Genomics Proteomics Bioinformatics 20 (2022) 765–779

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Defining A Global Map of Functional Group-based 3D Ligand-binding Motifs

Liu Yang [1,2,#], Wei He [1,2,*,#], Yuehui Yun [1,2], Yongxiang Gao [1,2], Zhongliang Zhu [1,2], Maikun Teng [1,2], Zhi Liang [1,2,*], Liwen Niu [1,2,*]

[1] *School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230026, China*
[2] *Division of Molecular and Cellular Biophysics, Hefei National Laboratory for Physical Sciences at the Microscale, Hefei 230026, China*

Received 22 October 2020; revised 30 June 2021; accepted 27 September 2021
Available online 11 March 2022

Handled by Xin Gao

KEYWORDS

Protein–ligand interaction;
Functional group;
Binding motif;
Computational method;
Drug design

Abstract Uncovering conserved 3D protein–ligand binding patterns on the basis of **functional groups** (FGs) shared by a variety of small molecules can greatly expand our knowledge of **protein–ligand interactions**. Despite that conserved binding patterns for a few commonly used FGs have been reported in the literature, large-scale identification and evaluation of FG-based 3D **binding motifs** are still lacking. Here, we propose a **computational method**, Automatic FG-based Three-dimensional Motif Extractor (AFTME), for automatic mapping of 3D motifs to different FGs of a specific ligand. Applying our method to 233 naturally-occurring ligands, we define 481 FG-binding motifs that are highly conserved across different ligand-binding pockets. Systematic analysis further reveals four main classes of binding motifs corresponding to distinct sets of FGs. Combinations of FG-binding motifs facilitate the binding of proteins to a wide spectrum of ligands with various binding affinities. Finally, we show that our FG–motif map can be used to nominate FGs that potentially bind to specific drug targets, thus providing useful insights and guidance for rational design of small-molecule drugs.

Introduction

Protein–ligand interactions play fundamental roles in many important cellular functions, including small molecule metabolism, enzymatic catalysis, and signal transduction and regulation. Comprehensive knowledge of protein–ligand interactions can not only provide important insights into

* Corresponding authors.
E-mail: hwkobe@mail.ustc.edu.cn (He W), liangzhi@ustc.edu.cn (Liang Z), lwniu@ustc.edu.cn (Niu L).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China

https://doi.org/10.1016/j.gpb.2021.08.014

biological functions of ligand-binding proteins but also be greatly benefit to drug discovery and development [1,2]. Many proteins that don't display overall sequence or structure similarities may share similar local 3D structures and can bind to same or similar ligands, thus identifying conserved binding patterns across different ligand-binding proteins at 3D level could facilitate a better understanding of protein–ligand recognition [3–5].

With the rapid accumulation of experimentally determined structures of protein–ligand complexes, it became possible for large-scale identification of conserved 3D binding motifs using computational approaches [6,7]. Current methods based on structural comparison or alignment of protein pockets have identified many well-defined 3D motifs that are conserved across different protein pockets and widely used for protein function annotation, pocket classification, and ligand-binding prediction [8–13]. However, these ligand-based 3D binding patterns are not applicable to large fraction of ligands especially small-molecule drugs due to the lack of reference 3D protein–ligand structures.

Despite the functional and structural diversity of different protein-binding ligands, many of them share same or similar functional groups (FGs) that mediate the interactions with the target proteins. Therefore, identification of conserved 3D binding motifs for FGs shared by different small molecules may extend our understanding of protein–ligand interactions to higher resolution and broader scope [14]. Previous studies have shown that conserved 3D motifs do exist in proteins binding different ligands with the same FG. For example, the phosphate-binding loop (P-loop) motif [15,16], in which the residues are highly conserved in terms of amino acid types as well as spatial positions among diverse phosphate-binding proteins. Conserved 3D motifs were also reported for other FGs such as adenine ring [17–19], heme group [20,21], and prosthetic groups [22]. However, these motifs were either uncovered through manual analysis of a small set of protein structures by an expert (*e.g.,* a crystallographer) or through structural alignment of proteins that bind FGs with rigid structures, which are subjected to limited FG types and/or biased datasets of 3D structures. Computational methods for automatic extraction of 3D binding motifs for a variety of FGs in large scale are still lacking.

To systematically identify and evaluate 3D binding motifs at FG level, we developed Automatic FG-based Three-dimensional Motif Extractor (AFTME), a computational method that automatically extract 3D FG-binding motifs. Our approach adopts commonly used strategy to describe 3D ligand-binding patterns [8,13,23], which quantitatively presents functional atoms (FAs) within certain distance of a specific ligand in 3D space. Explicit mapping of FAs to different FGs of the ligand is then achieved through two-dimensional clustering of the distance matrix. We applied our method to 233 natural ligands with abundant 3D protein–ligand structures and built an encyclopedia of 481 binding motifs for 160 different FGs, providing valuable resources for elucidating the mechanism of protein–ligand interactions as well as uncovering new rules for structure-based drug design.

Results

AFTME enables automatic extraction of FG-based 3D binding motifs

We have developed AFTME, a computational method to dissect protein pockets binding a specific ligand into sectors that interact with different functional groups (FGs). The basic assumption of this method is simple: if conserved binding pattern for a specific FG exists, the pattern-forming atoms should be spatially proximal to the corresponding FG and frequently co-appear, thus can be detected through clustering analysis of FAs from diverse protein pockets binding the same ligand. **Figure 1**A outlines the major steps of the method. 1) Given a set of protein pockets binding the same ligand, AFTME first parses all the FAs [24] that are considered to interact with the ligand atoms (LAs). 2) Then a distance matrix is constructed that evaluates the spatial distances between FAs and LAs. 3) Based on the distance matrix, a two-dimensional clustering algorithm is performed, through which LAs are clustered into different FGs at the first dimension and FAs are clustered into corresponding FG-binding motifs. 4) Each identified binding motif can be represented as a vector according to its chemical composition, which facilitates further analysis. The detailed description of each step is presented in the Materials and methods section.

Considering the abundance of studies on ATP-binding proteins, we first applied AFTME to a set of ATP-binding proteins as a proof of concept. As shown in Figure 1B, our method identified three FA clusters or binding motifs corresponding to triphosphate group, ribose, and adenine, respectively. The triphosphate-binding motif (M1) mainly consists of hydrophilic atoms from polar amino acids like Arg, Lys, Ser, *etc*., and the adenine-binding motif (M2) is enriched by atoms from hydrophobic and aromatic amino acids including Leu, Val, and Phe, whereas the ribose-binding motif (M3) contains both hydrophobic and hydrophilic amino acids (Figure 1C). Further investigation of these binding motifs indicated that the AFTME-identified FG-binding motifs are biological meaningful units. For instance, among the hydrophilic residues (rendered in red in Figure 1D and Figure S1) interacting with the triphosphate group, Lys and Ser are both well-known conserved residues in the P-loop, a common motif for phosphate-binding in ATP- and GTP-binding proteins, which is typically composed of a glycine-rich sequence followed by a conserved lysine and a serine or threonine [15]. It was found that the hydrophobic and/or aromatic residues (rendered in green in Figure 1D and Figure S1) making up the adenine-binding motif interact with adenine ring through C–H–π and/or π–π interactions. Notably, Moodie et al. described the recognition of adenine by proteins in terms of a fuzzy recognition template based on a sandwich-like structure formed by hydrophobic residues [25]. Denessiouk et al. also found that bulky hydrophobic residues can form a hydrophobic area by interacting with the adenine base [26]. The A-loop motif, which includes aromatic residues

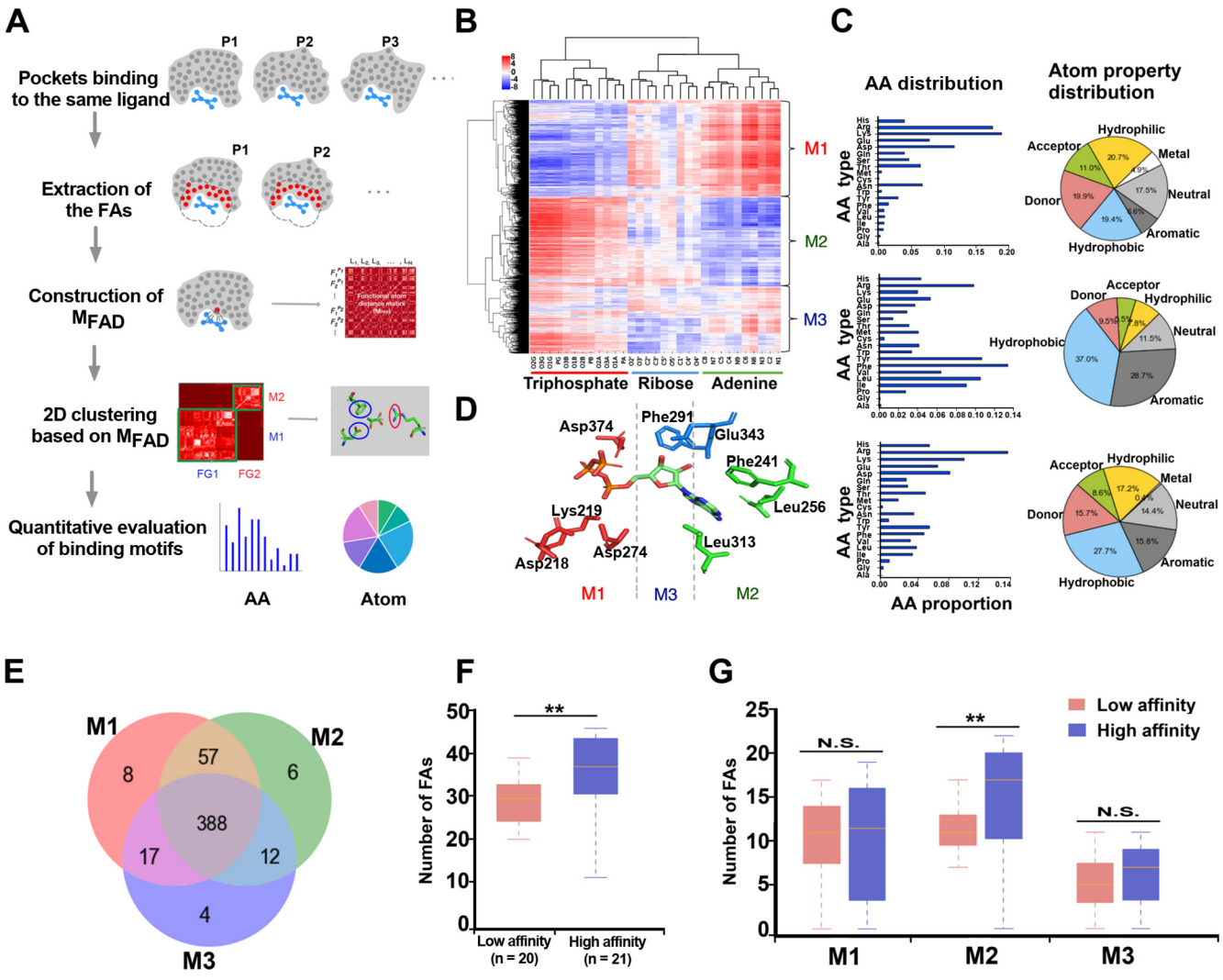

Figure 1 Workflow of AFTME and its application to ATP-binding proteins

A. A schematic view of major steps of the AFTME method. **B.** Two-dimensional clustering of the distance matrix for ATP-binding pockets. The vertical and horizontal axes correspond to FAs and LAs, respectively. The color encodes the distance between an FA and an LA. Three LA clusters corresponding to the triphosphate group, ribose, and adenine ring of ATP were identified, respectively. Three FA clusters or binding motifs (M1, M2, and M3) corresponding to the aforementioned three FGs were also obtained simultaneously. **C.** Distribution of amino acids and atom properties for M1 (top), M2 (middle), and M3 (bottom). **D.** An example (PDB: 1VJC) showing the spatial distribution of amino acids within each identified FG-binding motif. **E.** Different ATP-binding proteins use different combinations of FG-binding motifs M1, M2, and M3. **F.** Boxplot comparing the number of FAs within ATP-binding pockets with high affinity and those with low affinity. n refers to the number of pockets. **G.** Comparison of the FA numbers in FG-binding motifs of ATP-binding pockets with high and low affinities. The center line, bounds of box, and whiskers represent the median, interquartile range, and median ± 1.5 times interquartile range, respectively. The significant differences were calculated using Manney-Whitney test (**, $P < 0.01$; N.S., not significant). FA, functional atom; LA, ligand atom; FG, functional group; AA, amino acid; M_{FAD}, functional atom distance matrix.

forming π–π interactions with adenine ring was also reported [27]. These findings are largely consistent with the AFTME-identified adenine-binding motif. Although no well-defined motifs corresponding to the ribose-binding motif have been reported yet, it makes sense that hydrophobic/aromatic residues of the ribose-binding motif interact with five-carbon ring while hydrophilic residues interact with the extended hydroxyl groups through polar interactions.

We then set out to explore different roles played by the three identified motifs in the ATP-binding process. As we can see from the Venn diagram in Figure 1E, among the 492 ATP-binding pockets in the dataset, a majority (388, 75.9%) contain all the three binding motifs. Nevertheless, 86 (16.8%) pockets get two of them, among which 57 (66.3%) carry the M1 and M2 motifs, indicating that combination of two motifs, especially M1 and M2, is sufficient for ATP binding. Besides, we also noticed some cases in which only one binding motif together with one or more metal ions exists (Figure S2), indicating that metal ions may greatly affect the global binding profile. Next, we asked how different FG-binding

motifs contribute to the binding affinity. All the ATP-binding proteins with experimental affinity data available were collected and sorted from high to low affinities (Table S1). In general, protein pockets in high-affinity (top 1/3) group contain more FAs than those in low-affinity (bottom 1/3) group (Figure 1F, P = 7.4E−03, Mann-Whitney test). Interestingly, when looking into an individual binding motif, only M2 shows significant increment of FA numbers in high-affinity pockets (Figure 1G, P = 4.2E−03, Mann-Whitney test), suggesting that increase of hydrophobic interactions with the adenine ring makes major contributions to higher ATP binding affinity.

Taken together, the above results demonstrate the ability of AFTME to decompose ligand-binding sites into biological meaningful motifs, which are spatially well-defined to interact with different FGs and contribute unequally to protein–ligand binding affinity.

FG-based binding motifs are reused among different ligand-binding proteins

To see whether FG-based 3D binding motifs identified by our method are reused across different ligand-binding proteins, we applied AFTME to a few ligands sharing the same FGs with ATP including ADP, AMP, GTP, and UTP. Fine-mapped binding motifs for adenine and ribose were obtained from ADP- and AMP-binding proteins, and that for triphosphate was from GTP- and UTP-binding proteins (Figure S3A–D). We found that the chemical compositions of motifs binding the same FG are highly consistent although they were extracted from proteins binding different ligands (**Figure 2**A–C). For adenine and ribose, the binding motifs are universal among ATP-, ADP-, and AMP-binding pockets and have very similar distribution of amino acid types and atom categories (Figure 2A and B). Similarly, triphosphate-binding motifs extracted from GTP- and UTP-binding proteins also show consistent makeup with ATP-derived motifs (Figure 2C).

We then extended our evaluation to a wider range of ligands, among which 25 are adenine-containing, 38 are ribose-containing and 9 are triphosphate-containing. We described each FG-binding motif using a 26-dimensional vector representing the proportion of 20 types of amino acids and 6 categories of atoms, respectively (see Materials and methods). The vector representation of the FG-binding motifs enables correlation analysis of chemical composition for any pair of motifs. It was found that adenine-binding motif pairs showed significantly higher correlations than random FG-binding motif pairs (Figure 2D). Similar observations were obtained for ribose-binding (Figure 2E) and triphosphate-binding (Figure 2F) motif pairs, respectively. These results suggest high conservation of FG-binding motifs across a diversity of ligand-binding proteins.

Next, we performed a large-scale analysis of 3D FG-binding motifs for all the ligands with abundant 3D structures available. As shown in Figure 2G, we first derived all the 3D protein–ligand structures from BioLiP database [28]. Redundant proteins binding the same ligand that show over 50% sequence similarity were eliminated using CD-HIT [29]. 233 ligands with more than 5 structures available were kept for the following analysis. 481 FG-binding motifs corresponding to 160 unique FGs were identified using AFTME (Table S2), among which 39 FGs appeared in multiple (at least three) ligands. For each FG present in multiple ligands, a conservation score (CS) of the corresponding FG-binding motif was calculated as the average of pairwise Pearson's correlation coefficients among all the identified motifs binding this specific FG. And a corresponding P value was also calculated using permutation test (see Materials and methods). Most FG-binding motifs corresponding to FGs appeared in multiple ligands are highly conserved across different ligand-binding proteins (CS > 0.6, P < 0.05) (**Table 1**; Figure S4). Overall, two motifs binding the same FG show significantly higher composition correlations compared with two randomly selected motifs (Figure 2H), confirming the high conservation of FG-binding motifs. These lines of evidences showed that AFTME can be applied to detect binding motifs for a diversity of FGs. Importantly, the identified binding motifs are highly conserved among different ligand-binding pockets, laying the foundations for expanding limited 3D motifs to a broader range of ligands that are not suitable for AFTME analysis (*e.g.,* due to lack of structure data) but sharing same or similar FGs with applicable ligands.

Toward an encyclopedia of 3D binding motifs for a diversity of FGs

Given that 481 binding motifs for 160 different FGs have been identified using our method, we asked whether there are general interaction patterns between the identified motifs and the FGs they bind. We found that all the binding motifs could be clustered into 4 classes based on their physicochemical properties using k-means (Figure S4; see Materials and methods), which are well separated in the t-SNE plot shown in **Figure 3**A. Notably, the FG-binding motifs in different classes are featured with distinct physicochemical properties. The first class (red dots), denoted as the aromatic motif class, is enriched with atoms from aromatic amino acids like Trp, Tyr, and Phe. The second class (green dots), named the hydrophilic motif class, is mainly composed of hydrophilic, donor, and acceptor atoms from polar amino acids such as Arg, Lys, Asp, and Glu. The third class (blue dots), the mixed motif class, consists of both aromatic and hydrophilic atoms. The fourth class (purple dots), named the hydrophobic motif class, is dominated by atoms from hydrophobic amino acids including Leu, Ile, and Val (Figure 3A). We then sought to see the variability of data points within different clusters. We ranked all data points within each cluster based on their Euclidean distances to the k-center and calculated the Pearson's correlation between the mean of data points in top 10% (close to the k-center) and the bottom 10% (far away from the k-center). The results showed high correlation between marginal points and centered points in all the clusters: hydrophilic class (r = 0.824, P = 6.96E−08), hydrophobic class (r = 0.957, P = 9.89E−16), mixed class (r = 0.900, P = 6.87E−11), and aromatic class (r = 0.694, P = 4.20E05), suggesting that the data points are highly consistent within different clusters.

Next, we looked into the correspondence between different motif classes and the FGs they bind. As shown in Figure 3B, most of the FGs are uniquely mapped to a single motif class, indicating that different classes of motifs have their specific binding preference for FGs. Although a variety of FGs are involved, we found some dominating FGs in each motif class (Table S3).

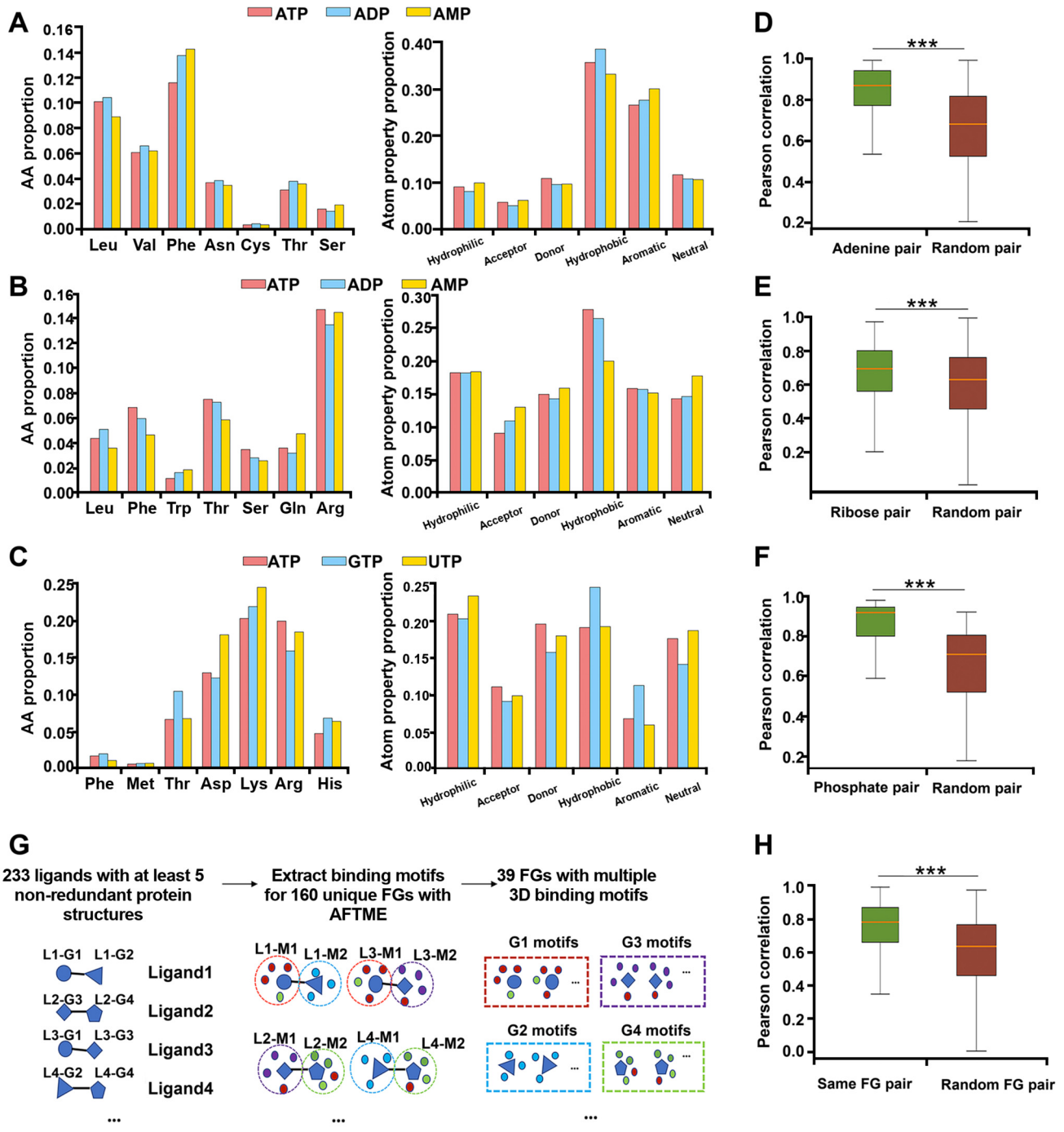

Figure 2 Conservation of FG-based 3D binding motifs

A.–C. The chemical compositions of adenine- (A), ribose- (B), and triphosphate-binding (C) motifs identified from proteins binding different ligands are similar. Adenine- and ribose-binding motifs are extracted from ATP-, ADP- and AMP-binding proteins, and triphosphate-binding motifs are obtained from proteins binding ATP, GTP, and UTP, respectively. **D.–F.** Pairs of adenine- (D), ribose- (E), and triphosphate-binding (F) motifs show significantly higher composition correlation than motif pairs binding random FGs. **G.** A schematic view of large-scale identification of FG-based binding motifs using AFTME. L, G, and M represent ligand FG, and motif, respectively. **H.** Correlation analysis of the 481 identified FG-binding motifs indicates that motifs binding the same FG are highly consistent in their composition. The center line, bounds of box, and whiskers represent the median, interquartile range, and median ± 1.5 times interquartile range, respectively. The significant differences were calculated using two-tailed Student's *t*-test (***, $P < 0.001$).

Table 1 Conservation evaluation of binding motifs for multi-ligand FGs

FG name	Frequency	Conservation score	*P* value
Ribose	38	0.665	1.00E−06
Phosphate	37	0.714	1.00E−06
Carboxyl	31	0.784	1.00E−06
Glucose	28	0.842	1.00E−06
Adenine	25	0.848	1.00E−06
Pyrophosphate	21	0.669	3.70E−05
Guanine	11	0.723	8.00E−06
Uracil	10	0.853	1.00E−06
Acetic acid	10	0.752	1.00E−06
Triphosphate	9	0.857	1.00E−06
Propylamine	8	0.713	4.46E−03
Alanine	8	0.856	1.00E−06
2-Deoxyribose	8	0.718	2.17E−03
Hydroxyethyl	7	0.678	7.78E−02
Acetamido group	6	0.776	2.90E−04
Glutamic acid	6	0.7	4.36E−02
Hexane group	6	0.907	1.00E−06
Hydroxymethyl	6	0.74	4.55E−03
Cytosine	5	0.646	3.16E−01
Pteridine rings	5	0.773	3.26E−03
Phenol	5	0.88	1.00E−06
Butyric acid	5	0.723	4.05E−02
Adenosine	5	0.845	3.00E−06
Glycine	4	0.748	4.27E−02
Ethylamine	4	0.756	3.70E−02
Para-amino benzoic acid	4	0.692	1.73E−01
Galactose	4	0.92	1.00E−06
Thymine	4	0.959	1.00E−06
Phenyl	4	0.726	1.27E−01
Glycolic acid	4	0.741	5.21E−02
Fructose	4	0.759	4.16E−02
Maltose	4	0.819	2.14E−03
Sulfo	4	0.636	4.11E−01
Glycerol	3	0.869	3.77E−03
Nicotinamide	3	0.817	2.63E−02
Propyl	3	0.748	1.38E−01
Ribose-3-phosphate	3	0.933	4.30E−05
Pyruvic acid	3	0.821	2.35E−02
Dimethylallyl	3	0.799	4.41E−02

Note: The frequency reported the number of ligands containing the FG in the dataset; the conservation score was calculated as the average of pairwise Pearson's correlations among all the FG-binding motifs for the specific FG; and *P* value was calculated using permutation test. FG, functional group.

Among FGs that interact with the aromatic motifs, two types of FGs are in the majority, one is with aromatic ring and the other is with non-aromatic ring. The former type, exemplified with the cytosine ring of cytidine-5′-monphosphate (C5P), interacts with the aromatic ring of Phe and Tyr through π-stacking (Figure 3C, left panel). The latter type, for example, the glucose ring in N-acetyl-D-glucosamine (NAG), of which the carbon atoms form hydrophobic interactions with the aromatic atoms of Tyr/Trp (Figure 3C, right panel).

In contrast, the hydrophilic motif class prefers to bind polar FGs through hydrogen bonds, among which carboxyl and phosphorus are the most prevalent ones. For instance, the carboxyl group in citric acid (CIT), forms N–H···O and O–H···O hydrogen bonds with N atom from imidazole of a HIS and O atom from hydroxyl of a Thr, respectively (Figure 3D, left panel). Four N–H···O hydrogen bonds are formed between O atoms of the phosphate group in adenosine-3′-5′-diphosphate (A3P) and N atoms of two basic amino acids (Lys and Arg) in the binding motif (Figure 3D, right panel).

There are over 10 FGs engaged in both the mixed and the aromatic motif classes, most of which are non-aromatic sugar rings. In addition to hydrophobic interactions between the sugar ring and the aromatic ring which are frequently used in the aromatic motif class, the mixed motif class also contains hydrophilic amino acids that form hydrogen bonds with the extended-out hydroxyl groups (Figure 3E, right panel). Besides, the mixed motif class is of high propensity to recognize carboxyl-amine, of which the amine group interacts with the aromatic ring through amide-π stacking and the carboxyl group interacts with hydrophilic amino acids via hydrogen bonds, respectively (Figure 3E, left panel).

The hydrophobic motif class also shares a major type of FG, the aromatic hetero ring, with the aromatic motif class. Instead of π–π interactions, the C–H–π interactions are the main driving force for hydrophobic-aromatic contacts. Another two major types of FGs involved in the hydrophobic motif class are alkene and alkane chains. As two examples showed in Figure 3F, the alkane (left) and the alkene (right) chains are well accommodated in protein pockets composed of hydrophobic residues.

Altogether, our systematic analysis suggested the existence of four classes of FG-binding motifs and their favored FGs. Deep investigations further revealed general interaction patterns between these functional motifs and the FGs they bind, thus build up a global map of 3D motif–FG interactions.

Motif combinations facilitate different modes of ligand binding

Having identified the corresponding relations between motif classes and FGs, we then asked how different motifs are combined to facilitate the binding of ligands that consist of different FGs. We found that the aforementioned four FG-binding motif classes are almost evenly distributed in protein pockets investigated in our analysis (**Figure 4**A), suggesting that all the identified motif classes are commonly used and important for protein–ligand recognition.

After careful inspection of the identified binding motifs and their host ligand-binding pockets, we found three distinct combination modes for motif classes in protein pockets (Figure 4B). 1) The single-class mode, which applies to nearly a quarter of investigated ligand-binding cases, combines only FG-binding motifs of the same class. 2) The double-class mode, which goes for more than 60% of the cases, integrates two different classes of FG-binding motifs. 3) The triple-class mode, which recognizes a smaller fraction of ligands, assembles three different classes of FG-binding motifs.

For the single-class mode, combinations of two mixed-class motifs are mostly observed, followed by hydrophobic-, aromatic-, and hydrophilic-class motif combinations (Figure 4C). For the double-class mode, there are 6 possible class-class combinations, among which the hydrophobic-hydrophilic combination applies to the greatest number of ligands, indicating a commonly used protein–ligand binding pattern in which the hydrophobic FG of the ligand interacts with a hydrophobic motif while another polar FG is oriented

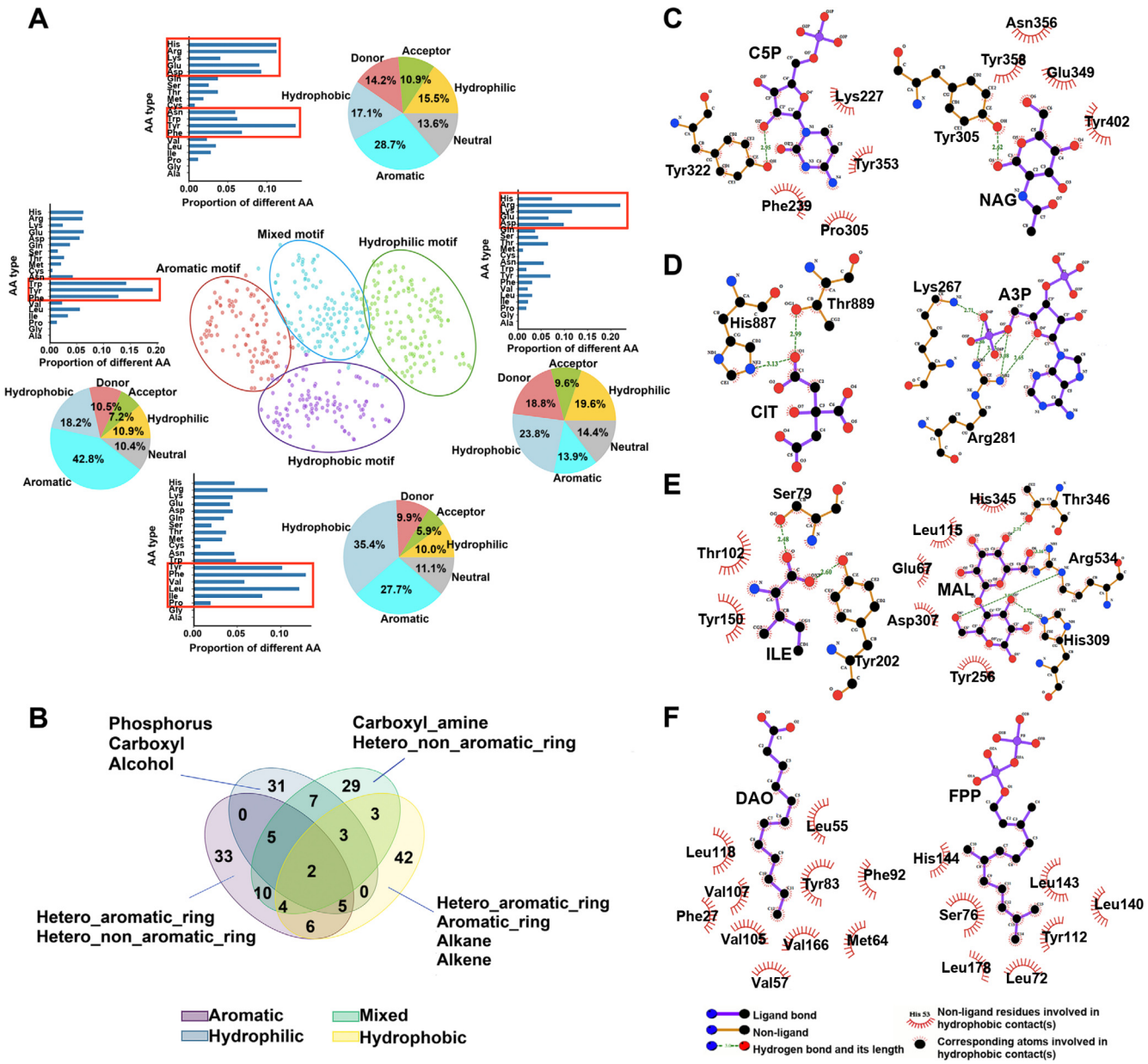

Figure 3 Systematic mapping of motif classes to different FG types

A. FG-binding motifs can be clustered into four well-separated classes, each of which has distinct distribution of amino acids (bar plot with the major amino acid types marked in red rectangular box) and atom types (pie plot, referring to the FA property proportion). **B.** Venn plot showing different FG-binding preferences for different motif classes. The numbers refer to the counts of FGs within each category. Dominant FG types for each motif class are denoted beside the plot. **C.–F.** Examples of 2D interaction map between FGs and identified motifs. The aromatic motifs identified for cytosine ring of cytidine-5′-monphosphate (PDB: 4G5T, left) and glucose ring in N-acetyl-D-glucosamine (PDB: 6EN3, right) (C). The hydrophilic motifs identified for the carboxyl group of citric acid (PDB: 6FXI, left) and the phosphate group of adenosine-3′-5′-diphosphate (PDB: 1KAI, right) (D). The mixed motifs identified for amino acid isoleucine (PDB: 1Z17, left) and two glucose rings in maltose (PDB: 1AHP, right) (E). The hydrophobic motifs identified for the hexane group of lauric acid (PDB: 2OVD, left) and the farnesyl group in farnesyl diphosphate (PDB: 2E90, right) (F). The 2D ligand–protein interactions were generated by LigPlot [49].

to a hydrophilic motif (Figure 4D, Figure S5). The triple-class mode also includes four different class–class combinations that are almost equally present for ligands they bind (Figure 4E).

To gain further insights, we investigated in greater detail of the ligands involved in different combination modes (Table S4). Notably, combinations of different classes of FG-binding motifs facilitate the binding of a vast diversity of ligands composed of FGs that are well mapped to the corresponding FG-binding motif classes.

Among ligands involved in the single-class mode, we outlined three for examples (Figure 4C, Figure S6A). 2-acetamido-2-deoxy-alpha-D-glucopyranose (NDG) is

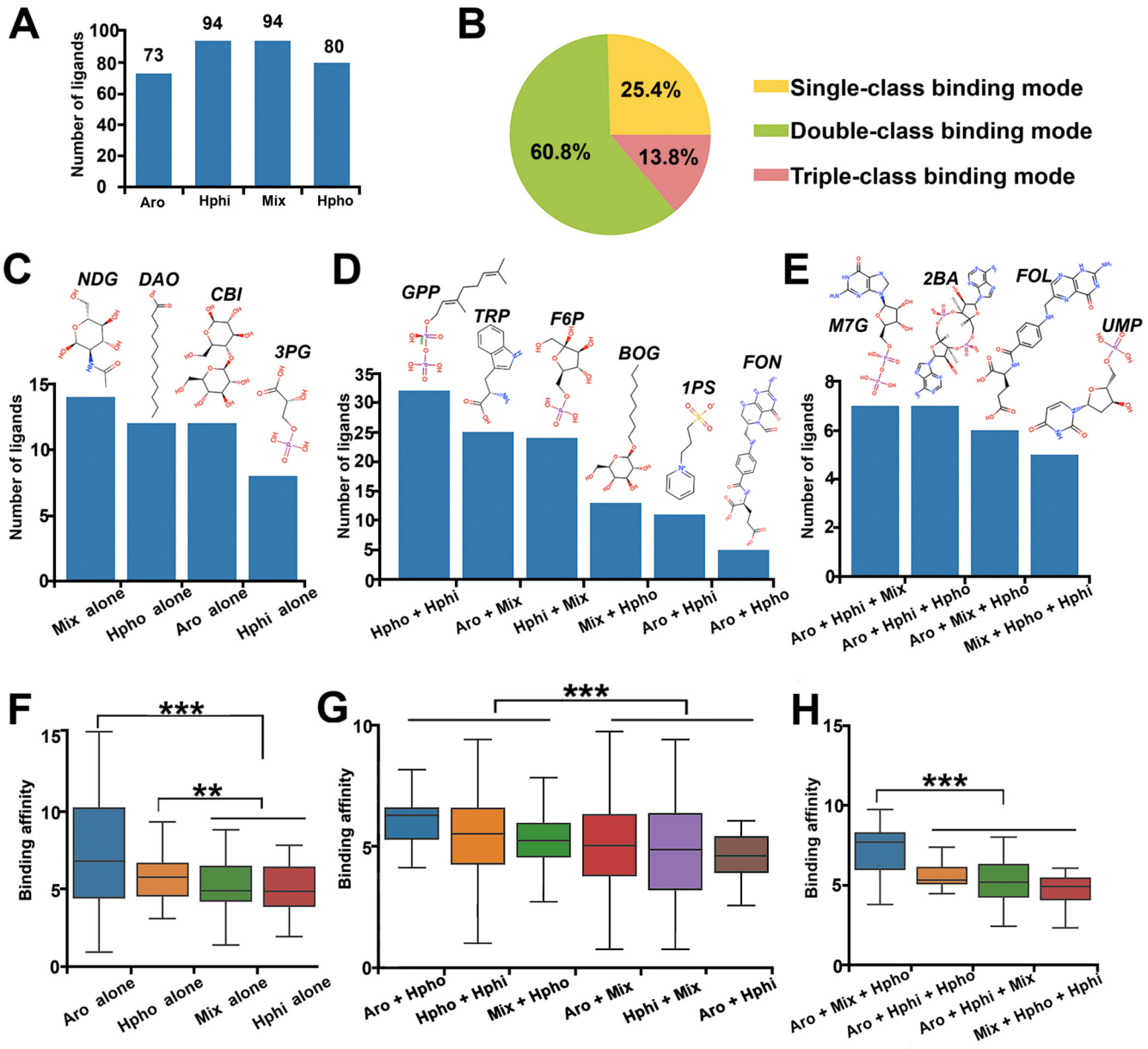

Figure 4 Combinations of FG-binding motif classes in a ligand-binding pocket

A. Distribution of the four classes of FG-binding motifs in ligands. The number above the blue rectangular box represents the counts of ligands in the corresponding FG-binding motif class. **B.** Proportion of the three different combination modes for motif classes. **C.–E.** Distribution of different FG-binding motif combinations and examples of ligands involved in single-class (C), double-class (D), and triple-class (E) modes. The ligand name and its 2D diagram are indicated above the corresponding blue rectangular box. **F.–H.** Ligand-binding affinity is affected by the combination of FG-binding motif classes for the single-class (F), double-class (G), and triple-class (H) combination modes. The center line, bounds of box, and whiskers represent the median, interquartile range, median ± 1.5 times interquartile range, respectively. The significant differences were calculated using Manney-Whitney test (**, $P < 0.01$; ***, $P < 0.001$). Mix, mixed-class; Hpho, hydrophobic-class; Aro, aromatic-class; Hphi, hydrophilic-class.

composed of two mixed-motif favored FGs including an acetamide group and a glucose ring, both being bound by FG-binding motifs of the mixed class. Cellobiose (CBI) is a disaccharide consisting of two glucoses and binds to proteins with two aromatic motifs. 3-phosphoglyceric acid (3PG) has two polar FGs, a phosphate group and a glyceric acid group, which are recognized by two hydrophilic motifs.

A greater number and higher variety of ligands were witnessed in the double-class mode (Figure 4D, Figure S6B). For instances, geranyl diphosphate (GPP) which is complexed with proteins comprising a hydrophobic and a hydrophilic motif, contains a hydrophobic-preferred alkene and a hydrophilic-preferred diphosphate group. For fructose-6-phospahte (F6P) proteins achieve ligand-binding with an aromatic motif to the sugar ring and a hydrophilic motif to the phosphate group, respectively. Other ligands such as tryptophan (TRP, with an aromatic motif to indole and a mixed motif to alanine), B-octyl glucoside (BOG, with a

hydrophobic motif to octyl and a mixed motif to glucose), 3-pyridinium-1-ylpropane-1-sulfonate (1PS, with an aromatic motif to pyridinium and a hydrophilic motif to sulfonate) all follow the general FG–motif interaction patterns we identified.

In the triple-class mode, the ligands have at least three FGs and thus are in relatively larger size (Figure 4E, Figure S6C). For examples, to bind 7N-methyl-8-hydroguanosine-5′-diphosphate (M7G), proteins adopt a binding pattern with an aromatic motif to dihydroguanine, a mixed motif to ribose, and a hydrophilic motif to diphosphate. Similarly, folic acid (FOL) contains a pteridine ring, a benzoic group, and a glutamic acid that are recognized by a hydrophobic, an aromatic, and a mixed motif, respectively.

In the example of ATP, we already showed the unequal contribution of different FG-binding motifs to the binding affinity. Here, we sought to explore how combinations of different classes of FG-binding motifs will affect the ligand-binding affinity. Experimental binding affinity for all the protein–ligand pairs in our analysis were retrieved from the PDBbind database (Table S5) [30]. In the single-class mode, both the aromatic and hydrophobic combinations show significantly higher affinity than the mixed and hydrophilic combinations, suggesting that general hydrophobic interactions including π–stacking, C-H–π interactions and interactions between two aliphatic carbons contribute more to high binding affinity than polar interactions such as hydrogen bonds and salt bridge (Figure 4F). Consistently, in the other two combination modes, the more hydrophobic interactions are involved, the higher affinity the ligand-binding achieve. For example, combination of the hydrophobic and aromatic motifs is the most efficient binding pattern in the double-class mode, and the hydrophobic motif involved modes get significantly higher affinity compared to combinations with non-hydrophobic motifs (Figure 4G). Moreover, non-hydrophilic combinations get significantly higher affinity compared to combinations with hydrophilic motifs for the triple-class modes (Figure 4H). The results further supplemented our observation from the ATP-binding motifs (Figure 1G) and confirmed the findings in the previous studies that hydrophobic interactions are a driving factor for the increased ligand efficiency [31,32].

Together, these evidences showed that FG-binding motifs are building blocks of ligand-binding sites, and combinations of different classes of FG-binding motifs facilitate the binding of proteins to a wide spectrum of ligands with various binding affinities.

Motif–FG binding can be used for drug design

We next asked whether the motif–FG map derived from naturally-occurring protein–ligand complexes can be used for the rational design of small-molecule drugs for given protein targets. To nominate FGs that potentially bind to a specific drug target, we adopted the following procedure. 1) We identified all the FA clusters from the ligand-binding pockets of the target protein. 2) For each FA cluster, we measured its similarity with each of the 481 FG-binding motifs obtained above by defining an FG-matching (FM) score. 3) The corresponding FGs were then ranked according to their FM scores from high to low and the enrichment of an FG or FG type in the top hits was evaluated by calculating a P value. FGs with higher FM scores are more likely to bind the specific target, FGs significantly enriched in the top hits are expected to be potential candidates (see Materials and methods).

We first applied this approach to DOT1L, which is the target of EPZ-5676, a small-molecule drug in clinical trial for the treatment of adult acute leukemia [33]. Three clusters of FAs were identified from the ligand-binding sites of DOT1L, which correspond to three different FGs of EPZ-5676, *i.e.*, adenine, ribose, and methionine, respectively (**Figure 5**A). We then computed the FM scores between the three FA clusters and all the 481 FG-binding motifs in the database and looked into the top 20% hits with the highest FM scores. For the first FA cluster (C1) which interacts with the adenine group of EPZ-5676, we found that adenine is also the most significantly enriched FG (Figure 5A and B, P = 8.47E−10). Consistent with the ribose group, 16 pentose groups from different ligands are among the top hits for the second FA cluster (C2) (Figure 5A and C, P = 1.22E−05). Although no specific FGs are enriched in the top hits for the third FA cluster (C3), 10 different amino acid groups (Figure 5A and D, P = 3.78E−03), including glutamic acid, glyoxylic acid, alanine, *etc.*, are ranked in the top, suggesting that FGs sharing partial similarities can also be used for prediction.

Next, we tried to nominate FGs that bind to the main protease (Mpro) of COVID-19 using the same strategy. We extracted three well-separated FA clusters in the ligand-binding sites (Figure 5E), which are located proximal to the indole-carboxamide, fluorophenyl, and Ala((2-oxopyrrolidin-3-yl))-al groups of 11b, a potent inhibitor of the protein [34]. Despite that the three FGs of the inhibitor are not included in the database, we observed similar FGs and/or same FG types enriched in the top hits regarding different FA clusters. For instances, hetero aromatic rings and sulfurs are enriched for the FA clusters binding indole-carboxamide (C1) (Figure 5E and F, P = 4.48E−06) and fluorophenyl (C2) (Figure 5E and G, P = 6.40E−03), respectively. Similar to the C3 of DOT1L, amino acid groups are among top hits for FA cluster binding Ala((2-oxopyrrolidin-3-yl))-al (C3) (Figure 5E and H, P = 4.30E−06).

Together, these examples showed that the global map of FG-binding motifs can be used to nominate specific FG and/or FG types that potentially bind to specific drug targets, thus providing important insights and guidance for rational design of small-molecule drugs.

Discussion

A classical assumption in structural biology is that the 3D structure of a protein determines its molecular function. However, many proteins that don't display overall sequence or structure similarities may share similar local 3D binding sites and can bind to same or similar ligands [35]. Thus, identifying conserved 3D patterns/motifs across different ligand-binding proteins serve as an efficient way to learn and predict protein–ligand interactions. Computational methods that rely on multiple structure alignments or pairwise pocket comparisons have identified many conserved 3D binding patterns across different protein pockets binding same or similar ligands [36]. Despite the validity and usefulness of these ligand-based binding patterns, they mainly go for naturally-occurring ligands with abundant protein–ligand 3D

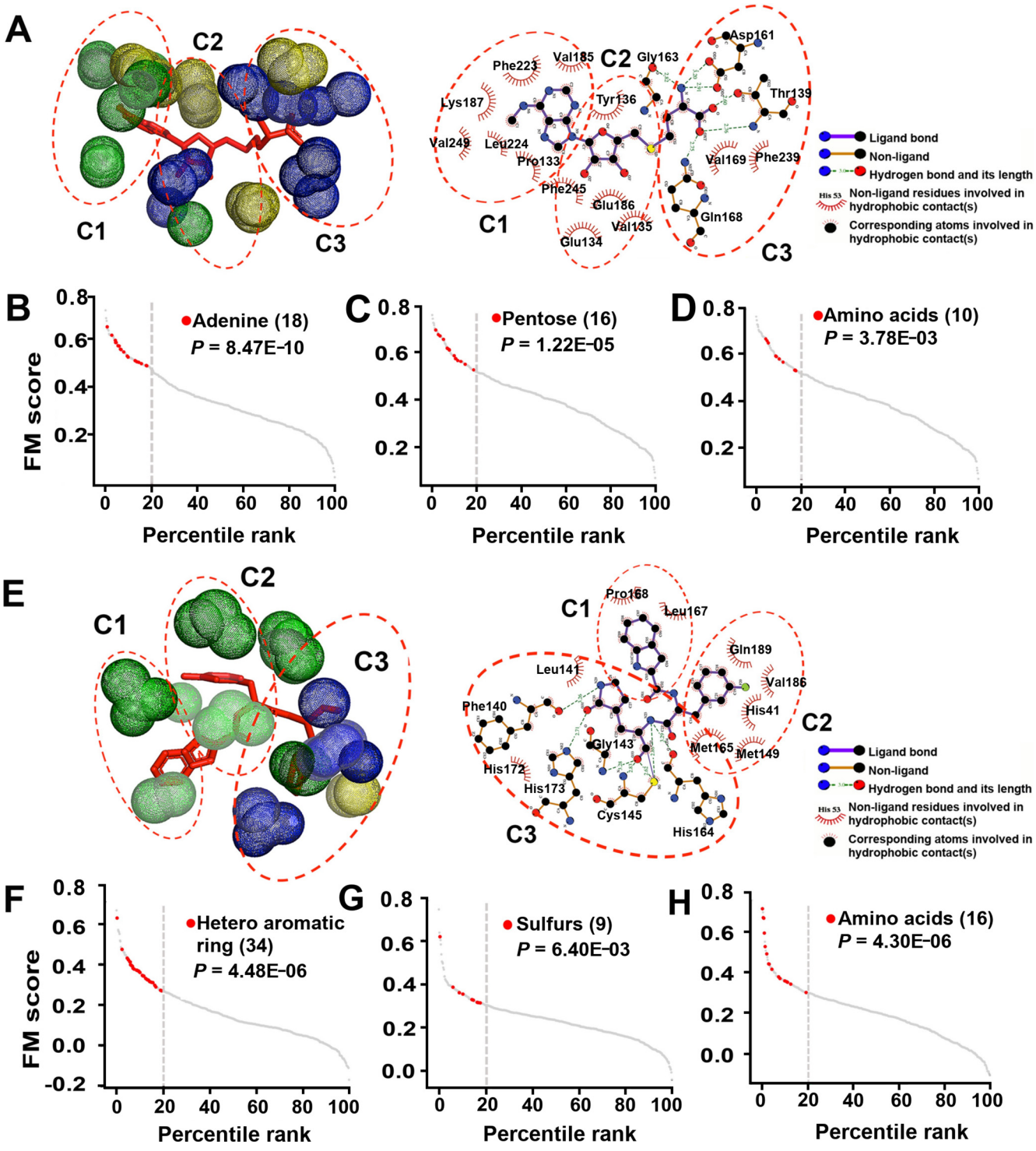

Figure 5 FG–motif map can be used for rational drug design

A. 3D map showing the distribution of FG-binding motifs relative to different FGs of small-molecule drugs (left) and the corresponding 2D ligand–protein interaction map (right) for DOT1L-EPZ5676 complex (PDB: 3SR4). C1, C2, and C3 refer to three FA clusters interacting with adenine, ribose, and methionine of EPZ5676, respectively. **B.**–**D.** Nomination of potential FG candidates that bind to C1 (B), C2 (C), and C3 (D) of DOT1L. **E.** 3D map showing the distribution of FG-binding motifs relative to different FGs of small-molecule drugs (left) and the corresponding 2D ligand–protein interaction map (right) for Mpro of COVID-19 in complex with 11b (PDB: 6M0K). C1, C2, and C3 refer to three FA clusters interacting with indole-carboxamide, fluorophenyl, and Ala((2-oxopyrrolidin-3-yl))-al of 11b, respectively. **F.**–**H.** Nomination of potential FG candidates that bind to C1 (F), C2 (G), and C3 (H) of Mpro. In (A) and (E), atoms with different physicochemical property are rendered in different colors: hydrophobic (green), polar (purple), and aromatic (yellow), and the 2D ligand–protein interaction maps are generated by LigPlot [49]. In (B–D) and (F–H), FGs are ranked based on the FM score; the dash line indicates the top 20% hits with highest probabilities to bind the specific target; and the red dots are specific FGs or FG types that significantly enriched in the top. The name of specific FG or FG type together with the frequency of its appearance in the top and the corresponding *P* value are displayed in the corner. The significant differences were calculated using Fisher's exact test. FM, FG-matching.

structures, thus limiting their application scope. Here, we proposed AFTME, a computational method for automatic identification of 3D binding motifs on the basis of FGs shared by different small molecules, which permits studying protein–ligand interactions in a wider scope and higher resolution. The application to ATP showed the feasibility and validity of our method to detect FG-based 3D binding motifs and confirmed the reusability of the motifs in different ligand-binding proteins. We further applied our method to 233 natural ligands and obtained 481 binding motifs for 160 unique FGs, providing useful resources for deep exploration of protein–ligand recognition.

Systematic investigation of FG-binding motifs identified by our method provides several important insights into protein–ligand interactions. First, ligand-binding sites of a protein can be dissected into independent sectors corresponding to different FGs of the binding ligand. These FG-based binding motifs are highly conserved among different ligand-binding pockets at both amino acid and atom level. Second, we found four classes of FG-binding motifs with distinct physicochemical properties and their own preference for FG binding. Moreover, the interactions between 3D motifs and FGs follow some general rules. For example, a hydrophobic motif is more likely to interact with a hydrophobic FG and a hydrophilic motif usually recognizes a polar FG. Third, following the general motif–FG recognition map, protein pockets consisting of different FG-binding motifs can bind to a wide spectrum of ligands through different motif combination modes. Of note is that protein pockets with more hydrophobic motifs tend to gain higher binding affinity.

Rapid development of high-throughput screening using CRISPR/Cas9 system has greatly accelerated the discovery of new cancer drug targets in recent years [37,38]. CRISPR screening with tiling-sgRNA designs can further infer essential protein domains that are suitable for drug targeting [39–41]. However, identifying effective small-molecule drugs for a specific target through high-throughput experimental screens is still expensive and inefficient [42]. Virtual screening using computational approaches has emerged as a starting point for identifying hit molecules for a given drug target [43]. 3D-based predictions of small molecules for a specific protein target with machine or deep learning approaches are of higher accuracy compared to sequence-based predictors. However, these ligand-based methods rely on multiple 3D structures binding to the ligands to learn the features, thus are limited to a small fraction of ligands. Our study showed that conserved 3D binding patterns can be obtained at FG level, which may expand the scope for ligand-binding prediction since many different ligands share same or similar FGs.

Although the FG-binding motifs are mainly derived from protein structures binding to natural ligands, we showed that our FG–motif map can also be used to infer FG candidates that potentially bind to specific protein targets, suggesting potential application in rational drug design. Meanwhile, we'd like to mention that it is still challenging to accurately predict the FGs bind to a specific drug target with current FG–motif datasets. First, the number of FG-binding motifs revealed in this study is still very limited due to the lack of protein–ligand structures. In most cases, we can only infer FG type(s) rather than a specific FG using our top hits enrichment analysis. Second, our current representations of FG-binding motifs are relatively simple and rough, a more refined and comprehensive description of the motifs may further improve the prediction. For example, the FEATURE framework proposed by Altman et al. represents local protein structures with more comprehensive feature sets regarding the number and types of physics-inspired quantities [44,45]. Third, protein–ligand interactions can be altered with even slight change of binding environment. For instance, it is frequently occurred in a protein family that similar pockets can bind to different ligands. Therefore, it's not enough to just predict common FG-binding patterns shared by the same or similar ligands. A comprehensive understanding of how slight differences can affect ligand binding is also required, which can be achieved through FG-based motif analysis at single protein pocket level instead of a group of similar pockets.

Materials and methods

Construction of the datasets

We collected all the protein–ligand complexes from the BioLiP database, a semi-manually curated database for biologically relevant ligand–protein interactions [28]. Proteins that only bind to metal ions were excluded. For each ligand, we removed the redundant proteins with more than 50% sequence similarity using CD-HIT [29]. Only ligands with at least 5 protein structures were kept for further analysis, producing a dataset containing 11,570 protein structures in complex with 233 ligands. The PDB codes for all the protein structures, as well as the information of all 233 ligands are available at https://github.com/MDhewei/AFTME.

We retrieved experimentally determined binding affinities of protein–ligand complexes from the latest version of the PDBbind database [30], resulting in binding affinity data for 599 complexes covering 158 out of the above 233 ligands. The affinity values of each complex used in the binding affinity analysis were pK_d, pK_i, and pIC_{50}, which were obtained by transforming raw affinity K_d, K_i, and IC_{50} as follows:

$$pX = -\log_{10}(raw\ affinity X)$$

A higher value of pX indicates a stronger binding affinity for the protein–ligand complex.

Ligand FG definition and classification

According to the knowledge of biochemistry, we manually defined the FGs of each ligand based on their shape, size, and physicochemical property. Given a ligand in the dataset, we firstly downloaded its 2D structure from PDB database [46], then scanned its structure and search for FGs in the following order: 1) ring structures with consistent physicochemical property, for instance adenine; 2) ring structures together with other polar groups, for instance ribose; 3) chain structures at the terminial or in the middle of the ligand, for instance alkane chain; 4) well-defined polar groups such as phosphate, carboxyl, and hydroxyl; and 5) other fragments that are close in size and composition with well-defined FGs. In general, we followed a basic principle that intra-ligand FGs should be different with each other in shape and physicochemical property but close in size, thus ensure their independency when interacting with the protein partner. For all the FGs, we further clas-

sified them into 21 major types referring to the classification in the previous study [47].

AFTME algorithm description

AFTME takes four major steps to extract FG-binding motifs: 1) extraction of ligand binding pockets, 2) construction of functional atom distance matrix, 3) two-dimensional clustering based on the distance matrix, and 4) identification and characterization of FG-binding motifs. Details of each step are described below.

Extraction of ligand binding pockets

A protein pocket is described as a set functional atoms situated around the bound ligand in 3D space [8]. Specifically, for each LA, we iterate to search for non-backbone heavy atoms that are within 5 Å in 3D space, all the atoms that meet the criteria are defined as FAs. For each FA, we defined an N-dimensional row vector (F) that describes its distance to each LA as:

$$F = (d_1, d_2, ..., d_N) \tag{1}$$

where d_l represents the distance between the FA and the l-th atom of the ligand and N is the number of atoms of the ligand. For a pocket composed of M FAs, we stacked the corresponding FA vectors to generate an $M \times N$ dimensional pocket matrix (P) that represents the geometrical configuration between the pocket and the ligand.

$$P = \begin{pmatrix} F_1 \\ F_2 \\ \cdot \\ \cdot \\ \cdot \\ F_M \end{pmatrix} = \begin{pmatrix} d_{1,1}, d_{1,2}, ..., d_{1,N} \\ d_{2,1}, d_{2,2}, ..., d_{2,N} \\ \cdot \\ \cdot \\ \cdot \\ d_{M,1}, d_{M,2}, ..., d_{M,N} \end{pmatrix} \tag{2}$$

where F_f denotes the FA vector for the f-th FA and $d_{f,l}$ represents the distance between the f-th FA and the l-th LA.

Construction of functional atom distance matrix

Given K pockets that bind to the same ligand, we calculate the pocket matrix for each pocket and stack them to form a functional atom distance matrix M_{FAD}.

$$M_{FAD} = \begin{pmatrix} P_1 \\ P_2 \\ \cdot \\ \cdot \\ \cdot \\ P_K \end{pmatrix} = \begin{pmatrix} F_{1,1} \\ ... \\ F_{1,M_1} \\ --- \\ F_{2,1} \\ ... \\ F_{2,M_2} \\ --- \\ \cdot \\ \cdot \\ \cdot \\ --- \\ F_{K,1} \\ ... \\ F_{K,M_K} \end{pmatrix} = \begin{bmatrix} d_{1,1,1}, ..., d_{1,1,N} \\ ... \\ d_{1,M_1,1}, ..., d_{1,M_1,N} \\ ------ \\ d_{2,1,1}, ..., d_{2,1,N} \\ ... \\ d_{2,M_2,1}, ..., d_{2,M_2,N} \\ ------ \\ \cdot \\ \cdot \\ \cdot \\ ------ \\ d_{K,1,1}, ..., d_{K,1,N} \\ ... \\ d_{K,M_K,1}, ..., d_{K,M_K,N} \end{bmatrix} \tag{3}$$

where P_p denotes the pocket matrix of the p-th pocket, $F_{p,f}$ is the FA vector for the f-th FA in the p-th pocket, M_p is the number of FAs in the p-th pocket and $d_{p,f,l}$ represents the distance between the f-th FA of the p-th pocket and the l-th LA. Note that despite of the same ligand, the number of FAs in different binding pockets may not be equal. Therefore, M_{FAD} is a matrix with $(M_1 + M_2 + ... + M_K)$ rows and N columns.

Two-dimensional clustering based on the distance matrix

Given a functional atom distance matrix (M_{FAD}) calculated for a specific ligand, we performed a two-dimensional hierarchical clustering on both the rows and columns, which represented the FAs and the LAs, respectively. Prior to this analysis, the matrix was element-wisely standardized by subtracting the minimum and dividing the maximum. The clustering ran by first considering each FA/LA as an individual cluster and calculating the Euclidean distances between any two FA/LA clusters. Then two clusters with the closest distance were merged, and the linkages were created using the Ward method to minimize the total within-cluster variance. The clustering was performed using "AgglomerativeClustering" module from the scikit-learn package in python [48]. We set the "n_clusters" parameter as the number of predefined FGs of the ligand.

Visualization and identification of FG-binding motifs

We used a heatmap with a row-oriented and a column-oriented dendrograms to visualize the hierarchical clustering results. Based on the heatmap, we could find explicit correspondence between different clusters of FAs and LAs. Specifically, the close connection between an FA cluster and its proximal LA cluster builds up a mapping between FA and LA clusters, where the LA clusters correspond to the FGs within the ligand, and the FA clusters correspond to the binding motifs for the specific FG matched. Following this step, we obtained different binding motifs for different FGs of the ligand they interact with. We filtered out clusters that are thought to be noises based on the following two criteria. 1) We reasoned that biological meaningful LA clusters should be largely consistent with manually defined FGs based on biochemical knowledge. Here, using a simple majority vote strategy, only LA clusters with more than half of its atoms overlapped with a predefined FG is thought to be biological meaningful. 90.6% (481 out of 531) of LA clusters aligned well with FGs in the predefined set, suggesting that most automatically extracted LA clusters are biological meaningful. 47 LA clusters have been removed in this step. 2) Since we only performed analysis to ligand with at least 5 protein structures, we considered a motif with less than 5 atoms (less than one atom per structure) to be meaningless. 3 motifs have been discarded in this step.

Quantitative representation of FG-binding motifs

To gain further insight of an identified FG-binding motif, we made a deep insight into its composition from both amino acid level and atom level. In terms of amino acid level, we counted the number of binding pockets for all motifs interacting with specific FGs of the ligand, and the presence of each type of 20 amino acids inside a binding motif. In particular, atoms were classified into 6 categories according to their biochemical properties [24], *i.e.*, hydrophilic, acceptor, donor, hydrophobic, aromatic, and neutral. By calculating frequency of

occurrence for each atom category within the motif, the motif could be expressed from the perspective of atom level. To make a quantitative evaluation of the binding motif, we expressed it using a 26-dimensional vector.

$$M = (X_1, X_2, X_3, \cdots, X_{19}, X_{20}, X_{21}, \cdots, X_{26}) \quad (4)$$

where the first 20 dimensions are used to compute the proportion of occurrence of the amino acid *aa* in a specific motif for each of the 20 types of amino acids, and the last 6 dimensions are used for the calculation of the proportion of occurrence of the atom properties *pp* of each category in the same motif for each of the 6 categories defined above. Particularly, the value in each dimension could be defined as follows:

$$X_i = \begin{cases} \frac{n_{aa,i}}{nresi}, if\ 1 \le i \le 20 \\ \frac{n_{pp,i}}{nprop}, if\ 21 \le i \le 26 \end{cases} \quad (5)$$

$n_{aa,i}$ and *nresi* are the number of residues of type *i* amino acid *aa* observed in the motif and total number of residues in that motif, respectively. And the 20 types amino acids are assigned to a fixed order from 1 to 20. Similarly, $n_{pp,i}$ and *nprop* are the number of properties of category *i* atom property *pp* and total number of properties, and the *i* ranging from 21 to 26 denotes the atom properties corresponding to the above 6 biochemical categories.

Conservation evaluation of FG-binding motifs

To quantitatively assess the reusability of two FG-binding motifs, we calculated the Pearson correlation coefficient (PCC) between two corresponding motif vectors as:

$$PCC(M_u, M_v) = \frac{\sum_{i=1}^{26}[(c_{u,i} - \overline{c_u}) \times (c_{v,i} - \overline{c_v})]/26}{\sqrt{\frac{\sum (c_{u,i} - \overline{c_u})^2}{26}} \times \sqrt{\frac{\sum (c_{v,i} - \overline{c_v})^2}{26}}} \quad (6)$$

where M_u and M_v denote the 26-dimensional vectors of the two motifs, $c_{u,i}$ and $\overline{c_u}$ are the elementwise and average physicochemical composition of motif *u* and *v*. Higher PCC value indicates stronger correlation between two binding motifs, indicating high reproducibility.

To systematically measure the conservation of motifs binding to the same FG, we calculated the pair-wise PCC among all the motifs binding to a specific FG, and evaluated the overall conservation score (CS) as the average of all the pair-wise PCC values:

$$\mathrm{CS} = \frac{\sum_{u=1}^{N}\sum_{v=1,u\neq v}^{N} PCC(M_u, M_v)}{N(N-1)} \quad (7)$$

In addition, a permutation test was used to evaluate the statistical significance of the CS. Specifically, for each FG-binding motif appeared in multiple ligands, we randomly selected same number of motifs from all the identified motifs 1,000,000 times and calculated their corresponding CS value, the *P* value was calculated as follows:

$$P = \frac{\sum_{i=1}^{1e-06}(CS_i > CS_0) + 1}{1000000 + 1} \quad (8)$$

where CS_i is the CS value of randomly selected motifs, CS_0 is the CS value of motifs binding to the same FG.

Clustering on the 3D binding motifs

FG-binding motifs were classified based on their physicochemical properties, specifically, we performed k-means clustering on the 26-dimension vectors representing all the motifs. To determine the optimal number of clusters, we used elbow method which follows the basic idea to minimize the total intra-cluster variation as much as possible. Concretely, we first computed the k-means clustering on the data consisting of 481 vectors for different numbers of clusters k, which is ranging from 1 to 20. Next the total intra-cluster variation was calculated for each k value, and the formula is defined as follows:

$$\sum_{i=1}^{N_C}\sum_{x\in C_i} d(x, \bar{x}_{C_i})^2 \quad (9)$$

where N_C is cluster number, C_i is the *i*-th cluster, $\bar{x}_{C_i}$ is the cluster centroid of C_i. Based on the computed variation under different values of k, a curve of the variation according to the number of clusters k could be plotted. Finally, the location of a bend (k = 4) in the plot was selected as the optimal number of clusters in our approach.

Nomination of FGs for a specific protein target

Given a FA cluster extracted from the ligand-binding pocket of a target protein and a reference FG-binding motif in the database constructed in this study, we defined the FM score as the Pearson's correlation coefficient between the 26-dimensional vectors of the FA cluster (C_u) and the reference motif (M_v):

$$FM = \frac{\sum_{i=1}^{26} \frac{[(c_{u,i} - \overline{c_u}) \times (m_{v,i} - \overline{m_v})]}{26}}{\sqrt{\frac{\sum (c_{u,i} - \overline{c_u})^2}{26}} \times \sqrt{\frac{\sum (m_{v,i} - \overline{m_v})^2}{26}}} \quad (10)$$

where $c_{u,i}$ and $\overline{c_u}$ are the elementwise and average physicochemical composition of the FA cluster C_u, and $m_{v,i}$ and $\overline{m_v}$ are the elementwise and average physicochemical composition of the reference motif M_v. Higher FM score indicate higher similarity between the FA cluster and the reference motif.

For any given FA cluster in the ligand-binding sites of a target protein, we calculated the FM scores against all the 481 FG-binding motifs identified with AFTME and ranked the corresponding FGs by their FM scores from high to low. Since we have multiple same or similar FGs in the dataset, we reasoned that FGs or FG types that are significantly enriched in the top hits should be potential candidates. Specifically, we looked into the top 20% hits and calculated a *p* value for each FG or FG type using hypergeometric test:

$$p(k, M, n, N) = \frac{\binom{n}{k}\binom{M-n}{N-k}}{\binom{M}{N}} \quad (11)$$

where *k* is the number of specific FG or FG types that appear in the top 20%, *M* is the total number of FGs in the dataset, *n* is the total number of a specific FG or FG types, and *N* is the number of FGs ranked in the top 20%. The value was calculated using "hypergeom" module in the SciPy package in python.

Code availability

The source codes of AFTME algorithm and the results of the large-scale functional group (FG)-motif analysis are available at https://github.com/MDhewei/AFTME.

CRediT author statement

Liu Yang: Conceptualization, Methodology, Software, Formal analysis, Data curation, Investigation, Visualization, Writing - original draft, Writing - review & editing. **Wei He:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Investigation, Visualization, Project administration, Writing - original draft, Writing - review & editing. **Yuehui Yun:** Data curation, Formal analysis, Investigation, Writing - review & editing. **Yongxiang Gao:** Investigation, Writing - review & editing. **Zhongliang Zhu:** Investigation, Writing - review & editing. **Maikun Teng:** Investigation, Writing - review & editing. **Zhi Liang:** Conceptualization, Methodology, Software, Investigation, Visualization, Project administration, Supervision, Writing - original draft, Writing - review & editing. **Liwen Niu:** Conceptualization, Methodology, Funding acquisition, Resources, Supervision, Writing - original draft, Writing - review & editing. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 31621002), the Ministry of Science and Technology of China (Grant No. 2017YFA0504903 to LN), and the Hefei National Science Center Pilot Project Funds, China (in part). We thank all the lab members in Niu lab for helpful discussion.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.08.014.

ORCID

ORCID 0000-0002-6549-8482 (Liu Yang)
ORCID 0000-0002-3088-5650 (Wei He)
ORCID 0000-0001-8904-4285 (Yuehui Yun)
ORCID 0000-0003-1214-889X (Yongxiang Gao)
ORCID 0000-0002-3500-0727 (Zhongliang Zhu)
ORCID 0000-0003-4621-826X (Maikun Teng)
ORCID 0000-0001-8055-7922 (Zhi Liang)
ORCID 0000-0003-1217-1782 (Liwen Niu)

References

[1] Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 2010;9:203–14.
[2] Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, et al. Protein function annotation by homology-based inference. Genome Biol 2009;10:207.
[3] Persson J, Beall B, Linse S, Lindahl G. Extreme sequence divergence but conserved ligand-binding specificity in *Streptococcus pyogenes* M protein. PLoS Pathog 2006;2:e47.
[4] Abrusan G, Marsh JA. Ligand binding site structure influences the evolution of protein complex function and topology. Cell Rep 2018;22:3265–76.
[5] Du X, Li Y, Xia YL, Ai SM, Liang J, Sang P, et al. Insights into protein-ligand interactions: mechanisms, models, and methods. Int J Mol Sci 2016;17:144.
[6] Kinjo AR, Nakamura H. Comprehensive structural classification of ligand-binding motifs in proteins. Structure 2009;17:234–46.
[7] Ribeiro VS, Santana CA, Fassio AV, Cerqueira FR, da Silveira CH, Romanelli JPR, et al. visGReMLIN: graph mining-based detection and visualization of conserved motifs at 3D protein-ligand interface at the atomic level. BMC Bioinformatics 2020;21:80.
[8] Hoffmann B, Zaslavskiy M, Vert JP, Stoven V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. BMC Bioinformatics 2010;11:99.
[9] Yeturu K, Chandra N. PocketMatch: a new algorithm to compare binding sites in protein structures. BMC Bioinformatics 2008;9:543.
[10] Gao M, Skolnick J. APoc: large-scale identification of similar protein pockets. Bioinformatics 2013;29:597–604.
[11] Pu L, Govindaraj RG, Lemoine JM, Wu HC, Brylinski M. DeepDrug3D: classification of ligand-binding pockets in proteins with a convolutional neural network. PLoS Comput Biol 2019;15:e1006718.
[12] Hwang H, Dey F, Petrey D, Honig B. Structure-based prediction of ligand-protein interactions on a genome-wide scale. Proc Natl Acad Sci U S A 2017;114:13685–90.
[13] Pires DE, de Melo-Minardi RC, da Silveira CH, Campos FF, Meira Jr W. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. Bioinformatics 2013;29:855–61.
[14] Guvench O. Computational functional group mapping for drug discovery. Drug Discov Today 2016;21:1928–31.
[15] Saraste M, Sibbald PR, Wittinghofer A. The P-loop–a common motif in ATP- and GTP-binding proteins. Trends Biochem Sci 1990;15:430–4.
[16] Via A, Ferre F, Brannetti B, Valencia A, Helmer-Citterich M. Three-dimensional view of the surface motif associated with the P-loop structure: *cis* and *trans* cases of convergent evolution. J Mol Biol 2000;303:455–65.
[17] Narunsky A, Kessel A, Solan R, Alva V, Kolodny R, Ben-Tal N. On the evolution of protein-adenine binding. Proc Natl Acad Sci U S A 2020;117:4701–9.
[18] Denessiouk KA, Rantanen VV, Johnson MS. Adenine recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins. Proteins 2001;44:282–91.
[19] Nebel JC, Herzyk P, Gilbert DR. Automatic generation of 3D motifs for classification of protein binding sites. BMC Bioinformatics 2007;8:321.
[20] Zubieta C, Krishna SS, Kapoor M, Kozbial P, McMullan D, Axelrod HL, et al. Crystal structures of two novel dye-decolorizing peroxidases reveal a beta-barrel fold with a conserved heme-binding motif. Proteins 2007;69:223–33.

[21] Ferousi C, Lindhoud S, Baymann F, Hester ER, Reimann J, Kartal B. Discovery of a functional, contracted heme-binding motif within a multiheme cytochrome. J Biol Chem 2019;294:16953–65.
[22] Nebel JC. Generation of 3D templates of active sites of proteins with rigid prosthetic groups. Bioinformatics 2006;22:1183–9.
[23] Tang GW, Altman RB. Knowledge-based fragment binding prediction. PLoS Comput Biol 2014;10:e1003589.
[24] He W, Liang Z, Teng M, Niu L. mFASD: a structure-based algorithm for discriminating different types of metal-binding sites. Bioinformatics 2015;31:1938–44.
[25] Moodie SL, Mitchell JB, Thornton JM. Protein recognition of adenylate: an example of a fuzzy recognition template. J Mol Biol 1996;263:486–500.
[26] Denessiouk KA, Johnson MS. When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families. Proteins 2000;38:310–26.
[27] Ambudkar SV, Kim IW, Xia D, Sauna ZE. The A-loop, a novel conserved aromatic acid subdomain upstream of the Walker A motif in ABC transporters, is critical for ATP binding. FEBS Lett 2006;580:1049–55.
[28] Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res 2013;41:D1096–103.
[29] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9.
[30] Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, et al. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics 2015;31:405–12.
[31] Ferreira de Freitas R, Schapira M. A systematic analysis of atomic protein-ligand interactions in the PDB. Medchemcomm 2017;8:1970–81.
[32] Young T, Abel R, Kim B, Berne BJ, Friesner RA. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. Proc Natl Acad Sci U S A 2007;104:808–13.
[33] Stein EM, Garcia-Manero G, Rizzieri DA, Tibes R, Berdeja JG, Savona MR, et al. The DOT1L inhibitor pinometostat reduces H3K79 methylation and has modest clinical activity in adult acute leukemia. Blood 2018;131:2661–9.
[34] Dai W, Zhang B, Jiang XM, Su H, Li J, Zhao Y, et al. Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. Science 2020;368:1331–5.
[35] Kahraman A, Morris RJ, Laskowski RA, Thornton JM. Shape variation in protein binding pockets and their ligands. J Mol Biol 2007;368:283–301.
[36] Dukka BK. Structure-based methods for computational protein functional site prediction. Comput Struct Biotechnol J 2013;8:e201308005.
[37] Jost M, Weissman JS. CRISPR approaches to small molecule target identification. ACS Chem Biol 2018;13:366–75.
[38] Fellmann C, Gowen BG, Lin PC, Doudna JA, Corn JE. Cornerstones of CRISPR-Cas in drug discovery and therapy. Nat Rev Drug Discov 2017;16:89–100.
[39] He W, Zhang L, Villarreal OD, Fu R, Bedford E, Dou J, et al. *De novo* identification of essential protein domains from CRISPR-Cas9 tiling-sgRNA knockout screens. Nat Commun 2019;10:4541.
[40] Shi J, Wang E, Milazzo JP, Wang Z, Kinney JB, Vakoc CR. Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nat Biotechnol 2015;33:661–7.
[41] Neggers JE, Kwanten B, Dierckx T, Noguchi H, Voet A, Bral L, et al. Target identification of small molecules using large-scale CRISPR-Cas mutagenesis scanning of essential genes. Nat Commun 2018;9:502.
[42] Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, et al. Impact of high-throughput screening in biomedical research. Nat Rev Drug Discov 2011;10:188–95.
[43] Lavecchia A, Di Giovanni C. Virtual screening strategies in drug discovery: a critical review. Curr Med Chem 2013;20:2839–60.
[44] Halperin I, Glazer DS, Wu S, Altman RB. The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. BMC Genomics 2008;9:S2.
[45] Lam JH, Li Y, Zhu L, Umarov R, Jiang H, Héliou A, et al. A deep learning framework to predict binding preference of RNA constituents on protein surface. Nat Commun 2019;10:1–13.
[46] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.
[47] Cai YD, Qian Z, Lu L, Feng KY, Meng X, Niu B, et al. Prediction of compounds' biological function (metabolic pathways) based on functional group composition. Mol Divers 2008;12:131–7.
[48] Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform 2014;8:14.
[49] Laskowski RA, Swindells MB. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. J Chem Inf Model 2011;51:2778–86.