JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

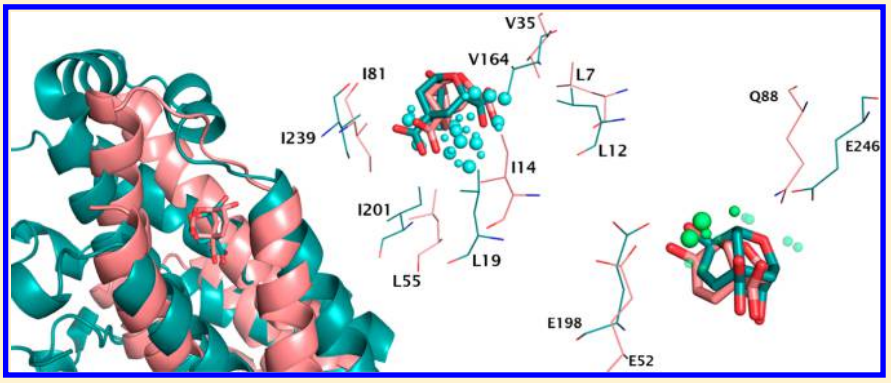
pubs.acs.org/jcim

Detection of Binding Site Molecular Interaction Field Similarities

Matthieu Chartier and Rafael Najmanovich*

Department of Biochemistry, Faculty of Medicine and Health Sciences, University of Sherbrooke, 12e Avenue Nord, Sherbrooke, J1H 5N4 Québec, Canada

Supporting Information



ABSTRACT: Protein binding-site similarity detection methods can be used to predict protein function and understand molecular recognition, as a tool in drug design for drug repurposing and polypharmacology, and for the prediction of the molecular determinants of drug toxicity. Here, we present IsoMIF, a method able to identify binding site molecular interaction field similarities across protein families. IsoMIF utilizes six chemical probes and the detection of subgraph isomorphisms to identify geometrically and chemically equivalent sections of protein cavity pairs. The method is validated using six distinct data sets, four of those previously used in the validation of other methods. The mean area under the receiver operator curve (AUC) obtained across data sets for IsoMIF is higher than those of other methods. Furthermore, while IsoMIF obtains consistently high AUC values across data sets, other methods perform more erratically across data sets. IsoMIF can be used to predict function from structure, to detect potential cross-reactivity or polypharmacology targets, and to help suggest bioisosteric replacements to known binding molecules. Given that IsoMIF detects spatial patterns of molecular interaction field similarities, its predictions are directly related to pharmacophores and may be readily translated into modeling decisions in structure-based drug design. IsoMIF may in principle detect similar binding sites with distinct amino acid arrangements that lead to equivalent interactions within the cavity. The source code to calculate and visualize MIFs and MIF similarities are freely available.

INTRODUCTION

The identification of similarities between proteins has many practical applications. Sequence similarity algorithms like BLAST[1] allow one to quickly retrieve homologous proteins and transfer potential functional annotations from the target to the query protein. Sequence similarity also enables the construction of phylogenetic trees grouping proteins into families, for example, with kinases,[2] to view functional information within an evolutionary context.[3] Whereas traditionally the function of a protein was well-characterized biochemically prior to the elucidation of its structure, with the advent of structural genomics projects in the past decade, there has been an influx of proteins for which the structure is known but not their function. Because sequence is less conserved than structure,[4] comparing two proteins using sequence alone ignores structural information that is relevant for function. A number of well-established methods such as DALI[5] measure global structural similarities, which coupled with databases like SCOP[6] or CATH[7] can be used to understand the cellular (biological processes) or molecular (molecular interactions) functions of a protein. Finally, meta-servers such as ProFunc[8] combine a number of sequence- and structure-based methods to predict protein function from structure.

Even when the function of a protein is known, the detection of similarities has important applications. From a drug design perspective, detecting proteins that are similar to the drug target is important to prevent cross-reactivity and the potential associated side effects. In such cases, sequence and structural similarities can fail for two reasons. First, divergent evolution may introduce mutations deemed minor at the level of sequence or structure, but with drastic effects locally in the binding site, altering the molecular function, e.g., shifting substrate specificity. Second, convergent evolution can bring proteins unrelated by sequence or structure to acquire the same molecular function. In these two cases, the similarities act at a more local scale and affect the physicochemical environment.

Received: March 9, 2015
Published: July 9, 2015

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

Therefore, a solution is to use methods that can detect local physicochemical similarities in binding pockets.

Such methods can help predict molecular function, identify cross-reactivity or polypharmacological targets, predict binding fragments, and repurpose existing drugs. For example, IsoCleft[9,10] is graph-matching-based method for the detection of 3D atomic binding site similarities that was used to predict function for structural genomics proteins,[11,12] reclassify members of the human cytosolic sulfotransferase family,[13,14] and analyze the druggability of histone methyltransferases[15] based on binding site similarities. SOIPPA[16] predicted off-targets for CETP inhibitors in agreement with experimental assays, and the authors explained adverse side effects observed in clinical trials using a systems biology approach.[17] FragFEATURE[18] uses the FEATURE microenvironments[19] combined with a knowledge-based approach to predict binding fragments for a target protein. CavBase detected similarities between COX-2 and carbonic anhydrase protein structures both known to bind COX-2 selective Celecoxib inhibitor.[20] The algorithms behind these and the numerous other methods for the prediction of binding site similarities can be divided into three components: representation, search, and scoring.

With respect to representations, many methods transform the protein into a simplified representation and sometimes incorporate additional information. Mostly, proteins are represented using C_α atoms, functional atoms, pseudocenters, or electrostatic surfaces, but some use all atoms. For example, SOIPPA[16] and Psilo[21] use C_α atoms, IsoCleft[9,10] uses a two-step C_α/all-atom process, CavBase[22] places pseudocenters in the vicinity of amino acids using distance and angle cutoffs, SiteEngine[23] proceeds similarly but without angle considerations for hydrogen bonds, and eF-site[24] uses a Connolly surface where the electrostatic potential is calculated. Other more unique forms of representation exist; pocketFEATURE[25] transforms important binding site atoms into micro environments that consist of physicochemical descriptors on concentric spherical shells, and siteAlign[26] uses an 80-face polyhedron onto which descriptors are projected. FuzCav[27] represents binding sites with 4833-long integer fingerprints. Proteins can also be represented with molecular interaction fields (MIFs) as in FLAP[28,29] used in BioGPS,[30] where protein MIFs are measured in binding sites using the GRID force field[31] and VolSite/Shaper.[32] FLAP filters out grid points other than those that represent local energetic minima (MINI points). VolSite/Shaper assigns pharmacophore properties to grid points based on the properties of the closest protein atom in the surface. For some methods, the representation is complemented with a degree of buriedness, curvature, or conservation scores. In all cases, as the task is to detect binding site similarities, only cavities are identified and subsequently used in search and scoring. Depending on the goal of the detection of binding site similarities, information about bound ligands in the query binding site can be used, but most generally, the definition of binding sites, unless for validation purposes, should be independent of bound ligands. Unfortunately this is not the case for all methods as recognized by others.[33]

The similarity search is generally based on graph-matching, geometric hashing, or downhill simplex minimization. Graph-matching requires building an association graph where nodes represent pairs of chemically similar atoms,[9] pairs of pseudocenters,[22] or pairs of triangles of atoms,[34] with each member of a given pair coming from the query and target binding sites, respectively. Edges in the association graph represent the relative geometric similarity. Finally, cliques in the association graph represent an entire subset of units (atoms, pseudocenters, triangles) in each of the binding sites that are chemically and geometrically equivalent. SiteEngine[23] builds a hash table of triangles of pseudocenters and KRIPO[35] lists in alphabetical representation 2, 3, and 4 point pharmacophores. These lists can be searched with a query key. FLAP[28,29] performs an exhaustive search among all pairs of similar 4 point pharmacophores between two MIFs. VolSite/Shaper relies on the proprietary use of OpenEye Chem and Shape toolkits to perform the superimposition of smoothed Gaussian functions representing shapes defined by each of the probe grids. Psilo[21] searches exhaustively all possible C_α superpositions. eSiteMatch[36] uses machine-learning to predict distances between pairs of C_α atoms of two cavities. The Kuhn−Munkres algorithm then finds the set of residue pairs that minimizes the overall distances between C_α atoms. The polyhedron representation of siteAlign[26] allows a search that rotates and translates the polyhedron into the query cavity. Search algorithms are often time consuming, and this must be considered for high-throughput applications. Krotzky et al.[37] decreased the run-time of cavBase by replacing the graph-matching algorithm by a linear comparison of weighted distance histograms.

The search aims to find the largest similarity between two binding sites pockets. Thus, the methods to score similarity impact the results. The scoring functions are as diverse as the methods to detect it. For example, similarity scores can be calculated as the absolute number of atoms (or pseudocenters) matched, RMSD, or fraction of binding site matched. The score can be weighted by various descriptors added to the representation. For example, SOIPPA uses the profile distances weighted by normal vector differences with a distance penalty. CavBase and SiteEngine rely on a measure of surface overlap. eMatchSite scores each C_α pairs using seven descriptors including secondary structure, hydrophobicity, and residue binding probability among others. A Tanimoto similarity is often used, usually using the number of atoms in common and initial search space of both cavities. PocketFEATURE measures a Tanimoto between microenvironments but using the presence/absence of similar properties and normalized with a background similarity distribution calculated with a non-redundant set of 3D structures. Psilo, while only using C_α in the search, measures an overlap score using C_β atoms. IsoCleft utilizes a Tanimoto score based on all binding site heavy atoms.

Ultimately, more important than the location of atoms or mapped physicochemical properties on the molecular surface of a binding site are the actual interactions that these atoms can participate on. For example, single mutations can drastically affect the binding affinity of ligands or not at all.[38] Here, we present IsoMIF, a method that calculates a molecular interaction field (MIF) within binding site volumes and detects pairwise MIF similarities between binding sites. MIFs are calculated using six chemical probes representing hydrophobic, aromatic, H-bond donor/acceptor, and positively/negatively charged interactions with a distance-dependent exponential function. Similarities are identified using an approximation of the Bron and Kerbosch[39] graph-matching algorithm and scored with a Tanimoto coefficient of the matched probes in the largest clique. The use of graph matching in IsoMIF makes it possible to detect MIF similarities without requiring any prior superimposition of the MIFs. To our knowledge, besides IsoMIF, FLAP is the only method that can detect MIF

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

similarities in the absence of prior superimposition. We validate the method using four data sets previously used to validate other methods and compare the performance of IsoMIF to a number of existing methods. We discuss how the performance of IsoMIF is affected by the definition of binding sites as well as molecular motions. We further validate IsoMIF with two larger nonredundant data sets derived from PDBbind[40] and sc-PDB[41] and show how IsoMIF can identify similarities across protein folds in the absence of sequence or structural similarities. The source code of IsoMIF can be freely obtained at http://bcb.med.usherbrooke.ca/isomif, where we also provide scripts to visualize MIFs and MIF similarities.

METHODS

Molecular Interaction Fields. MIFs encode the physicochemical environment of a protein. While MIFs can theoretically be calculated anywhere at the surface of proteins, the focus of this paper is on MIFs calculated in cavities relevant for small-molecule binding. A MIF can be seen as a unique signature of the ensemble of atoms of the protein onto a binding site cavity. Cavities are detected in a purely geometric fashion using GetCleft,[42] our *in house* implementation of the Surfnet algorithm.[43] Cavities identified with GetCleft are transformed into a regular grid of 0.50 Å within minimum and maximum distances to any non-hydrogen atom of the protein set to 2.5 and 4.0 Å, respectively (can be set by the user). Thus, cavities are identified without any consideration of a bound ligand but as will be shown in the validation, information relative to a bound ligand when present can, and for certain applications should, be used. These parameters are set to maximize the probability of considering spatial positions where noncovalently bound ligand atoms are likely to be found (Figure S1, Supporting Information) while minimizing the total size of the grid for computational efficiency. Protein PDB files are processed with Reduce[44] to add hydrogen atoms that are used as reference atoms to calculate angles for directional interactions (see below). During this step, OH, SH, NH_3^+, and methionine methyl groups as well aspargine, glutamine, and histidine side-chains are reoriented to optimize H-bonds and van der Waals overlaps.[44] The protonation state of histidine can be influenced by local electrostatics, which varies with the solvent pH, solvent accessibility, or bound ligands. Thus, the protonation state derived from the crystal might not represent the protonation state in other contexts. Thus, to account for all scenarios, IsoMIF considers both nitrogen atoms of the imidazole ring as H-bond donor or acceptors similar to cavBase.[22] For each vertex v, we calculate the interaction potential energy for a given probe p using all surrounding protein atoms as

$$\beta_{p,v} = \sum_{i=1}^{N} \varepsilon(p, T_i) e^{-d(i,v)/\alpha} \tag{1}$$

where p represents one of six chemical probes (hydrophobic, aromatic, H-bond donor/acceptor, positive/negative charge), v represents a grid vertex, $d(i,v)$ is the Cartesian distance in angstroms (Å) between the vertex v and a protein atom i (out of N atoms), and $\varepsilon(p,T_i)$ is a pairwise interaction matrix between the probe p and protein atoms i of type T_i. The length constant α is set to 1 Å for all probes (see Discussion). Protein atoms are classified into 13 atom types.[45] The original sybyl atom types were supplemented with the addition of the n.his atom type to represent the two nitrogen atoms of the imidazole ring of histidine that can be at times considered donors, acceptors, or charged. The interaction matrix ε is given in Table S1 of the Supporting Information. A threshold β_p^0 for each probe p (Table S2, Supporting Information) is used to determine if it is relevant to retain a given probe in a particular vertex given its interaction energy (eq 1). The final MIF at each position is given by

$$E_{p,v} = \delta(\beta_{p,v}, \beta_p^0)\delta(\theta, \theta_0) \tag{2}$$

where

$$\delta(\beta_{p,v}, \beta_p^0) = \begin{cases} 0 \text{ for } \beta_{p,v} > \beta_p^0 \\ \text{or} \\ 1 \text{ for } \beta_{p,v} \le \beta_p^0 \end{cases} \tag{3}$$

$$\delta(\theta, \theta_0) = \begin{cases} 0 \text{ for donor/acceptor interactions with } \theta > 60^\circ \\ \text{or} \\ 1 \text{ otherwise} \end{cases} \tag{4}$$

The angle θ is calculated between the atom-probe axis and a reference vector. These vectors are aligned with the axis drawn from the donor/acceptor atoms (O or N) and a reference atom, either the adjacent carbon or hydrogen. Vectors for donor/acceptor atoms of histidine and tryptophan rings are drawn in the axis of the exterior bisector of the two lines defined from the atom to two adjacent reference carbon atoms. Angles are calculated between aromatic probes and aromatic atoms. For this, two vectors are drawn in opposite directions perpendicular to the plane of the ring defined by two reference atoms. The vectors for each donor/acceptor atom and their reference atoms are depicted in Figure S2 of the Supporting Information. The list of reference atom(s) for each donor, acceptor, or aromatic atom is available in Table S3 of the Supporting Information. The vectors (colored arrows in Figure S2, Supporting Information) could represent the central axis of a cone extending outward with an aperture of 120° (60° on each side of the central axis), outside which probes are discarded during the calculation of the potential. The vectors are meant to define only approximately an optimal direction of hydrogen bond interactions. An angle threshold was tested for aromatic interactions to constrain probes in the face or in the edge regions relative to the ring's plane but was found to not significantly increase performance. Figure S3 of the Supporting Information shows the maximum probe−atom distance required to obtain the interaction energy of every probe threshold for an $\varepsilon(p,T_i)$ value of −1. Lower thresholds (more negative) require more or closer favorable atoms around a probe in order to reach the threshold and keep the probe at the vertex. For instance, because aromatic atoms come in groups in the rings of His, Trp, Tyr and Phe, the interaction energy of nearby aromatic probes will be lower than for other probes. Thus, the aromatic probe threshold is lowered. Charged amino acids contain two charged atoms (the two carboxyl oxygens of Asp and Glu and the two nitrogens of His) except for Lys and Arg that have one atom with a charged atom (N.4 and c.cat). On the other hand, amino acids that participate in H-bonds always have only one donor or acceptor atom. Because of this, the threshold for charged probes is lowered relative to the H-

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

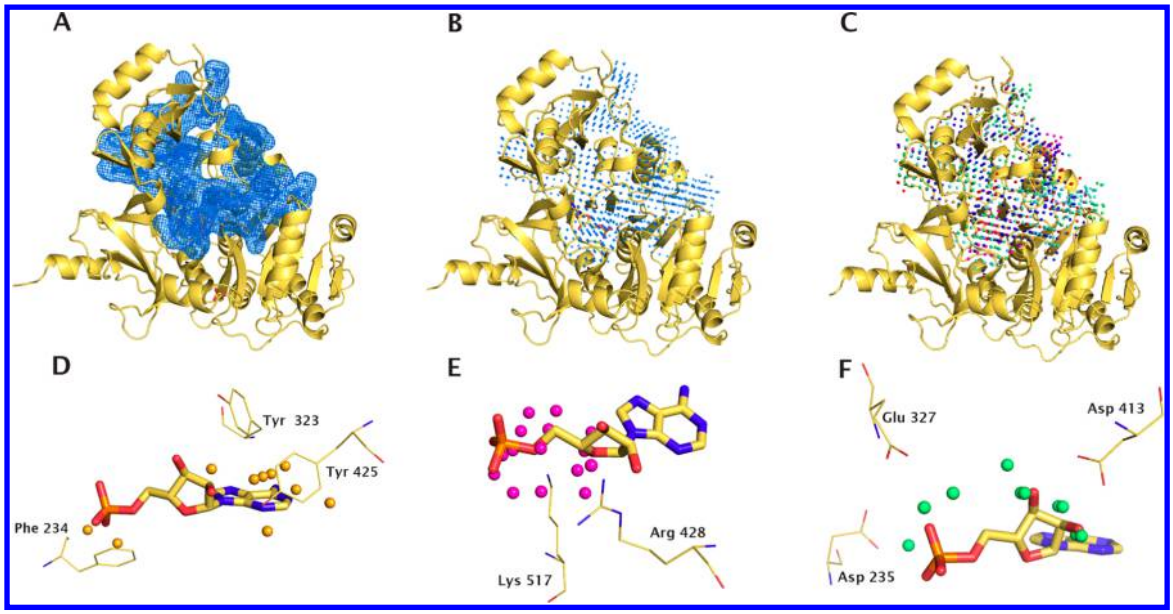


Figure 1. MIF representation for gramicidin synthetase 1 bound to AMP (PDB 1AMU). (A) The mesh encompasses the cavity identified by GetCleft. (B) The grid with 1.5 Å spacing built in the volume of the cavity. (C) The probes found present at each vertex represented in different colors. It is possible to find more than one probe at a given vertex. (D) Three aromatic residues lead to significant interaction energies for aromatic probes in the vicinity of the adenine ring of AMP. (E) Negatively charged probes with significant interaction energies near the PO_4 group of AMP caused by Lys 517 and Arg 428. (F) Positively charged probes with significant interaction energies found near the glucose moiety and near the PO_4 group. Positively charged amino acids in E and F also yield H-bond acceptor and donor probes, respectively (not shown in panels E and F).

bond threshold. It must be reminded that H-bond probes also have an additional constraint, the angle threshold, which further restricts their presence in the MIF. The hydrophobic probe has a relatively high threshold, but is the only probe with unfavorable atom types in the interaction matrix, set to keep hydrophobic probes away from polar atoms.

Equations 1−4 are applied for every probe in every grid vertex to produce the MIF for that cavity. Therefore, the MIF is the set of six-dimensional binary interaction vectors $\vec{E}_v = \{E_{1,v},...,E_{6,v}\}$ at all vertex positions v in the grid covering the volume of the binding site cavity. The MIF of individual cavities can be visualized using PyMol. Figure 1 shows a MIF for PDB structure 1AMU representing gramicidin synthetase 1 from *B. brevis* bound to AMP. The figure shows the mesh representation of the cavity found by GetCleft in blue, the constructed grid at 1.5 Å resolution in the cavity, and the MIF probes identified at each vertex of this grid. A close up of the probes identified near the bound AMP for negatively charged, positively charged, and aromatic probes, respectively, in magenta, green, and orange spheres and the corresponding amino acids responsible for the favorable interactions in these positions. Along with GRID and FLAP, IsoMIF differs from all other existing methods for the detection of binding site similarities in that it aims to detect similarities in MIFs within cavities.

Search and Scoring of the MIF Similarities. The level of MIF similarity between two cavities can be measured by finding the largest ensemble of vertices between two cavities that have corresponding interaction types and that are in geometrically equivalent positions. To find such ensemble of grid points, we use the Bron and Kerbosch (BK) graph-matching algorithm to detect maximum common subgraph (MCS) isomorphisms.[39] To detect the MCS, we create an association graph (Figure S4, Supporting Information). A node in the association graph is a pair of vertices, one from each of the two MIFs being compared that have at least one energetically significant common interaction; in other words, both have at least one of the six probes considered with significant interaction energies in both MIFs. An edge is drawn between two nodes in the association graph if the distances between the two corresponding vertices in each MIF are within 3.0 Å. This distance threshold allows accounting implicitly for geometric variability between similar binding sites that is the result of conformational flexibility.

The BK algorithm can be very time consuming, and we use two strategies to optimize the run-time. First, we make use of the observation by Bron and Kerbosch in their original paper[39] that the largest cliques tend to be identified first. This has the advantage of allowing to explore larger graphs and not drastically increasing run-time as the search can be stopped early while remaining confident that the largest clique is found. In IsoMIF, the search is stopped after the identification of the first 100 cliques (a parameter that can be defined by the user), and the best-scored clique is kept as a compromise between speed and accuracy. A 2-fold run-time increase is observed with doubling this threshold without significant improvements in the results, whereas a decrease to 50, while decreasing run-time, decreased the AUC of the steroid data set (0.77 to 0.75) as reported in Table S4 of the Supporting Information. Second, the grid spacing can be increased to decrease the number of vertices and run-time. IsoMIF can calculate MIFs with grid spacing of 2.0, 1.5, 1.0, or 0.5 Å. In the present work, we use a grid spacing of 1.5 Å (can be defined by the user) as this value yields an optimal speed/accuracy trade-off with respect to the large number of pairwise comparisons performed in the validation of the method and drastically reduces run-time compared to grids with smaller spacing. The probes distributed in the 1.5 Å grid spacing should not be interpreted as actual ligand atoms, rather it should be seen as a sampling of potential positions for intermolecular interactions in the binding site. The graph matching procedure can generate cliques that

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

represent two sets of corresponding vertices that are chiral. The MIF similarities of these cliques are not transposable in a biological context. Therefore, when identified during the Singular Value Decomposition (as described by Arun et al.[46]), such cliques are skipped. The MIF similarity score (MSS) of a clique is calculated as a Tanimoto score:

$$\mathrm{MSS} = \frac{N_c}{N_a + N_b - N_c} \tag{5}$$

where N_c is the sum of common probes in vertices belonging to the clique and represent the set of potential intermolecular interactions in equivalent geometric positions in the two MIFs. N_a and N_b represent the sum of energetically significant probes in each of the two MIFs under comparison.

Data Sets and Performance Evaluation. Unlike other active areas of research in structural computational biology such as small-molecule protein docking for the detection of protein binding site similarities, there is a lack of standard practices or benchmark data sets to evaluate methods. Here, we compare IsoMIF to a number of programs using the same data sets and the reported values for these methods using the Kahraman,[47] Homogenous,[48] Steroid,[36] and SOIPPA[16] data sets as well as two additional ones that we introduce derived from PDBbind[49] and sc-PDB.[41] The Kahraman data set contains 100 entries representing the structure of the biological unit bound to their cognate ligand and from different CATH[50] homologous superfamilies.[47] Removing entries of homologous superfamilies removes trivial similarities that result from divergent evolution. The homogeneous data set built by Hoffman et al.[48] also has 100 entries bound to 10 ligands. The authors did not filter for homologous superfamilies but rather ensured that bound ligands were of homogeneous size. The Steroid data set[36] contains 1853 entries containing eight true positives (pharmacologically relevant steroid-binding proteins), and remaining entries represent negative examples comprising ligands of equivalent size to the true positives. The SOIPPA data set[16] contains 319 proteins bound to adenine-containing ligands, among which 91 entries are thought not to bind adenine-containing ligands.

Furthermore, we introduce two larger data sets to evaluate the performance of IsoMIF derived from the PDBbind Refined[49] and sc-PDB data sets.[41] The 2014 release of the Refined PDBbind data set contains a total of 3446 structures. This data set comprises only crystal structures with overall resolution below or equal to 2.5 Å, with no covalently bound ligands and in a binary complex. All protein−ligand complexes have an experimentally measured K_d or K_i between 1pM and 10 mM. The sc-PDB data set contains 8077 structures and is a collection of proteins with druggable binding sites based on ligand molecular weight, buried surface area, volume of the cavities, and chemical structure of ligands. Unlike the PDBbind data set, complexes in the sc-PDB need not have experimentally measured binding affinities. For both data sets, the ligand Tanimoto coefficient of topological similarity between entries of both data sets was measured using Babel.[51] A Tanimoto coefficient of 1.0 was used to determine true positives. For the PDBbind and sc-PDB data sets, we define sequence identity thresholds of 15%, 20%, 25%, 30%, and 35% over at least 50 amino acids to remove redundancy.

An important scenario in which the detection of binding site similarities must be performed to predict the function of a protein is the case where there is no ligand bound within the cavity. In this case, the definition of the searchable area of the cavity can affect the results, most notably by increasing the search space. In order to estimate this effect, we utilize the bound ligand to define increasingly large grids encompassing increasing volumes around the bound ligands. Thus, for the Kahraman, Homogenous, Steroid, and SOIPPA data sets, the MIFs were generated for grid vertices within 3, 5, 10, and d_{max} Å to the bound ligands. The d_{max} distance is used to obtain the MIF calculated at all grid vertices contained in the cavity identified by GetCleft and is denoted as Original Model (OM). For the PDBbind and sc-PDB data set, the MIFs were calculated using the 3.0 Å threshold only. Most results, unless otherwise stated, are those for the 3.0 Å threshold as it produces comparable binding sites as those used in the validation of other methods. The comparison of results obtained with different methods is complicated by the distinct ways used by the different methods to define the searchable area of binding sites. When comparing IsoMIF to other methods, this is further complicated by the intrinsic difference in the nature of what is being searched, i.e., a volume in the case of IsoMIF and atoms/surfaces in the case of other methods (see Effect of Binding Site Definition). While MIFs are calculated at existing grid vertices (say at 3 Å from the ligand), all atoms of the protein may contribute to the MIF calculations according to eqs 1−4, but the closest atoms have the strongest influence. For computational efficiency, users may select probe-specific threshold distances above which to ignore the effect of protein atoms. All results presented here were calculated with a 3.5 Å threshold for all probes.

The performance of IsoMIF was evaluated using the Area Under the Curve (AUC) of Receiver-Operator (ROC) curves. For a given query, the entries are sorted by decreasing MSS. The ROC curve represents the fraction of true positives found as a function of the fraction of negatives found. If the area under the curve is 1.0, it means that all true positives are found before any negative, while an AUC of 0.5 means the true positives are randomly distributed and the method has no predicting power. The ability to distinguish true positives can also be measured with an enrichment factor. This measure tells how much top scored hits are over represented with true positives and is calculated for the top 10% as

$$\mathrm{EF}_{10} = \frac{\text{fraction of TP in top 10\%}}{\text{fraction of TP}} \tag{6}$$

Enrichment factors are more normally used in the evaluation of virtual screening, but this measure of prediction success is clearly complementary to AUC values. For example, for a set of 100 entries with 10 true positives, if four are within the top 10 ($\mathrm{EF}_{10} = 4$) and the other six are misclassified with less similarity than all true negatives, the AUC would be at best 0.4 (worst than random). Yet, despite the low AUC, having a number of true positives correctly classified among the most similar provides important clues about function. Furthermore, different binding sites that evolved independently to bind the same ligand may rely on distinct interactions to stabilize the interaction[9] and therefore be very different. Therefore, when validating a method for the detection of binding site similarities, there will always be true positive examples with very little similarity decreasing the AUC. Therefore, AUC and EF_{10} provide important and distinct information to judge if a method provides useful predictions.

Visualization of MIF Similarities. A singular value decomposition (SVD) allows one to find analytically[46] the

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

Table 1. AUCs for IsoMIF and Six Other Methods on Different Datasets

method	Kahraman[a]	Homogenous	Steroid	SOIPPA	mean AUC[b]
IsoMIF[c]	0.85 (4.4)	0.80 (4.0)	0.77 (3.6)	0.87 (1.3)	0.82 ± 0.04
eMatchSite[d]	0.69	0.92	0.66	0.94	0.80 ± 0.15
Sup-CK	0.90[e]	0.77	–	–	–
IsoCleft[d]	0.70[f]	–	–	–	–
Sup-CK[d]	0.66	0.81	–	–	–
PocketMatch[d]	0.52	0.74	0.53	0.60	0.60 ± 0.10
SiteEngine[d]	0.66	0.76	0.55	0.93	0.73 ± 0.16
PocketFeature[d]	–	–	–	0.85	–

[a]All AUC values exclude PO_4 entries unless otherwise stated. [b]Mean AUC was calculated for the methods with an AUC for the four data sets only. [c]This work. Enrichment factors at 10% (EF_{10}) are in parentheses. [d]Data obtained from the cited references as follows: eMatchSite,[36] Sup-CK,[36] IsoCleft,[9] PocketMatch,[36] SiteEngine,[36] and PocketFeature.[25] [e]Includes PO_4 entries; data taken from Hoffman et al.[48] [f]Subset of 72 nonhomologous entries across ligand classes and excludes PO_4 entries.

transformation matrix which best superimposes two MIFs using the detected MIF similarities. This transformation matrix is then applied to the protein structure and bound ligand coordinates. The superimposed protein structures and matched probes of each binding site can be visualized as a PyMol session using the scripts that are provided together with the code and executables of IsoMIF. The similarities for each probe type are represented by different colors: cyan (hydrophobic), orange (aromatic), blue (donor), red (acceptor), green (positive charge), and magenta (negative charge). They are represented as individual objects in a PyMOL session and can be visualized separately. The similarities for each probe are further decomposed: large spheres for the first protein and smaller spheres for the second protein. This allows a step-by-step inspection of the MIFs being compared and the common MIF between the two cavities, as well as to better understand what are the corresponding amino acids in each protein for the different types of interactions. All superimpositions of the structures and their bound ligands are performed this way.

RESULTS

Validation with the Kahraman Data Set. We compare here the performance of IsoMIF to that of other methods as reported in the literature. The AUCs of different methods on the data sets are summarized in Table 1. The Kahraman data set contains a collection of 100 proteins not evolutionarily related, bound to 10 cognate ligands, and with different CATH homologous superfamilies.[47] This data set was originally devised to study relationships between the shapes of binding pocket and ligand, and thus, ligands were chosen to be variable in size and flexibility. Binding site volume alone can be used to classify entries in this data set with high accuracy.[9,47,48] Nonetheless, the Kahraman data set is still commonly used to evaluate the performance of binding site similarity methods.[9,36,48] A detail often unmentioned when assessing the performance of a method on the Kahraman data set is the use of different subsets. Twenty percent of the original Kahraman data set (20 entries) are structures bound to PO_4. For the evaluation of Sup-CK,[48] those entries were kept but excluded for eMatchSite and the benchmark of other methods therein.[36] IsoCleft used a subset of 72 entries to further address redundancy of homologous superfamilies across ligand classes in addition to removing PO_4 entries.[9] We evaluated the performance of IsoMIF with and without PO_4 entries using the nonhomologous subset of Kahraman. When excluding the PO_4 entries, IsoMIF performs better than any other method with an AUC of 0.85, followed by eMatchSite (AUC 0.69). Including PO_4 entries in the benchmark may be questioned as the small size of the molecule restricts the possibilities for the evolution of conserved specific interactions that would be detected using binding site similarity methods. Most methods that do benefit from the introduction of PO_4 entries in this data set are likely benefiting as a result of the size differences between PO_4 binding sites and those of other ligand classes in the data set. For example, Sup-Ck obtains an AUC of 0.90 and 0.66 in the presence or absence of PO_4 entries, respectively. The inclusion of PO_4 entries for IsoMIF decreases its performance from an AUC of 0.85 to 0.79.

Dissecting the AUC into different ligand classes (Figure 2) indicates that IsoMIF more easily classified the binding sites of

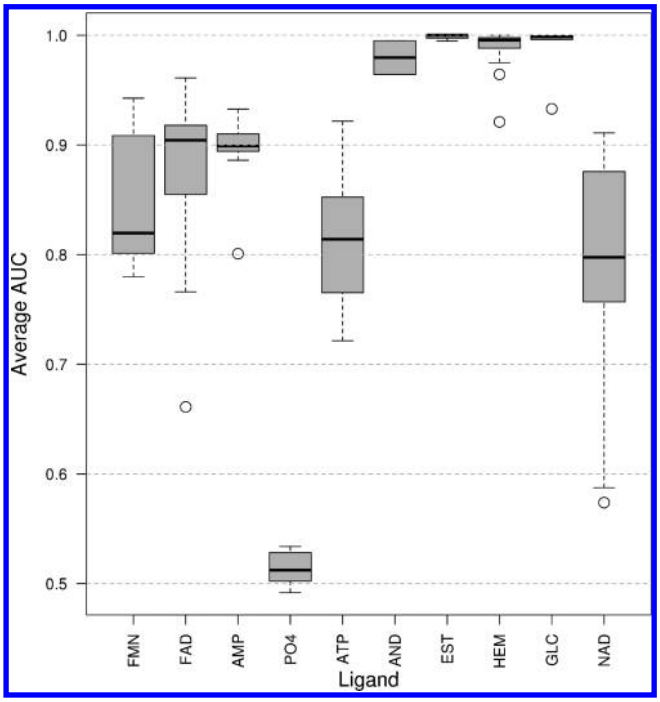


Figure 2. Boxplot showing the distribution of the AUCs obtained with IsoMIF for different ligand classes in the complete Kahraman data set.

dehydroepiandrosterone (AND), estradiol (EST), Heme (HEM), and Glucose (GLC). One notable property that distinguishes these ligands from the others in the data set is that these ligands are in general more rigid, but it remains to be tested if this ligand property is at the source of the differences in classifying distinct ligands. Excluding PO_4, the most difficult

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

ligand binding sites to predict are those of ATP, FMN, and NAD. As these ligands share a common substructure, we set out to determine if combining these three ligand classes into one meta-class could further increase AUCs. Interestingly, whereas this helps eMatchSite increasing its AUC from 0.69 (Table 1) to 0.79,[36] in the case of IsoMIF, it leads to a decrease to 0.81. Lastly, the AUC obtained with IsoMIF using the subset of 72 entries used to evaluate IsoCleft was calculated in order to permit a direct comparison between the performances of these two methods in the same data set. Interestingly, the obtained AUC of 0.85 was not affected by the removal of the 8 cases of proteins belonging to homologous families across ligand classes.

Validation with Homogenous, Steroid, and SOIPPA Data Sets. Hoffmann et al.[48] created the homogeneous data set to validate the Sup-CK program. The authors wish to address the size bias in the Kahraman data set by including only entries with bound ligands of approximately equal number of atoms. Among all methods, eMatchSite performs best on this data set (AUC 0.92), followed by IsoMIF (AUC 0.80) and sup-CK (AUC 0.77). Unfortunately, the data set contains two ligand classes representing molecules that are not cognate ligands. Specifically, the data set includes structures bound to pentaethylene glycol (three-letter PDB code 1PE), a precipitant, and to octyle glucoside (three-letter PDB code BOG), a detergent used in purifications. The concern relative to the inclusion of such molecules is the same as that of including PO_4 entries in the Kahraman data set. This is supported by the average AUCs for each ligand class shown in Figure 3. The boxplot of the average AUCs for each ligand class shows the averages of 1PE (AUC 0.72) and BOG (AUC 0.62) to be the lowest.

The steroid data set was built by Brylinski et al.[36] and contains eight steroid-binding proteins and 1854 other proteins bound to ligands of similar size (15−25 heavy atoms), all with a Tanimoto coefficient of topological similarity smaller than or equal to 0.1 with estradiol. One dimer PDB structure (2PU4) is present twice in the original data set, and here, only chain B was considered. IsoMIF performs best by far with an AUC of 0.77 followed by eMatchSite (AUC 0.66) and much better than siteEngine (AUC 0.55) or pocketMatch (AUC 0.53). As this data set contains only one ligand class, we do not dissect the performance of individual ligands.

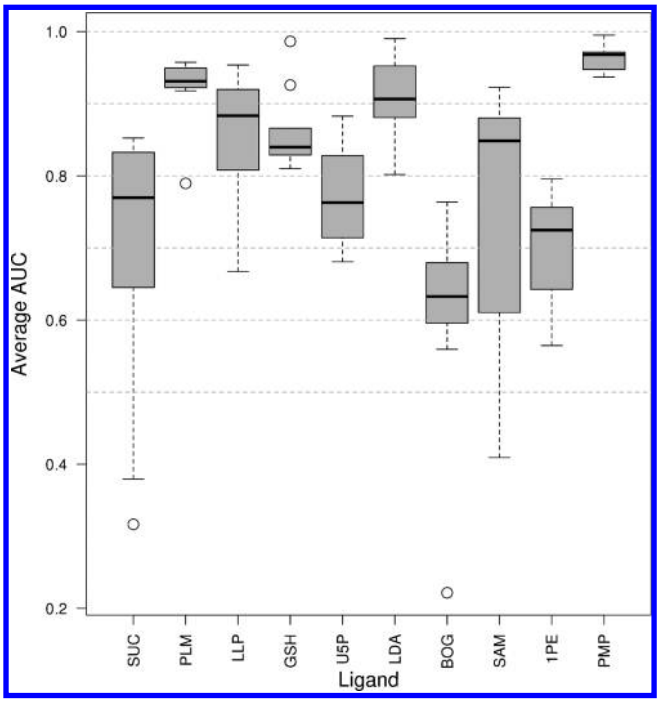


Figure 3. Boxplot showing the distribution of the AUCs obtained with IsoMIF for different ligand classes in the Homogeneous data set.

The SOIPPA data set was created by Xie et al.[16] to validate the SOIPPA method. The original data set contains 247 proteins bound to adenine containing ligands (ATP, ADP, NAD, FAD, SAM, and SAH) and 101 control structures bound to a ligand not containing either adenine, flavin, nicotinamide, or ribose. We utilize a subset of the SOIPPA data set as defined by Brylinski et al.[36] containing 228 true positives (bond to adenine containing ligands) and 91 true negatives. There are three entries (1HNN, 1V1A, and 1VQW) that are each originally represented with two chains and are here present only with one chain giving a total of 225 entries). Three structures (1HNN, 1V1A, 1VQW) were represented with two chains in the original data set and one chain was kept for each. We use this subset of SOIPPA in order to compare the results obtained with IsoMIF to those reported by Brylinski et al. for eMatchSite, Sup-CK, PocketMatch, and SiteEngine. The original use of this data set is to consider all the adenine-containing ligands as positives in a single class. The highest AUCs are obtained by eMatchSite and siteEngine with values of 0.94 and 0.93, respectively, followed by IsoMIF (AUC 0.87), pocketFeature (AUC 0.85), and pocketMatch (AUC 0.60).

Alternatively, one can treat each ligand independently as in the case of the Kahraman or Homogeneous data sets; in other words, only examples bound to the same ligand are defined as true positives and the rest are considered true negatives. In this case, we can either include or exclude the set of 91 examples originally defined as negatives. In the first case with the restricted set of negatives, IsoMIF obtains an average AUC of 0.79, whereas in the case of the extended negative set, the AUC is 0.84. In either cases, it is possible to dissect the performance for individual ligands as shown in Figure 4 using the restricted (Panel A) and extended (Panel B) negative sets. As expected, the inclusion of negatives that are so different from the positives contributes to augment the AUC. This is of course an effect that also contributes to increasing the AUC of different methods in other data sets where the ligands are diverse. Interestingly, comparing the average AUC of 0.84 when using the extended negative set with the original AUC of 0.87 when considering as true positives all adenine-containing ligands, suggests that most of the predictive power of IsoMIF within this data set comes from correctly predicting ligands in their own class. This is particularly interesting as the SOIPPA data set is nonredundant within a 30% sequence identity threshold.

Mean AUC Values. The above results show that the various methods perform differently in distinct data sets as a result of differences in the composition and design principles employed into building each data set. No single benchmark data set exists, and it is unlikely that a benchmark data set can be defined to account for all possible present and future requirements of data sets to validate methods for the prediction of binding site similarities given the distinct and ever growing applications of such methods. As a measure of the overall accuracy of the different methods, we opt to calculate the mean AUC for different methods across data sets (Table 1). We calculate such means only for methods for which we have the AUC across the four different data set: IsoMIF, eMatchSite, PocketMatch, and

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

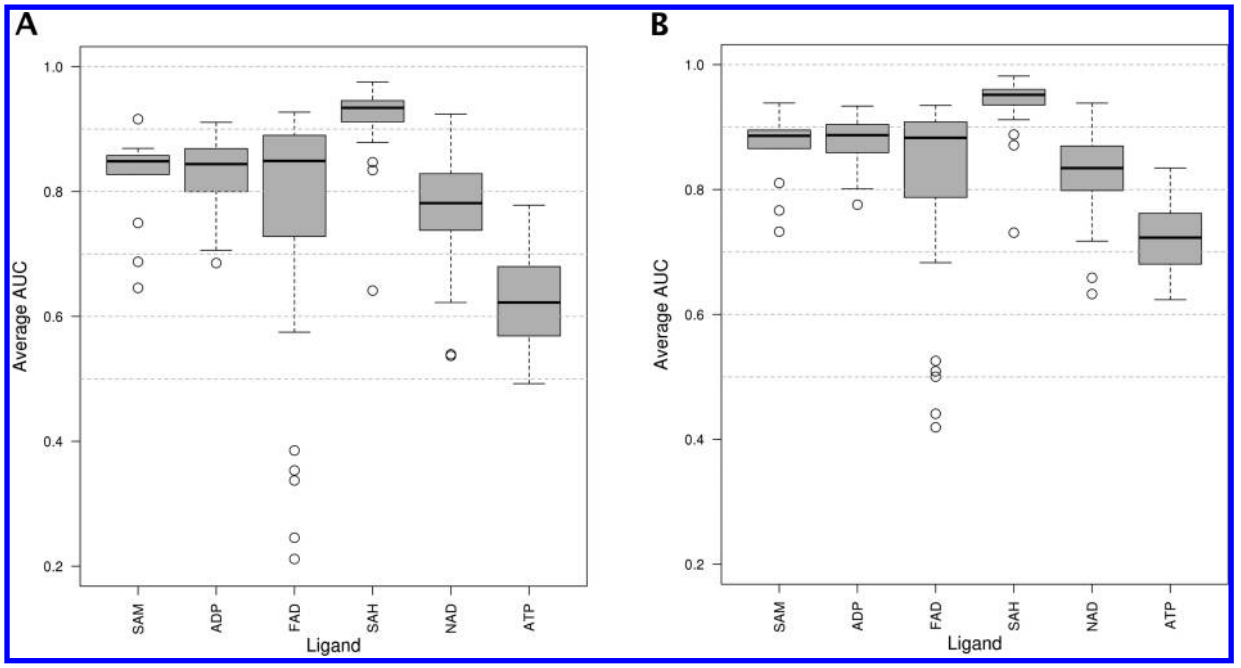


Figure 4. Boxplot showing the distribution of the AUCs obtained with IsoMIF for different ligand classes in the SOIPPA data set using the restricted (panel A) and extended (panel B) negative sets.

SiteEngine. IsoMIF is the best method based on mean AUC closely followed by eMatchSite with mean AUC 0.80. Additionally, IsoMIF performs consistently well across data sets as noted by the small standard deviation of 0.04 compared to 0.15 for eMatchSite. The other two methods, PocketMatch and SiteEngine, have considerably lower mean AUC values and large standard deviations.

Enrichment Factors. As discussed in the Methods section, a low AUC value does not necessarily mean that top scoring solutions are incorrect. When the goal of using a method for the detection of binding site similarity is the prediction of function, it is important to analyze a number of top hits. This situation is not unlike high throughput docking simulations in virtual screening. In both situations, an enrichment of top solutions with true positives is favorable. We calculated enrichment factors at 10% (EF_{10}) for IsoMIF for the four data sets. Values of EF_{10} report the level of enrichment of true positives in the first 10% most similar targets to the query. A value of 1 means the enrichment of true positives in the top 10% is no different from what is expected from a random classifier. We obtain values of 4.4, 4.0, and 3.6 for the Kahraman, Homogenous, and Steroid data sets, respectively, which follows the same trend as the calculated AUC values. The SOIPPA data set would appear to have an uncharacteristically low EF_{10} value of 1.3 despite having the highest AUC. As a matter of fact, with the SOIPPA data set having 228 true positives out of a total of 319 entries, the fraction of true positives in the entire data set is 0.72 (228/319). Even if all top 10% results were true positives, the maximum possible EF_{10} value that may be obtained is therefore 1.4 (1/0.72) according to eq 6. Thus, a value of 1.3 shows that 92% of the top 10% predictions obtained with the SOIPPA data set are correct. If we utilize the alternative definition of true positives in which each ligand is predicted independently, we obtain EF_{10} values of 3.17 and 3.91, respectively, when excluding the nonadenine binding proteins from the negative set, what was called above the restricted negative set or including these (extended negative set), respectively.

Extended Validation with the PDBbind and sc-PDB Data Sets. To further validate IsoMIF, we introduce two data sets derived from the PDBbind[49] 2014 release and sc-PDB.[41] For PDBbind, only the 3446 entries in the refined set were considered. Within those 3446 entries, 1415 (41%) had at least one true positive entry within the set, that is, at least one other example bound to the same ligand. In the sc-PDB data set, 3809 examples (47%) had at least one true positive. In what follows, we utilize only the subsets of 1415 and 3809 entries for the PDBbind and sc-PDB data sets, respectively. The list of entries can be downloaded from our website (http://bcb.med.usherbroke.ca/isomif). For each entry in each data set, the objective is to classify the true positive examples for that ligand (i.e., other entries bound to the same ligand) with the highest AUC. The mean AUC and EF_{10} for all ligands in this case are 0.93 and 8.08 for PDBbind and 0.87 and 6.40 for sc-PDB (Table S5, Supporting Information).

While it may at first appear that IsoMIF performs better in these data sets than in previous ones, considering sequence similarity as a classification measure alone produces AUC values (denoted as AUC_{seq} in Table S5, Supporting Information) of 0.81 and 0.68 for PDBbind and sc-PDB, respectively. Thus, we created different subsets of the original PDBbind and sc-PDB data sets above with decreasing thresholds of permitted pairwise sequence identity to remove trivial cases. Proteins with sequence identity above 35% are still likely to share structural similarities. A gray zone was shown to exist between 20% to 35% sequence identity where a certain number of proteins still share structural similarities.[4] We use the 15% sequence identity threshold to minimize the chances of structural similarity. Interestingly, at this threshold, there are still 414 and 2292 entries remaining for PDBbind and sc-PDB, respectively. As expected, the AUC decreases with a decrease in the threshold (Table S5, Supporting Information). At the lowest level of pairwise sequence identity threshold of 15%, we

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

obtain an AUC of 0.79 for both data sets and EF_{10} values of 4.97 and 4.46, respectively. At the same time, the AUC_{seq} values decrease as expected down to nearly 0.5 and EF_{10} of 1.0, comparable to a random predictor for a the sequence identity threshold of 15% in both data sets.

Effect of Binding Site Definition. Within the context of validating a method, the binding site definition relates to the extent with which information about a bound ligand is used to constrain the search space. Most methods were validated using a binding site definition that includes protein atoms within a threshold distance of around 5 Å from any ligand atom, and some methods cannot be applied to situations in which a bound ligand is not present (for a discussion, see ref 33). With such a definition, for the most part, only protein atoms that are directly interacting with the bound ligand are included. The results presented thus far for the validation of IsoMIF use grid vertices within 3.0 Å of the ligand in order to utilize a binding site definition equivalent to that used in the validation of other methods. With a 3.0 Å threshold, only grid points around a ligand and between a ligand and the residues directly interacting with the ligand are included. This threshold is sufficient to encompass a volume that represents what most users would define as the binding site by visual inspection as can be observed in the examples presented. For any given threshold distance from the ligand, the size of the search space for IsoMIF also depends on the choice of grid spacing. In this section, our goal is not to understand how computational performance is affected under conditions of increasing search spaces, but how decreasing information on the bound ligand affects the prediction of similarities. Therefore, we analyze the performance of IsoMIF using increasingly wider binding site definitions for the query binding sites in the Kahraman, Homogenous, Steroid, and SOIPPA data sets with a single grid spacing. As expected, there is a decrease in predictive power as the imprint of the bound ligand onto the volume used for the search decreases. Specifically, the predicting power (average AUC) decreases slightly between 3.0 and 5.0 Å thresholds and more steeply between 5.0 and 10.0 Å, where the average AUC is just over 0.65 for most data sets (Figure 5). OM stands for Original Model and designates the entire cavity predicted by GetCleft, solely based on geometric criteria and without any information from bound ligands used to define the grid. In this case, the average AUC for all data sets except SOIPPA further decreases to just above 0.60. Interestingly, the average AUC is the highest for the SOIPPA data set among all data sets up to 5.0 Å at which point there is a sharp loss in predictive power as the threshold continues to increase. It is important to stress that the OM model over represents the size of binding sites immensely (Figure 1A) and is only used here due to the convenience of calculating it automatically for the large number of queries across data sets. Even when a ligand is not bound, visual inspection of the cavities defined by GetCleft allows one to considerably trim their size. The NRGsuite[42] contains the mechanisms necessary to define cavities with GetCleft and trim their size manually in a convenient manner.

The effect of increasing the noise in the definition of binding site has not been reported for other methods with a few exceptions. IsoCleft uses a two-stage C_α/all-atom graph-matching algorithm and has an AUC of 0.70 at 5 Å that drops to AUC 0.64 for the OM model (defined with GetCleft) on a subset of 72 entries of the Kahraman data set. The SOIPPA algorithm does not require a predefined binding site and can easily compare whole protein structures. This is possible due to the C_α representation, which decreases the size of the search space. To compensate for the simplicity of the model, evolutionary information is added by assigning an amino acid conservation score. This allows an interesting speed/accuracy compromise. Shulman-Peleg et al.[23] compared binding sites defined with 4.0 Å around the ligand but also tested the comparison of binding sites against whole proteins and whole proteins against whole proteins in the validation of siteEngine. The protein representation in siteEngine consists of replacing amino acids with functional pseudocenters and the search uses a geometric hashing algorithm and a scoring that incorporates surface curvature. One concern about redundancy must be raised for the data set used by Shulman-Peleg et al. and in general for the data sets used to validate binding site similarity methods. Sequence identity thresholds of 30−40% are often the norm, and this threshold does not properly remove structurally redundant examples as shown above.

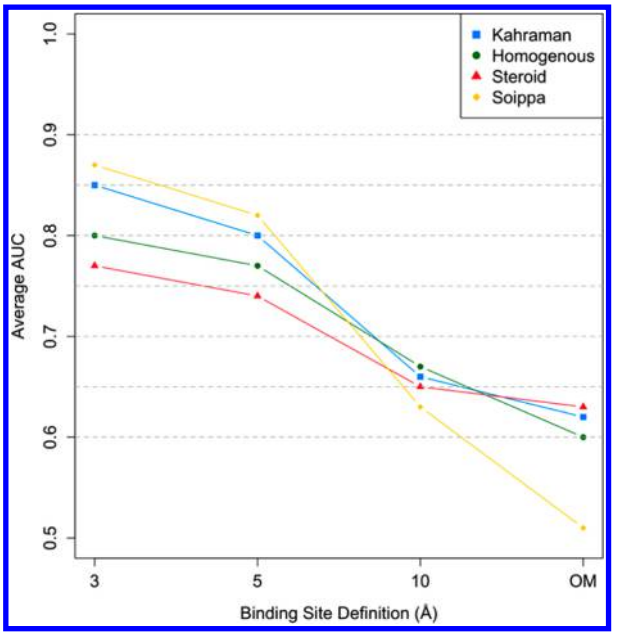


Figure 5. Performance of IsoMIF for Kahraman, Homogenous, Steroid, and SOIPPA data set at different binding site knowledge. The binding site knowledge refers to a distance in Å to the nearest ligand atom where the grid is built and the MIF calculated.

Adding additional information such as curvature or evolutionary conservation may help zoom in on areas of the cavity more likely to bind a ligand and effectively decrease the search space. Thus, while some methods integrate this information directly, this type of information could be used in IsoMIF as a preprocessing step during the definition of the binding site search area. One must be careful to not place too much weight to such measures and bias the definition of binding sites as this could lead to a decrease in performance. In the case of geometric measures such as curvature, this is due to the potential effect of flexibility. In the case of conservation, the patterns of conservation in cases of convergent evolution may vary widely (see Discussion section).

Effect of Structural Variability. Protein structures can exhibit a continuum of movements from small subrotameric and rotameric side-chain movements and small backbone and loop rearrangements to the movements of entire domains. For example, side-chain rotamer changes upon ligand binding occur in 90% of cases.[52] Flexibility is often related to function and

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

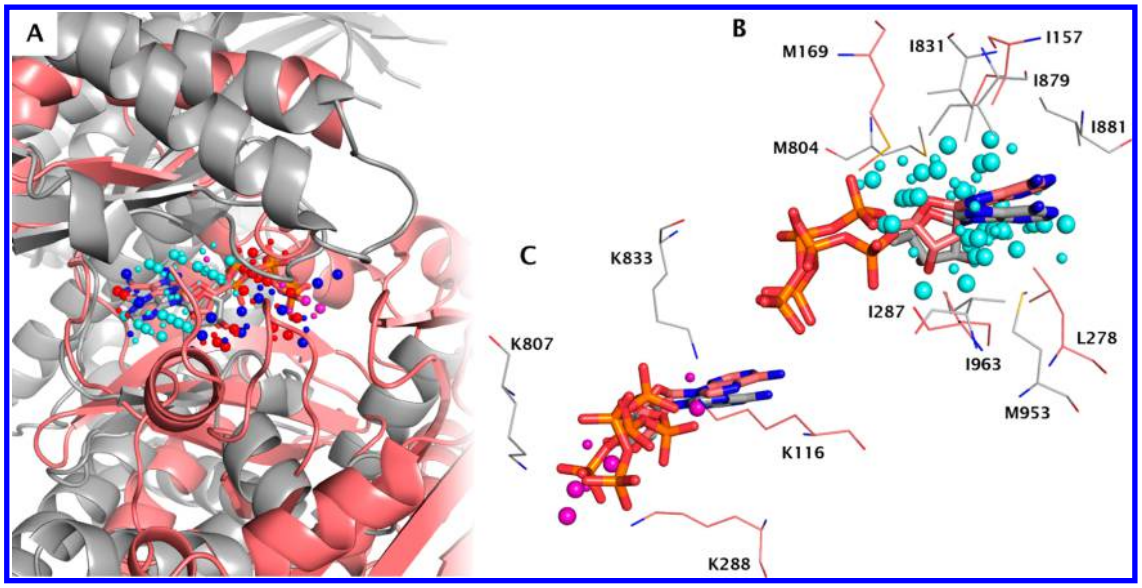


Figure 6. MIF similarities found between 1E8X (gray) and 1DV2 (light red). (A) MIF similarities (colored spheres) are used to perform the superimposition of the structures and bound ATP. (B) The adenine region is surrounded by hydrophobic similarities caused by the presence of methionine, isoleucine, and leucine residues in both binding sites. (C) Zoom on the bound ATP molecules (RMSD of 1.87 Å) showing two lysine pairs (K807−K288 and K833−K116) with terminal amino groups in the vicinity of phosphate groups yielding similar negatively charged probes (magenta spheres).

stability.[53] IsoMIF implicitly accounts for structural variability with the 3.0 Å user-defined distance threshold employed in the creation of edges in the association graph (Figure S4, Supporting Information). However, it is interesting to measure to what extent structural variability affects the detection of similarities.

Normal mode analysis (NMA) is particularly well suited to perform large-scale dynamic analyses as it is a computationally inexpensive technique. Our group recently introduced ENCoM, a coarse-grained NMA analysis method to account for the nature and magnitude of atomic interactions.[54] Among other applications, ENCoM is capable to model loop and domain movements upon ligand-binding with better accuracy than other NMA methods.[54] We utilized the ENCoM Server[55] to generate conformational ensembles with distortions in the original query binding sites in the Kahraman data set excluding PO_4. For each entry, we produced approximately 100 conformations with binding site (atoms within 6.0 Å of the ligand) RMSD values of up to 6.0 Å and minimum global RMSD of 0.5 Å between models. In what follows, we refer to the original PDB structures of the Kahraman data set as unperturbed and each model as a perturbed structure. Each perturbed structure is used as a query against the Kahraman data set to classify the remaining 79 unperturbed entries with respect to their MIF similarities to the perturbed model. Thus, for each model, we can calculate an AUC that can be compared to that of the parent unperturbed structure. In Table S6 of the Supporting Information, we show the mean AUC for perturbed models within each ligand class as a function of the range of binding site RMSD with respect to unperturbed structures for cases where the AUC of the unperturbed structure is equal or higher than 0.85. The reason to restrict the analysis to cases where the unperturbed query can be classified successfully is to study the effect of flexibility on cases where other confounding factors (see Discussion section) are less prominent. For relatively small ligands such as glucose (GLC), even the slightest perturbation of the binding site (RMSD up to 1.0 Å) leads to a drastic decrease in mean AUC. Interestingly, as the range of variation increases, the AUC steeply increases. We believe that the initial drop is an artifact related to relatively small binding site perturbations actually causing drastic changes in the volume of the cavity detected by GetCleft. For larger ligands, this effect seems to be attenuated. Medium-sized ligands seem to also be affected in this way but less so than GLC. Large ligands such as HEM or FAD do not seem to be affected at all by this and have a mean AUC similar to that in the absence of perturbations. In several cases (FMN, FAD, HEM, NAD), there is a small but steady pattern of decrease in average AUC. In other cases (ATP and AMP), the mean AUC seems to oscillate, whereas in the cases of EST and GLC the mean AUC increases with the increase in binding site RMSD. It is possible that increasing binding site structural diversity in the cases of EST and GLC permits a better sampling of conformational space translating into increases in AUC.

Lastly, we performed a related experiment in which instead of using as query a single binding site model (either unperturbed or perturbed) we utilize the entire ensemble. That is to say, for a given target binding site, we select the highest similarity value that can be obtained with any of the structures in the query ensemble (unperturbed or perturbed). As with the analysis above shown in Table S6 of the Supporting Information, we stratify the results according to the range of binding site RMSD variability in the models used to find the highest similarity. Overall, it seems to be slightly beneficial to consider a query ensemble but not for all binding site RMSD ranges (Table S7, Supporting Information).

The results in Tables S6 and S7 of the Supporting Information suggest that the consideration of structural variability in the query binding sites also increases the similarities between the query and negative examples.

Examples and Applications. The results above show that IsoMIF has superior performance on average compared to other methods. In this section, we present a few examples that highlight some of the reasons why this could be so.

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

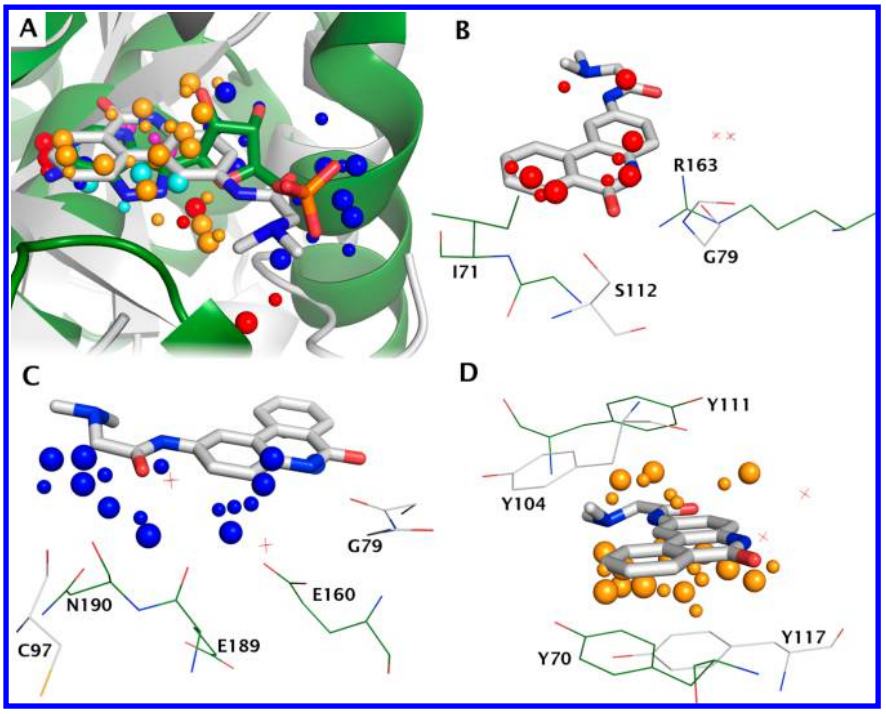


Figure 7. MIF similarities between poly [ADP-ribose] polymerase 15 (PDB 3GEY, gray) bound to inhibitor Pj34 and alpha-momorcharin (PDB 1AHB, green) bound to formycin-5′-monophosphate, respectively (A). A number of acceptor (B), donor (C), and aromatic (D) probes are responsible for the MIF similarities between these two proteins.

Figure 6 shows the result of the comparison of phosphoinositide 3-kinase (1E8X) with biotin carboxylase (1DV2), two proteins bound to ATP with different CATH topologies, with at most 23% sequence identity over a 74 residue overlap. The alignment of the structures with the transformation matrix that superimposes the MIF similarities yields an RMSD of 1.87 Å between the ATP molecules. There are abundant hydrophobic probes (cyan spheres) near the adenine ring with some donor and acceptor probes (blue and red, respectively) in the vicinity of nitrogen atoms of the purine indicating potential H-bonds.

The two structures display two lysine pairs (K807−K288 and K833−K116) that potentially contribute in stabilizing phosphate groups of the ATP. While the distances between the terminal amino groups of pairs K807−K288 and K833−K116 are 5.4 and 1.1 Å, respectively, the molecular interaction field projected into the binding site is found at corresponding positions and represented by negatively charged probes (magenta spheres). This observation highlights an advantage of detecting MIF similarities in the volume of the cavity in locations of potential ligand atoms rather than the position of atoms. Using the latter, the lysine pair K807−K288 could be missed by similarity algorithms that have stringent geometrical constraints or that rely on the position of C_α atoms. Thus, although a number of methods that are less sensitive to geometric constraints could potentially capture this similarity, it serves to show that IsoMIF can detect binding site similarities that are the result of distinct arrangements of amino acid residues within the binding site.

Drug repurposing is an exciting new way to exploit existing drug for new targets potentially decreasing the cost and time involved in drug development.[56] The protein Cryptogein (PDB 1BXM) from the plant pathogen *P. cryptogea* is a member of the elicitins, a highly conserved protein family responsible for tissue necrosis in host plants. The binding site is characterized by a long hydrophobic tunnel with potential H-bonds concentrated at one extremity of the pocket. Interestingly, all top hits found for 1BXM exhibit this characteristic sterol binding site despite the fact that these are embedded within proteins of different folds (Table S8, Supporting Information). Their functions vary from intracellular transport and transcriptional regulation to signaling pathways. DAF-12 (3GYU, rank 9) mediates steroid hormone−nuclear receptor signaling involved in the progression of larvae cycle associated with strongyloidiasis, a human parasitic disease. Dafachronic acids (DAs) are a potential new class of drugs for this disease,[57] and our results suggest that DAs could be potentially inhibitors of Cryptogein as well as other proteins in Table S8 of the Supporting Information. Figure S5 shows the similarities between 1BXM and four of the top hits found (1ZHZ, 3A51, 3GYU, and 1N83). While they all share the large hydrophobic pocket (cyan spheres) and a polar extremity (blue and red), DAF-12 (3GYU) seems to share more H-bond donor (blue) and less aromatic hotspots with 1BXM than the other top hits. This type of information can be used in a number of ways in drug design by focusing on regions of detected similarities or the remaining dissimilar regions. In a polypharmacological drug design context, similarities shared only between hits can be used to define spatial regions (and the corresponding binding site residues) in which to design favorable interactions to maximize the chance that the developed drug binds to the desired proteins. Likewise, regions without detected similarities between otherwise very similar proteins (either from the same family or across families) can be targeted to prevent binding to potential cross-reactivity targets.

Lastly, another application of methods for the detection of binding site similarities is to detect bioisosteric replacements[58] in which different ligands bound to very similar binding sites may help suggest modification to the ligand. The human poly [ADP-ribose] polymerase 15 bound to an inhibitor Pj34 (PDB

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

code 3GEY) from PDBbind was used as a query against 2343 nonredundant entries (sequence identity threshold 15%) among which one other protein (the Cholix toxin from *V. cholerae* PDB code 2Q6M), the true positive, was found as the third top ranked solution (AUC = 1.0, EF_{10} = 10). The second hit is eukaryotic α-momorcharin (PDB 1AHB), a ribosome-inactivating protein bound to formycin-5′-monophosphate (PDB three-letter code FMP). Figure 7A shows the superimposition of the 3GEY and 1AHB based on the MIF similarities found for five probes. Focusing on the H-bond acceptor probes (Figure 7B) shows that the H-bond made by 3GEY with the backbone amide of glycine 79 could be made by arginine 163 of 1AHB. While it does not seem to be used in binding Pj34, the donor hydroxyl of serine 112 seems to have a counterpart with the backbone amide of isoleucine 71. Similarly, there are many H-bond donor probes found, and Figure 7C shows the acceptors atoms in each structure that are responsible for these similarities. The ligand seems to bind 3GEY through hydrophobic (cyan spheres in 7A) and aromatic interactions. Shared aromatic interactions are the result of favorable interaction energies between aromatic probes and tyrosine 117 or tyrosine 70, respectively, that can make face-to-face aromatic interactions with the inhibitor (Figure 7D). With this example, we wish to illustrate two applications of IsoMIF. First, as with the previous example, it is possible that Pj34 binds α-momorcharin and therefore could be repurposed to inhibit it. Second, differences between the two ligands may be exploited to introduce bioisosteric modifications.

DISCUSSION

The identification of binding site similarities is not trivial. This is reflected in the number and variety of existing methods. No method obtains an AUC of 1.0 meaning that there are always query binding sites for which it is possible to find negative examples with higher similarity to the query than positive examples. A number of factors contribute to this: the design of methods, ways in which they are evaluated, and limitations observed in nature as a consequence of evolution.

The idea of calculating MIFs has been previously introduced in chemoinformatics in Comparative Molecular Field Analysis (CoMFA) methods. In particular, protein binding site MIFs were used in X-SITE[59] to predict favorable ligand interaction regions within binding sites and later in the AFMoC method[60] to predict binding affinities. These methods are limited by the availability of experimental data but also by the requirement of a reliable ligand or protein superimposition. This limits the comparison of proteins from different families or ligands with different scaffolds as they cannot be reliably superimposed. This limitation prevents the use of MIF-based methods for applications outlined in this paper. The graph matching procedure used by IsoMIF allows the comparison of cavities without requiring an initial superimposition, thus removing the limitation and allowing the comparison of unrelated proteins. In addition to IsoMIF, the previously described FLAP (Fingerprints for Ligands And Proteins) algorithm[28,29] introduced and applied in BioGPS[30] also permits the detection of MIF similarities without requiring prior superimposition. MIF-based methods have the advantage of having as output a set of common interactions that can be transposed into ligand features, i.e., pharmacophores, which can then be used as input in another algorithms. FLAP differs from IsoMIF in a number of points as discussed in the paragraphs that follow.

A fundamental difference between FLAP and IsoMIF is the extent to which binding-site flexibility is accounted for and the impact that this has on the size of the search space. In FLAP, all combinations of four pharmacophore probes are compared with those of a second cavity to find matches with distance differences below 1.0 Å. Clearly, increasing this threshold augments the search space. This low threshold can bias the search and prevent the selection of better matches with higher distance differences. In IsoMIF, a higher threshold value of 3.0 Å is used to implicitly account for geometric variability due to molecular movements. The performance of IsoMIF was evaluated with thresholds of 1−5 Å (Table S9, Supporting Information). While 4 Å gives better AUCs for the Kahraman, Homogenous, and Steroid data sets, the 3 Å threshold was chosen for the results in this manuscript due to the faster runtime. Beyond 4 Å, the performance seems to plateau, suggesting most effects of flexibility are captured at this threshold and beyond that nonrelevant similarities are identified.

Another difference between FLAP and IsoMIF lies in the force fields used. FLAP uses the GRID force field to calculate MIFs using its own set of probes and potential energy function whereas IsoMIF uses a decaying exponential function. GRID accounts for electrostatic, hydrogen bonds, Lennard−Jones, and entropic interactions in its potential energy function. In IsoMIF, we use a coarse-grained representation of the intermolecular interaction energy by using a single exponential term with a dimensionality (α parameter) set to 1 for all interaction types. The α parameter could be set individually for charged, H-bond, aromatic, and hydrophobic interactions, differently affecting the potential of each probe and on which a fixed threshold for all probes can be used. Instead, individually, each probe thresholds (Table S2, Supporting Information) are adjusted, and it is the α values that are fixed to 1. It would be naïve to believe that the potential used in IsoMIF describes intermolecular forces as accurately as more complex force fields do. In principle, IsoMIF could use a more complex force field with coulomb, Lennard−Jones, and other parameters assigned to each atom type using any reference force field (e.g., GRID, Amber,[61] CHARMM[62]), and intermolecular potential energies could be calculated with their appropriate energy functions. However, as the results demonstrate, even the simplistic representation of molecular interactions utilized in the IsoMIF force field is sufficient to identify biologically relevant similarities with an accuracy comparable or higher than that of existing methods. The consistency of IsoMIF as well as its performance across data sets may be due to the more realistic representation that focuses on interactions aggregating the effect of atomic positions rather than the positions themselves. Additionally, the Bron and Kerbosch approximation (also used in IsoCleft[9]) permits us to benefit from the powerful search algorithm that is graph matching but with reasonable computational cost. This allows us to include a large number of probes (potential interactions) when building the association graph and still perform high throughput calculations.

In the absence of benchmark data sets for the validation of methods for the detection of similarities, we opted to compare the performance of IsoMIF to that of a number of other methods on several data sets. While such data sets vary in the details of how they are created, the comparison across data sets makes it possible to have a more unbiased measure of the overall performance of the method. We compared IsoMIF to a number of methods for which such data exists. Unfortunately,

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

the FLAP method has not been tested in any of these data set, and the method is not freely available, thus making it impossible compare FLAP to IsoMIF or any of the other existing methods.

Using crystallized ligands to classify proteins as true positives may lead to artifacts in the validation of methods. Two proteins that were crystallized with different ligands could both bind other identical ligands. One alternative would be to use experimental binding profiles to define true positives more accurately, that is, a vector that defines the binding (or enzymatic activity) of proteins to a large number of ligands.[63] In this case, care must be taken to ensure that the data comes as a result of binding to the same binding site and not a mixture of binding sites (e.g., allosteric binding sites). One important caveat is that whereas one expects a correlation between binding site similarities and binding profile similarities in the case of cognate ligands as a result of divergent evolution, this does not have to be the case for designed molecules such as inhibitors. To illustrate this point, imagine two enzymes from the same family, say protein kinases. When developing an inhibitor, the most likely potential cross-reactivity targets that will be made sure not to bind the inhibitors are likely to be those most similar to the target of the inhibitor. Thus, when trying to correlate binding site similarities with binding profiles for inhibitors, the cases where the highest levels of similarity are expected to be found, particularly within members of the same protein family, are the ones likely to have the lowest levels of binding profile correlation.

Still in the case of divergent evolution, high levels of binding site similarity do not unequivocally mean that two proteins bind the same ligand. For example, mutations may lead to binding promiscuity[64] that can be then exploited through a different set of mutations to shift the affinity of the protein from one ligand to another. Such differences may be below the discrimination capabilities of existing methods that would classify the two proteins as having the same molecular function (i.e., binding the same ligand). In such cases, if these small-differences have drastic effects[9] then IsoMIF may be able to correctly decrease the relative similarity between the pair of proteins. Furthermore, in many cases, binding site mutations are also known to not disrupt ligand binding.[38,65] Such mutations may even occur in evolutionarily conserved positions.

In the case of proteins with different folds that as a result of convergent evolution bind similar or identical ligands, only a small and potentially different set of interactions may be responsible for binding. We previously observed that conservation is a very poor predictor of binding site similarities in these cases.[9] Thus, it may be more difficult to determine that two proteins bind such ligands. For example, NAD, ATP, and FAD can adopt highly diverse conformations.[66] It was shown that sequence motifs around the pyrophosphate atoms of FAD are more conserved than near the flavin or adenine ribose.[67] The latter two moieties are large and can participate in multiple combinations of hydrogen bonding, aromatic or hydrophobic interactions, with each combination yielding different binding modes. The problem is even more complex when considering the role of structural waters that mediate interactions between the ligand and the protein.

Thus, the intermolecular interactions within a binding site are complex, synergistic, environmentally dependent, and evolutionarily idiosyncratic. The above considerations led us to hypothesize[9] that binding site residues may be responsible for selectivity as much as specificity of ligand binding. It is necessary to take in consideration the spatial and temporal cellular contexts in which interactions between a protein and a ligand take place relative to the competing pool of proteins and ligands present to fully understand molecular recognition. Therefore, the large scale integration of systems and structural computational biology remains a challenge[68] that needs to be overcome to fully explain molecular recognition events.

CONCLUSIONS

In this work, we present IsoMIF, a tool to detect molecular interaction field similarities in protein binding sites. IsoMIF was extensively validated using a number of previously used data sets, thus making it possible to compare the performance of IsoMIF to that of other methods. The mean AUC obtained across data sets for IsoMIF is higher than those of other methods. Furthermore, while IsoMIF obtains consistently high AUC values across data sets, other methods perform more erratically across data sets. Like other methods for the prediction of binding site similarities, IsoMIF can be used to predict function from structure, to detect potential cross-reactivity or polypharmacology targets, and to help suggest bioisosteric replacements to known binding molecules. Given that IsoMIF detects spatial patterns of molecular interaction field similarities, its predictions are more directly related to pharmacophores and may be more readily translated into modeling decisions in structure-based drug design. Furthermore, unlike other methods, IsoMIF may in principle detect similar binding sites with distinct amino acid arrangements that lead to equivalent interactions within the cavity. It may also decrease the similarity between two binding sites that differ by a small but critical (in terms of their effect on the MIF) number of mutations. IsoMIF can be readily adapted to predict MIF similarities on protein surfaces to study protein−protein interactions as well as around small molecules. In the latter case, the Comparative Molecular Field Analysis (CoMFA) of sets of ensembles of small-molecule conformers is permitted to perform without requiring that they share a large common chemical scaffold for their superimposition to define a common grid. Finally, the freely available source code to calculate and visualize MIFs and detected MIF similarities make IsoMIF a welcome addition to the arsenal of methods for prediction of function of proteins, to understand molecular recognition, and to exploit similarities in drug design such as in drug repurposing and polypharmacology and the prediction of the molecular determinants of drug toxicity.

ASSOCIATED CONTENT

Supporting Information

Tables showing the terms of the interaction matrix; six probe energy thresholds; list of all amino acid atoms with corresponding atom types; reference atoms (for angle calculations) and interaction type; performance (AUC and EF_{10}) for different values of maximum cliques searched in Bron and Kerbosch for Kahraman, Homogenous, Steroid, and Soippa data sets; performance (AUC and EF_{10}) for PDBbind and scPDB at different sequence identity redundancy thresholds with number of entries for each subset; performance (AUC) for Kahraman data set for different levels of structural distortions using (1) the average AUC and (2) using the best AUC; top hits found for query 1BXM; and performance (AUC and EF_{10}) for different distance thresholds (applied when drawing edges of the association graph) for Kahraman, Homogenous, Steroid, and Soippa data sets. All entries of the four data sets (including

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

the PO4 entries for Kahraman) with the residue name, number, chain, and alternate location (when necessary) and their AUC. Distribution of distances between vertices and ligand atoms, figure showing the reference vectors and reference atoms used to calculate them for each amino acids, plot of the six probe thresholds as a function of distance, and figure of the procedure to build the association graph and figure of structural superimpositions of four top hits of 1BXM with MIF similarities. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00333.

Accession Codes

Our website (http://bcb.med.usherbrooke.ca/isomif) contains the IsoMIF source code and executables for Mac OS and Linux, scripts for the visualization of MIFs and MIF similarities with PyMOL, and detailed PDBbind and sc-PDB data sets introduced in this article. All the above can be obtained free of charge and without any registration required under the GNU General Public License 3.0.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rafael.najmanovich@usherbrooke.ca.

Author Contributions

M.C. and R.J.N. developed the method, designed the experiments, and wrote the manuscript. M.C. performed the calculations.

Funding

M.C. is the recipient of a Ph.D. fellowship from the Natural Sciences and Engineering Research Council of Canada (NSERC). This project was funded by a CQDM Explore grant and NSERC Discovery Grant RGPIN-2014-05766.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

R.J.N. is part of CR-CHUS, a member of the Institute of Pharmacology of Sherbrooke, PROTEO (the Québec network for research on protein function, structure and engineering) and GRASP (Groupe de Recherche Axé sur la Structure des Protéines).

REFERENCES

(1) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(2) Manning, G.; Whyte, D.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912−1934.

(3) Chartier, M.; Chénard, T.; Barker, J.; Najmanovich, R. J. Kinome Render: a Stand-Alone and Web-Accessible Tool to Annotate the Human Protein Kinome Tree. *PeerJ* **2013**, *1*, e126.

(4) Rost, B. Twilight Zone of Protein Sequence Alignments. *Protein Eng., Des. Sel.* **1999**, *12*, 85−94.

(5) Holm, L.; Sander, C. Dali/FSSP Classification of Three-Dimensional Protein Folds. *Nucleic Acids Res.* **1997**, *25*, 231−234.

(6) Andreeva, A.; Howorth, D.; Chothia, C.; Kulesha, E.; Murzin, A. G. SCOP2 Prototype: a New Approach to Protein Structure Mining. *Nucleic Acids Res.* **2014**, *42*, D310−D314.

(7) Sillitoe, I.; Cuff, A. L.; Dessailly, B. H.; Dawson, N. L.; Furnham, N.; Lee, D.; Lees, J. G.; Lewis, T. E.; Studer, R. A.; Rentzsch, R.; Yeats, C.; Thornton, J. M.; Orengo, C. A. New Functional Families (FunFams) in CATH to Improve the Mapping of Conserved Functional Sites to 3D Structures. *Nucleic Acids Res.* **2013**, *41*, D490−D498.

(8) Laskowski, R.; Watson, J.; Thornton, J. ProFunc: a Server for Predicting Protein Function From 3D Structure. *Nucleic Acids Res.* **2005**, *33*, W89−W93.

(9) Najmanovich, R.; Kurbatova, N.; Thornton, J. Detection of 3D Atomic Similarities and Their Use in the Discrimination of Small Molecule Protein-Binding Sites. *Bioinformatics* **2008**, *24*, i105−i111.

(10) Kurbatova, N.; Chartier, M.; Zylber, M. I.; Najmanovich, R. J. IsoCleft Finder - a Web-Based Tool for the Detection and Analysis of Protein Binding-Site Geometric and Chemical Similarities. *F1000Research* **2013**, *2*, 117.

(11) Han, G. W.; Bakolitsa, C.; Miller, M. D.; Kumar, A.; Carlton, D.; Najmanovich, R. J.; Abdubek, P.; Astakhova, T.; Axelrod, H. L.; Chen, C.; Chiu, H.-J.; Clayton, T.; Das, D.; Deller, M. C.; Duan, L.; Ernst, D.; Feuerhelm, J.; Grant, J. C.; Grzechnik, A.; Jaroszewski, L.; Jin, K. K.; Johnson, H. A.; Klock, H. E.; Knuth, M. W.; Kozbial, P.; Krishna, S. S.; Marciano, D.; McMullan, D.; Morse, A. T.; Nigoghossian, E.; Okach, L.; Reyes, R.; Rife, C. L.; Sefcovic, N.; Tien, H. J.; Trame, C. B.; van den Bedem, H.; Weekes, D.; Xu, Q.; Hodgson, K. O.; Wooley, J.; Elsliger, M.-A.; Deacon, A. M.; Godzik, A.; Lesley, S. A.; Wilson, I. A. Structures of the First Representatives of Pfam Family PF06938 (DUF1285) Reveal a New Fold with Repeated Structural Motifs and Possible Involvement in Signal Transduction. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **2010**, *66*, 1218−1225.

(12) Bakolitsa, C.; Kumar, A.; McMullan, D.; Krishna, S. S.; Miller, M. D.; Carlton, D.; Najmanovich, R.; Abdubek, P.; Astakhova, T.; Chiu, H.-J.; Clayton, T.; Deller, M. C.; Duan, L.; Elias, Y.; Feuerhelm, J.; Grant, J. C.; Grzechnik, S. K.; Han, G. W.; Jaroszewski, L.; Jin, K. K.; Klock, H. E.; Knuth, M. W.; Kozbial, P.; Marciano, D.; Morse, A. T.; Nigoghossian, E.; Okach, L.; Oommachen, S.; Paulsen, J.; Reyes, R.; Rife, C. L.; Trout, C. V.; van den Bedem, H.; Weekes, D.; White, A.; Xu, Q.; Hodgson, K. O.; Wooley, J.; Elsliger, M.-A.; Deacon, A. M.; Godzik, A.; Lesley, S. A.; Wilson, I. A. The Structure of the First Representative of Pfam Family PF06475 Reveals a New Fold with Possible Involvement in Glycolipid Metabolism. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **2010**, *66*, 1211−1217.

(13) Allali-Hassani, A.; Pan, P. W.; Dombrovski, L.; Najmanovich, R.; Tempel, W.; Dong, A.; Loppnau, P.; Martin, F.; Thornton, J.; Thonton, J.; Edwards, A. M.; Bochkarev, A.; Plotnikov, A. N.; Vedadi, M.; Arrowsmith, C. H. Structural and Chemical Profiling of the Human Cytosolic Sulfotransferases. *PLoS Biol.* **2007**, *5*, e97.

(14) Najmanovich, R. J.; Kuttner, J.; Sobolev, V.; Edelman, M. Side-Chain Flexibility in Proteins Upon Ligand Binding. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 261−268.

(15) Campagna-Slater, V.; Mok, M. W.; Nguyen, K. T.; Feher, M.; Najmanovich, R.; Schapira, M. Structural Chemistry of the Histone Methyltransferases Cofactor Binding Site. *J. Chem. Inf. Model.* **2011**, *51*, 612−623.

(16) Xie, L.; Bourne, P. E. Detecting Evolutionary Relationships Across Existing Fold Space, Using Sequence Order-Independent Profile-Profile Alignments. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 5441−5446.

(17) Xie, L.; Li, J.; Xie, L.; Bourne, P. E. Drug Discovery Using Chemical Systems Biology: Identification of the Protein-Ligand Binding Network to Explain the Side Effects of CETP Inhibitors. *PLoS Comput. Biol.* **2009**, *5*, e1000387.

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

(19) Halperin, I.; Glazer, D. S.; Wu, S.; Altman, R. B. The FEATURE Framework for Protein Function Annotation: Modeling New Functions, Improving Performance, and Extending to Novel Applications. *BMC Genomics* **2008**, *9* (Suppl2), S2.

(20) Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C.; Scozzafava, A.; Klebe, G. Unexpected Nanomolar Inhibition of Carbonic Anhydrase by COX-2-Selective Celecoxib: New Pharmacological Opportunities Due to Related Binding Site Recognition. *J. Med. Chem.* **2004**, *47*, 550−557.

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

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

(22) Schmitt, S.; Kuhn, D.; Klebe, G. A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.* **2002**, *323*, 387−406.
(23) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. Recognition of Functional Sites in Protein Structures. *J. Mol. Biol.* **2004**, *339*, 607−633.
(24) Kinoshita, K.; Nakamura, H. Identification of Protein Biochemical Functions by Similarity Search Using the Molecular Surface Database eF-Site. *Protein Sci.* **2003**, *12*, 1589−1595.
(25) Liu, T.; Altman, R. B. Using Multiple Microenvironments to Find Similar Ligand-Binding Sites: Application to Kinase Inhibitor Binding. *PLoS Comput. Biol.* **2011**, *7*, e1002326.
(26) Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A Simple and Fuzzy Method to Align and Compare Druggable Ligand-Binding Sites. *Proteins: Struct., Funct., Genet.* **2008**, *71*, 1755−1778.
(27) Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−135.
(28) Baroni, M.; Baroni, M.; Cruciani, G.; Cruciani, G.; Sciabola, S.; Sciabola, S.; Perruccio, F.; Perruccio, F.; Mason, J. S.; Mason, J. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands and Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **2007**, *47*, 279−294.
(29) Sciabola, S.; Stanton, R. V.; Mills, J. E.; Flocco, M. M.; Baroni, M.; Cruciani, G.; Perruccio, F.; Mason, J. S. High-Throughput Virtual Screening of Proteins Using GRID Molecular Interaction Fields. *J. Chem. Inf. Model.* **2010**, *50*, 155−169.
(30) Siragusa, L.; Cross, S.; Baroni, M.; Goracci, L.; Cruciani, G. BioGPS: Navigating Biological Space to Predict Polypharmacology, Off-Targeting, and Selectivity. *Proteins: Struct., Funct., Genet.* **2015**, *83*, 517−532.
(31) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.
(32) 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.
(33) Krotzky, T.; Rickmeyer, T.; Fober, T.; Klebe, G. Extraction of Protein Binding Pockets in Close Neighborhood of Bound Ligands Makes Comparisons Simple Due to Inherent Shape Similarity. *J. Chem. Inf. Model.* **2014**, *54*, 3229−3237.
(34) Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A New Bioinformatic Approach to Detect Common 3D Sites in Protein Structures. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 137−145.
(35) Wood, D. J.; de Vlieg, J.; Wagener, M.; Ritschel, T. Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and Its Application to Bioisostere Replacement. *J. Chem. Inf. Model.* **2012**, *52*, 2031−2043.
(36) Brylinski, M. eMatchSite: Sequence Order-Independent Structure Alignments of Ligand Binding Pockets in Protein Models. *PLoS Comput. Biol.* **2014**, *10*, e1003829.
(37) Krotzky, T.; Grunwald, C.; Egerland, U.; Klebe, G. Large-Scale Mining for Similar Protein Binding Pockets: with RAPMAD Retrieval on the Fly Becomes Real. *J. Chem. Inf. Model.* **2015**, *55*, 165−179.
(38) Eyal, E.; Najmanovich, R. J.; Sobolev, V.; Edelman, M. MutaProt: a Web Interface for Structural Analysis of Point Mutations. *Bioinformatics* **2001**, *17*, 381−382.
(39) Bron, C.; Kerbosch, J. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun. ACM* **1973**, *16*, 575−577.
(40) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(41) Meslamani, J.; Rognan, D.; Kellenberger, E. Sc-PDB: a Database for Identifying Variations and Multiplicity of "Druggable" Binding Sites in Proteins. *Bioinformatics* **2011**, *27*, 1324−1326.
(42) Gaudreault, F.; Morency, L.-P.; Najmanovich, R. J. NRGsuite: a PyMOL Plugin to Perform Docking Simulations in Real-Time Using FlexAID. *Bioinformatics* **2015**.
(43) Laskowski, R. Surfnet - a Program for Visualizing Molecular-Surfaces, Cavities, and Intermolecular Interactions. *J. Mol. Graphics* **1995**, *13*, 323−330.
(44) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-Chain Amide Orientation. *J. Mol. Biol.* **1999**, *285*, 1735−1747.
(45) Clark, M.; Cramer, R. D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982−1012.
(46) Arun, K.; Huang, T.; Blostein, S. Least-Squares Fitting of 2 3-D Point Sets. *IEEE Trans. Patt. Anal. Mach. Intell.* **1987**, *9*, 699−700.
(47) Kahraman, A.; Morris, R.; Laskowski, R.; Thornton, J. Shape Variation in Protein Binding Pockets and Their Ligands. *J. Mol. Biol.* **2007**, *368*, 283−301.
(48) Hoffmann, B.; Zaslavskiy, M.; Vert, J.-P.; Stoven, V. A New Protein Binding Pocket Similarity Measure Based on Comparison of Clouds of Atoms in 3D: Application to Ligand Prediction. *BMC Bioinf.* **2010**, *11*, 99.
(49) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics* **2015**, *31*, 405−412.
(50) Orengo, C.; Michie, A.; Jones, S.; Jones, D.; SWINDELLS, M.; Thornton, J. CATH - a Hierarchic Classification of Protein Domain Structures. *Structure* **1997**, *5*, 1093−1108.
(51) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: an Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33.
(52) Gaudreault, F.; Chartier, M.; Najmanovich, R. J. Side-Chain Rotamer Changes Upon Ligand Binding: Common, Crucial, Correlate with Entropy and Rearrange Hydrogen Bonding. *Bioinformatics* **2012**, *28*, i423−i430.
(53) Frappier, V.; Najmanovich, R. J. Vibrational Entropy Differences Between Mesophile and Thermophile Proteins and Their Use in Protein Engineering. *Protein Sci.* **2015**, *24*, 474−483.
(54) Frappier, V.; Najmanovich, R. J. A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations. *PLoS Comput. Biol.* **2014**, *10*, e1003569.
(55) Frappier, V.; Chartier, M.; Najmanovich, R. J. ENCoM Server: Exploring Protein Conformational Space and the Effect of Mutations on Protein Function and Stability. *Nucleic Acids Res.* **2015**, *43*, W395.
(56) Strittmatter, S. M. Overcoming Drug Development Bottlenecks with Repurposing: Old Drugs Learn New Tricks. *Nat. Med.* **2014**, *20*, 590−591.
(57) Wang, Z.; Zhou, X. E.; Motola, D. L.; Gao, X.; Suino-Powell, K.; Conneely, A.; Ogata, C.; Sharma, K. K.; Auchus, R. J.; Lok, J. B.; Hawdon, J. M.; Kliewer, S. A.; Xu, H. E.; Mangelsdorf, D. J. Identification of the Nuclear Receptor DAF-12 as a Therapeutic Target in Parasitic Nematodes. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 9138−9143.
(58) Wagener, M.; Lommerse, J. P. M. The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677−685.
(59) Laskowski, R.; Thornton, J.; Humblet, C.; Singh, J. X-SITE: Use of Empirically Derived Atomic Packing Preferences to Identify Favourable Interaction Regions in the Binding Sites of Proteins. *J. Mol. Biol.* **1996**, *259*, 175−201.
(60) Gohlke, H.; Klebe, G. DrugScore Meets CoMFA: Adaptation of Fields for Molecular Comparison (AFMoC) or How to Tailor Knowledge-Based Pair-Potentials to a Particular Protein. *J. Med. Chem.* **2002**, *45*, 4153−4170.
(61) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(62) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM General Force Field (CGenFF): a Force Field for Drug-Like Molecules Compatible with the CHARMM All-

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615

Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671−690.
(63) Najmanovich, R. J.; Allali-Hassani, A.; Morris, R. J.; Dombrovsky, L.; Pan, P. W.; Vedadi, M.; Plotnikov, A. N.; Edwards, A. M.; Arrowsmith, C.; Thornton, J. M. Analysis of Binding Site Similarity, Small-Molecule Similarity and Experimental Binding Profiles in the Human Cytosolic Sulfotransferase Family. *Bioinformatics* **2007**, *23*, e104−e109.
(64) Copley, S. D. An Evolutionary Biochemist's Perspective on Promiscuity. *Trends Biochem. Sci.* **2015**, *40*, 72−78.
(65) Eyal, E.; Najmanovich, R.; Edelman, M.; Sobolev, V. Protein Side-Chain Rearrangement in Regions of Point Mutations. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 272−282.
(66) Stockwell, G.; Thornton, J. Conformational Diversity of Ligands Bound to Proteins. *J. Mol. Biol.* **2006**, *356*, 928−944.
(67) Dym, O.; Eisenberg, D. Sequence-Structure Analysis of FAD-Containing Proteins. *Protein Sci.* **2001**, *10*, 1712−1728.
(68) Samish, I.; Bourne, P. E.; Najmanovich, R. J. Achievements and Challenges in Structural Bioinformatics and Computational Biophysics. *Bioinformatics* **2015**, *31*, 146−150.

DOI: 10.1021/acs.jcim.5b00333
J. Chem. Inf. Model. 2015, 55, 1600−1615