International Journal of
Molecular Sciences

MDPI

Review

Estimating the Similarity between Protein Pockets

Merveille Eguida and Didier Rognan *

Laboratoire d’Innovation Thérapeutique, UMR7200 CNRS-Université de Strasbourg, 67400 Illkirch, France
* Correspondence: rognan@unistra.fr; Tel.: +33-3-68-85-42-35

Abstract: With the exponential increase in publicly available protein structures, the comparison of protein binding sites naturally emerged as a scientific topic to explain observations or generate hypotheses for ligand design, notably to predict ligand selectivity for on- and off-targets, explain polypharmacology, and design target-focused libraries. The current review summarizes the state-of-the-art computational methods applied to pocket detection and comparison as well as structural druggability estimates. The major strengths and weaknesses of current pocket descriptors, alignment methods, and similarity search algorithms are presented. Lastly, an exhaustive survey of both retrospective and prospective applications in diverse medicinal chemistry scenarios illustrates the capability of the existing methods and the hurdle that still needs to be overcome for more accurate predictions.

Keywords: binding site; alignment; off-target; library design; virtual screening; machine learning; artificial intelligence drug design

check for updates

Citation: Eguida, M.; Rognan, D. Estimating the Similarity between Protein Pockets. *Int. J. Mol. Sci.* **2022**, *23*, 12462. https://doi.org/10.3390/ijms232012462

Academic Editor: Giulio Vistoli

Received: 6 October 2022
Accepted: 16 October 2022
Published: 18 October 2022

1. Introduction

In living organisms, biological processes are regulated through specific molecular recognition at local surfaces. Proteins, one of the major biomolecules composing our cells, interact with different partners: other proteins, peptides, nucleic acids, small molecules, and transition metals. The exploration of the proteome, considering amino acid sequences, makes it possible to rapidly compare proteins but does not necessarily indicate whether potential cavities at their surfaces will be conserved or not. Hence, structure conservation does not always mirror sequence homology [1]. Progress in molecular and structural biology have enabled us to uncover the three-dimensional (3D) structure of proteins, either by X-ray diffraction [2], nuclear magnetic resonance (NMR) [3], or more recently cryo-electron microscopy (cryo-EM) at an atomic scale [4], all approaches being now integrated [5]. Characterizing the binding cavities for small molecules has bolstered the rise of structure-based drug design [6]. Supported by the outlooks and successful case studies, many methods have been developed in the last three decades. The bottleneck of protein cavity comparison is common to all similarity estimates—similarity is a non-measurable characteristic that depends on the considered aspects. Instead, derived hypotheses (e.g., function, ligand binding) are further evaluated. This presents many challenges for benchmarking methods and highlights the importance of carefully designing datasets in retrospective studies. For users as well as developers, knowing where we come from and what has been achieved in the field enables realistic expectations and spot limitations to be addressed by future developments.

Structure-based algorithms for protein site comparison emerged in the 1970s, a decade marked by the establishment of the Protein Data Bank (PDB) [7] and the deposit of a few structures. Initially, efforts were made to compare protein 3D structural motifs independently of sequence order and gaps. Computer vision approaches [8] were applied in structural biology for similar substructure identification even in the absence of sequence homology via rigid body alignments. Protein functions could be predicted from a database of known 3D templates, by querying or inferring protein active sites [9]. Beyond the functional annotations, cavity alignment and comparison rapidly became promising for the rational design of proteins and ligands, since similar 3D arrangements of surface motifs may be involved in similar molecular recognition events [10].

The path from the earlier to the current site comparison methods involved several implementations. It was common for the user to define researched features (e.g., set of atom/residues distances defining a motif: catalytic triads, similar ligands) from prior knowledge to initiate the search [11]. The subsequent advantages are a better control of the comparison, an easier selection of relevant matches, and the reliability of the solutions. Progressively, methods allowing for the automatic identification of pockets [12–16] and of relevant matched patterns opened the doors to the analysis of the relationships between evolutionally and structurally remote members of an entire database, without any a priori judgment [17–21]. Such predictions led to unexpected findings with implications for drug design [22,23]. Screening large databases requires effective computing time. Together with the progress of computing technologies, fast methods were introduced, but often at the cost of interpretability [24–26].

The repertoire of possible comparison algorithms is tailored to the representation made of the pocket [27]. Pocket representation is a way to provide structured information to the algorithm for exploration. Once delimited in the protein, a pocket can be modeled as a list of amino acids, graphs, or unconnected pseudo atoms, among other possibilities. The geometrical constraints of alpha carbon tuples were extensively used to identify equivalenced areas [28–30]. Other cavity descriptors further encode the chemical properties of atoms or residues, hence reducing redundancy in the possible matches [17,31,32]. The intricacy of the representation lays in finding a good balance between fuzziness (with a risk of false positive matches) and exhaustiveness (with a risk of missing remote similarities). In any case, similarity can only be properly reported with a fair scoring function. The scoring scheme aims at quantifying how two pockets resemble or differ. Often, a score threshold is applied in screening campaigns for decision making. How to assign the value of that threshold and assess the significance of that similarity is a genuine question raised by earlier studies [24,33,34].

In practice, the variability of the pocketome [35] in terms of size, solvent accessibility, and flexibility constitutes an obstacle to the performance of binding site comparison methods, as it is for other structure-based approaches [36]. It is perceived that comparing subpockets instead of entire cavities might better handle the conformational variations, typically induced by ligand binding. Noteworthy, the ability to detect local and global similarities is suitable for different purposes. As the reader will notice, different parameters entail the success of protein cavity comparisons, as discussed by previous articles [23,37–39]. In this review, we will provide a most recent and broad overview of all stages involved in pocket comparison, from the prediction of ligand binding sites to the evaluation and prospective applications in drug design.

2. Pocket Detection and Druggability Estimate

The identification of potential interaction sites is crucial to structure-based approaches and constitutes the very first step of binding site comparison. Contact surfaces exhibit different geometric and physicochemical characteristics according to the nature of the binding partner (proteins, peptides, nucleic acids, small molecules, and transition metals). For example, small molecule interaction sites are buried clefts, while protein–protein interaction interfaces are rather flat and hydrophobic [40]. Although the available methods for binding site detection covers the different applications above, the majority relates to small molecule pocket identification as a testimony of efforts to structure-based drug design of small chemical entities in recent decades. The accessibility to binding site identification is possible via standalone tools [16], websites [41], or databases of precomputed sites [35,42].

The methods can be classified into three levels: (i) the genomic or 3D structure of the input, (ii) the dependency to bound ligands, and (iii) the class of the algorithm (Figure 1).

Template or sequence-based methods such as ConSeq [43] identifies functionally important residues in protein sequences by searching for evolutionary relations with other proteins. Another approach is 3DLigandSite, which takes a protein sequence as input, although it relies on homology models or de novo structure predictions [41]. Structure-based pocket identification uses only the 3D coordinates of the structures as input and benefits from the augmentation of structural data [7].

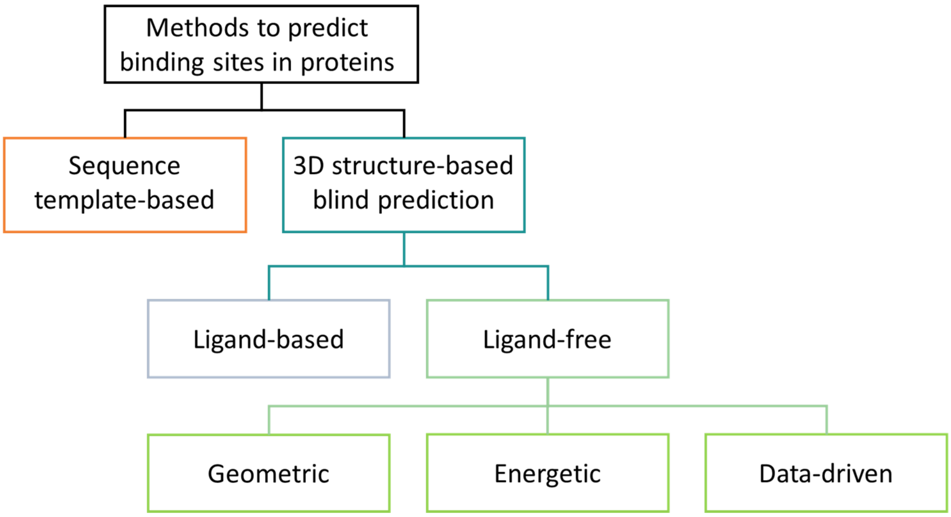

Methods to predict binding sites in proteins
Sequence template-based
3D structure-based blind prediction
Ligand-based
Ligand-free
Geometric
Energetic
Data-driven

Figure 1. Classification of binding site detection methods.

Ligand-centric methods are restricted to protein–ligand complexes and is a site delimitation rather than a prediction. Noticeably, the analysis of crystallization additives binding sites might suggest potential allosteric pockets [44,45]. Typically, a site is defined as all residues within a certain distance cutoff to the partner's heavy atoms, ca. 6 Å for protein–small molecule complexes. Alternatively, the set of residues can be restricted to those properly oriented toward the ligand, with the particularity that the distance cutoff varies according to the interaction type. These approaches are available through integrated environments, making it possible to manipulate protein structure coordinates and interactions such as Molecular Operating Environment (Chemical Computing Group, Montreal, Canada), or through independent tools for parsing protein 3D structure data [46].

Ligand-free approaches can operate on a larger range of structures, enabling the discovery of unprecedented sites. According to their search algorithm, they can be classified as geometric, energetic, or data-driven (Table 1). At first glance, all geometric methods aim at identifying sufficiently buried zones unoccupied by protein atoms, but they differ in their strategies to search for these areas. Grid-based methods place the protein into a cartesian grid and identify grid cells likely to be in a cleft by analyzing their neighborhood [13,14,47–58]. POCKET [13] and LIGSITE [12], two of the earliest methods, keep cells that correspond to a 'protein-solvent-protein' event by scanning in three and seven directions, respectively. VolSite annotates cavity points by pharmacophoric properties, complementary to that of the protein microenvironment [47]. Such algorithms are sensitive to grid resolution and orientation but are powerful to detect cavities of different sizes and curvatures.

Table 1. Common structure-based methods to predict ligand binding pocket in proteins.

Category	Search Approach	Methods
Geometric	Grid	CAVER [51], CAVIAR [49], DoGSite [14], ghecom [57], KVFinder [48], LIGSITE [12], LIGSITEcsc [53], McVol [58], POCKET [13], PocketDepth [55], PocketPicker [54] VICE [56], VOIDOO [52], VOLSITE [47]
	Alpha-shape	APROPOS [59], CAST [6], CASTp [60], Fpocket [16],
	Spherical probes	DEPTH [61], HOLE [62], HOLLOW [63], PHECOM [57], PASS [64], Roll [65], SURFNET [66], SURFNET-ConSurf [67], Xie and Bourne [33]
	Other	MSPocket [68], SplitPocket [69]
Energetic	Grid	AutoLigand [70], DrugSite [71], FTSite [72], PocketFinder [73], Q-SiteFinder [74], SITEHOUND [75], SiteMap [76], pocket-finder [77], GRID [78]
	Spherical probes	dPredGB [79], Morita et al. [80]
	Other	Gaussian Network Model [81]
Data-driven	Machine learning	GRaSP [82], MCSVMBs [83], P2Rank [15], PRANK [84], SCREEN [85]
	Deep learning	PoinSite [86], DeepPocket [87], PUResNet [88], DeepSurf [89], BiteNet [90], Jiang et al. [91], DeepSite [92], ISMBLab-LIG [91]

Contrarily, other methods process the protein coordinates directly and are not affected by the grid initialization phenomena. Based on the alpha-shape concept introduced by Edelsbrunner et al. [93], they circumvent protein cavities by connecting adequate adjacent Delaunay triangles via the 'discrete flow' method [6,59,60,69], or by clustering alpha spheres to satisfy pocket descriptors [16]. Alternative purely geometric approaches fill or coat the protein with spherical probes to delimit cavity void [61–67,94,95]. Finally, other concepts, such as monitoring the direction of surface normal vectors, were implemented [68].

The second category of ligand-free methods estimate favorable surfaces for protein–ligand contacts by calculating the potential energy of probes at different positions [78]. Generally, the Lennard-Jones potentials are used with hydrophobic probes. The nature and number of probes vary from a simple carbon probe in DrugSite [71] to 16 different ones in FTSite [72]. The potentials are either mapped to grid positions [70–78] or to probe the protein surface [79,80]. Evidently, the outputs of energy-based methods are influenced by the chosen force field.

The final class of methods uses supervised models, trained on the features of well-characterized ligand binding sites. Hence, they differ in the features' representation, training models, set of parameters, and datasets. P2RANK [84] is one of the examples based on classical machine learning models. The protein solvent-exposed atoms are processed into a topological and physicochemical feature vector which serves as input to a random forest classifier. Recently, many deep learning methods [86–92], majorly based on 3D-convolutional neural networks (CNNs) were introduced. CNNs have shown to be very powerful in image recognition problems [96] and were thus directly applied to protein binding sites represented as voxels with atomic attributes, while keeping the architecture of the CNNs previously used for other purposes. Another possibility to represent binding sites is used in PointSite [86], which addresses point clouds segmentation using sparse convolution. While these methods need to be challenged by prospective usages, recent advances in 3D point cloud deep learning [97] offer some wide perspectives for this type of problem.

Altogether, these methods have been evaluated on their performance to accurately predict binding pockets by comparing predictions on unbound proteins to ground truth ligand locations in their corresponding bound structures. Not only is the accuracy of the location analyzed, but also the delimitation or overlap with respect to the ligand [14]. The detected pockets might be too large or too small where a clustering is required. Thus, post-processing data generated by various tools may be useful [98]. Cleverly, meta-methods thrive to find consensus from different algorithms to increase the chances of correct predictions [99,100]. However, consensus might not always yield the right solution.

Indeed, all identified clefts do not necessarily correspond to the ability to accommodate a drug-like ligand (druggability). The concept of structural druggability [101] arose from observing the characteristics of pockets bound to pharmacological ligands: an average volume between 200 and 800 Å^3, a good balance of hydrophobic and polar atoms enabling some binding specificity, and sufficient buriedness. A few methods were developed to predict target druggability [14,47,102–106]. Consistently, the topological and physicochemical characteristics of the pocket sites are encoded into descriptors and trained on curated datasets to generate classification models (e.g., support vector machines, random forest, linear regression). Since pocket druggability does not guarantee that the bound ligand will also be druggable, the term may be replaced by ligand-ability [107] or bind-ability [104]. For more information, we refer the reader to a recent review [108]. Interestingly, some of the previously described methods have implemented both a pocket detection and a rule-based druggability prediction [14,16,47,76], thereby enabling a straightforward selection of the most interesting pockets, notably for supramolecular assemblies [40].

3. Comparing Pockets: A Multi-Step Procedure

The methods that compare protein cavities operate in three steps: (i) describing the cavity with a suitable representation, (ii) comparing these representations, (iii) scoring the proposed comparison. Hence, successful results reside in a coordinated performance of each of these tasks. Yet, cavity representation, which is the first step of the procedure, is crucial as it influences the later steps. Generally, a poor representation in which relevant characteristics are missing cannot be compensated by the most efficient comparison or scoring algorithm. State-of-the art methods to compare protein cavities are summarized in Table 2. In the following sections, we will discuss the different approaches to achieve this end.

Table 2. Methods to compare protein cavities.

Year	Name	Detection	Principle	Scoring	Evaluation Datasets
2002	CavBase [17]	LIGSITE [12]	Clique detection in graphs of pseudoatoms	Overlap of surface grid points, RMSD	Cofactor sites, kinases, serine proteases
2002	eF-site [109]	Ligand Databases	Clique detection in graph of surface normal vectors and electrostatic potentials	Normalized and weighed contributions of vectors angles, potentials, distances	Phosphate sites, antibodies, PROSITE classes
2003	SuMo [110]	Ligand	Incremental match of triplets of pseudocenters	Count of matches, RMSD, composite of Euclidean and density distances	Protease catalytic sites, lectines
2004	SiteEngine [18]	Ligand	Match of triplets of points by hashing	Hierarchical scoring: count of matches, RMSD, overlap of patches, local shape	Cofactors, steroids, fatty acid sites, catalytic triad in proteases
2004	SitesBase [111]	Ligand	Match of triplets of points	Count of matches, RMSD	Cofactors, phosphate sites
2007	Ramensky et al. [112]	Ligand	Clique detection in graph of atoms	Dice similarity of matches	Diverse
2008	Binkowski et al. [113]	CAST [6] Ligand	Comparison of pairwise distance histograms	Kolmogorov–Smirnov divergence, overlap of volume, RMSD	Cofactor sites, HIV proteases
2008	PocketMatch [19]	Ligand	Comparison of sorted pairwise distances	Normalized count of matches	Diverse, SCOP classes

Table 2. *Cont.*

Year	Name	Detection	Principle	Scoring	Evaluation Datasets
2008	SiteAlign [20]	Ligand	Alignment of polyhedron fingerprints	Normalized distances of fingerprints	Proteases, kinases, estrogen receptors, GPCRs
2008	SOIPPA [114]	Ligand	Clique detection in graphs of atoms	Composite weighted by frequencies, PSSM, distances	Cofactor sites, SCOP classes
2009	SMAP [33]	Ligand	Clique detection in graphs of atoms	Gaussian densities from distances, angles of normal vectors, BLOSSUM weights	Cofactor sites
2010	BSSF [25]	PASS [64]	Comparison of fingerprints of binned distances and properties	Canberra distances of fingerprints	Diverse, synthetic data, SCOP classes
2010	Feldman et al. [30]	Ligand	Match of subsets of Cα atoms	Probabilistic score from distances between matches	Diverse, kinases
2010	FuzCav [24]	Ligand	Fingerprints of triplets of atom features	Maximal proportion of matches	Diverse, functional groups, 8 difficult cases
2010	Milletti et al. [115]	Ligand	Comparison of 3 concentric spheres fingerprints encoding neighborhood for each point, solving linear assignment	Composite of fingerprint distances and RMSD	ATP sites, kinases
2010	P.A.R.I.S (sup-CK) [116]	Ligand	Initial alignment optimized by gradient ascent to maximize a Gaussian kernel	Gaussian kernel	Cofactor sites
2010	ProBiS [31]	Ligand	Maximum clique detection in graphs of surface atoms	Count of Matches, RMSD, angle between vectors	Cofactor/metal sites, protein–protein interfaces, protein–DNA complexes
2011	PocketAlign [117]	Ligand	Initial pairs from sorted lists of atom distances, then extend	Count of matches, RMSD	Cofactor sites, SCOP classes
2011	PocketFEATURE [118]	Ligand	Comparison of 7 concentric spheres fingerprints encoding neighborhood for each microenvironment	Normalized Tanimoto similarity of fingerprints	Kinases
2012	KRIPO [21]	Ligand	Fingerprints of triplets of pharmacophore	Modified Tanimoto of fingerprints	Diverse, fragments subpockets, search of bioisosteric substructures
2012	Patch-Surfer [119]	Ligand LIGSITE [12]	Comparison of 3D Zernike of surface patches solving a weighted bipartite matching	Composite of surface match distances and size differences	Cofactor sites

Table 2. *Cont.*

Year	Name	Detection	Principle	Scoring	Evaluation Datasets
2012	Shaper [47]	VolSite [47]	Comparison of cloud of points by Gaussian shapes matching	Tanimoto, Tversky of matches	Diverse, GPCRs, proteases
2012	TIPSA [120]	Ligand	Match of quadruplets of points, iterative refinement by Hungarian algorithm	Tanimoto of matches, overlap of volume, normalized RMSD	Cofactor sites
2013	Apoc [28]	Ligand CAVITATOR [28], LIGSITE [12]	Seed alignment by comparing secondary structures, optimized by solving linear assignment problem	Composite of vector orientation, distance, properties	Diverse, similar ligand recognition sites
2013	TrixP [121]	DoGSite [122]	Search for common shape and triplets of points by bitmap indexing	Composite of matches count, angle between vectors, mismatches penalty	Diverse, 8 difficult cases, protease, estrogen receptor, HIV reverse transcriptase
2014	*e*MatchSite [29]	*e*FindSite [123]	Template-based alignment optimized by Hungarian algorithm	Machine learning score: RMSD, residue, properties	Cofactors, steroid sites
2014	RAPMAD [26]	LIGSITE [12]	Comparison of 14 pairwise distance histograms, one for each property	Jensen–Shannon divergence of histograms	Cofactor sites, proteases, diverse
2015	IsoMIF [124]	GetCleft [125]	Clique detection in graphs of interaction grid points	Tanimoto of descriptors of matched points	Cofactors, steroid sites
2016	G-LoSA [126]	Ligand	Clique detection in graphs of atoms	Feature-weighted count of matches	Diverse, Ca+ sites, similar ligands recognition sites, protein–protein interfaces
2016	SiteHopper [127]	Ligand	Comparison of surface atoms by Gaussian shapes matching	Weighted combination of shape and color Tanimoto	Diverse using binding affinities
2019	DeepDrug3D [128]	Ligand	Convolutional neural network model	Binary classification	Cofactors, steroids sites, proteases
2020	DeeplyTough [32]	Fpocket [16] Ligand	Convolutional neural network model	Binary classification	Cofactor sites, diverse and using binding affinities
2020	ProCare [129]	VolSite [47]	Match of randomly sampled quadruplets refined by iterative closest point	Tversky of matched pharmacophoric properties	Diverse, using functional annotation, fragments subpockets, search of bioisosteric structures
2021	PocketShape [130]	Ligand	Initial alignment optimized by Hungarian algorithm	Composite of matches, orientation of residues	Diverse SCOP classes, kinases
2021	Site2Vec [131]	Ligand	Random forest model on autoencoder-generated descriptors	Binary classification	Cofactors, steroid sites, diverse

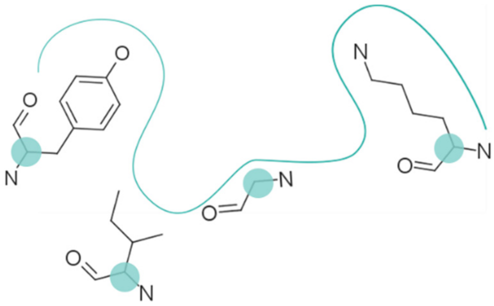
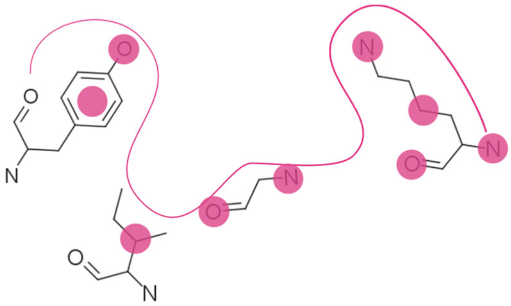
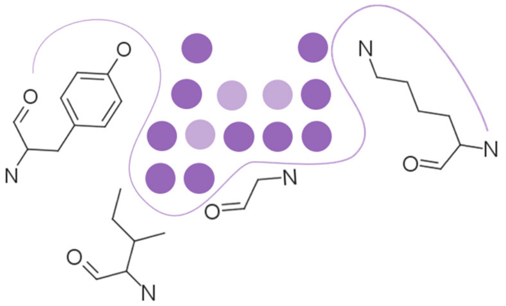
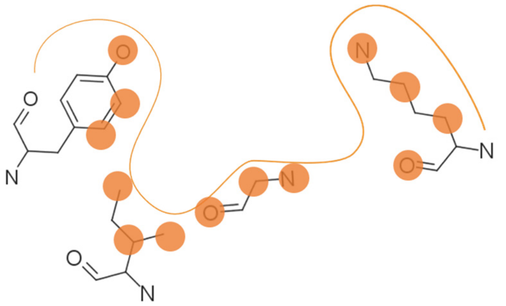
3.1. Pocket Representation

Once the pockets are delimited, the features are selected by considering different aspects. This step aims at focusing on the relevant characteristics that explain ligand recognition, while discarding unnecessary information. Our brain performs the same exercise on everyday life's objects, for example, if we are asked to compare two cars, we might decompose the information into major aspects such as the brand, design, color, motor, etc. Interestingly, different people will focus on different combinations of these aspects, resulting in different decision making. For pocket modeling, there is the general knowledge that the attributes (size, physicochemical properties, flexibility) of residues flanking the site and their relative 3D location explain the specific recognition of ligands [17,27,132]. Therefore, pocket comparison methods approximate these residues into various representations which differ at three levels: (i) the discretization of the residues, (ii) the viewpoint, and (iii) the chemical features.

First, possible representations (Table 3), from coarse-grained to more detailed, can be a representative atom (typically the Cα or Cβ atom) describing an entire residue (e.g., Apoc), a group of pseudocenters or vectors associated with residue fragments (e.g., CavBase), a cloud of atoms (e.g., VolSite), or 3D voxels (e.g., DeepSite). The resolution of the representation determines how local the subsequent comparison can be. For example, the rigid matching of atoms which are 7 Å apart in a query pocket can only be associated with similarly spaced atoms in the reference pocket, therefore excluding a pertinent association of smaller areas. Resolution also influences sensitivity to chemical and coordinate variations (Figure 2). Coarse-grained representations are less sensitive to variations in atomic coordinates but are more perceptive of changes in chemical properties such as single residue mutations. They offer a better signal-to-noise ratio at the cost of information. In grid-based approaches, the grid resolution (often 0.5 to 1.5 Å) is adjusted to capture the shape of the site, while compromising between precision and computing [47,124]. Although small changes of residues are reflected in detailed representations, they can be perceived to a lesser extent since drowned in many other information elements. The detection of such details is highly influenced by the assignment of chemical features and the performance of the search algorithm. Noticeably, some methods have adopted a mix representation scheme, wherein gross representations are used for a faster search and finer representations are involved in the scoring [17].

Secondly, most methods adopt the protein perspective by considering atoms or pseudocenters at the protein surface (e.g., FuzCav, SMAP). A few stand out by projecting these protein patterns into the ligand space, wherein polyhedron, voxels, or points are annotated with the properties of nearest or well-oriented protein features (e.g., IsoMIF, SiteAlign) (Table 3). Such discretization aims at offering a good balance between information completeness while handling variations in atomic coordinates and features. However, it is important to recall that grid-based representations are affected by the centroid location and axes orientation during the grid initialization. As a result, the distribution of feature types might change (a protein feature might move in adjacent voxels or not be represented at all), particularly when a voxel is associated with only one feature at a time. The same representation (e.g., cloud of points) can be applied to either key protein atoms [116] or grid points delimiting the accessible cavity space [129], thereby offering the possibility to mirror an imaginary ligand viewpoint and providing an alternative comparison approach (Figure 3).

Table 3. Discretization of the residues to represent a protein cavity.

Representation	Illustration [a]	Methods
Single points		APoc, eMatchSite, FuzCav, G-LoSA, PocketAlign [b], SiteAlign [b], SMAP, SOIPPA,
Pseudocenters		BSSF, CavBase [b], KRIPO, PocketAlign [b], PocketMatch, RAPMAD, Site2Vec, SiteEngine, SuMo, TrixP [b]
Surface points, surface patches, volume points, polyhedron		CavBase [b], DeepDrug3D, DeeplyTough, IsoMiF, Patch-Surfer, ProCare, Shaper, SiteAlign [b], TrixP [b]
All heavy atoms		Binkowski et al., Brakoulias et al., Milletti et al., P.A.R.I.S, ProBiS, SiteHopper, TIPSA

[a] The protein cavity is delimited by a few residues. Representative points at different resolutions are depicted as colored spheres. [b] Some methods use mixed representations.

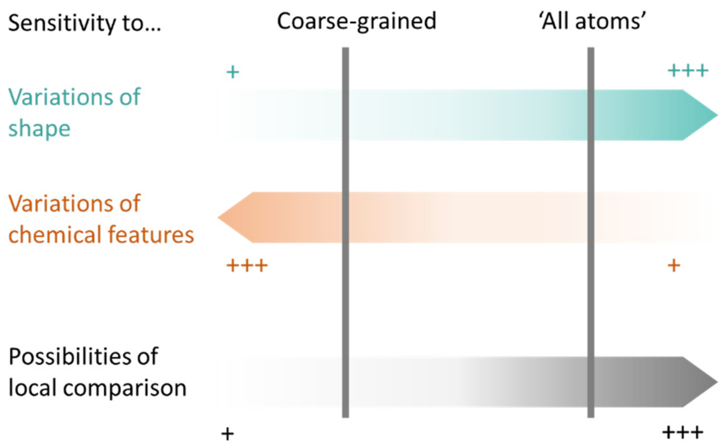


Figure 2. Sensitivity of coarse-grained to all-atom cavity representations to variations in atomic coordinates, chemical features, and subsequent applications (+: low, +++: high).

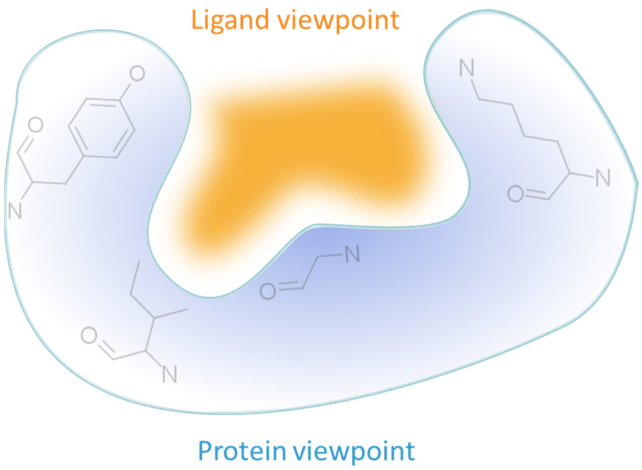


Figure 3. Protein cavity representation according to the protein or the ligand perspective.

Finally, besides the two aspects described above, the methods differ in their definition of chemical and geometric features. For example, Binkowski et al. do not consider the chemical type of atoms but showed that the relative position of the surface atoms describing the shape of the pocket already contains some discriminative information [113]. However, shape information alone is insufficient; hence, it is not surprising that almost all the state-of-the-art site comparison methods annotate surface coordinate atoms with pharmacophoric features to improve the discrimination between redundant areas. In coarse-grained representations, $C\alpha/C\beta$ atoms are annotated according to the chemical groups of their residues. For instance, APoc defined eight exclusive chemical groups, allowing for a residue to belong to only one [28]. Searching for the identity of chemical features between the query and reference pockets with such representations does not account for the interchanging role that fragments in different amino acids may have. For example, the hydroxyl group of serine and tyrosine can be a hydrogen bond donor or acceptor, as tyrosine additionally displays an aromatic feature; yet serine and tyrosine belong to different classes. To correct this effect, residues can be assigned multiple classes (e.g., SiteAlign). Alternatively, single or groups of atoms defining pseudocenters are annotated according to their interaction capacities (e.g., a histidine side chain is represented by a hydrogen-bond donor–acceptor feature and aromatic pseudocenters in CavBase). Commonly, five to eight pharmacophoric features are defined (KRIPO, SiteEngine, VolSite), as well as up to more than 40 atom types (e.g., PocketFEATURE). Other possible chemical attributes are partial charges used in P.A.R.I.S (sup-CK) or SiteEngine scoring, atomic density (SuMo), or atom types (e.g., SitesBase). The definition of many feature types might improve the description of the site with precision but might at the same time hinder remote similarity detection by narrowing the applicability domain of the method. Aside from the chemical features, the geometrical patterns are sometimes considered: CavBase and RAPMAD indicate the directionality of polar features by vectors, SuMo considers the directionality of the patterns toward the cavity by scalar triple product, SOIPPA assigns normal vectors to local surfaces, TrixP and Sitelign consider distances to fixed points.

In a nutshell, there are various ways to represent a protein cavity. Challenges reside in finding a good balance between comprehensive representation of features to ensure reliability and loose representation making it possible to detect remote similarities. While the absence of pocket attributes cannot be recovered at the later comparison step, too many attributes may constitute difficulties for the search algorithm in separating the signal from the noise.

3.2. Similarity Search

Following the selection of features characterizing the cavities, similarity is estimated by algorithms that search for common patterns shared between two sites. First, the representations of the protein cavities are converted or organized into comparable and computer-friendly objects that can be processed automatically. There are a variety of search algorithms to this end, which can be categorized according to their inputs, procedure, and visual interpretability (Figure 4).

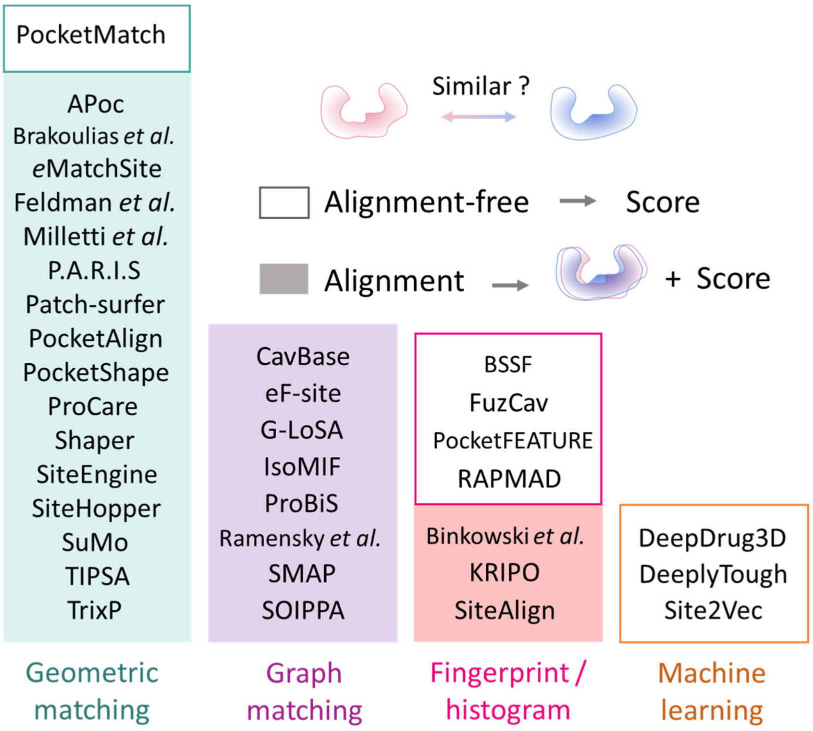


Figure 4. Classification of state-of-the-art methods for protein pockets comparison. Alignment-based methods compute a transformation (rotation, translation) to superpose the query to the target site.

The first category of algorithms searches for geometric (e.g., pairwise distances, angles, shape) and chemical (identical or compatible types) constraints to match. It is not safe to expect a perfect match, given the errors in 3D structure resolution, the flexibility nature of proteins, and the aim to find unobvious similarities. Therefore, a certain margin of geometric errors is always tolerated. PocketMatch compares set of distances belonging to 90 combinations of atom types and properties to establish correspondences between two pockets and keeps the solution maximizing the number of correspondences [19]. Global alignment methods (P.A.R.I.S, SiteHopper, Shaper) try to maximize the overlap between two cavities. A seed alignment is initialized, for example, by superposing the centroids or principal axes of the two sites, which are then optimized [47,116,133]. SiteHopper and Shaper rely on the OpenEye's ROCS (OpenEye Scientific Software, Santa Fe, NM, USA), wherein atoms/points are represented by smooth Gaussians to enable fuzzy shape comparisons [47,127]. A different approach for global optimization is to establish seed correspondences. APoc compares local protein fragments [28], Milletti et al. associate points based on their circular fingerprint similarity [115], eMatchSite assigns seven residue-level scores at selected Cα atoms [29], Patch-Surfer compares the patch surface properties by 3D functions [119]. The next alignment is solved by the Hungarian algorithm or other combinatorial optimization algorithms [8,22,102,115]. PocketAlign uses a similar approach using BLOSSUM62 weights when generating local seed alignments, which are later extended to the full structures [117]. Alternatively, some methods partition the pocket by considering a few points each time. Given that at least three points are necessary to superpose two objects without ambiguity, those methods enumerate triplets or quadruplets of feature points in the query to iteratively search for equivalent cliques in the target [17,30,110,111,120,121,129]. The formation of the n-tuples can be customized to avoid promiscuous sets. In TrixP, triangles solely made of hydrophobic features are not considered [121]. A match can signify a simple correspondence of identical chemical types and pairwise distances (SiteEngine, TIPSA) or of additional properties such as vector angles, local shape (ProCare, TrixP). ProCare relies on a 41-bin histogram describing each point, accounting for both shape and pharmacophoric features [129]. The alignment is performed in two steps, first by finding equivalent pocket points using a random sample consensus algorithm [134], then iteratively refining the preliminary alignment by the iterative closest point (ICP) method [135]. Aligning all possible combinations is costly in time, hence SiteEngine and TrixP, employ hashing and bitmap indexing, respectively, allowing for a 'search IN' for the faster identification of similar patterns.

In the second category, selected points form the nodes of a graph. According to the cavity representation, each node is annotated by a property and the edges by their lengths. Comparing two cavities results in comparing two graphs to extract the maximum common subgraphs. To achieve this end, a product graph is built by associating similar nodes (property comparison) and edges of almost equal distances, tolerating a certain deviation. Cliques are identified in this association graph to derive pairs of equivalent points that can be used to superpose the two cavities. CavBase, G-LoSA, ProBiS, etc. (Figure 4) are based on this principle. Differences between methods arise from the graph construction (minimal and maximal distances to consider adjacent nodes), distance tolerances, and the definition of a property match (identity or compatibility). For example, G-LoSA explores three different distance deviations (1.5, 2.0 and 2.5 Å) and further evaluates the alignment of local triangles within each clique of at least four nodes [126]. Clique detection is computationally expensive, particularly with dense graphs (e.g., 0.5 Å grid, [124]). Therefore, it requires practically efficient solutions such as the Bron–Kerbosch algorithm [136] and improved variants [137].

Methods in the third category generally adopt a global vision of the protein cavity. They consider a pocket as a fixed-length fingerprint or histogram, and comparing two pockets is amounts to calculating the similarity or distances between their fingerprints/histograms. BSSF, FuzCav, and KRIPO compute couples or triplets of pharmacophoric features separated by binned distances [21,24,25]. While the two former count the number of occurrences of each combination, bits are activated in KRIPO when a combination occurs. Then, KRIPO fuzzifies its fingerprints to account for the neighborhood phenomena [21]. SiteAlign also compares fingerprints but, contrarily to the other methods, the fingerprint of the query pocket is iteratively generated, as it derives from properties of the cavity projected on a rotated/translated 80-face polyhedron [20]. Since the binding site is discretized and a finite number of geometric transformations are sampled, the performance of the search depends on the resolution of the steps, at the cost of computing time. Finally, Binkowski et al. [113] and RAPMAD [26] compare the distributions of pairwise distances between the pocket features. RAPMAD generates 14 histograms, one for each of the seven pharmacophoric features, considering two centroids. The idea behind these implementations is that similar binding sites exhibit similar sets of distances. However, these methods may suffer from matching redundant distances that do not superpose geometrically. The advantage of fingerprints/histograms is to enable a faster comparison, without the computationally expensive alignment. Still, KRIPO and Binkowski et al. generate an alignment independently of the comparison procedure for visual inspections, with SiteAlign as part of its search procedure.

Finally, the recent regain of interest for deep neural networks on chemical information favored the emergence of data-driven methods for binding site comparison. Typically, binary classification models are created to discriminate between similar and dissimilar pairs of pockets. Site2Vec transforms the features representing a cavity into a fixed-length vector that can feed a random forest classifier [131]. DeepSite, DeepDrug3D, and DeeplyTough discretize the 3D space of the pocket as voxels, and logically train a convolutional neural network (CNN) binary classification model [32,92,128]. Besides the dependency on sufficiently diverse training datasets for a generalized model, these approaches suffer from the interpretability of the predictions. Interestingly, DeepDrug3D exploits the activation map to highlight areas that largely contribute to the classification.

All the above summarized methods use the protein information only for comparison. Provided that a pocket is delimited, they have a larger scope that may reach target deorphanization [36]. When bound ligands are available, comparing the protein–ligand interactions can be an efficient alternative, particularly when the goal is to reproduce existing binding modes. Likewise, dedicated methods aimed at aligning interaction patterns are based on graph alignment or fingerprint matching [138].

3.3. Local Comparison of Protein Cavities

Looking for an average match that maximizes the overlap between entire cavities is not necessarily the right solution to similarity estimation. Local comparison is a popular term, often used to differentiate full protein structural comparison from protein site comparison. Here, we refer to the truly local comparison of protein pockets (Figure 5), i.e., subpockets of approximately 3 to 4 Å radius (for reference, approximately the shortest distance between a chain of four atoms connected by simple bonds).

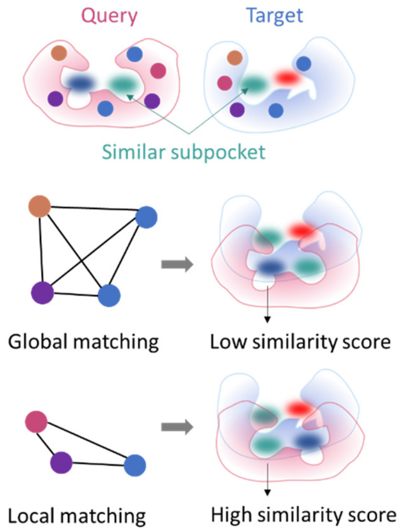


Figure 5. Global versus local pattern comparison.

Enabling local similarity detection is relevant for drug design applications since a few similar subpockets between two targets may suffice for a same ligand to bind. This observation was applied to explain the binding of cyclooxygenase type 2 inhibitors to carbonic anhydrase [22]. Local comparison is notably suitable to handle cases of protein/ligand conformational change upon ligand binding [115].

Logically, methods that can operate locally have implemented detailed site representation and/or adequate algorithms that partition the cavity during the search. In G-LoSA, global matches are decomposed into local subsites to generate other solutions [139]. Local comparison can also be achieved by providing subpockets as input into the search algorithm. KRIPO [21] and ProCare [129] make it possible to compare subpockets delimited by fragmented ligands. While the search algorithms are crucial for identifying zones of similar patterns in two pockets, how these similarities are quantified is equally important, since generalizing the scoring over the full pockets might hinder any local similarity as well. By analogy with ligand versus fragment promiscuity, comparing smaller cavity regions is likely to be more redundant at the proteome scale than comparing full cavities, enabling to catch similarities between remote proteins but at the same time yielding possible unspecific matches that need to be discarded by robust scoring functions to quantify pocket similarity.

3.4. Scoring Functions

Scoring functions serve two purposes: (i) guiding the alignment by discarding unrealistic solutions and prioritizing the best matches, (ii) quantifying the estimated similarity between the pair of pockets to consider. It is not uncommon to use distinct scoring functions for the alignment search and its final quantification [18]. Consequently, a method may implement an accurate representation and efficient search algorithm but may fail to accurately predict similarity levels if the scoring function is incorrectly calibrated. Some analogy can be made with the problem of pose sampling and ranking in docking, leading to rescoring efforts.

Aspects to consider when defining a scoring function for binding site comparison are (i) the discriminative potential, (ii) the minimal and maximal boundaries, (iii) the broadness, (iv) the sensitivity to the size of the cavities, and (v) the interpretability. The very simple and intuitive scoring scheme counts the number of common patterns between two pockets (Brakoulias et al.) [111]. However, bigger pockets may tend to score higher as the chances for a match increase. To avoid this bias, methods account for the size of the pockets using metrics such as the proportion of aligned features with respect to the query/target size (FuzCav, PocketMatch), Tanimoto indices (IsoMIF, KRIPO, TIPSA, Shaper), and Tversky indices (ProCare, Shaper). SiteHopper adopts a linear combination of Tanimoto measures for shape and chemical features matching. Almost all alignment-based geometric matching methods aim at minimizing the root mean square deviation (RMSD) of superposition candidates or with respect to a cutoff (Brakoulias et al., SuMo, etc.). In some cases, the RMSD is also a composite of the final score (Milletti et al., PocketAlign). In the same way, the CavBase R2 score accounts for the RMSD of pseudocenters when scoring the overlap of the surface grid points. Implementing successive scores (Binkowski et al., ProBiS) allows the user to apply a custom filter according to the desired application or suggests a hierarchical scoring. For instance, SiteEngine proposes a workflow wherein a gross evaluation makes it possible to rapidly filter out bad solutions before applying a finer rescoring on promising matches. Instead of reporting similarities, some methods measure the distances between pockets (SiteAlign) instead—the lower, the better. BSSF and RAPMAD, which compare histograms, respectively report the Kolmogorov–Smirnov and the Jensen-Shannon divergences. Scoring functions can be more complex, often at the cost of interpretability (Feldman et al., eMatchSite, P.A.R.I.S).

Weights are used to give more or less importance to different variables (types of features, geometric patterns) but their assignment is at best subjective [119,121,139], intuitive such as inverse of feature frequency, or adapted from sequence alignment methods (BLOSSUM, PSSM) [114,117,140]. Proportioning penalties of mismatches with respect to the positive contributions of the matches (e.g., TrixP) is tricky and may ameliorate or worsen the discrimination performance in context-dependent noisy representations. In fingerprint comparisons, bins are populated with counts or integer descriptors with variable ranges.

The descriptors are normalized [20] or the scores are corrected to account for the increase in activated bits with respect to the size of the cavity [21]. Finally, the commutativity of the score should be regarded to ensure a consistent output whatever the reference/query order.

A few studies [30,47,129,141] have assessed the significance (Z-score, *p*-values) of the scoring function by analyzing random distributions or robustness to variations in the cavities (simulated data, molecular dynamic simulations). While these studies offer a certain overview on possible scoring thresholds in screening settings, we draw attention to potential bias in setting up calibration datasets.

4. Retrospective Evaluations and Datasets

To demonstrate their applicability, the methods for comparing protein binding sites have been evaluated for their ability to (i) discriminate between similar and dissimilar binding sites (classification), (ii) retrieve similar pairs seeded in decoys (enrichment), and (iii) cluster proteins belonging to the same families according to other classifications (e.g., SCOP, functional annotations). The availability of structural data impacts the design of the evaluation datasets.

As for any benchmarking study, the quality of the dataset is instrumental to the reliability of the conclusions. Ligand-based and structure-based virtual screening benefit from well-established standards and datasets [142,143]. Predicting the binding affinity of molecules to a target can be directly verified by experimental measures in many circumstances. Contrarily, pocket similarity cannot be measured experimentally. Instead, similarity prediction suggests hypotheses such as the recognition of similar ligands or the catalysis of the same reaction, which are then confronted with in vitro experiments. What is conveyed here is that there is not a straight line between predictions and verifications since ligand recognition involves other parameters likely not evaluated by binding site comparison methods, such as the pocket flexibility, the influence of disregarded parts of the protein, and the ligand conformations and energetics. Indeed, the ligand may bind to different proteins in different conformations and use different interaction patterns [144].

Nevertheless, many available datasets [24,28,116,128,144–146] have been set-up with the assumption that similar pockets bind to identical or similar ligands, and vice versa (Table 4). These include proteins belonging to the same family for the easiest ones, and unrelated proteins for the most difficult datasets. In these cases, unrelated proteins are predicted by other computational approaches (sequence alignment, global structural comparison). Besides the discussions above, one issue encountered with these definitions is how to set the similarity cutoff to cluster binding sites and ligands.

Chen et al. defines similar pairs as pockets in proteins sharing at least three submicromolar ligands, while dissimilar pairs share at least three ligands large affinity variations going from one target to the other [133]. However, from a medicinal chemistry perspective, this dataset is imbalanced as the number of similar pairs largely exceeds that of dissimilar pairs (Table 4). Still, a main concern is that structural data evidencing that the proposed pair of binding sites effectively accommodating the same ligand are usually missing. Generally, datasets relying on ligand binding information suffer from data incompleteness [147]. Dissimilar pairs are based on limited available/accessible binding information, while all ligands have not been tested against all targets. Otherwise, pairs labeled as 'dissimilar' might have fallen into the 'similar' classes.

Table 4. Common datasets used in benchmarking studies for pocket comparisons.

Purpose	Name	Content	# Positive (# Negatives)
Pairs of cavities from dissimilar proteins binding identical or similar ligands (positives) and dissimilar ligands (negatives)	APoc set [28]	Diverse	38,066 (38,066)
	Barelier et al. [144]	Diverse	62
	Homogeneous [116]	Diverse	100
	Kahraman [146]/extended [116]	Cofactor sites	100/972
	sc-PDB subset [47]	Diverse	1070
	TOUGH-M1 [145]	Diverse	505,116 (556,810)
	TOUGH-C1 [128]	Nucleotides, heme, steroid sites	2218
Pairs of proteins sharing 3 high affinity ligands (potency < 100 nM) vs. pairs of proteins sharing 3 ligands with divergent affinities	Vertex [133]	Diverse	6598 (379)
	Vertex refined [129]	Diverse	338 (338)
Pairs of cavities of associated with the same (positives) or different (negatives) functions and fold class	sc-PDB subset [24]	Diverse	769 (769)
	sc-PDB subset [121]	Diverse	766 (766)
	sc-PDB subset [129]	Diverse	383 (383)
Intra-family classification	Proteases, kinases, GPCRs, Estrogen receptors [17,20,47,115,148]		-
Difficult cases	Difficult cases [19,24]	Diverse from experimental validations	8
Successful applications	ProSPECCTs D7 [38]	Diverse from experimental validations	115 (56,284)
Structures of identical sequences	ProSPECCTs D1 [38]	Diverse	13,430 (92,846)
	ProSPECCTs D1.2 [38]	Diverse	241 (1784)
NMR structures	ProSPECCTs D2 [38]	Diverse	7729 (100,512)
Artificial sets: random mutations	ProSPECCTs D3 and D4 [38]	Diverse	13,430 (67,150)

Given the bias in the PDB dataset towards some protein–cofactors complexes and well-studied protein families, methods have been extensively evaluated on nucleotide-binding pockets [146], although such test cases are quite specific or far too easy to be really predictive of real drug discovery scenarios. Similarly, the capacity of binding site comparison tools to cluster together binding sites originating from the same protein family (e.g., proteases, kinases, or steroid-binding sites) have been widely studied [17,20,115,148]. Alternatively, other datasets proposed pairs of similar and dissimilar sites based on functional annotations [149] or folds [150,151]. Starting from really druggable protein–ligand complexes [152] is often advised in the case of medicinal chemistry applications [20,24,129,138]. Due to the increasing accuracy of deep learning methods [153,154] to predict protein structures with near-atomic resolution, the druggable pocketome is predicted to significantly expand in the next years [155]. Therefore, clear guidelines, as those recently proposed in ProSPECCTs [38], are welcome. Many artificially built datasets are too easy or do not correspond to realistic challenges. Compilations of difficult cases drawn from experimental observations are provided, but such examples are rare [19,24,121].

5. Prospective Applications

The best possible validation method of any binding site comparison tool is indeed to experiment. True prospective validations (Table 5) are still rare for several reasons:

- Fragment/ligand promiscuity towards unrelated targets of known 3D structure remains are a rare event [156];
- Direct drug repurposing from in silico [157] or in vitro screening strategies have not yet yielded any success in terms of new indication approvals [158], as recently exemplified by the COVID-19 pandemic;
- The experimental validation of putative binding site similarities is not as straightforward as testing many compounds on a single target. For every putative off-target, a suitable assay has to be used if available, or more likely needs to be developed on purpose. In vitro biophysical assays (e.g., NMR, thermal shift) give a direct answer of shared ligand binding to two different targets [159,160] but do not necessarily evidence the binding site location, by opposition to enzymatic assays [40,161–164] or binding competitions experiments for which the binding site is usually unambiguous [165–167]. If not possible otherwise, functional and/or in vivo assays [168,169] can be used but are more difficult to interpret since the examined function might be biased by binding to another target.

  Known success stories (Table 5) have notably enabled:
- The explanation of target-mediated side effects and guidelines to optimize the ligand selectivity by suitable structural modifications [168,169];
- The explanation of off-target beneficial effects [165];
- The validation of cross-docking data for repurposing hypotheses [162,166];
- The confirmation of ligand 2D and 3D shape similarities [164];
- The serendipitous discovery of remote similarities across totally unrelated targets during code benchmarking and validation [159,167].

The above-cited examples share common characteristics. First, the repurposed ligands usually exhibit (very) weak affinities towards the secondary target, notably when the on- and off-targets are unrelated. In all cases, the studied ligand needs to be optimized for potency and selectivity towards the secondary target, thereby abolishing the benefits of immediate in silico-guided drug repurposing [159–167,169]. Second, and in relation to the first observation, the noticed pocket similarity is usually local and not global. In other words, only the subpockets of the two targets under investigation account for the shared ligand binding. This explains why some targets, notably those exhibiting hydrophobic subpockets (COX-1, HIV-1 RT, PPARγ, ER-α) are frequently observed among the protein pairs cited below (Table 5). The conservation of shared polar and apolar pocket features is a rarer event but leads to higher affinities of the corresponding complexes [164,167].

Table 5. Examples of small molecular weight ligand-binding site comparisons relevant to medicinal chemistry.

Method	On-Target	Secondary Target	Ligand	Secondary Target Affinity	Ref.
SOIPPA	Estrogen receptor alpha	SERCA Ca^{2+} ion channel ATPase	Tamoxifen	IC_{50} = 5 µM	[168]
CPASS	Bcl-2 apoptosis protein Bcl-xL	Type III SS Needle Protein (PrgI)	Chelerythrine	N/A [a]	[160]
SOIPPA	Catechol-O-methyltransferase	Enoyl-acyl carrier protein reductase	Entacapone	IC_{50} = 80 µM	[162]
SiteAlign	Pim-1 kinase	Synapsin I	Quercetagetin	IC_{50} = 0.15 µM	[167]
SMAP	HIV-1 protease	ErbB2 receptor tyrosine kinase	Nelfinavir	N/A [b]	[163]

Table 5. *Cont.*

Method	On-Target	Secondary Target	Ligand	Secondary Target Affinity	Ref.
PSSC	Monoamine oxidase	Lysine-specific demethylase 1	Namoline	IC_{50} = 51 μM	[166]
SMAP	Epidermal growth factor	β-secretase	Gefitinib	IC_{50} = 20 μM	[165]
KRIPO	Cannabinoid type 1 receptor	Adenine nucleotide translocase 1	Ibipinabant	N/A [c]	[169]
PSIM	PPAR gamma	Cyclooxygenase type 1	Fenofibrate	IC_{50} = 950 μM	[161]
TM-align	Receptor Tyrosine kinases	Acetylcholinesterase	Pazopanib Sunitinib	IC_{50} = 0.93 μM IC_{50} = 5.87 μM	[164]
Shaper	Cyclooxygenase type 1	Cinnamoylesterase	Flurbiprofen	IC_{50} = 400 μM	[40]
ProCare	HIV-1 reverse transcriptase	TNF-α trimer	Efavirenz Delavirdine	Kd = 24 μM Kd = 49 μM	[159]

[a] binding evidenced by ^{15}N-^{1}H NMR-HSQC spectra; [b] 15% inhibition at 20 μM in a kinase activity assay; [c] 30% Inhibition of ANT-dependent mitochondrial ADP uptake at a concentration of 100 μM.

6. Conclusions

This review presents the current state of ligand-binding site comparison applied to small molecule drug design. As computer-aided drug design strategies, assessing the similarity of protein pockets constitutes a unique way to analyze structural information, as they complement other well-spread approaches. The repertoire of available methods is diverse with respect to the detection and representation of cavities, the search algorithms, and the scoring functions. All of these aspects must somehow be coordinated to achieve the best performance. Still, the limitation of experimental data and biases in datasets represent major obstacles to properly evaluate such methods. In reality, estimating protein site similarity is always context-dependent. The importance of matched features is influenced by the chemical context and physicochemical considerations of the targets, making it hard to predict subtle and specific similarities from generalized principles.

One holy grail of computational chemists is to repurpose existing drugs proposed by structure-based experiments. Although this pursuit appears at best hardly probable due to the optimization of drugs for their on-targets [157,158], we believe that binding site comparisons are the most useful in finding not global but local similarities, and therefore to repurpose fragments [22] and not full ligands, provided that the selected fragments can be grown or linked to enumerate full ligands or target-focused libraries [170].

Binding sites comparisons have demonstrated an effective contribution to medicinal chemistry projects, from the elucidation of previous biological observations to the generation of new hypotheses supported by experimental validation. The majority of the state-of-the-art methods are based on the superposition of the compared structures. The alignment allows for a visual inspection and increases the possibilities of applications. Typically, pocket-bound ligands in the reference frame can be transposed into the target pocket and serve as a starting point for ligand generation. The improvement of the algorithmic efficiency of the methods alongside with technological progress may enable to better follow the current growth of publicly available protein structures, determined experimentally or predicted at near-atomic resolution [171].

Author Contributions: Conceptualization, M.E. and D.R.; validation, M.E. and D.R.; writing—original draft preparation, M.E.; writing—review and editing, D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Illergard, K.; Ardell, D.H.; Elofsson, A. Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins* **2009**, *77*, 499–508. [CrossRef] [PubMed]
2. McCoy, A.J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr. D Biol. Crystallogr.* **2007**, *63 Pt 1*, 32–41. [CrossRef] [PubMed]
3. Cavalli, A.; Salvatella, X.; Dobson, C.M.; Vendruscolo, M. Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 9615–9620. [CrossRef] [PubMed]
4. Renaud, J.P.; Chari, A.; Ciferri, C.; Liu, W.T.; Remigy, H.W.; Stark, H.; Wiesmann, C. Cryo-EM in drug discovery: Achievements, limitations and prospects. *Nat. Rev. Drug Discov.* **2018**, *17*, 471–492. [CrossRef] [PubMed]
5. Shimada, I.; Ueda, T.; Kofuku, Y.; Eddy, M.T.; Wuthrich, K. GPCR drug discovery: Integrating solution NMR data with crystal and cryo-EM structures. *Nat. Rev. Drug Discov.* **2019**, *18*, 59–82. [CrossRef] [PubMed]
6. Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7*, 1884–1897. [CrossRef] [PubMed]
7. Burley, S.K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G.V.; Christie, C.H.; Dalenberg, K.; Di Costanzo, L.; Duarte, J.M.; et al. RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **2021**, *49*, D437–D451. [CrossRef] [PubMed]
8. Nussinov, R.; Wolfson, H.J. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 10495–10499. [CrossRef]
9. Russell, R.B.; Sasieni, P.D.; Sternberg, M.J. Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **1998**, *282*, 903–918. [CrossRef] [PubMed]
10. Fischer, D.; Wolfson, H.; Lin, S.L.; Nussinov, R. Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: Potential implications to evolution and to protein folding. *Protein Sci.* **1994**, *3*, 769–778. [CrossRef] [PubMed]
11. Wallace, A.C.; Borkakoti, N.; Thornton, J.M. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **1997**, *6*, 2308–2323. [CrossRef] [PubMed]
12. Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **1997**, *15*, 359–363. [CrossRef]
13. Levitt, D.G.; Banaszak, L.J. Pocket—A Computer-Graphics Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino-Acids. *J. Mol. Graph. Model.* **1992**, *10*, 229–234. [CrossRef]
14. Volkamer, A.; Kuhn, D.; Rippmann, F.; Rarey, M. DoGSiteScorer: A web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics* **2012**, *28*, 2074–2075. [CrossRef] [PubMed]
15. Krivak, R.; Hoksza, D. P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminform.* **2018**, *10*, 39. [CrossRef] [PubMed]
16. Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinform.* **2009**, *10*, 168. [CrossRef] [PubMed]
17. 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. [CrossRef]
18. Shulman-Peleg, A.; Nussinov, R.; Wolfson, H.J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607–633. [CrossRef]
19. Yeturu, K.; Chandra, N. PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinform.* **2008**, *9*, 543. [CrossRef]
20. Schalon, C.; Surgand, J.S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71*, 1755–1778. [CrossRef] [PubMed]
21. 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. [CrossRef] [PubMed]
22. Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C.T.; 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. [CrossRef] [PubMed]
23. Ehrt, C.; Brinkjost, T.; Koch, O. Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. *J. Med. Chem.* **2016**, *59*, 4121–4151. [CrossRef] [PubMed]
24. Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123–135. [CrossRef] [PubMed]
25. Xiong, B.; Wu, J.; Burk, D.L.; Xue, M.; Jiang, H.; Shen, J. BSSF: A fingerprint based ultrafast binding site similarity search and function analysis server. *BMC Bioinform.* **2010**, *11*, 47. [CrossRef]
26. 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. [CrossRef]
27. Kellenberger, E.; Schalon, C.; Rognan, D. How to Measure the Similarity Between Protein Ligand-Binding Sites? *Curr. Comput. Aided Drug Des.* **2008**, *4*, 209–220. [CrossRef]

28. Gao, M.; Skolnick, J. APoc: Large-scale identification of similar protein pockets. *Bioinformatics* **2013**, *29*, 597–604. [CrossRef]
29. Brylinski, M. eMatchSite: Sequence Order-Independent Structure Alignments of Ligand Binding Pockets in Protein Models. *PLoS Comput. Biol.* **2014**, *10*, e1003829. [CrossRef]
30. Feldman, H.J.; Labute, P. Pocket Similarity: Are alpha Carbons Enough? *J. Chem. Inf. Model.* **2010**, *50*, 1466–1475. [CrossRef]
31. Konc, J.; Janezic, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26*, 1160–1168. [CrossRef] [PubMed]
32. Simonovsky, M.; Meyers, J. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *J. Chem. Inf. Model.* **2020**, *60*, 2356–2366. [CrossRef] [PubMed]
33. Xie, L.; Xie, L.; Bourne, P.E. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics* **2009**, *25*, i305–i312. [CrossRef]
34. Levitt, M.; Gerstein, M. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 5913–5920. [CrossRef] [PubMed]
35. Bhagavat, R.; Sankar, S.; Srinivasan, N.; Chandra, N. An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure. *Structure* **2018**, *26*, 499–512.e2. [CrossRef] [PubMed]
36. Rognan, D. Structure-Based Approaches to Target Fishing and Ligand Profiling. *Mol. Inform.* **2010**, *29*, 176–187. [CrossRef] [PubMed]
37. Nisius, B.; Sha, F.; Gohlke, H. Structure-based computational analysis of protein binding sites for function and druggability prediction. *J. Biotechnol.* **2012**, *159*, 123–134. [CrossRef]
38. Ehrt, C.; Brinkjost, T.; Koch, O. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). *PLoS Comput. Biol.* **2018**, *14*, e1006483. [CrossRef]
39. Naderi, M.; Lemoine, J.M.; Govindaraj, R.G.; Kana, O.Z.; Feinstein, W.P.; Brylinski, M. Binding site matching in rational drug design: Algorithms and applications. *Brief. Bioinform.* **2019**, *20*, 2167–2184. [CrossRef]
40. Da Silva, F.; Bret, G.; Teixeira, L.; Gonzalez, C.F.; Rognan, D. Exhaustive Repertoire of Druggable Cavities at Protein-Protein Interfaces of Known Three-Dimensional Structure. *J. Med. Chem.* **2019**, *62*, 9732–9742. [CrossRef]
41. McGreig, J.E.; Uri, H.; Antczak, M.; Sternberg, M.J.E.; Michaelis, M.; Wass, M.N. 3DLigandSite: Structure-based prediction of protein-ligand binding sites. *Nucleic Acids Res.* **2022**, *50*, W13–W20. [CrossRef] [PubMed]
42. Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. sc-PDB: A 3D-database of ligandable binding sites-10 years on. *Nucleic Acids Res.* **2015**, *43*, D399–D404. [CrossRef] [PubMed]
43. Ben Chorin, A.; Masrati, G.; Kessel, A.; Narunsky, A.; Sprinzak, J.; Lahav, S.; Ashkenazy, H.; Ben-Tal, N. ConSurf-DB: An accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. *Protein Sci.* **2020**, *29*, 258–267. [CrossRef] [PubMed]
44. Fogha, J.; Diharce, J.; Obled, A.; Aci-Seche, S.; Bonnet, P. Computational Analysis of Crystallization Additives for the Identification of New Allosteric Sites. *ACS Omega* **2020**, *5*, 2114–2122. [CrossRef] [PubMed]
45. Drwal, M.N.; Jacquemard, C.; Perez, C.; Desaphy, J.; Kellenberger, E. Do Fragments and Crystallization Additives Bind Similarly to Drug like Ligands? *J. Chem. Inf. Model.* **2017**, *57*, 1197–1209. [CrossRef]
46. Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem* **2018**, *13*, 507–510. [CrossRef]
47. 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. [CrossRef]
48. Oliveira, S.H.P.; Ferraz, F.A.N.; Honorato, R.V.; Xavier-Neto, J.; Sobreira, T.J.P.; de Oliveira, P.S.L. KVFinder: Steered identification of protein cavities as a PyMOL plugin. *BMC Bioinform.* **2014**, *15*, 197. [CrossRef]
49. Marchand, J.R.; Pirard, B.; Ertl, P.; Sirockin, F. CAVIAR: A method for automatic cavity detection, description and decomposition into subcavities. *J. Comput.-Aided Mol. Des.* **2021**, *35*, 737–750. [CrossRef]
50. Fathi, S.M.S.; Tuszynski, J.A. A simple method for finding a protein's ligand-binding pockets. *BMC Struct. Biol.* **2014**, *14*, 18. [CrossRef]
51. Petrek, M.; Otyepka, M.; Banas, P.; Kosinova, P.; Koca, J.; Damborsky, J. CAVER: A new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinform.* **2006**, *7*, 316. [CrossRef] [PubMed]
52. Kleywegt, G.J.; Jones, T.A. Detection, Delineation, Measurement and Display of Cavities in Macromolecular Structures. *Acta Crystallogr. Sect. D* **1994**, *50*, 178–185. [CrossRef] [PubMed]
53. Huang, B.D.; Schroeder, M. LIGSITE(csc): Predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6*, 19. [CrossRef]
54. Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: Analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **2007**, *1*, 7. [CrossRef]
55. Kalidas, Y.; Chandra, N. PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins. *J. Struct. Biol.* **2008**, *161*, 31–42. [CrossRef] [PubMed]
56. Tripathi, A.; Kellogg, G.E. A novel and efficient tool for locating and characterizing protein cavities and binding sites. *Proteins* **2010**, *78*, 825–842. [CrossRef]
57. Kawabata, T. Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins* **2010**, *78*, 1195–1211. [CrossRef]

58. Till, M.S.; Ullmann, G.M. McVol—A program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm. *J. Mol. Model.* **2010**, *16*, 419–429. [CrossRef]
59. Peters, K.P.; Fauck, J.; Frommel, C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **1996**, *256*, 201–213. [CrossRef]
60. Binkowski, T.A.; Naghibzadeh, S.; Liang, J. CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res.* **2003**, *31*, 3352–3355. [CrossRef]
61. Tan, K.P.; Varadarajan, R.; Madhusudhan, M.S. DEPTH: A web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Res.* **2011**, *39* (Suppl. 2), W242–W248. [CrossRef] [PubMed]
62. Smart, O.S.; Neduvelil, J.G.; Wang, X.; Wallace, B.A.; Sansom, M.S.P. HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph. Model.* **1996**, *14*, 354–360. [CrossRef]
63. Ho, B.K.; Gruswitz, F. HOLLOW: Generating accurate representations of channel and interior surfaces in molecular structures. *BMC Struct. Biol.* **2008**, *8*, 49. [CrossRef] [PubMed]
64. Brady, G.P.; Stouten, P.F.W. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383–401. [CrossRef]
65. Yu, J.; Zhou, Y.; Tanaka, I.; Yao, M. Roll: A new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics* **2010**, *26*, 46–52. [CrossRef] [PubMed]
66. Laskowski, R.A. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* **1995**, *13*, 323–330. [CrossRef]
67. Glaser, F.; Morris, R.J.; Najmanovich, R.J.; Laskowski, R.A.; Thornton, J.M. A method for localizing ligand binding pockets in protein structures. *Proteins* **2006**, *62*, 479–488. [CrossRef] [PubMed]
68. Zhu, H.; Pisabarro, M.T. MSPocket: An orientation-independent algorithm for the detection of ligand binding pockets. *Bioinformatics* **2011**, *27*, 351–358. [CrossRef]
69. Tseng, Y.Y.; Dupree, C.; Chen, Z.J.; Li, W.H. SplitPocket: Identification of protein functional surfaces and characterization of their spatial patterns. *Nucleic Acids Res.* **2009**, *37*, W384–W389. [CrossRef]
70. Harris, R.; Olson, A.J.; Goodsell, D.S. Automated prediction of ligand-binding sites in proteins. *Proteins* **2008**, *70*, 1506–1517. [CrossRef]
71. An, J.; Totrov, M.; Abagyan, R. Comprehensive identification of "druggable" protein ligand binding sites. *Genome Inform.* **2004**, *15*, 31–41. [PubMed]
72. Ngan, C.H.; Hall, D.R.; Zerbe, B.; Grove, L.E.; Kozakov, D.; Vajda, S. FTSite: High accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics* **2012**, *28*, 286–287. [CrossRef] [PubMed]
73. An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteom.* **2005**, *4*, 752–761. [CrossRef] [PubMed]
74. Laurie, A.T.; Jackson, R.M. Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **2005**, *21*, 1908–1916. [CrossRef]
75. Ghersi, D.; Sanchez, R. EasyMIFS and SiteHound: A toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics* **2009**, *25*, 3185–3186. [CrossRef]
76. Halgren, T.A. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377–389. [CrossRef]
77. Ruppert, J.; Welch, W.; Jain, A.N. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.* **1997**, *6*, 524–533. [CrossRef]
78. Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857. [CrossRef]
79. Schneider, S.; Zacharias, M. Combining geometric pocket detection and desolvation properties to detect putative ligand binding sites on proteins. *J. Struct. Biol.* **2012**, *180*, 546–550. [CrossRef]
80. Morita, M.; Nakamura, S.; Shimizu, K. Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. *Proteins* **2008**, *73*, 468–479. [CrossRef]
81. Tuzmen, C.; Erman, B. Identification of Ligand Binding Sites of Proteins Using the Gaussian Network Model. *PLoS ONE* **2011**, *6*, e16474. [CrossRef] [PubMed]
82. Santana, C.A.; Silveira, S.D.; Moraes, J.P.A.; Izidoro, S.C.; de Melo-Minardi, R.C.; Ribeiro, A.J.M.; Tyzack, J.D.; Borkakoti, N.; Thornton, J.M. GRaSP: A graph-based residue neighborhood strategy to predict binding sites. *Bioinformatics* **2020**, *36*, I726–I734. [CrossRef] [PubMed]
83. Wong, G.Y.; Leung, F.H.F.; Ling, S.S.H. Identification of Protein-Ligand Binding Site Using Multi-Clustering and Support Vector Machine. In Proceedings of the IECON 2016—42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, Italy, 24–27 October 2016; pp. 939–944.
84. Krivak, R.; Hoksza, D. Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features. *J. Cheminform.* **2015**, *7*, 12. [CrossRef] [PubMed]
85. Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins* **2006**, *63*, 892–906. [CrossRef]

86. Yan, X.; Lu, Y.F.; Li, Z.; Wei, Q.; Gao, X.; Wang, S.; Wu, S.; Cui, S.G. PointSite: A Point Cloud Segmentation Tool for Identification of Protein Ligand Binding Atoms. *J. Chem. Inf. Model.* **2022**, *62*, 2835–2845. [CrossRef] [PubMed]
87. Aggarwal, R.; Gupta, A.; Chelur, V.; Jawahar, C.V.; Priyakumar, U.D. DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks. *J. Chem. Inf. Model.* **2021**. [CrossRef] [PubMed]
88. Kandel, J.; Tayara, H.; Chong, K.T. PUResNet: Prediction of protein-ligand binding sites using deep residual neural network. *J. Cheminform.* **2021**, *13*, 65. [CrossRef]
89. Mylonas, S.K.; Axenopoulos, A.; Daras, P. DeepSurf: A surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics* **2021**, *37*, 1681–1690. [CrossRef]
90. Kozlovskii, I.; Popov, P. Spatiotemporal identification of druggable binding sites using deep learning. *Commun. Biol.* **2020**, *3*, 618. [CrossRef]
91. Jiang, M.; Li, Z.; Bian, Y.; Wei, Z. A novel protein descriptor for the prediction of drug binding sites. *BMC Bioinform.* **2019**, *20*, 478. [CrossRef]
92. Jimenez, J.; Doerr, S.; Martinez-Rosell, G.; Rose, A.S.; De Fabritiis, G. DeepSite: Protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **2017**, *33*, 3036–3042. [CrossRef] [PubMed]
93. Edelsbrunner, H.; Kirkpatrick, D.G.; Seidel, R. On the Shape of a Set of Points in the Plane. *IEEE Trans. Inf. Theory* **1983**, *29*, 551–559. [CrossRef]
94. Kawabata, T.; Go, N. Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins* **2007**, *68*, 516–529. [CrossRef]
95. Xie, L.; Bourne, P.E. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinform.* **2007**, *8*, S9. [CrossRef] [PubMed]
96. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
97. Qi, C.R.; Su, H.; Mo, K.C.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 77–85. [CrossRef]
98. Degac, J.; Winter, U.; Helms, V. Graph-Based Clustering of Predicted Ligand-Binding Pockets on Protein Surfaces. *J. Chem. Inf. Model.* **2015**, *55*, 1944–1952. [CrossRef]
99. Zhang, Z.M.; Li, Y.; Lin, B.Y.; Schroeder, M.; Huang, B.D. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* **2011**, *27*, 2083–2088. [CrossRef]
100. Huang, B. MetaPocket: A meta approach to improve protein ligand binding site prediction. *OMICS* **2009**, *13*, 325–330. [CrossRef]
101. Hajduk, P.J.; Huth, J.R.; Tse, C. Predicting protein druggability. *Drug Discov. Today* **2005**, *10*, 1675–1682. [CrossRef]
102. Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.* **2012**, *52*, 360–372. [CrossRef]
103. Perola, E.; Herman, L.; Weiss, J. Development of a rule-based method for the assessment of protein druggability. *J. Chem. Inf. Model.* **2012**, *52*, 1027–1038. [CrossRef] [PubMed]
104. Sheridan, R.P.; Maiorov, V.N.; Holloway, M.K.; Cornell, W.D.; Gao, Y.D. Drug-like Density: A Method of Quantifying the "Bindability" of a Protein Target Based on a Very Large Set of Pockets and Drug-like Ligands from the Protein Data Bank. *J. Chem. Inf. Model.* **2010**, *50*, 2029–2040. [CrossRef] [PubMed]
105. Krasowski, A.; Muthas, D.; Sarkar, A.; Schmitt, S.; Brenk, R. DrugPred: A Structure-Based Approach To Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *J. Chem. Inf. Model.* **2011**, *51*, 2829–2842. [CrossRef] [PubMed]
106. Borrel, A.; Regad, L.; Xhaard, H.; Petitjean, M.; Camproux, A.C. PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties. *J. Chem. Inf. Model.* **2015**, *55*, 882–895. [CrossRef] [PubMed]
107. Edfeldt, F.N.; Folmer, R.H.; Breeze, A.L. Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discov. Today* **2011**, *16*, 284–287. [CrossRef]
108. Abi Hussein, H.; Geneix, C.; Petitjean, M.; Borrel, A.; Flatters, D.; Camproux, A.C. Global vision of druggability issues: Applications and perspectives. *Drug Discov. Today* **2017**, *22*, 404–415. [CrossRef]
109. Kinoshita, K.; Furui, J.; Nakamura, H. Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics* **2002**, *2*, 9–22. [CrossRef]
110. Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137–145. [CrossRef]
111. Brakoulias, A.; Jackson, R.M. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: An automated all-against-all structural comparison using geometric matching. *Proteins* **2004**, *56*, 250–260. [CrossRef]
112. Ramensky, V.; Sobol, A.; Zaitseva, N.; Rubinov, A.; Zosimov, V. A novel approach to local similarity of protein binding sites substantially improves computational drug design results. *Proteins* **2007**, *69*, 349–357. [CrossRef]
113. Binkowski, T.A.; Joachimiak, A. Protein functional surfaces: Global shape matching and local spatial alignments of ligand binding sites. *BMC Struct. Biol.* **2008**, *8*, 45. [CrossRef] [PubMed]
114. Xie, L.; Bourne, P.E. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 5441–5446. [CrossRef] [PubMed]

115. Milletti, F.; Vulpetti, A. Predicting polypharmacology by binding site similarity: From kinases to the protein universe. *J. Chem. Inf. Model.* **2010**, *50*, 1418–1431. [CrossRef] [PubMed]
116. 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 Bioinform.* **2010**, *11*, 99. [CrossRef]
117. Yeturu, K.; Chandra, N. PocketAlign a novel algorithm for aligning binding sites in protein structures. *J. Chem. Inf. Model.* **2011**, *51*, 1725–1736. [CrossRef]
118. Liu, T.Y.; Altman, R.B. Using Multiple Microenvironments to Find Similar Ligand-Binding Sites: Application to Kinase Inhibitor Binding. *PLoS Comput. Biol.* **2011**, *7*, e1002326. [CrossRef]
119. Sael, L.; Kihara, D. Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins* **2012**, *80*, 1177–1195. [CrossRef]
120. Ellingson, L.; Zhang, J. Protein surface matching by combining local and global geometric information. *PLoS ONE* **2012**, *7*, e40540. [CrossRef]
121. von Behren, M.M.; Volkamer, A.; Henzler, A.M.; Schomburg, K.T.; Urbaczek, S.; Rarey, M. Fast protein binding site comparison via an index-based screening technology. *J. Chem. Inf. Model.* **2013**, *53*, 411–422. [CrossRef]
122. Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the topology of active sites: On the prediction of pockets and subpockets. *J. Chem. Inf. Model.* **2010**, *50*, 2041–2052. [CrossRef]
123. Brylinski, M.; Feinstein, W.P. eFindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 551–567. [CrossRef]
124. Chartier, M.; Najmanovich, R. Detection of Binding Site Molecular Interaction Field Similarities. *J. Chem. Inf. Model.* **2015**, *55*, 1600–1615. [CrossRef] [PubMed]
125. Gaudreault, F.; Morency, L.P.; Najmanovich, R.J. NRGsuite: A PyMOL plugin to perform docking simulations in real time using FlexAID. *Bioinformatics* **2015**, *31*, 3856–3858. [CrossRef] [PubMed]
126. Lee, H.S.; Im, W. G-LoSA: An efficient computational tool for local structure-centric biological studies and drug design. *Protein Sci.* **2016**, *25*, 865–876. [CrossRef]
127. Batista, J.; Hawkins, P.C.D.; Tolbert, R.; Geballe, M.T. SiteHopper—A unique tool for binding site comparison. *J. Cheminform.* **2014**, *6*, P57. [CrossRef]
128. Pu, L.; Govindaraj, R.G.; Lemoine, J.M.; Wu, H.C.; Brylinski, M. DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network. *PLoS Comput. Biol.* **2019**, *15*, e1006718. [CrossRef] [PubMed]
129. Eguida, M.; Rognan, D. A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design. *J. Med. Chem.* **2020**, *63*, 7127–7142. [CrossRef]
130. Li, S.; Cai, C.; Gong, J.; Liu, X.; Li, H. A fast protein binding site comparison algorithm for proteome-wide protein function prediction and drug repurposing. *Proteins* **2021**, *89*, 1541–1556. [CrossRef]
131. Bhadra, A.; Yeturu, K. Site2Vec: A reference frame invariant algorithm for vector embedding of protein-ligand binding sites. *Mach. Learn. Sci. Technol.* **2020**, *2*, 015005. [CrossRef]
132. Haupt, V.J.; Daminelli, S.; Schroeder, M. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLoS ONE* **2013**, *8*, e65894. [CrossRef]
133. Chen, Y.C.; Tolbert, R.; Aronov, A.M.; McGaughey, G.; Walters, W.P.; Meireles, L. Prediction of Protein Pairs Sharing Common Active Ligands Using Protein Sequence, Structure, and Ligand Similarity. *J. Chem. Inf. Model.* **2016**, *56*, 1734–1745. [CrossRef]
134. Fischler, M.A.; Bolles, R.C. Random Sample Consensus—A Paradigm for Model-Fitting with Applications to Image-Analysis and Automated Cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
135. Besl, P.J.; Mckay, N.D. A Method for Registration of 3-D Shapes. *IEEE Trans. Pattern Anal.* **1992**, *14*, 239–256. [CrossRef]
136. Bron, C.; Kerbosch, J. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun. ACM* **1973**, *16*, 575–577. [CrossRef]
137. Johnston, H.C. Cliques of a Graph-Variations on the Bron-Kerbosch Algorithm. *Int. J. Comput. Inf. Sci.* **1976**, *5*, 209–238. [CrossRef]
138. Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding protein-ligand interaction patterns in fingerprints and graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623–637. [CrossRef] [PubMed]
139. Lee, H.S.; Im, W. G-LoSA for Prediction of Protein-Ligand Binding Sites and Structures. *Methods Mol. Biol.* **2017**, *1611*, 97–108. [CrossRef] [PubMed]
140. Henikoff, S.; Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 10915–10919. [CrossRef] [PubMed]
141. Davies, J.R.; Jackson, R.M.; Mardia, K.V.; Taylor, C.C. The Poisson Index: A new probabilistic model for proteinligand binding site similarity. *Bioinformatics* **2007**, *23*, 3001–3008. [CrossRef]
142. Mysinger, M.M.; Carchia, M.; Irwin, J.J.; Shoichet, B.K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594. [CrossRef]
143. Tran-Nguyen, V.K.; Jacquemard, C.; Rognan, D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60*, 4263–4273. [CrossRef] [PubMed]
144. Barelier, S.; Sterling, T.; O'Meara, M.J.; Shoichet, B.K. The Recognition of Identical Ligands by Unrelated Proteins. *ACS Chem. Biol.* **2015**, *10*, 2772–2784. [CrossRef] [PubMed]
145. Govindaraj, R.G.; Brylinski, M. Comparative assessment of strategies to identify similar ligand-binding pockets in proteins. *BMC Bioinform.* **2018**, *19*, 91. [CrossRef] [PubMed]

146. Kahraman, A.; Morris, R.J.; Laskowski, R.A.; Thornton, J.M. Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.* **2007**, *368*, 283–301. [CrossRef]
147. Mestres, J.; Gregori-Puigjane, E.; Valverde, S.; Sole, R.V. Data completeness—The Achilles heel of drug-target networks. *Nat. Biotechnol.* **2008**, *26*, 983–984. [CrossRef] [PubMed]
148. Kuhn, D.; Weskamp, N.; Hullermeier, E.; Klebe, G. Functional classification of protein kinase binding sites using cavbase. *Chemmedchem* **2007**, *2*, 1432–1447. [CrossRef] [PubMed]
149. UniProt, C. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515. [CrossRef]
150. Andreeva, A.; Kulesha, E.; Gough, J.; Murzin, A.G. The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* **2020**, *48*, D376–D382. [CrossRef]
151. Sillitoe, I.; Bordin, N.; Dawson, N.; Waman, V.P.; Ashford, P.; Scholes, H.M.; Pang, C.S.M.; Woodridge, L.; Rauer, C.; Sen, N.; et al. CATH: Increased structural coverage of functional space. *Nucleic Acids Res.* **2021**, *49*, D266–D273. [CrossRef]
152. sc-PDB: An Annotated Database of Druggable Binding Sites from the Protein DataBank. Available online: http://bioinfo-pharma.u-strasbg.fr/scPDB/ (accessed on 15 October 2022).
153. Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Zidek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; et al. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590–596. [CrossRef]
154. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876. [CrossRef]
155. Wang, S.; Lin, H.; Huang, Z.; He, Y.; Deng, X.; Xu, Y.; Pei, J.; Lai, L. CavitySpace: A Database of Potential Ligand Binding Sites in the Human Proteome. *Biomolecules* **2022**, *12*, 967. [CrossRef]
156. Kalliokoski, T.; Olsson, T.S.; Vulpetti, A. Subpocket analysis method for fragment-based drug discovery. *J. Chem. Inf. Model.* **2013**, *53*, 131–141. [CrossRef] [PubMed]
157. Lewis, R.A. Best practices for repurposing studies. *J. Comput. Aided Mol. Des.* **2021**, *35*, 1189–1193. [CrossRef] [PubMed]
158. Edwards, A. What Are the Odds of Finding a COVID-19 Drug from a Lab Repurposing Screen? *J. Chem. Inf. Model.* **2020**, *60*, 5727–5729. [CrossRef]
159. Eguida, M.; Rognan, D. Unexpected similarity between HIV-1 reverse transcriptase and tumor necrosis factor binding sites revealed by computer vision. *J. Cheminform.* **2021**, *13*, 90. [CrossRef] [PubMed]
160. Shortridge, M.D.; Powers, R. Structural and Functional Similarity between the Bacterial Type III Secretion System Needle Protein PrgI and the Eukaryotic Apoptosis Bcl-2 Proteins. *PLoS ONE* **2009**, *4*, e7442. [CrossRef]
161. Cleves, A.E.; Jain, A.N. Chemical and protein structural basis for biological crosstalk between PPAR alpha and COX enzymes. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 101–112. [CrossRef]
162. Kinnings, S.L.; Liu, N.; Buchmeier, N.; Tonge, P.J.; Xie, L.; Bourne, P.E. Drug Discovery Using Chemical Systems Biology: Repositioning the Safe Medicine Comtan to Treat Multi-Drug and Extensively Drug Resistant Tuberculosis. *PLoS Comput. Biol.* **2009**, *5*, e1000423. [CrossRef]
163. Xie, L.; Evangelidis, T.; Xie, L.; Bourne, P.E. Drug Discovery Using Chemical Systems Biology: Weak Inhibition of Multiple Kinases May Contribute to the Anti-Cancer Effect of Nelfinavir. *PLoS Comput. Biol.* **2011**, *7*, e1002037. [CrossRef] [PubMed]
164. Yang, Y.; Li, G.; Zhao, D.; Yu, H.; Zheng, X.; Peng, X.; Zhang, X.; Fu, T.; Hu, X.; Niu, M.; et al. Computational discovery and experimental verification of tyrosine kinase inhibitor pazopanib for the reversal of memory and cognitive deficits in rat model neurodegeneration. *Chem. Sci.* **2015**, *6*, 2812–2821. [CrossRef] [PubMed]
165. Niu, M.S.; Hu, J.; Wu, S.J.; Zhang, X.E.; Xu, H.X.; Zhang, Y.W.; Zhang, J.; Yang, Y.L. Structural Bioinformatics-Based Identification of EGFR Inhibitor Gefitinib as a Putative Lead Compound for BACE. *Chem. Biol. Drug. Des.* **2014**, *83*, 81–88. [CrossRef] [PubMed]
166. Willmann, D.; Lim, S.; Wetzel, S.; Metzger, E.; Jandausch, A.; Wilk, W.; Jung, M.; Forne, I.; Imhof, A.; Janzer, A.; et al. Impairment of prostate cancer cell growth by a selective and reversible lysine-specific demethylase 1 inhibitor. *Int. J. Cancer* **2012**, *131*, 2704–2709. [CrossRef] [PubMed]
167. De Franchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of Protein Kinase Inhibitors to Synapsin I Inferred from Pair-Wise Binding Site Similarity Measurements. *PLoS ONE* **2010**, *5*, e12214. [CrossRef]
168. Xie, L.; Wang, J.; Bourne, P.E. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comput. Biol.* **2007**, *3*, e217. [CrossRef] [PubMed]
169. Schirris, T.J.; Ritschel, T.; Herma Renkema, G.; Willems, P.H.; Smeitink, J.A.; Russel, F.G. Mitochondrial ADP/ATP exchange inhibition: A novel off-target mechanism underlying ibipinabant-induced myotoxicity. *Sci. Rep.* **2015**, *5*, 14533. [CrossRef] [PubMed]
170. Eguida, M.; Rognan, D. Fragment-based and pocket-focused library design by protein-applied computer vision and deep generative linking. *J. Med. Chem.* 2022, *in press*. [CrossRef]
171. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]