Title

Clustering Protein Binding Pockets and Identifying Potential Drug Interactions: A Novel Ligand-Based Featurization Method.

Journal

Journal of chemical information and computer sciences, 63(21)

Authors

Stevenson, Garrett

Kirshner, Dan

Bennion, Brian

et al.

Publication Date

2023-11-13

DOI

10.1021/acs.jcim.3c00722

Peer reviewed

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

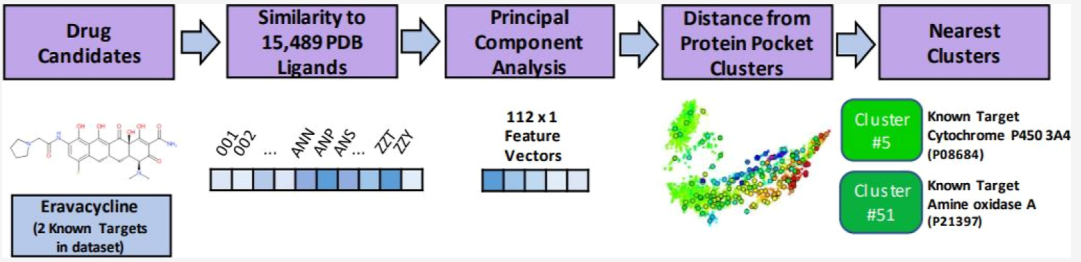
pubs.acs.org/jcim
Article

Clustering Protein Binding Pockets and Identifying Potential Drug Interactions: A Novel Ligand-Based Featurization Method

Garrett A. Stevenson,* Dan Kirshner, Brian J. Bennion, Yue Yang, Xiaohua Zhang, Adam Zemla, Marisa W. Torres, Aidan Epstein, Derek Jones, Hyojin Kim, W. F. Drew Bennett, Sergio E. Wong, Jonathan E. Allen, and Felice C. Lightstone*

Cite This: J. Chem. Inf. Model. 2023, 63, 6655−6666
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Protein−ligand interactions are essential to drug discovery and drug development efforts. Desirable on-target or multitarget interactions are the first step in finding an effective therapeutic, while undesirable off-target interactions are the first step in assessing safety. In this work, we introduce a novel ligand-based featurization and mapping of human protein pockets to identify closely related protein targets and to project novel drugs into a hybrid protein−ligand feature space to identify their likely protein interactions. Using structure-based template matches from PDB, protein pockets are featured by the ligands that bind to their best co-complex template matches. The simplicity and interpretability of this approach provide a granular characterization of the human proteome at the protein-pocket level instead of the traditional protein-level characterization by family, function, or pathway. We demonstrate the power of this featurization method by clustering a subset of the human proteome and evaluating the predicted cluster associations of over 7000 compounds.

INTRODUCTION

Whether a drug candidate is targeted at a single protein or multiple proteins, the candidate must also be tested for potential adverse (off-target) effects and toxicity. Targeted assays are the conventional method to verify the interaction or lack thereof between a new drug and subsets of specific human pathways and proteins.[1] Using several narrowly focused assays to assess a new drug's safety is a response to the complexity of the human proteome. The expanse of proteins coupled with the diversity of their roles and functions is a problem domain too large for drug candidates to be tested in vitro against all possibilities. On the computational side, screening large numbers of drug−protein interactions necessitates high-performance computers.[2,3] However, even with large amounts of computational power and specialized toxicity assays, unforeseen drug−target interactions and their adverse reactions are frequently detrimental to investigational new drugs in clinical trials.[4]

In this work, we demonstrate, using the human proteome, that a pocket-characterization-based approach can facilitate the identification of a drug's most likely targets. We break up 4331 human proteins into clusters of their pockets using similarities in detected protein−ligand pockets instead of grouping proteins by other characteristics such as function or sequence similarity. That is, we focus on evaluating compounds against protein pocket groups, formed by commonalities among the ligands found to bind to those pockets, rather than whole-protein groups, typically formed by commonalities among pathways, families, or functions. Our method is designed to receive a new compound and identify which groups of protein pockets appear to be likely to interact with the compound. The output is potentially useful on several fronts, from prioritizing experimental assays to informing in silico drug design optimization with regard to potential off-target interactions.

BACKGROUND

Describing ligands as keys that match a lock—a protein's binding pocket—is an often-used paradigm in computational

Received: May 11, 2023
Published: October 17, 2023

https://doi.org/10.1021/acs.jcim.3c00722
J. Chem. Inf. Model. 2023, 63, 6655−6666

chemistry.[5] Drug repurposing, multitarget drugs, and off-target binding analyses are areas that extrapolate from a single ligand and protein pocket to a ligand binding to multiple proteins or different ligands binding to the same protein.[6,7] For multiple ligands to bind to a similar protein pocket, they must share some essential, core configuration of features.[8] Conversely, for a ligand to bind to multiple protein pockets, the pockets must share some commonality favorable to that ligand. Such cases are often termed "cross-reactivities" or "cross-sensitivities", where protein pockets that bind similar compounds will in turn have a correlated likelihood of interacting with a similar new drug.[9] The traditional place to search for cross-reactivities is in similar families, pathways, or proteins with similar functions.[10] In this work, however, we focus on individual pockets, which we believe provide greater specificity with regard to relevant protein−ligand interactions while also recognizing relevant commonalities across protein families, pathways, or function categorizations.

To the best of our knowledge, this approach is the first endeavor in the ligand-based clustering of human protein pockets for the identification of potential small-molecule interactions. This work is inspired by our previous study,[11] which demonstrated that PDBspheres, a strictly structure-based approach, can help protein function annotation efforts as well as guide the inference of binding affinity scores from one pocket−ligand pair to another pocket−ligand pair within certain boundaries. Additional related previous work, categorized by methods used for protein- and ligand-based clustering, is summarized below.

Structure/Property-Based Protein Pocket Clustering. Analyses based on individual pockets are more granular since a single protein may have multiple pockets. Significant work has been performed in this area, especially in clustering protein pockets. Weskamp et al. leveraged shape and physiochemical properties to create a global mapping of the cavity (pocket) space and found that similarities in the cavity space are best mapped to ligand binding similarities in comparison to mapping proteins by amino acid sequence or by fold.[12] Note that the analysis depends only on pocket characteristics; ligand characteristics (binding similarities) are used only for validation. Our approach fuses the ligand and protein pocket feature spaces.

Cavbase, a significant advance in protein pocket clustering, provides a means for comparing pockets based on "pseudocenters", which are projections of descriptors in 3D space.[13] Kuhn et al. also applied principal component analysis (PCA) to their Cavbase similarity matrix for clustering selected MAP kinases. Our approach also applies PCA as a preprocessing step but to the fused protein pocket and ligand feature space mentioned above.

The CavitySpace database of potential ligand binding sites in the human proteome includes binding site clusters created with a "PMSmax" similarity, which captures shape and chemical similarities between pockets.[14,15] The authors iteratively applied the Butina clustering algorithm[16] at different PMSmax thresholds. This approach places 31.6% of their database's 111,330 potential binding sites into one cluster, with the conclusion that "the cavities cannot be classified well".

It appears that the vast majority of protein pocket featurization and clustering methods rely, as do these studies, on protein features based on pocket structure and/or physicochemical properties[17−21]—to the exclusion of features based on known interacting ligands. A recent analog to this work is the ProBis-Score[22] scoring function's identification of template ligands for binding pockets. ProBis-Dock[22−24] identifies template ligands using local surface similarities from other binding sites in PDB. Similarly, in this work, we gather PDB ligand IDs from structurally matched templates in PDB. The ProBis-Dock software utilizes this information to label 3D points in a binding site and as input for modeling protein binding site flexibility in docking. On the other hand, we use template-matched ligands to create a clustered protein−ligand feature space which can be queried with either a protein binding site or a small molecule of interest.

Pharmacophore Modeling. Pharmacophore modeling includes a spectrum of approaches with applications in drug discovery, lead optimization, target identification, and toxicity prediction.[25,26] Pharmacophores represent the chemical interactions between a specific ligand or ligands and binding site(s). Pharmacophores can be feature-based[27] or molecular field-based[28] but generally fall into ligand- and structure-based approaches, both of which are related to our approach. For example, protein structure-based methods might infer ideal ligand features from the coordinates of pocket protein residues.[29] The spirit of this approach is analogous to our aggregation of ligands bound to high-scoring template matches in a pocket, where the similarities of the features in the set of co-complex templates emphasize the important components of the underlying pocket structure.

In ligand-based pharmacophore models, known active—and sometimes known inactive—compounds are often used to make inferences or sometimes create training data to virtually screen for new inhibitors.[30,31] Some ligand-based pharmacophores use clustering on known active and/or inactive compounds to group training sets by similarity.[32] Our approach also forms clusters using a similarity metric; however, a major difference arises between what is known about the ligands being clustered. Pharmacophores leverage known binders/nonbinders for clustering, where our approach clusters known and hypothetical binders based on a pocket's structure. We cluster protein pockets featured by ligands from known and predicted template matches. This creates a traversable feature space of protein pocket clusters, while pharmacophore models typically seek to create an in-depth characterization of a particular protein pocket or group of ligands.[33]

Predicting Toxicity, Off-Target Interactions, and Adverse Drug Reactions. A variety of in silico approaches exist for toxicity, adverse drug reaction (ADR), and off-target binding prediction; essentially, whether a candidate drug will bind to proteins that are not the drug's intended target and what the consequences might be. Methods vary from relation extraction from clinical notes[34] to hybrid computational pipelines using molecular docking and machine learning.[35] Pharmacophore modeling is also used, where toxicity and off-target concerns for specific receptors are modeled.[36,37] ADR and toxicity databases like SIDER,[38] T3DB,[39] and DrugBank,[40] have enabled the development and evaluation of numerous machine learning methods. A few notable approaches include REMAP,[41] ToxiM,[42] TargeTox,[43] and eToxPred.[44] Recent advances in the toxicity prediction space also include neural fingerprinting,[45] conformal prediction,[46] and several others.[47−49] While our approach does not delve into predicting ADRs, we expect that our simple, granular featurization approach will be useful in multitarget and toxicity prediction.

https://doi.org/10.1021/acs.jcim.3c00722
J. Chem. Inf. Model. 2023, 63, 6655−6666

METHODS

Pocket Featurization. Our featurization begins with human proteins from the AHA Atlas database,[50] for which high-confidence homology-based structural models have been constructed. Of the 20,375 proteins comprising the human proteome (reviewed reference set UniProt[51] ver. 2020.08.19), the AHA Atlas provides 11,681 structural models created by the homology-based modeling system AS2TS.[52] The Atlas includes structural models based on matches to PDB templates that meet criteria for sequence similarity to and coverage of the reference sequences. In this regard, the AHA Atlas models represent a conventional homology modeling approach; the models do not rely on ab initio (e.g., Rosetta[53]) or AI-informed (e.g., AlphaFold,[54] ESMFold[55]) methods. Potential binding cavities on these structures are identified by structural matches to ligand binding sites in the whole of PDB[56] using the PDBspheres library[11] (Figure 1). While binding sites for individual proteins are sometimes known, many of the structural models in the AHA Atlas do not have solved structures from experiments (i.e., the exact structures or complexes with ligands may not be deposited in the PDB). In this work, we treat pockets identified by PDBspheres as binding sites. Doing so allows for generalization of our approach to any available structural models of a given protein: experimentally solved protein structures, constructed homology models, or structure predictions from methods like AlphaFold.[54]

As mentioned above, the PDBspheres library of binding sites is based on experimentally resolved structures of protein−ligand co-complexes extracted from the PDB database. In this work, we associate those ligands with the binding pockets identified by protein structural matches between the Atlas human proteome structures and the PDBspheres library entries. This may associate multiple ligands with a particular cavity; the matches depend only on protein−protein comparisons; therefore, an Atlas structure cavity may match multiple cavities from different PDB structures having different ligands. We expect that structurally similar pockets will bind similar ligands, which is a concept explored in our previous work.[11] As a corollary, our hypotheses are that (1) ligand similarity provides additional information to characterize pockets (that is, an indication of pocket properties such as charge, hydrophilicity, hydrophobicity, polarity, etc.) and (2) ligand similarity measures can provide a basis for grouping pockets across proteins. We expect that basing our feature vectors on bound ligands (i.e., not focusing on the pocket structure similarity scores) will avoid possible discrepancies in pocket−ligand clustering, which arise from imperfections in the structural conformations of constructed models. Avoiding such pitfalls yields a more reliable and robust model for protein pocket−ligand clustering.

The PDBspheres library[11] is a comprehensive dataset of all experimentally solved protein−ligand co-complexes that can be extracted from PDB. For this work, we start by excluding from the PDB database (in this case from the PDBspheres library) all entries that may not be relevant to noncovalent binding (principally covalently bound branched oligosaccharides), may not be biologically relevant (for example, surfactants used in crystallization), or are antibody co-complexes. We filter out nonrelevant ligands in three ways. Ligands that overwhelmingly appear as crystallization buffers are outright removed (see Supporting Information files for ignored antibody structures and crystallization buffers). Covalently bound oligosaccharides are identified using PDB metadata. Surfactants are identified as ligands that are "surface-bound", that is, not substantially within a protein cavity. This is accomplished by constructing a set of spheres tangent to the calculated protein surface[57] to identify the cavity around the ligand in the PDB biological assembly of the structure used as a basis for the PDBspheres library entry. After removing pockets based on nonrelevant

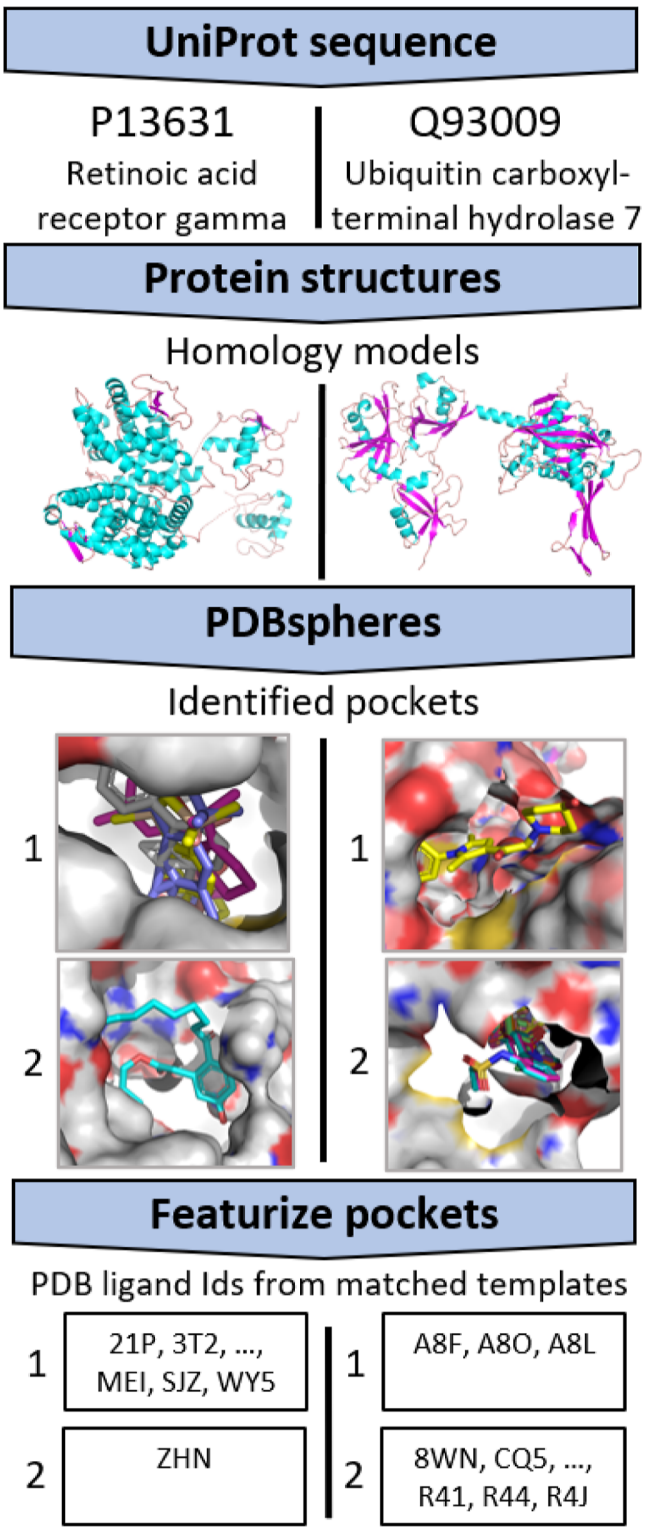


Figure 1. Protein pocket processing and featurization.

https://doi.org/10.1021/acs.jcim.3c00722
J. Chem. Inf. Model. 2023, 63, 6655−6666

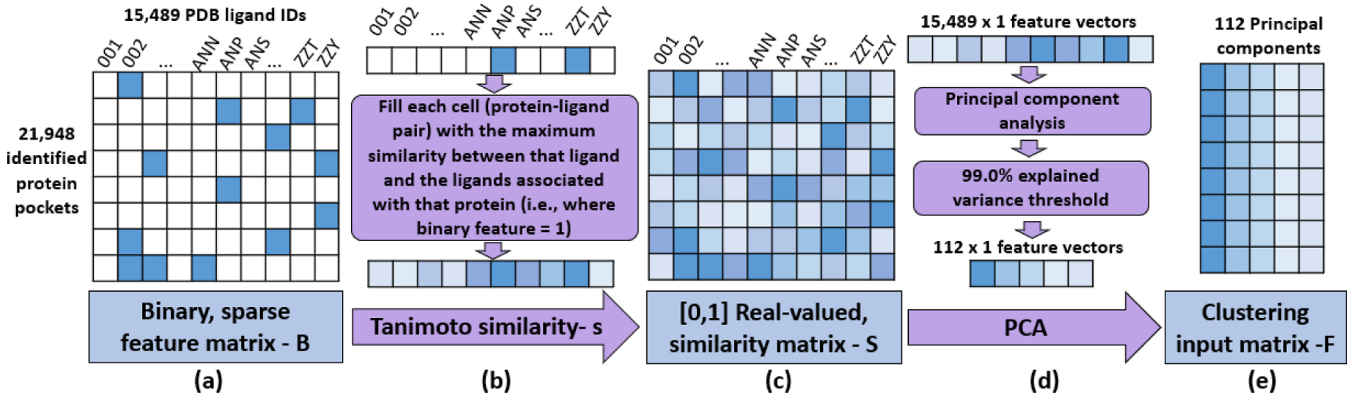


Figure 2. Preparing pocket featurizations for input to clustering.

ligands, each detected Atlas-protein pocket is associated with one or more ligands; the full set includes 21,948 pockets from the 11,681 proteins, where a single protein might have multiple pockets, and 15,489 different ligands. For an individual protein pocket, we retained a maximum of 20 ligands. In cases where more than 20 template matches are found by PDBspheres, we use a global distance calculation to select the best co-complex matches.[58] Consistent with our hypotheses mentioned above, we use the collection of ligands associated with a pocket as a feature vector or "profile", characterizing that pocket. Figure 1 illustrates different ligand associations for two pockets on two different proteins: retinoic acid receptor gamma (left) and ubiquitin carboxyl-terminal hydrolase 7 (right).

As described below, we represent each pocket by a real-valued feature vector derived from the list of ligands associated with the pocket. This serves two purposes. First, we reduce the high dimensionality of both the protein pocket's structure and properties, as well as its associated ligands' structures and properties, to a simple, one-dimensional vector. Second, in characterizing pockets by the small molecules associated with their template matches, our feature space represents a combination of both the ligand binding and pocket structural feature spaces for each entry. Of course, more detailed chemical properties, geometric information, and descriptors may prove beneficial, but we found this foundational method to show significant predictive power on its own.

We assemble our set of pocket−ligand pairs in a 21,948 × 15,489 indicator (binary-valued) matrix, **B**, as shown in Figure 2a. This is a sparse matrix where each row represents a human protein pocket and each column is a PDB ligand ID. The columns (15,489 PDB ligand IDs) are provided as a Supporting Information file. Note that this matrix is predominantly filled with zeros as 80% of individual pockets are associated with five or fewer PDB ligand IDs.

Sparse feature matrices are known to be notoriously difficult for modeling approaches—from logistic regression to deep learning.[59] Sparse PCA[60] was the first approach we tried to directly address starting with a binary-valued, sparse matrix. However, due to the lack of an orthogonality constraint on the fit components, the Scikit-learn[61] implementation we used was unable to provide an explained variance. Therefore, as described in the following paragraphs, we substituted a real-valued similarity measure for the zero-one binary values. The real values are conducive to conventional PCA. The 112 most significant components of these new, real-valued feature vectors across proteins (that is, across the 21,948 vectors of real values, each 15,489 × 1) explain 99% of the variance of the feature vectors. These reduced-dimension (i.e., 112-component) feature vectors are sorted in descending order of explained variance and form the basis for clustering proteins into groups (Figure 2e).

To create the real-valued similarity measure, we first calculate the Tanimoto similarity, which is a [0, 1] continuous value[62] between all ligand pairs. Then, for each pocket (row), for each matrix element in that row we use RDKit[63] to calculate the maximum similarity between that ligand and all the ligands having a matrix element value of 1 in that row; that is, the maximum similarity between that ligand and any of the ligands that have been associated with that pocket, as eq 1 describes

$$f_{ij} = \max(s_{jk}),\ \{k | b_{ik} = 1\} \tag{1}$$

where f_{ij} is the real-valued feature [0,1] for pocket i and ligand j. s_{jk} is the Tanimoto similarity between ligand j and ligand k. b_{ik} is the binary-valued indicator of whether ligand k is associated with pocket i.

Equation 1 provides a fast, deterministic computation that introduces more information into the feature matrix, such that every column is no longer an independent feature. This procedure can be thought of as "filling in missing values". For example, consider the (unrealistic) case where a ligand is duplicated in the set of PDB ligand IDs, that is, the same ligand with different IDs. Assume ligand "*A*" is associated with pocket *i*—either because the protein *i*-ligand "*A*" complex is in PDB or because a structurally matching pocket in another PDB protein is in complex with ligand "*A*". It could be the case that identical ligand "*B*" is not associated with protein *i* because ID "*B*" is not in co-complex with any matching pocket in PDB. In other words, b_{iA} is 1 and b_{iB} is 0. Equation 1 sets $T_{iB} = T_{iA} = 1.0$ since the Tanimoto similarity of a ligand to itself is the maximum value, 1.0.

Using the maximum similarity value between a particular ligand j and all ligands associated with a particular pocket i also accounts for the case where, for example, two very different ligands m and n nevertheless bind to the same pocket. If ligand j is very similar to one or the other of m and n, then T_{ij} will be set to the greater of the two similarity values between ligand j and ligands m and n.

Finally, if ligand j has very low similarity with any of the ligands associated with pocket i, then T_{ij} will be given a low similarity value (albeit the greatest of its low similarities with the associated ligands). This could be a "false negative"; it could be the case that no ligand similar to ligand j has ever

https://doi.org/10.1021/acs.jcim.3c00722
J. Chem. Inf. Model. 2023, 63, 6655−6666

been seen in a PDB co-complex having a pocket matching pocket *i*, but nevertheless such a ligand would bind to pocket *i*. The working assumption of PDBspheres is that the set of pockets identified throughout PDB by proximity to a ligand covers, at least with a high degree of similarity, the full universe of protein binding pockets.

Each pocket has 15,489 (now real-valued) features; that is, the similarities calculated using eq 1 yield the similarity matrix **S** (Figure 2c). We implement feature reduction by performing PCA on this matrix. This provides both the ability to make an interpretable reduction in dimensionality and quantifies the explained variance associated with each component produced. While exploring the linear combinations of singular value decompositions that make up each component is outside the scope of this work, using PCA maintains the possibility of doing so.

While the five most significant components explain 95% of the variance among the original columns, we use the 112 most significant components; these explain 99% of the variance. We substitute the elements of these components for the original 15,489 features for each pocket. The resulting matrix is suitable for clustering.

Pocket Clustering. Since we do not have pre-existing estimates of the number of groups of pockets that will best characterize the human proteome or the subset of the human proteome in the AHA Atlas—we use clustering methods that do not need a target number (prior) of groups to produce. Density-based spatial clustering of applications with noise (DBSCAN[64]) is a clustering algorithm that does not need a number of clusters prior and does not constrain clusters to a specific size. DBSCAN uses the concept of core samples (dense areas) and a distance threshold to form clusters from core samples and nearby noncore samples. The minimum number of points forming a core sample defines how densely populated a neighborhood must be and is a reflection of the noise in the dataset. A distance threshold, the maximum distance between cluster members and the core-sample area, controls the size of clusters, therefore influencing the number of clusters and how many points cannot be clustered within the minimum-size constraint; such points are labeled "outliers" (rather than, say, "monad clusters"). The simplicity of DBSCAN is attractive for interpretability. DBSCAN retains the concept of Euclidean distance between clusters and points, which is captured by the eigenvalues of the principal components. Therefore, the output of DBSCAN clustering on our feature matrix of components can be analyzed in terms of the 112-dimensional Euclidean distance in feature space. It is important to note that by thresholding at 99% of the explained variance, examining the distance between points in our feature matrix is an approximation and is not numerically exact.

Our goal in choosing values for these hyperparameters was to strike a balance between fewer large clusters, where some very large clusters may be dominant, and more small clusters, creating an unnecessarily large number of small clusters to review. We set the minimum number of pockets necessary to create a cluster in a local neighborhood of feature spaces to 10. This helps ensure that clusters are sufficiently populated for the examination and evaluation of their members' characteristics. The other major DBSCAN parameter, the core-to-member maximum distance, has a significant impact on the DBSCAN output; we experimented with a range of values for the distance threshold, examining the results in terms of the number of clusters, number of outliers, and largest cluster size, as shown in Table 1. Between large and small distance thresholds, DBSCAN marks pockets as outliers when they are far from other pockets or in a feature-space region without the minimum 10 nearby pockets necessary to form a cluster. It should be noted that the DBSCAN algorithm allows the 10-member minimum constraint to be violated in cases where the core sample is sufficiently dense.

Table 1. DBSCAN Cluster Results on 21,948 Protein Pockets with Varying Maximum Core-to-Member Distance Threshold

distance threshold	# of clusters	% of outliers (%)	largest cluster size
1.0	194	55.4	1186
1.5	185	49.4	3370
1.9	178	41.1	3640
2.0	**167**	**38.7**	**3696**
2.1	152	33.9	3768
2.5	86	25.9	12,675

Table 1 displays the clustering results achieved by varying the distance threshold from 1 to 2.5, which spans the extremes of many clusters with many outliers to a few clusters with few outliers. A plateau seems to appear in the 1.9−2.1 range where most pockets are in-distribution, the largest cluster sizes are around 4000, and the number of clusters is relatively steady at approximately 160. While the percentage of outliers is larger than desired, the 10-member minimum constraint and the fact that the data are from a subsample of the human proteome both contribute to pockets appearing anomalous. Focusing on the core functionality of the approach and the possibility of adding data in the future, a distance threshold of 2.0 was selected as a reasonable balance among the various factors considered.

RESULTS

The DBSCAN results with a distance threshold of 2.0 yield 167 clusters of 13,455 protein pockets; the remaining 8493 pockets are outliers. The first three principal components of the feature space capture 93.4% of the variance across pockets. Figure 3 visualizes these three components spatially, and the size of each cluster is illustrated by the size of its scatter point. The 167 clusters vary in size from seven to 3696 pockets and as mentioned above, the 10-pocket lower bound discussed can be violated in dense regions. The population of 167 clusters has a median of 16 and a mean of 81 pockets per cluster, a right-skewed distribution. As an initial check of the clustering, we examined the clusters in terms of the unique proteins (UniProt IDs) in each cluster. The statistics for unique proteins are similar to those for pockets; the minimum number of unique proteins in a cluster is six and the maximum number of unique proteins in a cluster is 1768. This indicates that even small clusters group different proteins; thus, small clusters are not simply repeated instances of the same pocket from a single protein biological assembly (as in a homodimer, for example). On average, 52 different proteins make up a cluster, where the median is 14 unique proteins, giving a skewness to the population of pockets.

In accord with previous protein-based clustering efforts,[11] we note several indications that the clusters form biologically meaningful groups. First, we look at cluster members' Enzyme Commission (EC) numbers,[65] which classify enzymes by the

https://doi.org/10.1021/acs.jcim.3c00722
J. Chem. Inf. Model. 2023, 63, 6655−6666

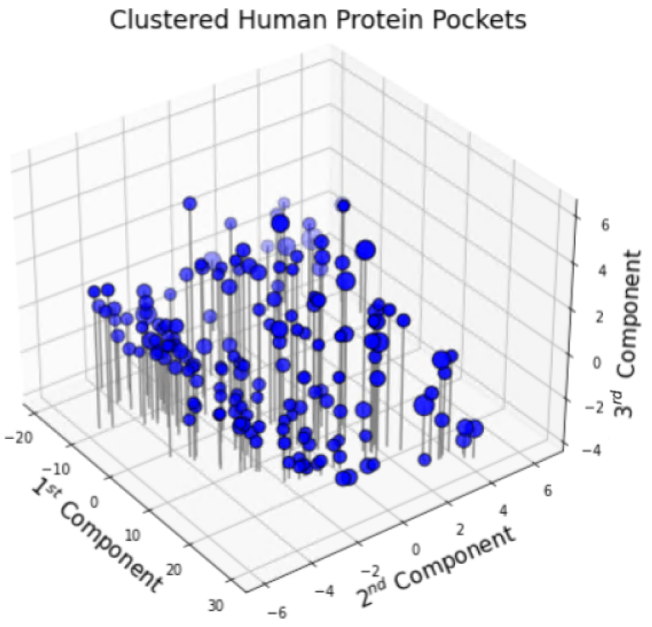


Figure 3. Protein pocket cluster centers (core sample centroids) and 167 clusters visualized by their first three principal components and size.

type of reaction they catalyze. Of the 167 clusters, 54 (32%) consist of proteins having the same major EC class. While typically not all proteins in a cluster have an EC number, nevertheless, a significant number of proteins share the same major class in a significant number of clusters. Furthermore, 36 of the 54 single-major-number clusters consist of proteins also sharing EC subclass and subsubclass numbers, despite the fact that there is no cluster having fewer than six proteins. To distinguish our clustering approach from a classification of binding sites by sequence similarity, we also performed intracluster similarity analysis using the amino acid sequences from the PDB chain associated with each binding site. Clustal Omega[66] executed multiple sequence alignments in each cluster and pairwise similarities between all cluster members were measured with respect to sequence length. Across all clusters, we observed a median of 21.3% intracluster similarity with a range from 3.0 to 70.1%. Clearly, the median within-cluster sequence similarity of 21.3% in our clusters is much lower than one would expect if clusters were formed based on sequence similarity. Thus, our clusters have a makeup significantly different from what a sequence similarity method would construct.

Next, we look at cluster members' appearance in the signaling pathways identified by the small molecule pathway database (SMPDB[67]). There are 17 clusters in which all of the members are associated with proteins in a single pathway; seven of the 17 clusters have more than one protein associated with that same pathway. The pathways associated with those seven clusters are ubiquitin-proteasome, Rac 1 cell motility signaling, and the GnRH signaling pathway. Notably, five of these seven clusters are associated with the ubiquitin-proteasome pathway, which consists of 28 different proteins. In SMPDB, a single protein is often associated with multiple pathways, which makes searching for homogeneous cluster−pathway associations difficult. Filtering clusters for those with multiple proteins associated with the same set of pathways yields 12 clusters of interest. Cluster #29 is one such example; it has 11 pockets from 10 different proteins. The proteins in cluster #29 found in SMPDB are each associated with the same three pathways: folate metabolism, methotrexate action, and methylenetetrahydrofolate reductase deficiency. This association makes sense as the leading PDB ligand IDs associated with the pockets in cluster #29 are folic acid and methotrexate. While observations such as these based on UniProt IDs, EC numbers, and SMPDB associations are interesting, more quantitative validation of the clusters can be achieved.

Comparison of Clustering with Known Protein−Ligand Interactions in DrugBank. One method of judging the accuracy and potential usefulness of our constructed protein clusters is to use the associated data to see which clusters are predicted to be associated with a given ligand and to compare these predictions with known protein interactions for that ligand. Here, we present a method for making such predictions and compare the results with known interactions in the DrugBank database.[40] Our clustered protein set includes proteins for which DrugBank shows interactions with 4827 compounds. Of these 4827 compounds having protein interactions in DrugBank, 2232 were successfully clustered and exist in our feature set (matrix **B**); that is, our approach used these compounds as protein pocket features. The remaining 2595 of the 4827 compounds are "novel" to our method.

We first looked at the DrugBank compounds found in our feature set. Since these compounds are part of the "training" set (e.g., used in our featurization for unsupervised clustering), they do not provide a basis for unbiased tests of the clustering. Certainly, this is the case if the DrugBank interacting protein−ligand pair appears in PDB as a co-complex. In cases where that particular protein−ligand pair does not exist in PDB, it is included in our feature set as a result of pocket similarities. Therefore, the "training" compounds can help test our approach.

We perform our comparison with DrugBank by treating the "training set" compounds as novel ligands and predicting which protein clusters these ligands would be associated with. Our prediction method is this: (1) measure each DrugBank ligand's similarity with the 15,489 ligands in our feature vector; (2) apply the prefit PCA to render 112 × 1 feature vectors used for clustering; (3) calculate the feature-space distance between each ligand and every DBSCAN core sample; and (4) label each DrugBank compound with the number of that nearest core sample's cluster. This creates a projection of each compound in the protein pocket feature space and allows for the creation of an ordered list of the protein pocket clusters nearest each compound.

Table 2. DrugBank Compound Results Are for Known Interactions in Nearby Clusters

DrugBank		one or more interactions in			all known activity in:		
set	count	10 nearest	5 nearest	nearest	10 nearest	5 nearest	nearest
"train"	2232	1129 (51%)	983 (44%)	727 (33%)	726 (33%)	617 (28%)	456 (20%)
test	2595	839 (32%)	763 (29%)	642 (25%)	529 (20%)	423 (16%)	294 (11%)

https://doi.org/10.1021/acs.jcim.3c00722
J. Chem. Inf. Model. 2023, 63, 6655−6666

The first row of Table 2 shows the results for this "training" group of ligands. Of the 2232 DrugBank compounds that are in our feature set, 1129 (51%) have a known interaction with a protein in one or more of the nearest 10 clusters from the field of 167 clusters. Furthermore, 726 of those (33%) have all known interactions within their 10 nearest clusters. These results are positive and demonstrate significant predictive power. Among the five nearest pocket clusters, 44% of the compounds have a known interaction in one or more of the nearest five clusters, and 28% have all known targets within those five nearest clusters. Finally, 456 of the 2232 ligands in this DrugBank subset (20%) are nearest to a cluster which contains all their known interactions. The presence of a protein known to interact with a ligand in the cluster nearest that ligand is preliminary evidence that our approach is functioning properly.

Note that there are several reasons why a ligand may not be projected near its known interacting proteins. First, PDBspheres may not have identified a pocket, meaning it is absent from the dataset in the first place. Second, even for all identified pockets, their predicted template matches may be inaccurate. Third, even if a pocket is identified and well-characterized by the predicted templates, it can be poorly captured by the PCA or, fourth, thrown out as an outlier by the clustering method. Finally, ambiguity from the simple featurization of that pocket may also lead to errors in ligand-to-pocket associations. These possible failures may be ameliorated as new PDB entries can substitute for homology models and allow additional human proteins to be included in the dataset.

The remaining 2595 compounds of the DrugBank subset are not part of our feature set and therefore are not associated with any pocket among our proteins' clusters; thus, they are "novel" to our method. We call these compounds the "test set".

The second line of Table 2 shows the results of applying to the "test" set the same procedure as that applied to the "train" set compounds. Of these 2595 compounds, 839 (32%) are placed with a known target in one or more of the nearest 10 clusters, and 20% have all known targets within the 10 nearest clusters, among our global set of 167 possible clusters. For the five nearest pockets, 763 (29%) of the compounds are placed with one or more known targets and 16% include all known targets. A final point of comparison between the two sets is that 294 drugs are placed in a cluster, which includes all their known targets. These percentages are lower across the board in comparison with those of the "training" set of ligands in DrugBank, which is to be expected. However, this set of compounds is a fair test of the generalization of our method, as the "test set" compounds on average have a maximum similarity of 0.64 to any of the ligands used as features.

Illustration of Predicting Potential On- and Off-Target Interactions. The comparison of predicted ligand−cluster associations with known ligand−protein interactions in DrugBank provides a macrolevel confirmation of the featurization and clustering. Microlevel analysis is possible with the Drug Repurposing Hub database.[68] Of the 6550 Drug Repurposing Hub compounds that are in the AHA Atlas, the Drug Repurposing Hub indicates there are 2899 compounds that have a known interaction with a protein present in one of our clusters but are not included in the ligand feature matrix, **B**. Thus, these compounds are "novel" to our method. Using the criteria from the rightmost columns in Table 2, 405 (14%) have all known interactions within the ten nearest clusters, 334 (12%) are within the nearest five clusters, and 255 (9%) have all known interactions contained by their nearest cluster. The AHA Atlas database provides machine learning (coherent fusion[2]), molecular docking (AutoDock Vina[69]), and MM/GBSA scores for many combinations of protein and ligand.[3,70] Information from these scoring functions allows for a deeper understanding of which ligands and pockets are well represented and why. Among the results from the 9% of compounds with any known activity exclusively in their nearest cluster, we filtered the compounds to those predicted in the smallest clusters to extract a list of five examples.

In the first example, two preclinical compounds (Figure 4, molecules A[71] and B[72]) are present in PDB as co-complexes that match the same protein pocket (in the structural model of UniProt ID Q93009) and were assigned to cluster #58. These compounds are both ubiquitin-specific peptidases. Cluster #58 consists of 67 pockets from 52 different proteins; the most

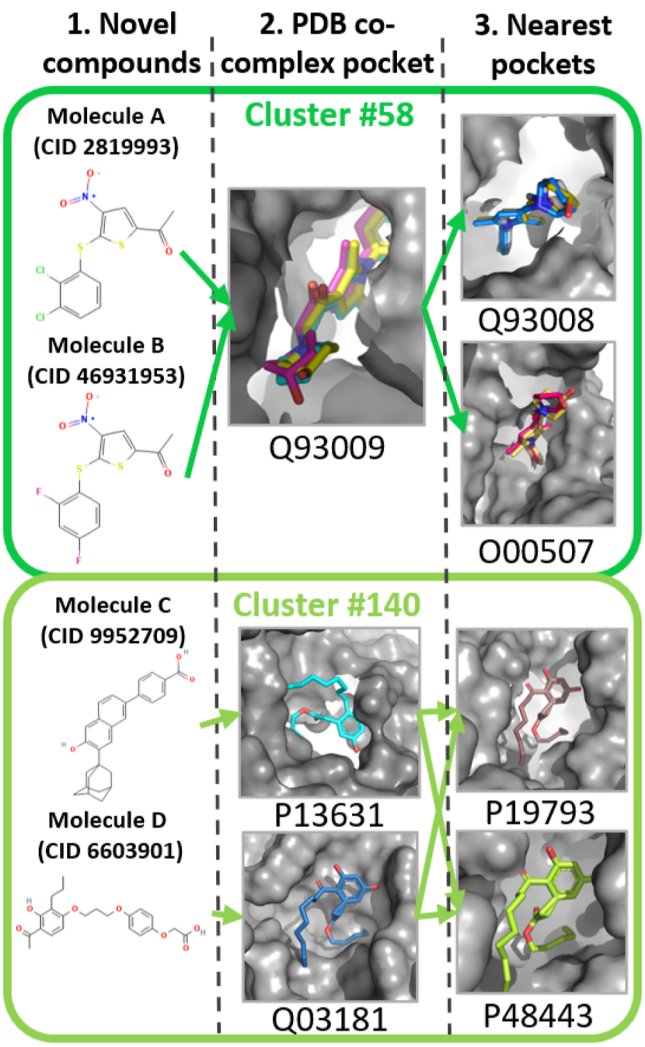


Figure 4. Sample compounds (CID: PubChem compound identifier) that are not in clustering feature matrix B (panel 1); their pose in the protein structural model of the UniProt reference sequence included in clusters #58 and #140—the poses are derived by alignment of the PDB co-complex pockets that matched the structural model to the structural model of the UniProt sequence (panel 2); and the nearest—in feature space—ligands/structural models, again in poses derived by aligning the PDB co-complex to the structural model alignment (panel 3).

https://doi.org/10.1021/acs.jcim.3c00722
J. Chem. Inf. Model. 2023, 63, 6655−6666

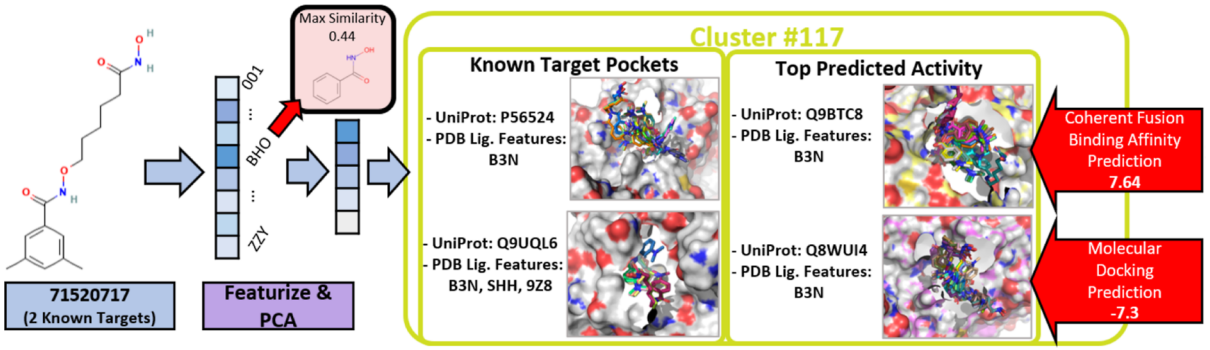


Figure 5. Compound 71520717 in cluster #117 with known and top-predicted activity.

common PDB ligand IDs that feature these pockets are N5S, N6J, HBI, N5D, and A8O. The pocket included in cluster #58, that matches the pocket of the structural model for UniProt human proteome reference sequence Q93009, has three PDBsphere-predicted ligands that featurize it: PDB ligand IDs A8O,[73] A8L,[74] and A8F.[75] Interestingly, molecules A[71] and B[72] (Figure 4) do not have high similarity to any of the 15,489 compounds in our dataset; the dataset compound having the greatest similarity to these two is PDB ligand ID D3U ("2-PCPA derivative"), which has Tanimoto similarities of 0.41 and 0.43 to each, respectively. Cluster #58's pockets represent the most likely candidates with which these two preclinical drugs have interactions. The protein pockets in that cluster that are most similar to the "spheres" (PDB co-complexes) from which molecules A and B were derived are structural models of UniProt IDs (1) Q93008 (probable ubiquitin carboxyl-terminal hydrolase FAF-X), (2) O00507 (probable ubiquitin carboxyl-terminal hydrolase FAF-Y), and (3) Q9UPU5 (ubiquitin carboxyl-terminal hydrolase 24). Among the receptors in cluster #58, our method highlights these three ubiquitin carboxyl-terminal hydrolases as the most likely proteins to interact with molecules A and B. This information is more specific than finding similar proteins through protein family associations. The proteins of the co-complex pocket (Q93009) and all three potential interacting pockets (Q93008, O00507, and Q9UPU5) belong to the protein family (peptidase C19) and have the same EC classification (3.4.19.12). The peptidase C19 family[76] contains 133 different human proteins and there are 782 proteins under the ubiquitinyl hydrolase 1 EC serial number (12), of the omega peptidases subclass (19), in the peptidases subclass, of the hydrolases EC class (3).[77] Instead of generally associating hundreds of proteins with the known target receptor, our approach provides quantified, granular information about similar receptors without limiting the results to ubiquitin carboxyl-terminal hydrolases.

Another pair of preclinical compounds (Figure 4, molecules C[78] and D[79]) were assigned to cluster #140 but have known activity with different proteins. Molecule C is a RARγ antagonist of UniProt ID P13631 (retinoic acid receptor gamma).[80] Molecule D binds to UniProt ID Q03181 (peroxisome proliferator-activated receptor delta).[81] Cluster #140 consists of only 11 pockets from 10 unique proteins, where the predicted co-complex pockets most common for featurizing the cluster #140 pockets have PDB ligands ZHN (pentyl (3,5-dihydroxy-2-nonanoylphenyl)acetate) and VIT (vitamin e). Also in cluster #140 are pockets on UniProt IDs P19793 (retinoic acid receptor RXR-alpha) and P48443 (retinoic acid receptor RXR-gamma). While our initial knowledge of interactions shows molecule D having activity only against Q03181, PubChem bioassay results indicate that molecule D has activity against RARA and RXRA, with potencies of 7.7 and 15.5 μM, respectively.[79,81] This is a significant finding that demonstrates a new use case, where activity not previously in our dataset was found by looking at activity for pockets in the same cluster. Both compounds were most similar to PDB ligands in our feature matrix **B** that were not used to characterize any of the pockets in this cluster. Molecule D was most similar to the PDB ligand JNM (Tanimoto similarity 0.66); molecule C was most similar to E9T (Tanimoto similarity 0.81). The absence of JNM and E9T as features in any of the cluster #140 pockets and the 39% pairwise sequence similarity between their targets indicate the predictive power of the information captured by this method for pocket featurization.

As mentioned above, the AHA Protein Atlas includes docking scores and coherent fusion[2,82] machine learning scores for 6550 compounds from the drug repurposing hub. This allows for quantification of how pockets in the same cluster might interact with a new compound. Using the same four-step procedure described above with regard to placing DrugBank compounds with their nearest cluster for all of the drug repurposing hub compounds, PubChem compound 71520717[83] (*N*-[[6-(hydroxyamino)-6-oxohexyl]oxy]-3,5-dimethylbenzamide) is placed closest to cluster #117. Displayed in Figure 5, compound 71520717 has two known interactions, histone deacetylases 4 (P56524) and 5 (Q9UQL6), both of which have a pocket in cluster #117.[84] Cluster #117 is made up of 10 total pockets from 9 different proteins, and every pocket has PDB ligand ID B3N associated with its co-complex pocket matches except for Q9UQL6, which is characterized by PDB ligands SHH, 9Z8, and B3N. Compound 71520717, like molecules A and B above, is significantly different from other ligands in the feature vector, where its best match is PDB ligand ID BHO (benzhydroxamic acid), having a Tanimoto similarity of 0.44.

Among the 10 pockets that make up cluster #117, the AHA Atlas's coherent fusion machine learning model and AutoDock Vina calculations highlight, with scores better than the Atlas's threshold for more detailed binding analyses, two different proteins likely to interact with compound 71520717. UniProt ID Q8WUI4 (histone deacetylase 7) has the best docking score among the cluster's pockets. PubChem's bioassay results report a 0.17 μ M "Active" result for compound 71520717.[85] Activity at various levels is also shown for the other histone deacetylases (2,6,8) in cluster #117. This evidence of cross-

https://doi.org/10.1021/acs.jcim.3c00722
J. Chem. Inf. Model. 2023, 63, 6655−6666

sensitivities in the cluster is indicative of the similarity of the pockets and the value of the information that they are grouped together. The machine learning model's leading prediction highlights a pocket from the structural model of UniProt Q8WUI4 (metastasis-associated protein MTA3) as likely to interact with compound 71520717. Unlike the docking prediction, the coherent fusion model's prediction does not have evidence in the literature or from PubChem. Instead, assuming the biological relevance of the interaction, it highlights a novel protein that might be examined for interaction.

DISCUSSION

These use cases illustrate how our ligand-based featurization and clustering approach may reveal useful information about candidate drugs and protein pocket interactions in the human proteome. While this simple approach is not perfect, it shows promise as a powerful foundational approach to improving algorithmically. As the number of solved structures in the PDB continues to grow, template-based binding site identification methods such as PDBspheres will become more accurate. The 4331 human proteins used in this work, which are provided in a Supporting Information file, can also be expanded beyond the AHA Atlas's existing subset. Adding data points to this approach will serve to improve the results in Table 2. DrugBank compounds result in known interactions in nearby clusters and increase granularity. In its current state, the featurization/clustering often associates new compounds with the larger clusters. Despite controlling for cluster size in choosing the DBSCAN clustering parameters, a few clusters stand out as dominant. Cluster #1 is made up of 1324 pockets which are characterized by template matches to NHE, HEM, 1MK, ARG, and HEA. Cluster #3 contains 653 pockets which are generally branched oligosaccharides made up of GLC, BGC, GAL, and UMQ. Finally, cluster #5 contains 1547 pockets often associated with HEM, HEC, RLZ, JNI, and FMN. While predicting ligands to be associated with these clusters is not outright wrong, their sizes make analysis more difficult. Additional data will naturally serve to reduce this occurrence, but other clustering methods such as OPTICS[86] may also provide an approach to keeping small clusters and breaking up larger ones without creating unreasonable numbers of divisions. In fact, early efforts in clustering data from the AHA Atlas revealed some valuable subgroups in the larger clusters.

Nevertheless, only a subset of the possible use cases for this approach have been covered. The protein pocket clusters may have uses spanning from high-throughput screening to better understanding a proteome. In drug development, the feature space created here can aid in ligand optimization, suggest other pockets to target, and provide lead pockets for toxicity concerns or assay prioritization. From a proteomic standpoint, small-molecule pathways can be considered in the context of the clusters in which their proteins and pockets are, and vice versa. Because the feature space is traversable in terms of coordinates and distance, measurements among pockets, clusters, and regions might reveal similarities and differences in a quantifiable way. Additionally, extending this method to other proteomes might reveal the nearest analogs between humans and animals, such as nonhuman primates, mice, rats, etc. As a template-based method, recent advances in protein structure prediction will have a significant effect on our approach's utility for proteins without known binding sites or activities.

CONCLUSIONS

In this work, we describe and demonstrate a novel ligand-based featurization of protein pockets and clustering. The latent space captured shows strong evidence of accuracy by nominating potential multitarget candidate pockets for either designing multimodal drugs or testing in toxicity assays. The straightforward, data-driven approach developed is able to associate previously unseen drugs with their known target proteins and pockets while also suggesting protein/ligand interactions that are denoted as "Active" in PubChem. The AHA Protein Atlas and binding affinity prediction methods serve to confirm the validity of the clusters formed. Our approach should have widespread utility, and in future work, we plan to explore additional use cases, clustering approaches, and improvements to pocket featurization.

ASSOCIATED CONTENT

Data Availability Statement

The protein pockets clustered in this work come from the AHA Atlas structural models (https://doi.org/10.11578/1969730) based on the UniProt human proteome (https://www.uniprot.org/uniprotkb?query=reviewed%3Atrue%20AND%20proteome%3Aup000005640) reviewed reference set version 2020.08.19. The structural models were created using the AS2TS system (http://proteinmodel.org/). The compounds and associated protein targets used to evaluate this approach come from DrugBank version 5.1.9 (https://go.drugbank.com/releases/5-1-9) and the drug repurposing hub version 3/24/2020 (https://clue.io/repurposing). The source code and data for PDBspheres is available at https://github.com/LLNL/PDBspheres. The PCA and DBSCAN clustering methods used come from Scikit-learn (https://scikit-learn.org/) and the molecular docking scores were acquired via ConveyorLC (https://github.com/XiaohuaZhangLLNL/conveyorlc).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00722.

PDB identifiers used to filter out crystallization buffer compounds and antibody structures, PDB identifiers for the 15,489 ligands used for featurization, and a list of UniProt identifiers for proteins (XLSX)

Feature vector of ligands (XLSX)

Ignored buffer compound (XLSX)

Removed antibody structures (XLSX)

Clustered human protein pockets (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Garrett A. Stevenson − *Computational Engineering Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States;* orcid.org/0000-0001-7085-8334; Email: stevenson32@llnl.gov

Felice C. Lightstone − *Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*; Email: felice@llnl.gov

https://doi.org/10.1021/acs.jcim.3c00722
J. Chem. Inf. Model. 2023, 63, 6655−6666

Authors

Dan Kirshner − *Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*

Brian J. Bennion − *Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*

Yue Yang − *Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*

Xiaohua Zhang − *Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*

Adam Zemla − *Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*

Marisa W. Torres − *Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*

Aidan Epstein − *Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*

Derek Jones − *Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States; Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California 92093, United States;* orcid.org/0000-0002-9510-6662

Hyojin Kim − *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*

W. F. Drew Bennett − *Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States;* orcid.org/0000-0003-3993-9077

Sergio E. Wong − *Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*

Jonathan E. Allen − *Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States;* orcid.org/0000-0002-4359-8263

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00722

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the American Heart Association Cooperative Research and Development Agreement TC02274 and by the Defense Threat Reduction Agency under grant HDTRA1242044. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344.

■ REFERENCES

(1) Jeong, J.; Kim, D.; Choi, J. Application of ToxCast/Tox21 Data for Toxicity Mechanism-Based Evaluation and Prioritization of Environmental Chemicals: Perspective and Limitations. *Toxicol. In Vitro* **2022**, *84*, 105451.

(2) Stevenson, G. A.; et al. High-Throughput Virtual Screening of Small Molecule Inhibitors for SARS-CoV-2 Protein Targets with Deep Fusion Models. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, New York*: NY, USA, 2021.

(3) Lau, E. Y.; Negrete, O. A.; Bennett, W. F. D.; Bennion, B. J.; Borucki, M.; Bourguet, F.; Epstein, A.; Franco, M.; Harmon, B.; He, S.; et al. Discovery of Small-Molecule Inhibitors of SARS-CoV-2 Proteins Using a Computational and Experimental Pipeline. *Front. Mol. Biosci.* **2021**, *8*, 678701.

(4) Sun, D.; Gao, W.; Hu, H.; Zhou, S. Why 90% of Clinical Drug Development Fails and How to Improve It? *Acta Pharm. Sin. B* **2022**, *12*, 3049−3062.

(5) Tutone, M.; Almerico, A. M. *Targeting Enzymes for Pharmaceutical Development: Methods and Protocols*; Labrou, N. E., Ed.; Springer US: New York, NY, 2020; pp 29−39.

(6) Löscher, W. Single-Target Versus Multi-Target Drugs Versus Combinations of Drugs With Multiple Targets: Preclinical and Clinical Evidence for the Treatment or Prevention of Epilepsy. *Front. Pharmacol* **2021**, *12*, 730257.

(7) Makhoba, X. H.; Viegas, C., Jr; Mosa, R. A.; Viegas, F. P.; Pooe, O. J. Potential Impact of the Multi-Target Drug Approach in the Treatment of Some Complex Diseases . *Drug Des., Dev. Ther.* **2020**, *14*, 3235−3249.

(8) Talevi, A. Multi-Target Pharmacology: Possibilities and Limitations of the "Skeleton Key Approach" From a Medicinal Chemist Perspective. *Front. Pharmacol* **2015**, *6*, 205.

(9) March-Vila, E.; Pinzi, L.; Sturm, N.; Tinivella, A.; Engkvist, O.; Chen, H.; Rastelli, G. On the Integration of In Silico Drug Design Methods for Drug Repurposing. *Front. Pharmacol* **2017**, *8*, 298.

(10) Múnera, M.; Martínez, D.; Labrada, A.; Caraballo, L.; Puerta, L. Identification of B Cell Epitopes of Blo T 13 Allergen and Cross-reactivity With Human Adipocytes and Heart Fatty Acid Binding Proteins. *Int. J. Mol. Sci.* **2019**, *20*, 6107.

(11) Zemla, A. T.; Allen, J. E.; Kirshner, D.; Lightstone, F. C. PDBspheres: A Method for Finding 3D Similarities in Local Regions in Proteins. *NAR: Genomics Bioinf.* **2022**, *4*, lqac078.

(12) Weskamp, N.; Hüllermeier, E.; Klebe, G. Merging Chemical and Biological Space: Structural Mapping of Enzyme Binding Pocket Space. *Proteins: Struct., Funct., Bioinf.* **2009**, *76*, 317−330.

(13) Kuhn, D.; Weskamp, N.; Hüllermeier, E.; Klebe, G. Functional Classification of Protein Kinase Binding Sites Using Cavbase. *ChemMedChem* **2007**, *2*, 1432−1447.

(14) 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.

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

(16) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Model.* **1999**, *39*, 747−750.

(17) Stegemann, B.; Klebe, G. Cofactor-Binding Sites in Proteins of Deviating Sequence: Comparative Analysis and Clustering in Torsion Angle, Cavity, and Fold Space. *Proteins: Struct., Funct., Bioinf.* **2012**, *80*, 626−648.

(18) Guo, Z., Chen, B. Y. Variational Bayesian Clustering on Protein Cavity Conformations for Detecting Influential Amino Acids. *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, **2014**; pp 703−710.

(19) Desdouits, N.; Nilges, M.; Blondel, A. Principal Component Analysis Reveals Correlation of Cavities Evolution and Functional Motions in Proteins. *J. Mol. Graphics Modell.* **2015**, *55*, 13−24.

(20) Kupas, K.; Ultsch, A.; Klebe, G. Large Scale Analysis of Protein-Binding Cavities Using Self-organizing Maps and Wavelet-Based Surface Patches to Describe Functional Properties, Selectivity Discrimination, and Putative Cross-reactivity. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 1288−1306.

(21) Derry, A.; Altman, R. B. COLLAPSE: A Representation Learning Framework for Identification and Characterization of Protein Structural Sites. *Protein Sci.* **2023**, *32*, No. e4541.
(22) Konc, J.; Lešnik, S.; Škrlj, B.; Sova, M.; Proj, M.; Knez, D.; Gobec, S.; Janežič, D. ProBiS-Dock: A Hybrid Multitemplate Homology Flexible Docking Algorithm Enabled by Protein Binding Site Comparison. *J. Chem. Inf. Model.* **2022**, *62*, 1573−1584.
(23) Konc, J.; Lešnik, S.; Škrlj, B.; Janežič, D. ProBiS-Dock Database: A Web Server and Interactive Web Repository of Small Ligand−Protein Binding Sites for Drug Design. *J. Chem. Inf. Model.* **2021**, *61*, 4097−4107.
(24) Konc, J.; Janežič, D. ProBiS-Fold Approach for Annotation of Human Structures from the AlphaFold Database with No Corresponding Structure in the PDB to Discover New Druggable Binding Sites. *J. Chem. Inf. Model.* **2022**, *62*, 5821−5829.
(25) Choudhury, C.; Narahari Sastry, G. *Structural Bioinformatics: Applications in Preclinical Drug Discovery Process*; Mohan, C. G., Ed.; Springer International Publishing: Cham, 2019; pp 25−53.
(26) Schaller, D.; Šribar, D.; Noonan, T.; Deng, L.; Nguyen, T. N.; Pach, S.; Machalz, D.; Bermudez, M.; Wolber, G. Next Generation 3D Pharmacophore Modeling. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, No. e1468.
(27) Salam, N. K.; Nuti, R.; Sherman, W. Novel Method for Generating Structure-Based Pharmacophores Using Energetic Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 2356−2368.
(28) Mortier, J.; Dhakal, P.; Volkamer, A. Truly Target-Focused Pharmacophore Modeling: A Novel Tool for Mapping Intermolecular Surfaces. *Molecules* **2018**, *23*, 1959.
(29) Tran-Nguyen, V.-K.; Da Silva, F.; Bret, G.; Rognan, D. All in one: Cavity detection, druggability estimate, cavity-based pharmacophore perception, and virtual screening. *J. Chem. Inf. Model.* **2019**, *59*, 573−585.
(30) Sangande, F.; Julianti, E.; Tjahjono, D. H. Ligand-Based Pharmacophore Modeling, Molecular Docking, and Molecular Dynamic Studies of Dual Tyrosine Kinase Inhibitor of EGFR and VEGFR2. *Int. J. Mol. Sci.* **2020**, *21*, 7779.
(31) Castleman, P.; Szwabowski, G.; Bowman, D.; Cole, J.; Parrill, A.; Baker, D. Ligand-Based G Protein Coupled Receptor Pharmacophore Modeling: Assessing the Role of Ligand Function in Model Development. *J. Mol. Graphics Modell.* **2022**, *111*, 108107.
(32) Kutlushina, A.; Khakimova, A.; Madzhidov, T.; Polishchuk, P. Ligand-Based Pharmacophore Modeling Using Novel 3D Pharmacophore Signatures. *Molecules* **2018**, *23*, 3094.
(33) Pascual, R.; Almansa, C.; Plata-Salamán, C.; Vela, J. M. A New Pharmacophore Model for the Design of Sigma-1 Ligands Validated on a Large Experimental Dataset. *Front. Pharmacol* **2019**, *10*, 519.
(34) Florez, E.; Precioso, F.; Pighetti, R.; Riveill, M. Deep Learning for Identification of Adverse Drug Reaction Relations. *Proceedings of the 2019 International Symposium on Signal Processing Systems*, 2019; pp 149−153.
(35) LaBute, M. X.; Zhang, X.; Lenderman, J.; Bennion, B. J.; Wong, S. E.; Lightstone, F. C. Adverse Drug Reaction Prediction Using Scores Produced by Large-Scale Drug-Protein Target Docking on High-Performance Computing Machines. *PLoS One* **2014**, *9*, No. e106298.
(36) Schieferdecker, S.; Vock, E. Development of Pharmacophore Models for the Important Off-Target 5-HT2B Receptor. *J. Med. Chem.* **2023**, *66*, 1509−1521.
(37) Pal, S.; Kumar, V.; Kundu, B.; Bhattacharya, D.; Preethy, N.; Reddy, M. P.; Talukdar, A. Ligand-Based Pharmacophore Modeling, Virtual Screening and Molecular Docking Studies for Discovery of Potential Topoisomerase I Inhibitors. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 291−310.
(38) Kuhn, M.; Letunic, I.; Jensen, L. J.; Bork, P. The SIDER Database of Drugs and Side Effects. *Nucleic Acids Res.* **2016**, *44*, D1075−D1079.
(39) Lim, E.; Pon, A.; Djoumbou, Y.; Knox, C.; Shrivastava, S.; Guo, A. C.; Neveu, V.; Wishart, D. S. T3DB: A Comprehensively Annotated Database of Common Toxins and Their Targets. *Nucleic Acids Res.* **2010**, *38*, D781−D786.
(40) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074−D1082.
(41) Lim, H.; Poleksic, A.; Yao, Y.; Tong, H.; He, D.; Zhuang, L.; Meng, P.; Xie, L. Large-Scale Off-Target Identification Using Fast and Accurate Dual Regularized One-Class Collaborative Filtering and Its Application to Drug Repurposing. *PLoS Comput. Biol.* **2016**, *12*, No. e1005135.
(42) Sharma, A. K.; Srivastava, G. N.; Roy, A.; Sharma, V. K. ToxiM: A Toxicity Prediction Tool for Small Molecules Developed Using Machine Learning and Chemoinformatics Approaches. *Front. Pharmacol* **2017**, *8*, 880.
(43) Lysenko, A.; Sharma, A.; Boroevich, K. A.; Tsunoda, T. An Integrative Machine Learning Approach for Prediction of Toxicity-Related Drug Safety. *Life Sci. Alliance* **2018**, *1*, No. e201800098.
(44) Pu, L.; Naderi, M.; Liu, T.; Wu, H.-C.; Mukhopadhyay, S.; Brylinski, M. eToxPred: A Machine Learning-Based Approach to Estimate the Toxicity of Drug Candidates. *BMC Pharmacol. Toxicol.* **2019**, *20*, 2.
(45) Dey, S.; Luo, H.; Fokoue, A.; Hu, J.; Zhang, P. Predicting Adverse Drug Reactions Through Interpretable Deep Learning Framework. *BMC Bioinf.* **2018**, *19*, 476.
(46) Lampa, S.; Alvarsson, J.; Arvidsson Mc Shane, S.; Berg, A.; Ahlberg, E.; Spjuth, O. Predicting Off-Target Binding Profiles With Confidence Using Conformal Prediction. *Front. Pharmacol* **2018**, *9*, 1256.
(47) Mohsen, A.; Tripathi, L. P.; Mizuguchi, K. Deep Learning Prediction of Adverse Drug Reactions in Drug Discovery Using Open TG−GATEs and FAERS Databases. *Front. Drug Discovery* **2021**, *1*, 768792.
(48) Sachdev, K.; Gupta, M. K. A Comprehensive Review of Computational Techniques for the Prediction of Drug Side Effects. *Drug Dev. Res.* **2020**, *81*, 650−670.
(49) Vo, A. H.; Van Vleet, T. R.; Gupta, R. R.; Liguori, M. J.; Rao, M. S. An Overview of Machine Learning and Big Data for Drug Toxicity Evaluation. *Chem. Res. Toxicol.* **2020**, *33*, 20−37.
(50) American Heart Association. *AHA Protein Atlas Database*. 2023; , (accessed July 28, 2023).
(51) Coudert, E.; Gehant, S.; de Castro, E.; Pozzato, M.; Baratin, D.; Neto, T.; Sigrist, C. J.; Redaschi, N.; Bridge, A.; Bridge, A. J.; et al. Annotation of Biologically Relevant Ligands in UniProtKB Using ChEBI. *Bioinformatics* **2023**, *39*, btac793.
(52) Zemla, A.; Zhou, C. E.; Slezak, T.; Kuczmarski, T.; Rama, D.; Torres, C.; Sawicka, D.; Barsky, D. AS2TS System for Protein Structure Modeling and Analysis. *Nucleic Acids Res.* **2005**, *33*, W111−W115.
(53) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. *Methods in Enzymology*; Elsevier, 2004; Vol. *383*, pp 66−93.
(54) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction With AlphaFold. *Nature (London, U.K.)* **2021**, *596*, 583−589.
(55) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Language models of protein sequences at the scale of evolution enable accurate structure prediction. **2022**, bioRxiv:2022.07.20.500902.
(56) 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.
(57) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

https://doi.org/10.1021/acs.jcim.3c00722
J. Chem. Inf. Model. 2023, 63, 6655−6666

(58) Keedy, D. A.; Williams, C. J.; Headd, J. J.; Arendall, W. B., III; Chen, V. B.; Kapral, G. J.; Gillespie, R. A.; Block, J. N.; Zemla, A.; Richardson, D. C.; et al. The other 90% of the protein: Assessment beyond the Cαs for CASP8 template-based and high-accuracy models. *Proteins: Struct., Funct., Bioinf.* **2009**, *77*, 29−49.
(59) Kuss, O. Global Goodness-of-Fit Tests in Logistic Regression With Sparse Data. *Stat. Med.* **2002**, *21*, 3789−3801.
(60) Jenatton, R.; Obozinski, G.; Bach, F. Structured Sparse Principal Component Analysis. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010; pp 366−373.
(61) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(62) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Model.* **1998**, *38*, 983−996.
(63) Landrum, G. *RDKit: Open-Source Cheminformatics*, 2022. https://www.rdkit.org (accessed 14 April, 2023).
(64) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996; pp 226−231.
(65) Webb, E. C.; et al. *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*; Academic Press, 1992.
(66) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539.
(67) Jewison, T.; Su, Y.; Disfany, F. M.; Liang, Y.; Knox, C.; Maciejewski, A.; Poelzer, J.; Huynh, J.; Zhou, Y.; Arndt, D.; et al. SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database. *Nucleic Acids Res.* **2014**, *42*, D478−D484.
(68) Corsello, S. M.; Bittker, J. A.; Liu, Z.; Gould, J.; McCarren, P.; Hirschman, J. E.; Johnston, S. E.; Vrcic, A.; Wong, B.; Khan, M.; et al. The Drug Repurposing Hub: A Next-Generation Drug Library and Information Resource. *Nat. Med.* **2017**, *23*, 405−408.
(69) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **2021**, *61*, 3891−3898.
(70) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the Performance of the Molecular Mechanics/Poisson Boltzmann Surface Area and Molecular Mechanics/Generalized Born Surface Area Methods. II. The Accuracy of Ranking Poses Generated From Docking. *J. Comput. Chem.* **2011**, *32*, 866−877.
(71) National Center for Biotechnology Information, PubChem. Compound Summary for CID 2819993. 2023. https://pubchem.ncbi.nlm.nih.gov/compound/p005091 (accessed April 17, 2023).
(72) National Center for Biotechnology Information, PubChem. Compound Summary for CID 46931953. 2023. https://pubchem.ncbi.nlm.nih.gov/compound/p22077 (accessed April 17, 2023).
(73) RCSB, PDB. A80 Ligand Summary Page. 2023. https://www.rcsb.org/ligand/A80 (accessed July 17, 2023).
(74) RCSB, PDB. A8L Ligand Summary Page. 2023. https://www.rcsb.org/ligand/A8L (accessed July 17, 2023).
(75) RCSB, PDB. A8F Ligand Summary Page. 2023. https://www.rcsb.org/ligand/A8F (accessed July 17, 2023).
(76) The UniProt Consortium. UniProt Summary for Pepsidase C19 Family Human Proteins. 2023. https://www.uniprot.org/uniprotkb?facets=model_organism%3A9606&query=%28family%3A%22peptidase%20C19%20family%22%29 (accessed April 19, 2023).
(77) Fleischmann, A.; Darsow, M.; Degtyarenko, K.; Fleischmann, W.; Boyce, S.; Axelsen, K. B.; Bairoch, A.; Schomburg, D.; Tipton, K. F.; Apweiler, R. IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.* **2004**, *32*, D434−D437.
(78) National Center for Biotechnology Information, PubChem. Compound Summary for CID 9952709. 2023. https://pubchem.ncbi.nlm.nih.gov/compound/cd-1530 (accessed April 17, 2023).
(79) National Center for Biotechnology Information, PubChem. Compound Summary for CID 6603901. 2023. https://pubchem.ncbi.nlm.nih.gov/compound/l-165041 (accessed April 17, 2023).
(80) Thacher, S. M.; Vasudevan, J.; Chandraratna, R. A. Therapeutic Applications for Ligands of Retinoid Receptors. *Curr. Pharm. Des.* **2000**, *6*, 25−58.
(81) Berger, J.; Leibowitz, M. D.; Doebber, T. W.; Elbrecht, A.; Zhang, B.; Zhou, G.; Biswas, C.; Cullinan, C. A.; Hayes, N. S.; Li, Y.; et al. Novel Peroxisome Proliferator-Activated Receptor (PPAR) γ and PPARδ Ligands Produce Distinct Biological Effects. *J. Biol. Chem.* **1999**, *274*, 6718−6725.
(82) Jones, D.; Kim, H.; Zhang, X.; Zemla, A.; Stevenson, G.; Bennett, W. F. D.; Kirshner, D.; Wong, S. E.; Lightstone, F. C.; Allen, J. E. Improved Protein−Ligand Binding Affinity Prediction With Structure-Based Deep Fusion Inference. *J. Chem. Inf. Model.* **2021**, *61*, 1583−1592.
(83) National Center for Biotechnology Information, PubChem. Compound Summary for CID 71520717. 2023. https://pubchem.ncbi.nlm.nih.gov/compound/lmk-235 (accessed April 19, 2023).
(84) Marek, L.; Hamacher, A.; Hansen, F. K.; Kuna, K.; Gohlke, H.; Kassack, M. U.; Kurz, T. Histone Deacetylase (HDAC) Inhibitors With a Novel Connecting Unit Linker Region Reveal a Selectivity Profile for HDAC4 and HDAC5 With Improved Activity Against Chemoresistant Cancer Cells. *J. Med. Chem.* **2013**, *56*, 427−436.
(85) Asfaha, Y.; Schrenk, C.; Alves Avelar, L. A.; Lange, F.; Wang, C.; Bandolik, J. J.; Hamacher, A.; Kassack, M. U.; Kurz, T. Novel Alkoxyamide-Based Histone Deacetylase Inhibitors Reverse Cisplatin Resistance in Chemoresistant Cancer Cells. *Bioorg. Med. Chem.* **2020**, *28*, 115108.
(86) Ankerst, M.; Breunig, M. M.; Kriegel, H.-P.; Sander, J. OPTICS: Ordering Points to Identify the Clustering Structure. *SIGMOD Rec.* **1999**, *28*, 49−60.

https://doi.org/10.1021/acs.jcim.3c00722
J. Chem. Inf. Model. 2023, 63, 6655−6666