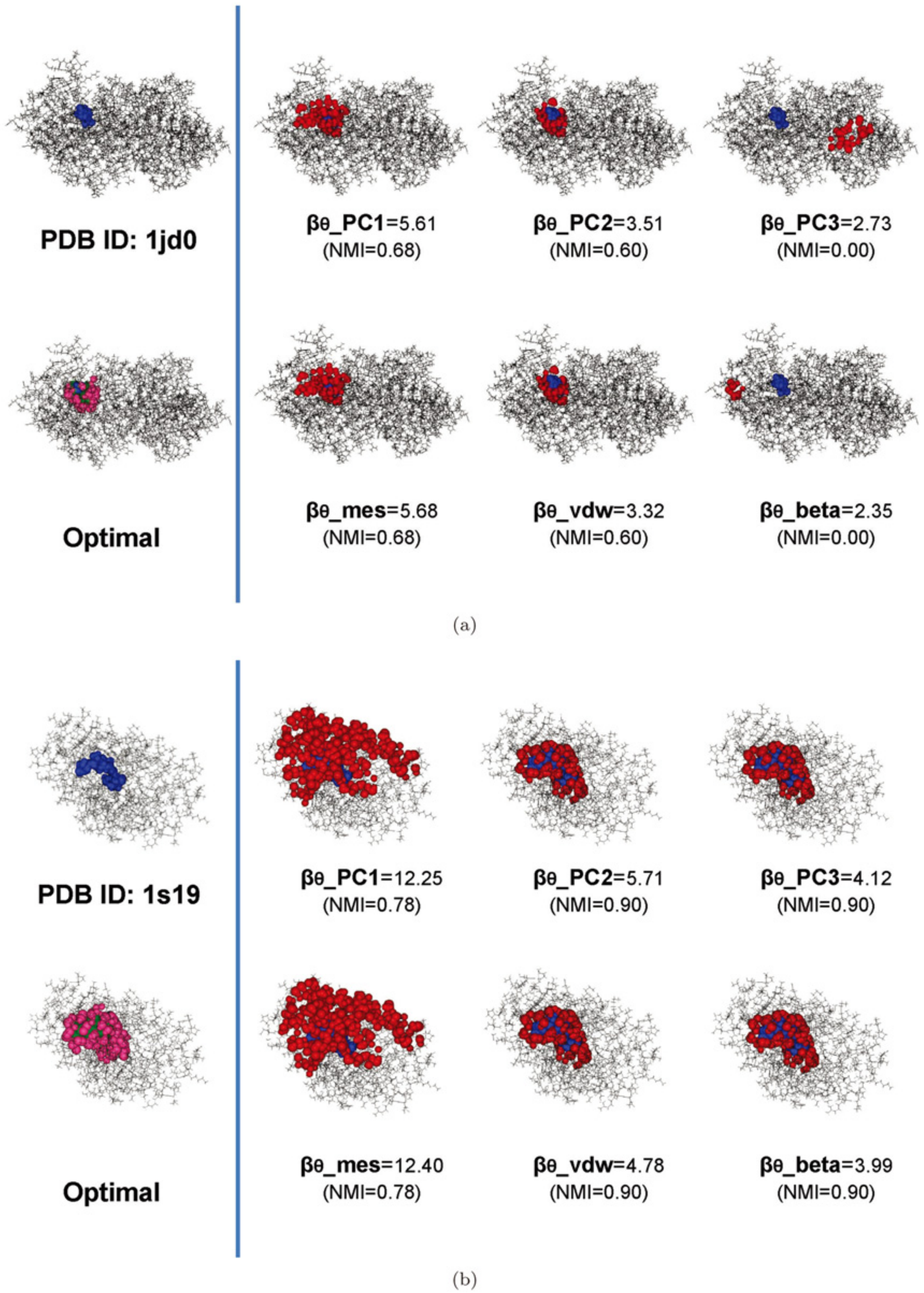
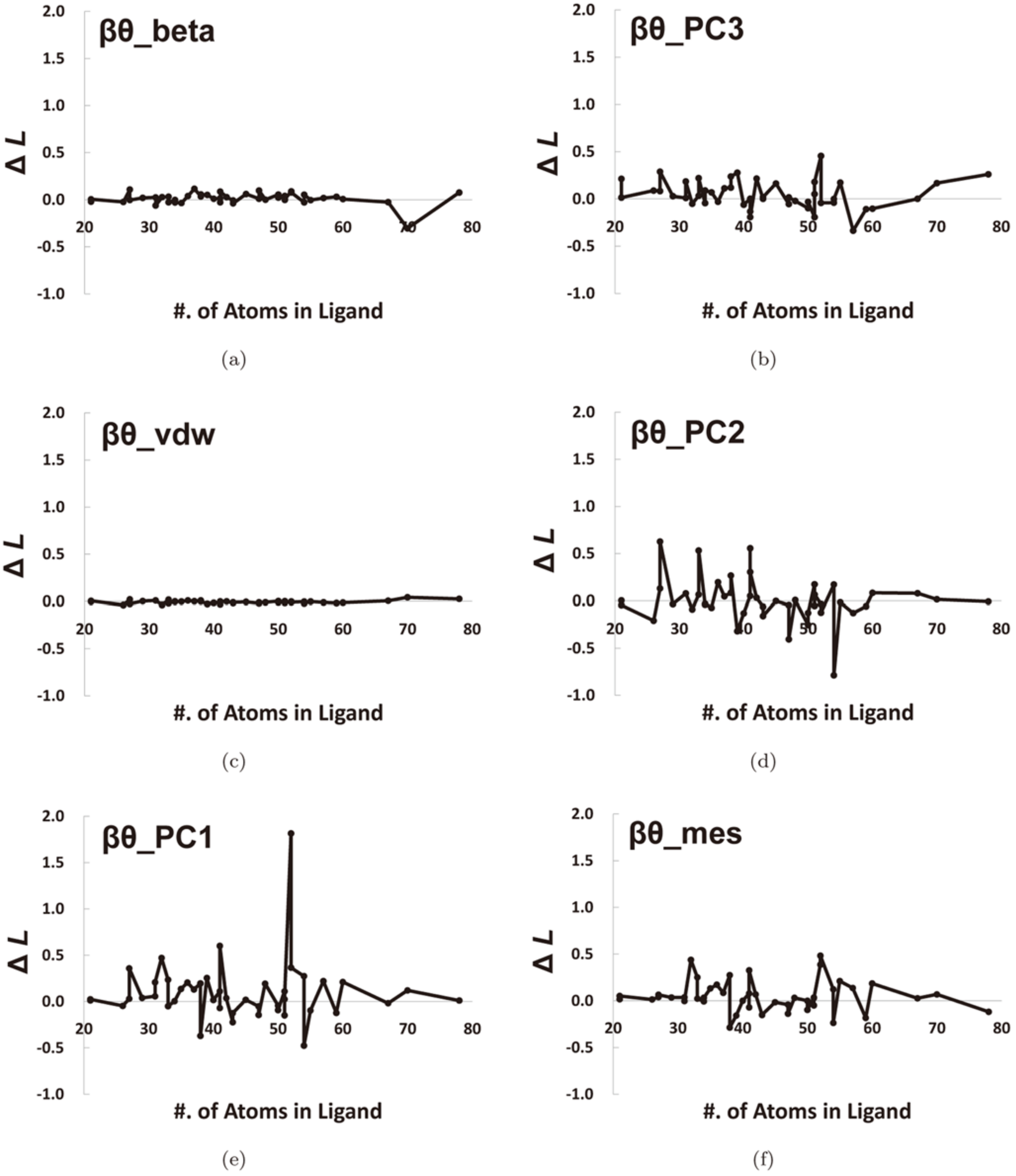


Fig 14. The optimal and recognized pockets of the PDB models. (a) PDB ID: 1jd0 (carbonic anhydrase XII—acetazolamide(18 atoms) complex) (b) PDB ID: 1s19 (vitamin D nuclear receptor-calcipotriol(70atoms) complex). The atoms are the colored receptor in black, the ligand in blue, the optimal pocket in pink, and the recognized pocket in red.

doi:10.1371/journal.pone.0122787.g014



(a)

(b)

(c)

(d)

(e)

(f)

Fig 15. Difference in the β_θ values by change of the ligand conformation. $\Delta L = \beta_{\theta_X}^{bound}$ - $\beta_{\theta_X}^{opt}$ (ie, ΔL = (β_θ of the bound ligand)–(β_θ of the ligand with minimum energy)).

doi:10.1371/journal.pone.0122787.g015

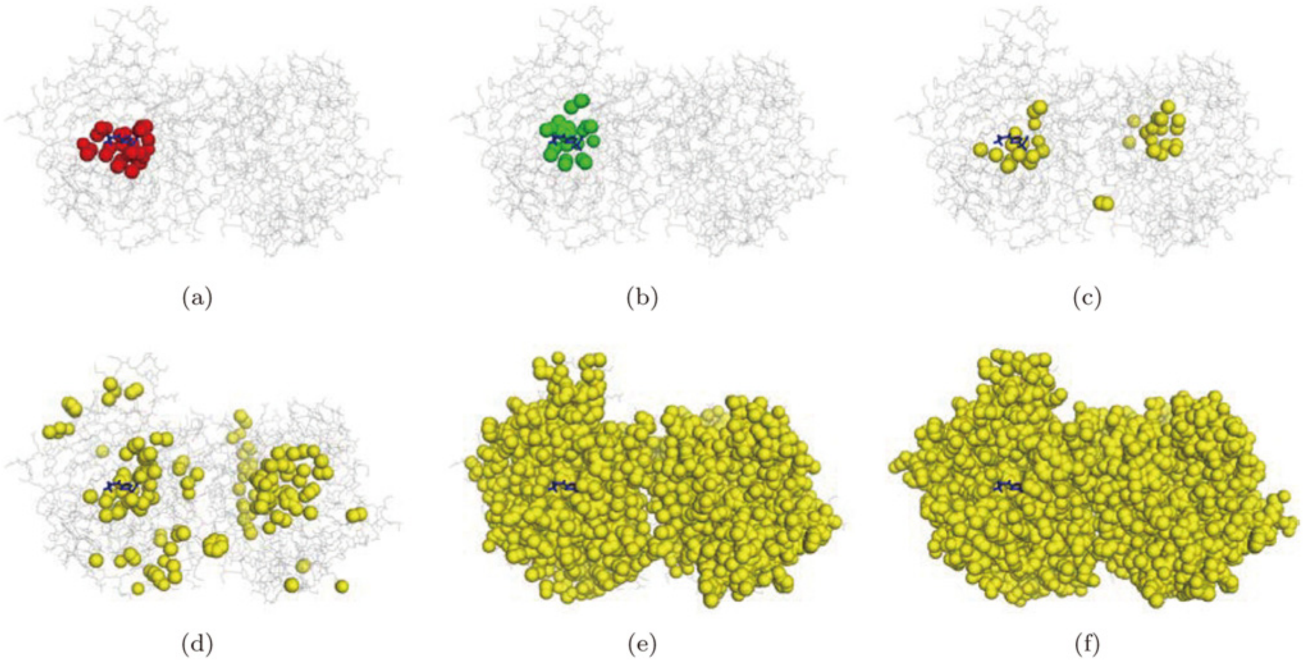


Fig 16. The visualization of pockets (PDB accession code: 1jd0). (a) The optimal pocket, (b) the best matched component produced by the proposed method, (c), (d), (e), and (f) are the atoms recognized by the STP method for the threshold values 80, 60, 40, and 20, respectively.

doi:10.1371/journal.pone.0122787.g016

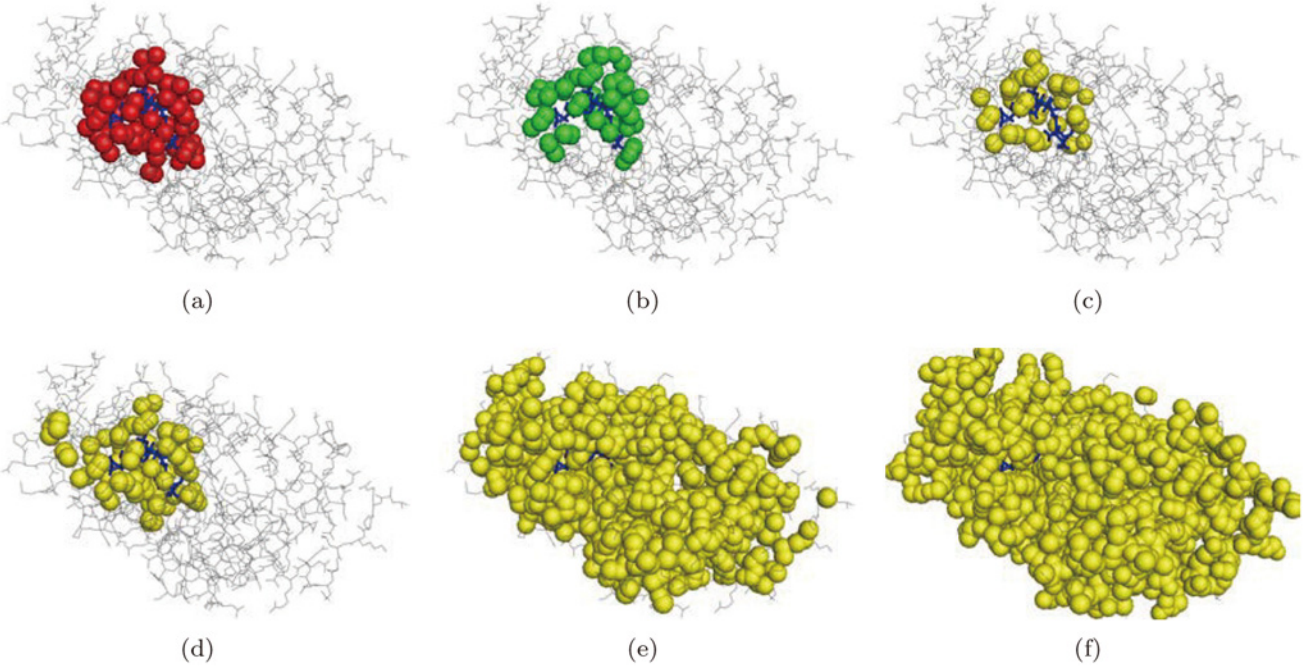


Fig 17. The visualization of pocket (PDB accession code: 1s19). (a) The optimal pocket, (b) the best matched component produced by the proposed method, (c), (d), (e), and (f) are the atoms recognized by the STP method for the threshold values 80, 60, 40, and 20, respectively.

doi:10.1371/journal.pone.0122787.g017

benchmark consisting of two types of tests. The first test type is the following. The proposed method selects the best five pocket candidates and the STP method selects atoms based on a threshold. We also select atoms at random for the reference where each random atom set has the size identical to the set produced by the STP method for each threshold value. Then, all atoms of each method forms one set, without processing to identify components where a "component" is a cluster of molecular boundary atoms which are topologically connected to each other. In this regard, we refer to this test type as "Without (component)."

The second test type is identical to the first except that the atoms in the atom set of each method are clustered together by the connectivity between the atoms. Then, the best matched component is used for the test. In this regard, we refer to this test type as "With (component)."

The following notations are for the "Without" case:

- A^{Beta}: The set of atoms in the five largest candidate sets by the proposed method.
- A^{STP}: The set of atoms by the STP method corresponding to each threshold τ whose value is determined from 0 to 95 by the increment of 5.
- A^{Random}: The set of randomly selected atoms where the $n(A^{Random}) = n(A^{STP})$ where $n(A)$ is the number of elements of A.

The following notations are for the "With" case:

- A^{Beta^*}: The best matched atom set to the optimal pocket by the proposed method.
- A^{STP^*}: The best matched component (of atom set) defined by clustering the atoms in A^{STP}.
- A^{Random^*}: The best matched component of A^{Random}.

We computed the five measures: The precision P (Fig. 18), the specificity SP (Fig. 19), the accuracy AC (Fig. 20), the sensitivity S (Fig. 21), and the normalized likelihood ratio LR (Fig. 22).

Fig. 18(a) shows the graphs of the precision for the three methods for "Without." The horizontal axis denotes the threshold and the vertical axis the computed precision value. Note the the proposed method, shown by the red solid circle labeled by "Beta," is constant, independent of the threshold. On the other hand, the STP (the black triangle) and the Random (the blue rectangle) methods heavily depends on the threshold value. It seems that the STP method

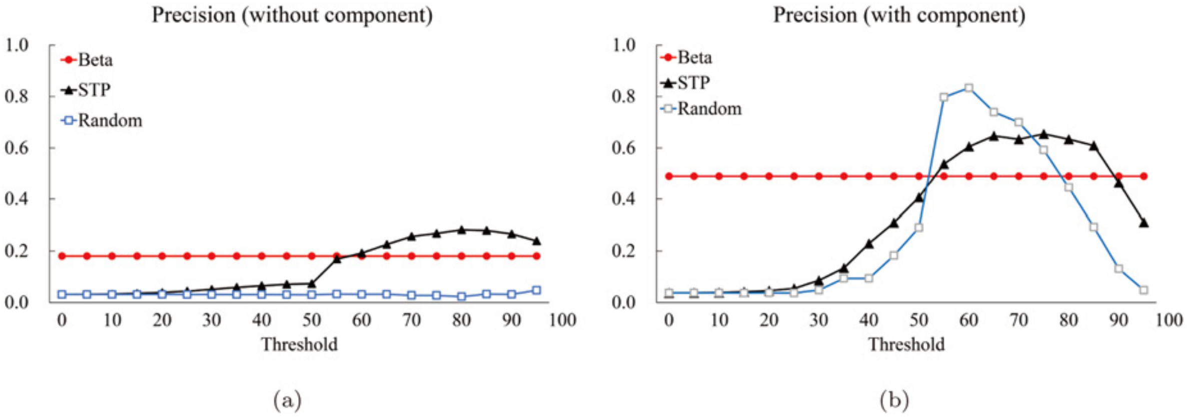


Fig 18. The precision graphs. The red circle corresponds to the proposed method. The black triangle and blue square correspond to the average value (of the 85 structures of the Astex Diverse Set) for the STP and Random methods for each threshold value, respectively. The horizontal and the vertical axes denote the thresholds and the computed values of precision, respectively. (a) Precision for "Without (component)" and (b) one for "With (component)."

doi:10.1371/journal.pone.0122787.g018

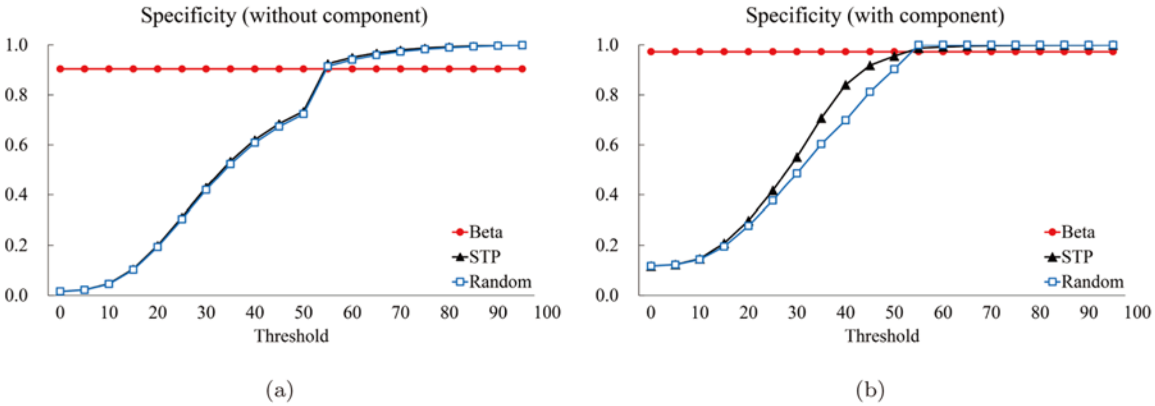


Fig 19. The specificity graphs. The red circle corresponds to the proposed method. The black triangle and blue square correspond to the average value (of the 85 structures of the Astex Diverse Set) for the STP and Random methods for each threshold value, respectively. The horizontal and the vertical axes denote the thresholds and the computed values of specificity, respectively. (a) Specificity for "Without (component)" and (b) one for "With (component)."

doi:10.1371/journal.pone.0122787.g019

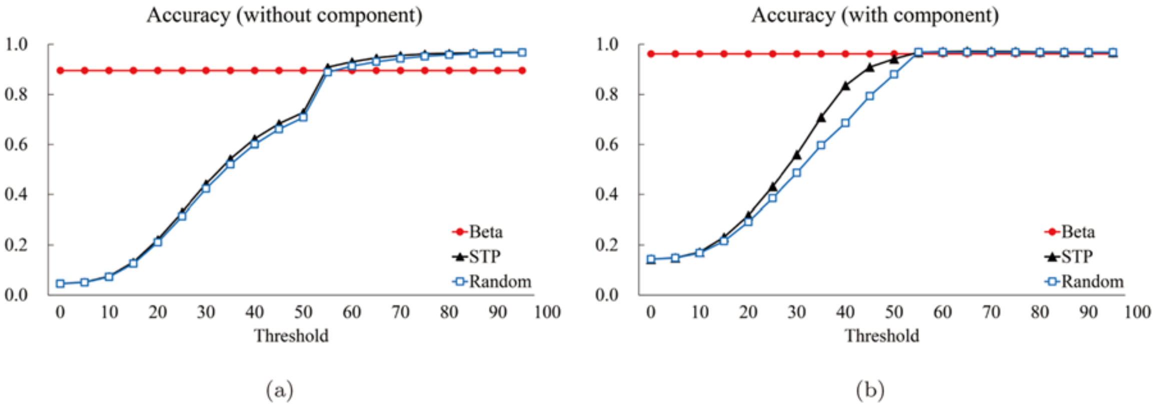


Fig 20. The accuracy graphs. The red circle corresponds to the proposed method. The black triangle and blue square correspond to the average value (of the 85 structures of the Astex Diverse Set) for the STP and Random methods for each threshold value, respectively. The horizontal and the vertical axes denote the thresholds and the computed values of accuracy, respectively. (a) Accuracy for "Without (component)" and (b) one for "With (component)."

doi:10.1371/journal.pone.0122787.g020

behaves better than the proposed method if the threshold is sufficiently big, say ≥ 60. No surprise to see the Random method behaves the worst.

Fig. 18(b) shows the precision graph for "With" component case. It is interesting to see that both STP and Random behave very well from the precision point of view if the threshold is big enough. Surprisingly the Random method shows the best precision for the range approximately between 55 and 70: It seems that this is because the Random method forms several component where each consists of relatively few atoms than the other two methods and some of the member atoms belong to the true pocket.

Fig. 19(a) shows the graphs for the specificity for the "Without" case. It is interesting that the STP and Random methods are surprisingly close and both produces slightly higher values than the proposed method where the threshold is bigger than (approximately) 60. The "With"

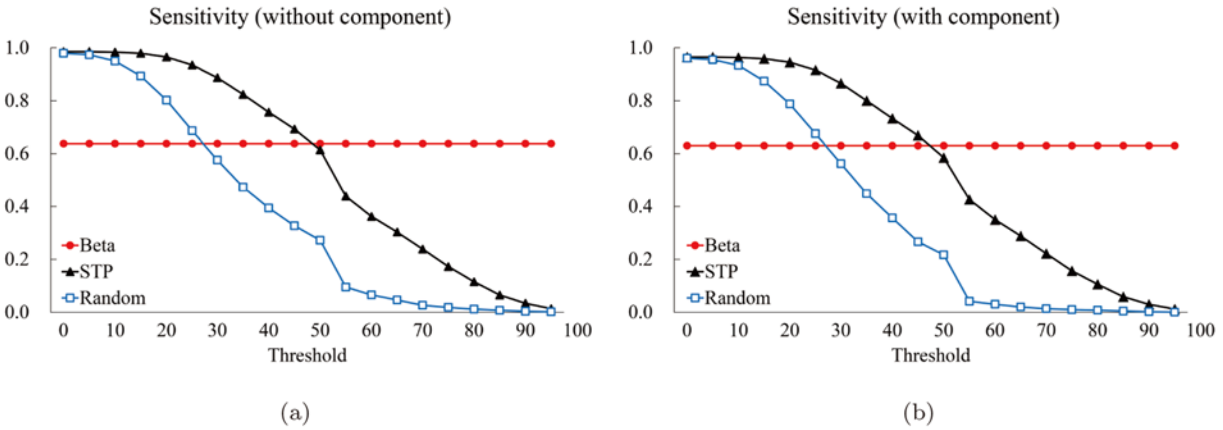


Fig 21. The sensitivity graphs. The red circle corresponds to the proposed method. The black triangle and blue square correspond to the average value (of the 85 structures of the Astex Diverse Set) for the STP and Random methods for each threshold value, respectively. The horizontal and the vertical axes denote the thresholds and the computed values of sensitivity, respectively. (a) Sensitivity for "Without (component)" and (b) one for "With (component)."

doi:10.1371/journal.pone.0122787.g021

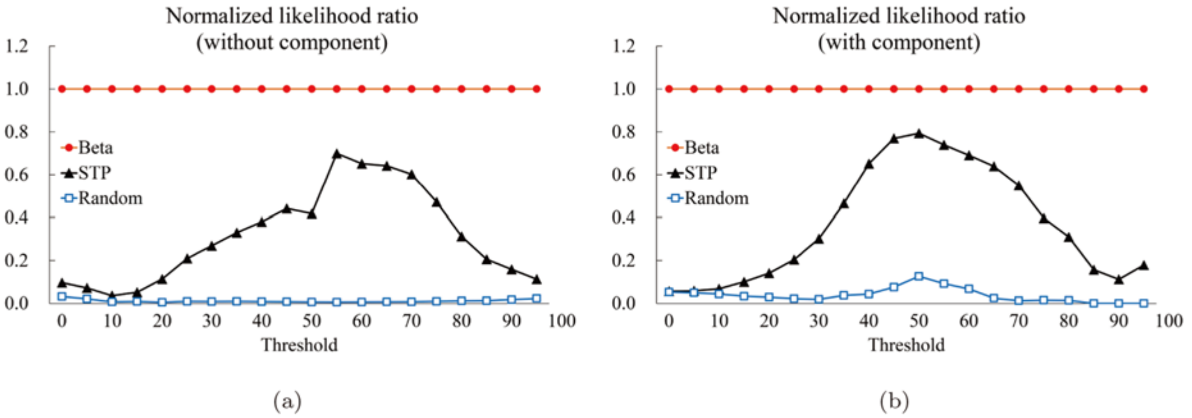


Fig 22. The normalized likelihood ratio graphs. The red circle corresponds to the proposed method. The black triangle and blue square correspond to the average value (of the 85 structures of the Astex Diverse Set) for the STP and Random methods for each threshold value, respectively. The horizontal and the vertical axes denote the thresholds and the computed values of likelihood ratio, respectively. (a) The normalized likelihood ratio for "Without (component)" and (b) one for "With (component)."

doi:10.1371/journal.pone.0122787.g022

case, Fig. 19(b), shows a similar behavior but all three methods are similar for bigger threshold values. Fig. 20 are the accuracy graphs which show patterns very similar to the specificity graphs. The similarity between the specificity and the accuracy is because there are significantly more atoms not belonging to the true pocket than the number of atoms belonging to the true pocket.

Fig. 21 shows the sensitivity graphs. While the proposed method (the red circle) shows a constant behavior, the STP method shows a decreasing pattern as the threshold increases and the two curves crosses approximately at the threshold of 50. It is obvious that the STP curve is monotonic because $A^{STP}(\tau = \tau_1) \subseteq A^{STP}(\tau = \tau_2)$, $\tau 1 > \tau_2$. As is expected, the graph of Random method is lower than the STP method. It is important to note that both Fig. 21(a) and (b) are

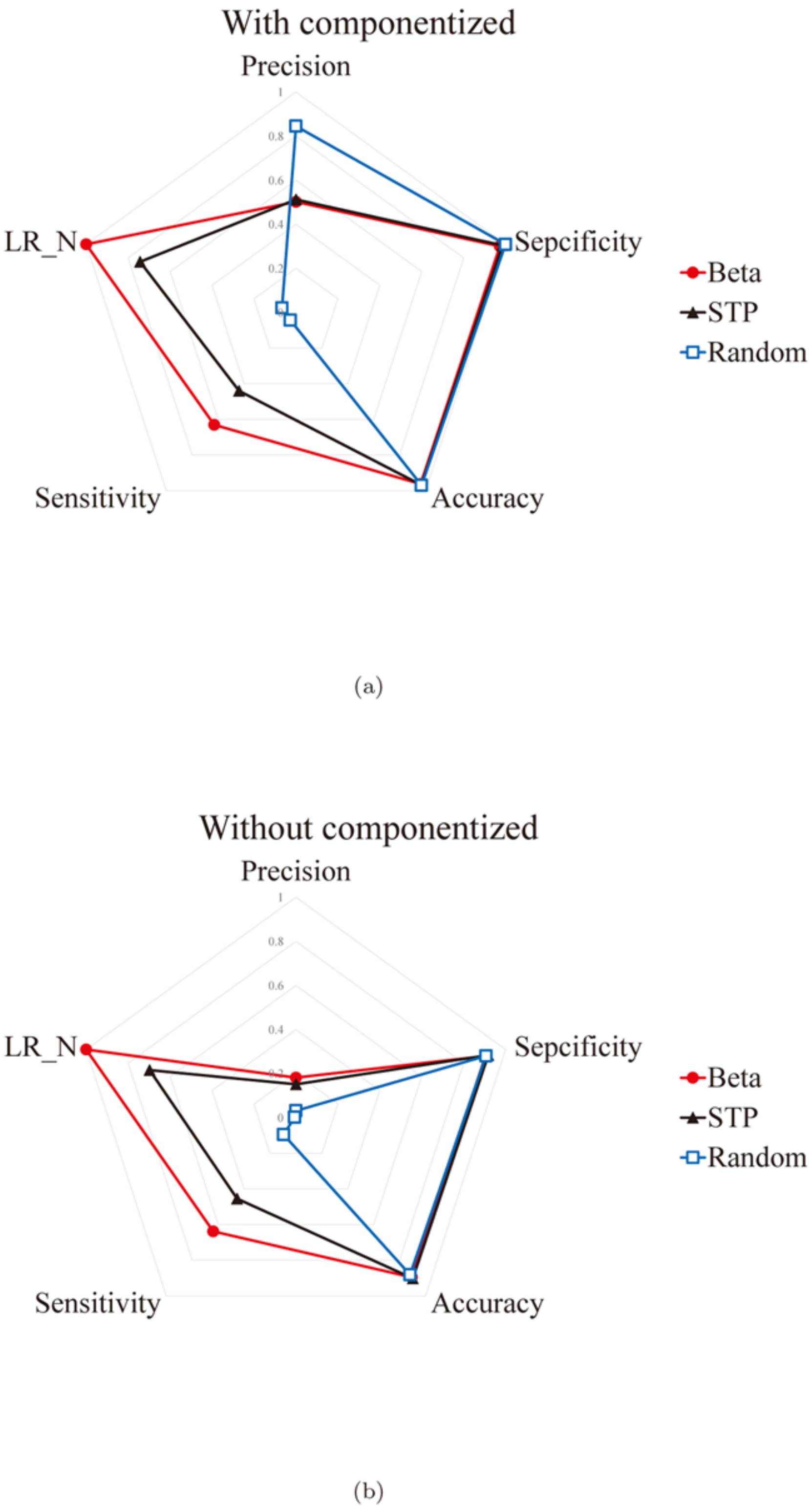


Fig 23. The radar charts of the proposed algorithm, the STP algorithm, and the Random method for the five statistical measures. (a) The case corresponding to the five best pockets recognized by the proposed algorithm, and (b) the case corresponding to the best pocket recognized by the proposed algorithm.

doi:10.1371/journal.pone.0122787.g023

very close to each other. This is because, regardless which method is used, the best matching component contains most of the atoms of the optimal pocket.

Fig. 22 shows the normalized likelihood graphs. Note that the proposed method outperforms the others independent of the threshold value.

We performed another test as follows. Let A^{Beta} be the set of all atoms belonging to the best five pockets recognized by the proposed algorithm. Let $A^{STP'}$ be the set of $n(A^{Beta})$ atoms recognized by the STP method. This means that we collect the best $n(A^{Beta})$ atoms from the one with the highest patch score to the ones with lower score, without considering the threshold. Let $A^{Random'}$ be the set of $n(A^{Beta})$ atoms randomly selected. Fig. 23(a) shows the distribution of the five statistical measures for the three methods. Suppose that we find the best matching component among the five pockets recognized by the proposed algorithm and let A^{Beta^*} be the set of the atoms belonging to this pocket. Let $A^{STP'^*}$ and $A^{Random'^*}$ be the sets of $n(A^{Beta^*})$ atoms recognized by the STP and the Random methods, respectively. Fig. 23(b) shows the distribution of the five statistical measures for the three methods with the three atom sets A^{Beta^*}, $A^{STP'^*}$ and $A^{Random'^*}$.

From the analysis above, we claim that the proposed method is better than the STP method in that it produces better quality pocket and is more robust.

Conclusion

This paper proposes a parameter optimization for a pocket recognition algorithm based on the recent theory of the beta-shape, which is a derivative structure of the Voronoi diagram of atoms in a molecule. The parameter optimization was done by considering the ligand shape, thus called the L-descriptor, in the pocket recognition process so that the recognized pocket is ligand-specific.

We examined six types of L-descriptor for ligands: the minimum enclosing sphere, the three principal axes of the principal component analysis, the van der Waals volume, and the beta-shape volume. From the experiment using the Astex Diverse Set containing 85 complexes of proteins with ligands and various statistical measures based on the confusion matrix, the L-descriptor based on the van der Waals volume showed the best and consistent performance throughout the entire range of the ligand size. The van der Waals volume also showed a consistent result over different ligand conformations. In conclusion, we claim that the van der Waals volume is the optimal shape descriptor of ligands for pocket recognition algorithms based on the beta-shape using a spherical probe representing the ligands. The claim is verified by a benchmark test against the STP algorithm using the Astex Diverse Set. The code for the proposed pocket algorithm will be included in the powerful `BetaVoid` program for extracting void features of molecules [68].

Supporting Information

S1 Table. The definition of symbols.
(PDF)

Acknowledgments

J.-K. Kim, C.-I. Won, J. Cha, and D.-S. Kim were supported by the National Research Foundation grant funded by MSIP (No. 2012R1A2A1A05026395), Republic of Korea. K. Lee was supported by the grant (201400000002667) funded by Small and Medium Business Administration (SMBA), Republic of Korea.

Author Contributions

Conceived and designed the experiments: J-KK C-IW D-SK. Performed the experiments: J-KK C-IW JC. Analyzed the data: J-KK C-IW JC KL. Wrote the paper: D-SK.

References

1. Ghosh S, Nie A, An J, Huang Z. Structure-based virtual screening of chemical libraries for drug discovery. Current Opinion in Chemical Biology. 2006; 10: 194–202. doi: 10.1016/j.cbpa.2006.04.002 PMID: 16675286
2. Jorgensen WL. The many roles of computation in drug discovery. Science. 2004; 303: 1813–1818. doi: 10.1126/science.1096361 PMID: 15031495
3. McInnes C. Virtual screening strategies in drug discovery. Current Opinion in Chemical Biology. 2007; 11: 494–505. doi: 10.1016/j.cbpa.2007.08.033 PMID: 17936059
4. Taha MO, Tarairah M, Zalloum H, Abu-Sheikha G. Pharmacophore and qsar modeling of estrogen receptor ligands and subsequent validation and in silico search for new hits. Journal of Molecular Graphics and Modelling. 2010; 28: 383–400. doi: 10.1016/j.jmgm.2009.09.005 PMID: 19850503
5. Politi A, Durdagi S, Moutevelis-Minakakis P, Kokotos G, Mavromoustakos T. Development of accurate binding affinity predictions of novel renin inhibitors through molecular docking studies. Journal of Molecular Graphics and Modelling. 2010; 29: 425–435. doi: 10.1016/j.jmgm.2010.08.003 PMID: 20855222
6. Dutta S, Burkhardt K, Young J, Swaminathan GJ, Matsuura T, Henrick K, et al. Data deposition and annotation at the worldwide protein data bank. Molecular Biotechnology. 2009; 42: 1–13. doi: 10.1007/s12033-008-9127-7 PMID: 19082769
7. Pieper U, Eswar N, Webb BM, Eramian1 D, Kelly L, Barkan DT, et al. MODBASE, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Research. 2009; 37: D347–D354. doi: 10.1093/nar/gkn791 PMID: 18948282
8. Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T. The SWISS-MODEL repository and associated resources. Nucleic Acids Research. 2009; 37: D387–D392. doi: 10.1093/nar/gkn750 PMID: 18931379
9. Castrignanó T, Meo PDD, Cozzetto D, Talamo IG, Tramontano A. The PMDB protein model database. Nucleic Acids Research. 2006; 34: D306–D309. doi: 10.1093/nar/gkj105 PMID: 16381873
10. Schindler T, Bornmann W, Pellicena P, Miller WT, Clarkson B, Kuriyan J. Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. Science. 2000; 289: 1938–1942. doi: 10.1126/science.289.5486.1938 PMID: 10988075
11. Hardy LW, Malikayil A. The impact of structure-guided drug design on clinical agents. Current Drug Discovery. 2003; 3: 15–20.
12. Alvarez JC. High-throughput docking as a source of novel drug leads. Current Opinion in Chemical Biology. 2004; 8: 365–370. doi: 10.1016/j.cbpa.2004.05.001 PMID: 15288245
13. Blundell TL, Patel S. High-throughput X-ray crystallography for drug discovery. Current Opinion in Pharmacology. 2004; 4: 490–496. doi: 10.1016/j.coph.2004.04.007 PMID: 15351354
14. Cherfils J, Janin J. Protein docking algorithms: Simulating molecular recognition. Current Opinion in Structural Biology. 1993; 3: 265–269. doi: 10.1016/S0959-440X(05)80162-9
15. Finn PW, Kavraki LE. Computational approaches to drug design. Algorithmica. 1999; 25: 347–371. doi: 10.1007/PL00008282
16. Campbell SJ, Gold ND, Jackson RM, Westhead DR. Ligand binding: Functional site location, similarity and docking. Current Opinion in Structural Biology. 2003; 13: 389–395. doi: 10.1016/S0959-440X(03)00075-7 PMID: 12831892
17. Coleman RG, Sharp KA. Protein pockets: Inventory, shape, and comparison. Journal of Chemical Information and Modeling. 2010; 50: 589–603. doi: 10.1021/ci900397t PMID: 20205445
18. Schulz-Gasch T, Stahl M. Binding site characteristics in structure-based virtual screening: Evaluation of current docking tools. Journal of Molecular Modeling. 2003; 9: 47–57. PMID: 12638011
19. Nayal M, Honig B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. Proteins: Structure, Function, and Bioinformatics. 2006; 63: 892–906. doi: 10.1002/prot.20897
20. Ho CM, Marshall GR. Cavity search: an algorithm for the isolation and display of cavity-like binding regions. Journal of Computer-Aided Molecular Design. 1990; 4: 337–354. doi: 10.1007/BF00117400 PMID: 2092080

21. Levitt D, Banaszak L. POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. Journal of Molecular Graphics. 1992; 10: 229–234. doi: 10.1016/0263-7855(92)80074-N PMID: 1476996

22. Voorintholt R, Kosters MT, Vegter G, Vriend G, Hol WG. A very fast program for visualizing protein surfaces, channels and cavities. Journal of Molecular Graphics. 1989; 7: 243–245. doi: 10.1016/0263-7855(89)80010-4 PMID: 2486827

23. Durrant JD, de Oliveira CAF, McCammon JA. POVME: An algorithm for measuring binding-pocket volumes. Journal of Molecular Graphics and Modelling. 2011; 29: 773–776. doi: 10.1016/j.jmgm.2010.10.007 PMID: 21147010

24. Brady GP Jr, Stouten PFW. Fast prediction and visualization of protein binding pockets with PASS. Journal of Computer-Aided Molecular Design. 2000; 14: 383–401. doi: 10.1023/A:1008124202956 PMID: 10815774

25. Kuntz ID, Blaney FM, Oatley SJ. A geometric approach to macromolecule-ligand interactions. Journal of Molecular Biology. 1982; 161: 269–288. doi: 10.1016/0022-2836(82)90153-X PMID: 7154081

26. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Science. 1996; 5: 2438–2452. PMID: 8976552

27. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. Protein Science. 1998; 7: 1884–1897. doi: 10.1002/pro.5560070905 PMID: 9761470

28. Peters KP, Fauck J, Frömmel C. The automatic search for ligand binding sites in protein of known three dimensional structure using only geometric criteria. Journal of Molecular Biology. 1996; 256: 201–213. doi: 10.1006/jmbi.1996.0077 PMID: 8609611

29. Kim D, Cho CH, Cho Y, Ryu J, Bhak J, Kim DS. Pocket extraction on proteins via the Voronoi diagram of spheres. Journal of Molecular Graphics & Modelling. 2008; 26: 1104–1112. doi: 10.1016/j.jmgm.2007.10.002

30. Yin X, Giap C, Lazo JS, Prochownik EV. Low molecular weight inhibitors of myc-max interaction and function. Oncogene. 2003; 22: 6151–6159. doi: 10.1038/sj.onc.1206641 PMID: 13679853

31. Hammoudeh DI, Follis AV, Prochownik EV, Metallo SJ. Multiple independent binding sites for small-molecule inhibitors on the oncoprotein c-Myc. Journal of the American Chemical Society. 2009; 131: 7390–7401. doi: 10.1021/ja900616b PMID: 19432426

32. Scheswohl DM, Harrell JR, Rajfur Z, Gao G, Campbellb SL, Schaller MD. Multiple paxillin binding sites regulate FAK function. Journal of Molecular Signaling 2008;3.

33. Sitry D, Seeliger MA, Ko TK, Ganoth D, Breward SE, Itzhaki LS, et al. Three different binding sites of Cks1 are required for p27-ubiquitin ligation. Journal of Biological Chemistry. 2002; 277: 42233–42240. doi: 10.1074/jbc.M205254200 PMID: 12140288

34. Kim DS, Ryu J. Side-chain prediction and computational protein design problems. Biodesign. 2014; 2: 26–38.

35. Edelsbrunner H, Mücke EP. Three-dimensional alpha shapes. ACM Transactions on Graphics. 1994; 13: 43–72. doi: 10.1145/174462.156635

36. Edelsbrunner H. Weighted alpha shapes. Technical Report UIUCDCS-R-92–1760, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL; 1992.

37. Kim DS, Cho Y, Sugihara K, Ryu J, Kim D. Three-dimensional beta-shapes and beta-complexes via quasi-triangulation. Computer-Aided Design. 2010; 42: 911–929. doi: 10.1016/j.cad.2010.06.004

38. Kim DS, Kim CM, Won CI, Kim JK, Ryu J, Cho Y, et al. BetaDock: Shape-priority docking method based on Beta-complex. Journal of Biomolecular Structure & Dynamics. 2011; 29: 219–242. doi: 10.1080/07391102.2011.10507384

39. Shin WH, Kim JK, Kim DS, Seok C. GalaxyDock2: protein-ligand docking using beta-complex and global optimization. Journal of Computational Chemistry. 2013; 34: 2647–2656. doi: 10.1002/jcc.23438 PMID: 24108416

40. Cho Y, Kim JK, Ryu J, Won CI, Kim CM, Kim D, et al. BetaMol: a molecular modeling, analysis and visualization software based on the beta-complex and the quasi-triangulation. Journal of Advanced Mechanical Design, Systems, and Manufacturing. 2012; 6: 389–403. doi: 10.1299/jamdsm.6.389

41. Kim DS, Cho Y, Kim D. Euclidean Voronoi diagram of 3D balls and its computation via tracing edges. Computer-Aided Design. 2005; 37: 1412–1424. doi: 10.1016/j.cad.2005.02.013

42. Kim DS, Kim D, Cho Y, Sugihara K. Quasi-triangulation and interworld data structure in three dimensions. Computer-Aided Design. 2006; 38: 808–819. doi: 10.1016/j.cad.2006.04.008

43. Kim DS, Cho Y, Sugihara K. Quasi-worlds and quasi-operators on quasi-triangulations. Computer-Aided Design. 2010; 42: 874–888. doi: 10.1016/j.cad.2010.06.002

44. Kim JK, Cho Y, Kim D, Kim DS. Voronoi diagrams, quasi-triangulations, and beta-complexes for disks in $\mathbb{R}^2$: The theory and implementation in BetaConcept. Journal of Computational Design and Engineering. 2014; 1: 79–87. doi: 10.7315/JCDE.2014.008

45. Lee, C. Manifoldization of Beta-shapes and Extraction of Pocket on Proteins. Ph.D. thesis, Hanyang University, Seoul, Korea. 2010.

46. Ghose AK, Viswanadhan VN, Wendoloski JJ. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. a qualitative and quantitative characterization of known drug databases. Journal of Combinatorial Chemistry. 1999; 14: 55–68.

47. Saranya N, Selvaraj S. Variation of protein binding cavity volume and ligand volume and ligand volume in protein-ligand complexes. Bioorganic & Medicinal Chemistry Letters. 2009; 19: 5769–5772. doi: 10.1016/j.bmcl.2009.07.140

48. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins: Structure, Function, and Genetics. 2002; 47: 409–443. doi: 10.1002/prot.10115.abs

49. Jolliffe I. Principal Component Analysis. Springer, second edition; 2002.

50. Kim DS, Ryu J, Shin H, Cho Y. Beta-decomposition for the volume and area of the union of three-dimensional balls and their offsets. Journal of Computational Chemistry. 2012; 33: 1252–1273. doi: 10.1002/jcc.22956

51. Kim CM, Won CI, Cho Y, Kim D, Lee S, Bhak J, et al. Interaction interfaces in proteins via the Voronoi diagram of atoms. Computer-Aided Design. 2006; 38: 1192–1204. doi: 10.1016/j.cad.2006.07.007

52. Kim CM, Won CI, Ryu J, Cho CH, Bhak J, Kim DS. Parameter selection of pocket extraction algorithm using interaction interface. Journal of Zhejiang University - Science A. 2006; 7: 1492–1499. doi: 10.1631/jzus.2006.A1492

53. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: In ICML 06: Proceedings of the 23rd international conference on Machine learning. 2006;233–240. doi: 10.1145/1143844.1143874

54. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27: 861–874. doi: 10.1016/j.patrec.2005.10.010

55. Wickens TD. Multiway Contigency Tables Analysis for the Social Sciences. Taylor & Francis, Inc; 1989.

56. Bush WS, Edwards TL, Dudek SM, McKinney BA, Ritchie MD. Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. BMC Bioinformatics. 2008; 9: 238. doi: 10.1186/1471-2105-9-238 PMID: 18485205

57. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. Biometrika. 1928; 20A: 175–240.

58. Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, et al. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic Epidemiology. 2007; 31: 306–315. doi: 10.1002/gepi.20211 PMID: 17323372

59. Kubat M, Holte RC, Matwin S. Machine learning for the detection of oil spills in satellite radar images. Machine Learning. 1998; 30: 195–215. doi: 10.1023/A:1007452223027

60. Youden WJ. Index for rating diagnostic tests. Cancer. 1950; 3: 32–35.

61. Lewis DD, Gale WA. A sequential algorithm for training text classifiers. In: In SIGIR 94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrie. 1994;3–12.

62. Linn S, Grunau PD. New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. Epidemiologic Perspectives & Innovations. 2006;3.

63. Goodman LA, Kruskal WH. Measures of association for cross classifications. Journal of the American Statistical Association. 1954; 49: 732–764. doi: 10.1080/01621459.1954.10501231

64. Kendall M A new measure of rank correlation. Biometrika. 1938; 30: 81–89. doi: 10.1093/biomet/30.1-2.81

65. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, et al. Diverse, high-quality test set for the validation of protein-ligand docking performance. Journal of Medicinal Chemistry. 2007; 50: 726–741. doi: 10.1021/jm061277y PMID: 17300160

66. CambridgeSoft, Chem3D user's guide revision 9.0.1. Technical report, CambridgeSoft; 2004.

67. Mehio W, Kemp GJ, Taylor P, Walkinshaw MD. Identification of protein binding surfaces using surface triplet propensities. Bioinformatics. 2010; 26: 2549–2555. doi: 10.1093/bioinformatics/btq490 PMID: 20819959

68. Kim JK, Cho Y, Laskowski RA, Ryu SE, Sugihara K, Kim DS. BetaVoid: molecular voids via beta-complexes and Voronoi diagrams. Proteins: Structure, Functions, and Bioinformatics. 2014; 82: 1829–1849. doi: 10.1002/prot.24537