JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

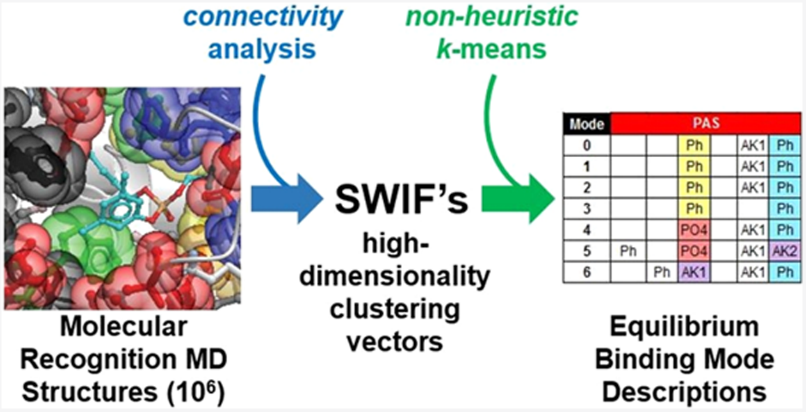
# Overcoming the Heuristic Nature of *k*-Means Clustering: Identification and Characterization of Binding Modes from Simulations of Molecular Recognition Complexes

Parker Ladd Bremer, Danna De Boer, Walter Alvarado, Xavier Martinez, and Eric J. Sorin*

Cite This: *J. Chem. Inf. Model.* 2020, 60, 3081−3092

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information


connectivity analysis
non-heuristic k-means
Molecular Recognition MD Structures (10⁶)
SWIF's high-dimensionality clustering vectors
Equilibrium Binding Mode Descriptions

| Mode | PAS | | | | | |
|---|---|---|---|---|---|---|
| 0 | | | Ph | | AK1 | Ph |
| 1 | | | Ph | | AK1 | Ph |
| 2 | | | Ph | | AK1 | Ph |
| 3 | | | Ph | | | Ph |
| 4 | | | PO4 | | AK1 | Ph |
| 5 | Ph | | PO4 | | AK1 | AK2 |
| 6 | | Ph | AK1 | | AK1 | Ph |

**ABSTRACT:** The accurate and reproducible detection and description of thermodynamic states in computational data is a nontrivial problem, particularly when the number of states is unknown *a priori* and for large, flexible chemical systems and complexes. To this end, we report a novel clustering protocol that combines high-resolution structural representation, brute-force repeat clustering, and optimization of clustering statistics to reproducibly identify the number of clusters present in a data set ($k$) for simulated ensembles of butyrylcholinesterase in complex with two previously studied organophosphate inhibitors. Each structure within our simulated ensembles was depicted as a high-dimensionality vector with components defined by specific protein−inhibitor contacts at the chemical group level and the magnitudes of these components defined by their respective extents of pair-wise atomic contact, thus allowing for algorithmic differentiation between varying degrees of interaction. These *surface-weighted interaction fingerprints* were tabulated for each of over 1 million structures from more than 100 $\mu$s of all-atom molecular dynamics simulation per complex and used as the input for repetitive $k$-means clustering. Minimization of cluster population variance and range afforded accurate and reproducible identification of $k$, thereby allowing for the characterization of discrete binding modes from molecular simulation data in the form of contact tables that concisely encapsulate the observed intermolecular contact motifs. While the protocol presented herein to determine $k$ and achieve non-heuristic clustering is demonstrated on data from massive atomistic simulation, our approach is generalizable to other data types and clustering algorithms, and is tractable with limited computational resources.

## 1. INTRODUCTION

Molecular recognition (MR) processes involve noncovalent interactions between two or more molecules of complementary size, shape, and chemistry.[1] Such interactions are ubiquitous in chemical, biochemical, and pharmaceutical processes including, but certainly not limited to, analytical[2] and chromatographic techniques,[3] protein synthesis,[4] the self-assembly of proteins[5] and nucleic acids,[6] and ligand binding.[7−9] Still, many questions remain regarding the interactions that dominate such recognition, as determined by the chemistry of the complementary species involved, as well as how to best describe the dynamic complexes that result.[10,11] Nevertheless, the structures of MR complexes, and the physicochemical properties that result from the more dominant intermolecular forces at play, are of great interest across myriad disciplines within the scientific community.[12]

In biochemical systems, it is often the case that MR complexes, such as enzyme−substrate or enzyme−inhibitor

**Received:** December 9, 2019
**Published:** May 8, 2020

https://dx.doi.org/10.1021/acs.jcim.9b01137

pairs, inherently include sets of distinctly different mean structures, each referred to herein as a binding mode, with set members having populations that are in equilibrium with one another within some broader distribution about the ensemble-average structure (rather than a single energetically best structure).[10,13] Two commonly employed computational techniques used to collect sets of mean structures are docking calculations[14−16] and molecular dynamics (MD) simulation.[13,17] In this study, we employ rigorous all-atom explicit-solvent MD simulations and massive sampling to obtain the most accurate representation of MR complex structural ensembles possible, where binding modes for each ensemble are then identified as groupings of individual structures within our massive MD data sets via the implementation of a commonly employed heuristic clustering algorithm.[18,19]

Clustering is an unsupervised machine learning approach, which implies that we have no prior knowledge of the number of data clusters present in the data set or the content of those distinct groupings.[20,21] Many clustering algorithms can be easily used within modern open-source modules,[22] each offering various advantages and disadvantages and each potentially resulting in qualitatively and quantitatively different groupings of a given data set.[18] While these algorithms can also vary greatly in their computational efficiency and theoretical groundings, we opted to employ the $k$-means algorithm within the Python-based scikit-learn machine learning library, selected for its superior speed in clustering the immense lists of values that are used to describe each structure.[22,23]

There are drawbacks to $k$-means clustering, however, the most notable of which is that $k$-means algorithms require the total number of clusters present within the data ($k$) to be given as an input parameter.[24,25] In practice, this parameter is seldom known *a priori* and it is thus highly desirable to have an efficient and reproducible method of identifying the total number of clusters that are present in a given data set. Additionally, the qualitative and quantitative results of independent invocations of $k$-means clustering can vary significantly due to the heuristic nature of this algorithm, depending largely on the initial placement of cluster centers.[24−26] The primary goal of this study is, therefore, to address these issues such that the identification of $k$ and the grouping of a data set into $k$ clusters are accurate and reproducible, both qualitatively and quantitatively.

To this end, we first consider the nature of MR in the absence of any knowledge or assumptions regarding the physicochemical properties of the complementary species involved: to best capture the essence of the MR complex, be it rigid and relatively static or flexible and highly dynamic, we describe each structure in our data set using only the observed intermolecular contacts as the basis of MR interactions (without geometric or spatial considerations other than proximity). The resulting surface-weighted interaction fingerprint (SWIF) that depicts a given structure then contains a complete accounting of intermolecular contacts between the MR complex pair and thus encapsulates distinct interaction motifs that will distinguish binding modes from one another,[20] not unlike the well-known bag-of-words or bag-of-features used in document and image clustering, respectively.[27,28] This informatics-based SWIF approach, in tandem with the clustering protocol reported below, in which cluster population range and variance are minimized, overcomes the heuristic nature of $k$-means, yielding a single optimal value of $k$ and clustering results that are statistically reproducible.

As proof of concept, our protocol was applied to the butyrylcholinesterase (BChE) enzyme in complex with two alkyl aryl phosphate inhibitors, thereby building on our colleagues' previous experimental work[9,29] and our earlier MD/clustering-based analyses.[13,30] Those previous reports, which included a color-coded visual representation of BChE similar to that shown in Figure 1, focused on the biochemical description of the protein active site and its interactions with chemical groups of the inhibitor. In this report, following significant improvements to procedural and sampling methods, we emphasize the application of this new information-based surface-weighted interaction fingerprint approach in tandem with a non-heuristic protocol for accurate $k$ selection. While our SWIF approach is generally applicable to molecular recognition and similar studies of molecular structure, our non-heuristic $k$-means protocol is expected to be generalizable to any application of $k$-means.

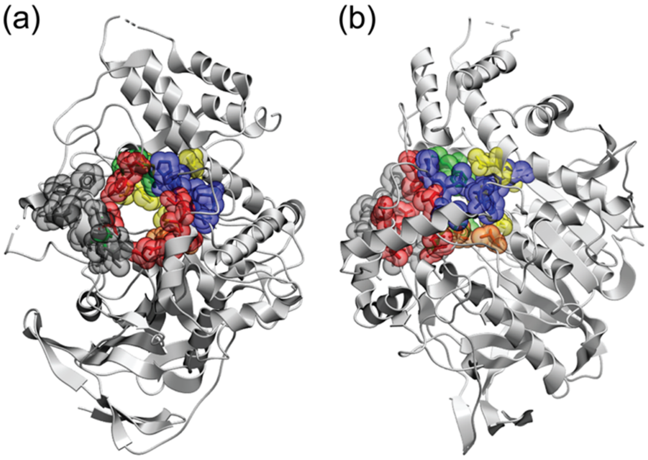
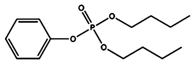
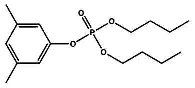


**Figure 1.** Visualization of 529-residue BChE in the grayscale ribbon mode with active site residues shown as semitransparent van der Waals surfaces (a) facing into the active site pocket from the gorge entrance and (b) rotated 90° about the vertical axis. Color-coded regions of the active site gorge include the peripheral anionic site (PAS, red), the catalytic triad (CAT, yellow), the oxyanion hole (OAH, orange), the choline-binding site (CBS, green), the acyl-binding site (ABS, blue), and the omega loop (OML, charcoal).

The two inhibitors to which the reported protocol is applied herein, as shown in Table 1, represent the weakest and strongest inhibitors that were assayed in recent studies: dibutyl phenyl phosphate (DAP4) and dibutyl 3,5-dimethylphenyl phosphate (DIM5), respectively.[9,30] In an effort to verify the efficacy of the reported protocol, this procedure was applied to both inhibitors in complex with the enzyme and the results are presented below.

**Table 1. Sampling and Assay Results[9,30] for Di-*n*-butyl Phenyl Phosphate and the 3,5-Dimethylphenyl Analog**

| Inhibitor | Code | Structure | Time (μs) | $K_I$ (μM) |
|---|---|---|---|---|
| di-*n*-butyl phenyl phosphate | DAP4 |  | 110.0 | 94.5 (±9) |
| di-*n*-butyl-3,5-dimethylphenyl phosphate | DIM5 |  | 106.8 | 1 (±0.4) |

## 2. METHODS

A model of human BChE was prepared by removing all water molecules, ions, and ligands from the crystal structure (PDB ID: 1P0I), inserting missing atoms and sidechains (none of which were near to, or part of, the active site gorge of the enzyme), and performing geometry optimization on these regions using Accelrys Discovery Studio.[31] The resulting structure was then energy-minimized, including sidechain rotamer relaxation, using the SwissPDB software.[32] The ICM Pro computational suite[33,34] was used to perform 10 000 molecular docking trials of each inhibitor within the BChE active site gorge, with the best scoring docked structure taken as the MD starting conformation for each inhibitor. Inhibitor molecules were modeled using the General AMBER Force Field (GAFF),[35] which was designed in tandem with partial charge calculation via the semiempirical (AM1) method with bond charge correction (BCC) to approximate the molecular electrostatic potential computed at the Hartree−Fock 6-31G* theory level.[36,37] Partial charges were calculated using the Quacpac Tool Kit from OpenEye Scientific[38] (see the Supporting Information).

All-atom molecular dynamics simulations of BChE in complex with each inhibitor were performed using the GROMACS 5.0.4 software suite.[39] The protein and counterions were modeled using the AMBER03 force field[40] ported to the GROMACS suite[41] and solvated with the TIP3P explicit water model.[42] To optimize simulation time, a periodic octahedral box was used, yielding a total system size of approximately 72 350 atoms. All simulations were performed in the NPT ensemble at 1.0 atm and 300 K using the Berendsen and modified-Berendsen barostat and thermostat, respectively,[43,44] with a 2.0 fs timestep using the LINCS algorithm[45] to constrain bonds involving hydrogen atoms. A switching function from 7 to 9 Å and a standard long-range correction term were applied to van der Waals interactions, and electrostatic interactions beyond 9 Å employed a reaction-field treatment with a dielectric coefficient of 80.

To build on our previous computational studies of BChE inhibition,[13,30] in which limited sampling was reported, and with the goal of reaching structural equilibrium across large simulated ensembles, 1000 independent simulations of each BChE−inhibitor complex were initiated from the best scoring docked structure of each complex using random seeds to assign unique initial atomic velocities. With an average simulation time of nearly 110 ns, yielding over 100 $\mu$s of total sampling per enzyme−inhibitor complex (216.8 $\mu$s of total sampling) and with structures stored every 100 ps, the full data set for each inhibitor is composed of over 1 000 000 structures, thereby highlighting the need for a high-efficiency clustering algorithm that is capable of processing very large data sets.

As detailed below, the evolution of cluster populations over time was monitored for numerous values of $k$ to identify the equilibration period for each simulated ensemble. Only data collected after this equilibration time were used for the ensemble (thermodynamic) analyses presented below. Indeed, to use all of the collected data in our clustering and characterization of binding modes would greatly bias our results toward both the starting structures obtained via docking and the binding modes within each data set that are most kinetically convenient to sample.

The clustering protocol followed, being one of the primary subjects of this report, is addressed in detail throughout the following section.

## 3. RESULTS AND DISCUSSION

### 3.1. Surface-Weighted Interaction Fingerprints.

As expected,[46] all-atom simulations prove the protein−ligand complex to be highly flexible. We therefore sought, first and foremost, to describe the recognition interface based not on spatial or geometric positioning but rather on distinct chemical contacts between active site amino acids and functional groups within the inhibitor. To this end, we employed a variation of the structural interaction fingerprint (SIFt) approach[47] to characterize the simulated structural data. SIFt contains contact information in binary bit strings, where each residue of interest is represented by its own string and each entry of 1 or 0 in that string indicates the presence or absence of an interaction between that residue and the ligand, respectively.[47,48] SIF't can be used as the input for softwares such as Ligplot+[49] or PLIP[50] to assess the types of interaction that exist between each residue and the ligand. Moreover, SIF't collected across large, rich data sets are natural candidates for clustering, making this an attractive approach with respect to our goal of identifying and characterizing discrete binding modes in simulated ensembles.

Our approach first identifies contacts between active site amino acids and an inhibitor as intermolecular atomic pairs separated by 5.0 Å or less. Where a contact is identified, our information strings describe the interaction as (a) existing between either the sidechain or backbone region of the specified residue and (b) a specific chemical group within the inhibitor (in the present study, per Table 1, this would be one of four possible groups: the central phosphate group, the phenyl moiety, or one of the two aliphatic chains). In addition, within our information strings, the binary bits used in SIF't are replaced with integer values representing the total number of atoms in the specified amino acid that are in contact with the specified inhibitor chemical group, providing a simple metric to represent the magnitude of surface contact between the specified pair.

With these differences in mind, we label our information strings as surface-weighted interaction fingerprints (SWIF's) and emphasize that the nonbinary weighting of contacts in this manner allows for simple algorithmic differentiation between varying degrees of intermolecular contact for each component (or contact pair). For example, the contact between a given active site aromatic sidechain and the phenyl moiety of our inhibitors may occur in one binding mode as significant $\pi-\pi$ stacking while occurring solely as edge-to-edge contact (or similarly less significant contact) in a different binding mode. Such differences, though seemingly minor, could constitute meaningful structural and chemical variations between binding modes and our contact identification and weighting scheme. Furthermore, our SWIF information strings allow a single binding mode to be highly flexible about a mean structure while maintaining a consistent contact motif, thereby contributing to the end goal of characterizing the specific interactions that define each binding mode.

As is common in clustering protocols, the weight of each SWIF component was independently min−max normalized over all SWIF's to give equal significance to all enzyme−inhibitor interactions during clustering and the final implemented list of components contributing to the SWIF for a given protein−inhibitor structure will therefore depend on the complete list of contacts between each observed interaction pair across all structures in that protein−inhibitor data set. In addition, we restrict SWIF components to include only intermolecular contacts, as defined above, that are observed in greater than 1% of the simulated structures for each complex. This requirement reduces the list of components used to characterize our 1000-simulation data sets from approximately 650(±50) to approximately 150(±20), making the clustering of over $10^6$ structures (SWIF's) per data set tractable on a single desktop workstation.

**3.2. Equilibration Assessment.** At equilibrium, the net population exchange rates between all pairs of binding modes (energetic minima) are zero, with populations in each energy basin constant. At long simulation times, the data beyond some initial equilibration period will exhibit ergodicity, where time-averaged properties of the system are equivalent to their analogous ensemble-averaged (thermodynamic) properties. Approximating the equilibration time is therefore critical to the analysis of binding mode equilibria. However, without *a priori* knowledge of the number of clusters (binding modes) present in each data set, it is not possible to identify structural clusters or to accurately track cluster populations over time to establish equilibration. To address this issue, the list of all SWIF's representing all structures within each data set was used as the input to *k*-means clustering with values of *k* ranging from 2 (assumed low) to 28 (assumed very high) and the populations of all clusters for each *k* value were monitored over time.

Examples of the resulting kinetics plots for the BChE−DAP4 and BChE−DIM5 complexes are displayed in Figure 2 for the representative *k* values of 4 and 12. As shown in the figure, cluster populations for these *k* values (and others, not shown) were approximately constant beyond 80.0 ns. Cluster kinetics for all *k* values demonstrated similar behavior, and we thus took 80.0 ns as a lower cutoff on approximating structural equilibrium, with only structures recorded after this 80.0 ns equilibration period used in the analyses that follow. Although this relatively conservative equilibration time results in a much smaller subset of the data contributing to our cumulative analysis, adoption of a uniform equilibration time facilitates this analysis by removing a variable from our protocol while also ensuring that we include only data that has reached, or very closely approximates, structural equilibrium, regardless of *k*.

It should be noted that when low values of *k* are chosen, neighboring clusters (binding modes) will be necessarily merged and identified as a single cluster, as demonstrated in Figure 3. When this occurs, such as the *k* = 4 examples in the insets of Figure 2, equilibration must then take place between more distant "superclusters" that are, on average, significantly farther apart than the (sub)clusters from which they are composed and this may increase the time required for equilibration. While the *k* = 12 kinetics plots show well-established structural equilibria, those for *k* = 4 demonstrate only approximately constant cluster populations beyond the 80.0 ns equilibration point, making this an approximation to the actual equilibration point for that *k* value. Hence, we emphasize that caution must be taken with assessing a given clustering result for proper equilibration and that this process will be dependent on both the system studied and the modeling technique(s) being used.

**3.3. Non-heuristic Identification of *k*.** As noted above, the primary drawback to using *k*-means algorithms is the need to provide the number of clusters as input, requiring knowledge of *k* prior to using the algorithm for accurate analyses. As illustrated in Figure 3, the use of *k* values that are too low in numerous successive *k*-means invocations will result in artificial cluster merging and a coarse view of the data, while the use of *k* values that are too high will lead to artificial cluster splitting, inconsistent clustering, and cluster populations with larger variance across successive implementations.

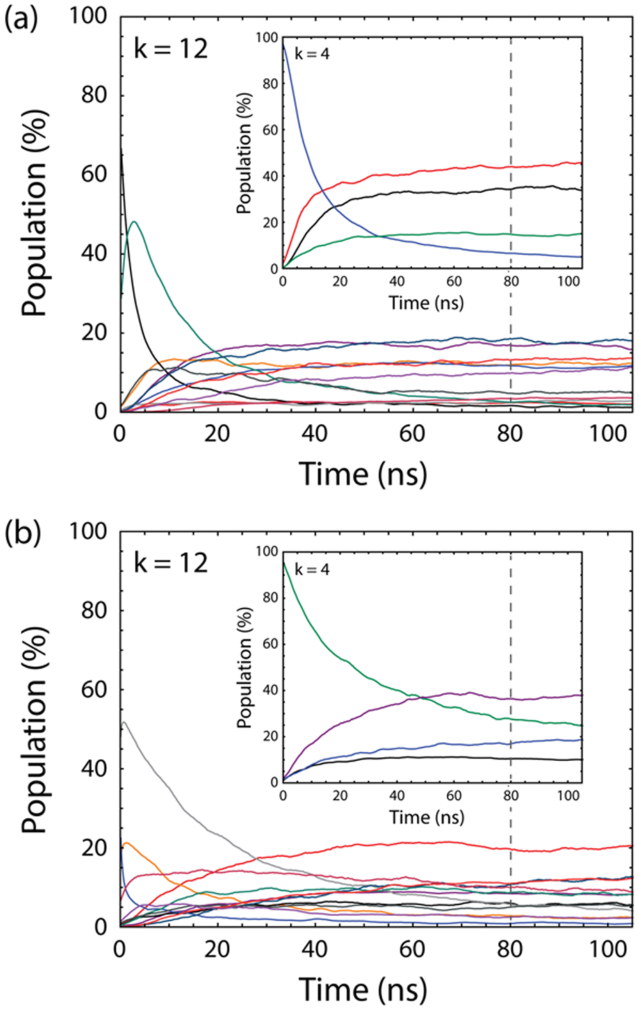


Figure 2. Kinetics plots for (a) DAP4 and (b) DIM5, with *k* = 4 (inset) and *k* = 12, respectively. Each line represents the population of a cluster obtained through the implementation of *k*-means. For all values of *k* beyond those pictured, 80.0 ns, marked by the vertical dashed line, was identified as the approximate time point at which to assume structural equilibrium.

While many philosophies and approaches exist for accurately choosing *k*, drawbacks to these methods have long been known.[51] Two such approaches are the elbow method and the merging method. In the former, the inertia of each *k* value (similar to a residual sum of squares taken across all data with respect to their assigned clusters) is plotted versus *k* with the expectation that a sharp change will reveal *k* accurately.[52] Our attempts at employing this method, however, failed to generate elbows (Figure S1 in the Supporting Information). In the merging method,[41] the value of *k* is largely overestimated initially and clusters are then merged or deleted until convergence to a final *k* value is reached with further iterations resulting in no change in cluster assignments. The large size of our data sets and our implementation of extremely information-rich (high-dimensionality) surface-weighted interaction fingerprints, however, made this method computationally intractable due to both memory and time limitations.

Numerous previous efforts to accurately identify *k* have incorporated logic regarding the statistics of data distribution

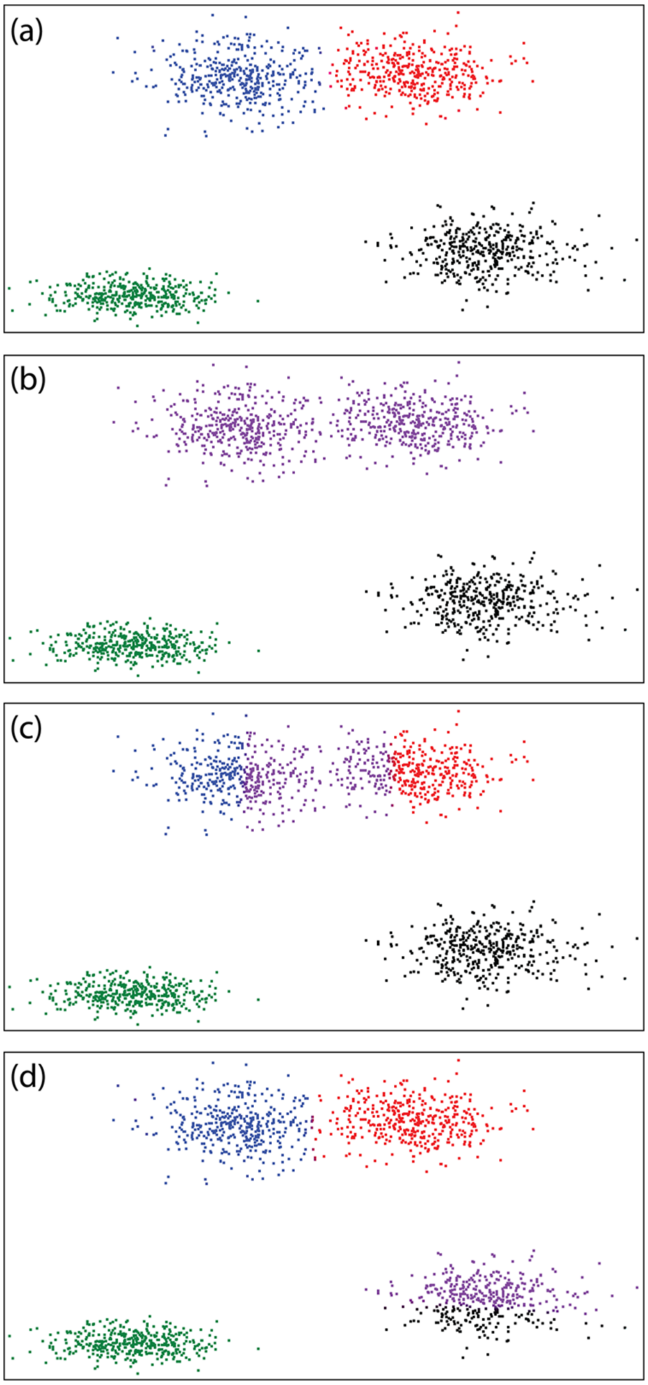


**Figure 3.** Visual representation of a hypothetical two-dimensional data set with four clusters clearly apparent. Cluster assignments for each datum are indicated by the color assigned to that point. In (a), the use of the correct value of $k = 4$ leads to consistent assignment of each point to the same cluster. In (b), a smaller value of $k = 3$ may lead to consistent assignment for repeated invocations of *k*-means but also results in the merging of neighboring clusters and thereby misses the finer structure within the data. In (c) and (d,) the use of too large a choice of $k = 5$ leads to inconsistent clustering and highly variant cluster populations over repeated *k*-means invocations.

within the clustering results including, but certainly not limited to, closeness, betweenness, withinness (or compactness), cluster distortion, and the gap statistic.[21,51,53−55] While this philosophy is sound, these statistics can be conceptually abstract, algorithmically complex, or both, making them unintuitive to the chemical or biological researcher. Our approach to identify proper *k* values herein is founded instead on a single, simple assumption: based on the discussion above and presented in Figure 3, we define the correct *k* value as the largest *k* value that minimizes both the statistical variance and range observed in cluster populations across successive invocations of *k*-means, thereby providing the most consistent and resolved view of the data possible using intuitive and expedient cluster metrics. The method employed here thus unites atomic-level chemical information in the form of high-dimensionality SWIF's; brute-force repetitive *k*-means clustering, shown to improve the richness of possible clustering solutions;[26,56] and simple statistical logic aimed at optimizing cluster reproducibility by minimizing cluster variance and range. This approach depends solely on the resolution of the clustering input but not the nature of the system being studied and is therefore expected to be valid across myriad applications.

To succinctly capture the range and standard deviation of cluster populations identified by the *k*-means algorithm, population distribution matrices (PDM's) were generated following 10 independent implementations of *k*-means for each *k* value ranging from 2 (assumed low) to 28 (assumed very high). As shown in the PDM's presented in Figure 4, clusters were sorted by population in descending order for each *k* value. In accordance with the above philosophy, the largest *k* value with minimal variance in cluster population and minimal population range was selected for each inhibitor, with $k = 8$ for DAP4 and $k = 7$ for DIM5 (green boxes in Figure 4). As shown by their respective PDM's in Figure 4, all larger *k* values demonstrated higher population ranges and standard deviations. Following the identification of these *k* values, a subsequent set of 10 additional *k*-means trials was collected for each BChE−inhibitor complex and the *k* value selections noted above were validated via population distribution matrices analogous to those shown in Figure 4 (see Figure S2 in the Supporting Information), thereby demonstrating that high-resolution clustering input can be reproducibly clustered within a limited set of repeat invocations using our minimization approach.

While population distribution matrices and the minimization of cluster population range and variance provide a means to reproducibly identify *k* with high confidence that the resulting clusters are well defined, it is important to emphasize that clustering solutions obtained via repeat clustering invocations using the same *k* value may differ to some degree based not only on the initialization of the clustering algorithm,[26] but also on the structure of the data set being analyzed. For example, where clusters occupy similar regions of the phase space defined by the input, small differences in final cluster center positions could lead to different assignments of a given datum among neighboring clusters. While it is beyond the scope of the current study, we expect such differences to be negligible for large data sets with high-resolution input vectors. Moreover, depending on the nature of one's data set and the desired resolution of the resulting model of that data, post-clustering analysis may suggest that specific clusters are close enough in the high-dimensionality space occupied by the data to warrant cluster merging in order to provide a more coarse or minimalistic view of that data.

### 3.4. Clustering Efficacy.

There are many metrics for evaluating clustering results, particularly in the context of molecular dynamics simulation data, many of which examine the extent of similarity or dissimilarity between elements in a cluster.[18] One such approach that directly compares the clustered surface-weighted interaction fingerprints to one

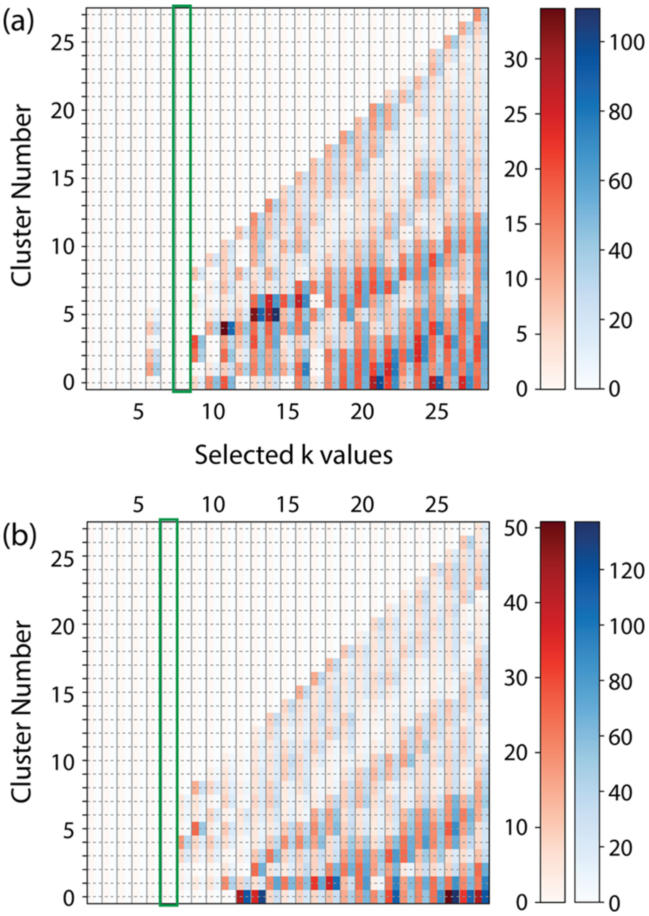


**Figure 4.** Population distribution matrices displaying the standard deviation (red) and range (blue) in cluster populations for (a) DAP4 and (b) DIMS after 10 independent *k*-means invocations of each data set using *k* values from 2 to 28. Green boxes indicate the largest *k* value that minimizes population variance and range, as discussed in the text. The streaks of color that appear across these PDM's, which lower in population rank with increasing *k* value, are due to nonresolved cluster structures for some *k* values: as additional centroids are added to the algorithm with higher *k*, data in that region will be more likely to be consistently assigned to a unique cluster and the portion of unresolved structures will decrease.

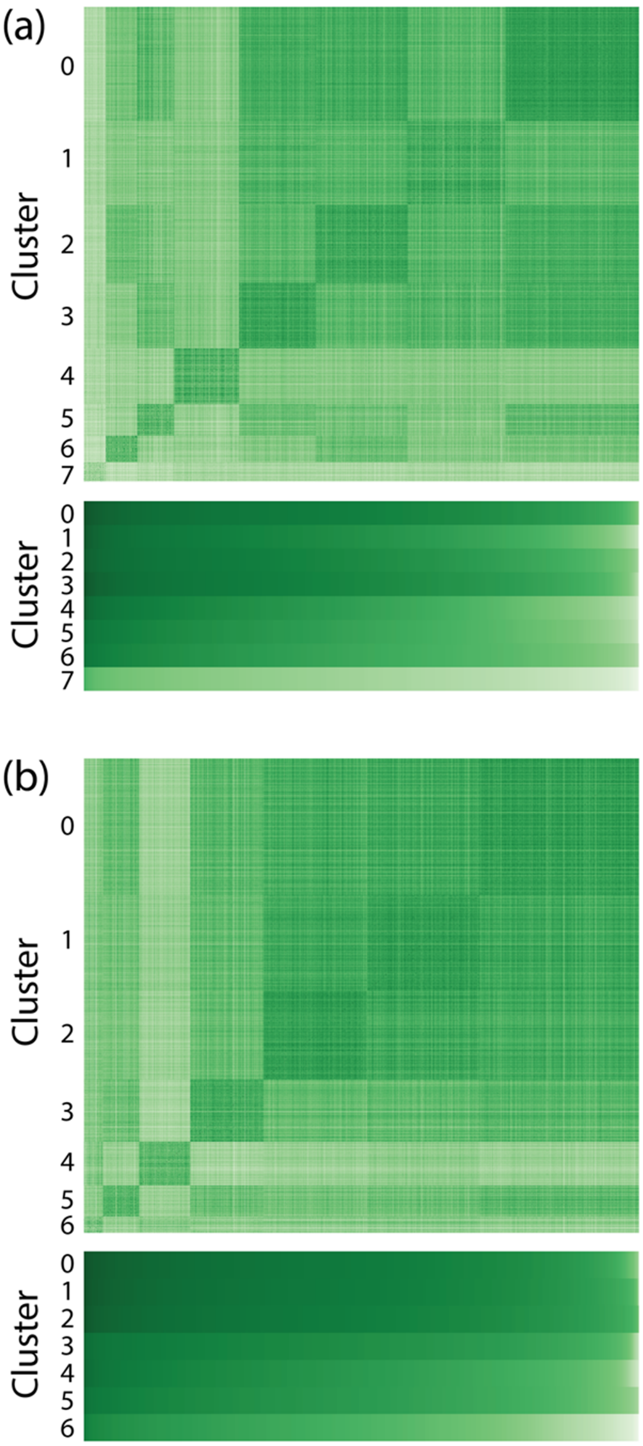


**Figure 5.** Similarity matrices and centroid-similarity matrices are shown for (a) DAP4 and (b) DIMS. In the similarity matrices in the top panels for each inhibitor, blocks along the diagonal represent intracluster similarities, while off-diagonal blocks represent intercluster similarities. In the centroid-similarity matrices shown in the lower panels for each inhibitor, each structure within a cluster was compared to the centroid, or cluster center, of the cluster to which it is assigned. The same color scheme is employed in all panels, with darker shades indicating greater similarity.

another is the use of similarity matrices in a manner akin to that described by Cheatham and co-workers.[18] In deriving a similarity matrix, we calculate the root-mean-squared difference (RMSD) between all pairs of SWIF's within each data set, yielding a numeric measure of the difference between the recognition contacts found in that pair of MR complex structures, where structures with highly similar intermolecular contacts will have low RMSD and those of differing contact motifs will have large RMSD. SWIF's are then grouped by cluster number, from most populated (cluster 0) to least populated (cluster $k - 1$), and ordered symmetrically in rows $i$ and columns $j$ to produce the color-coded similarity matrix, as shown in Figure 5 (top panels). Consequently, blocks of color along the diagonal represent intracluster similarities, while off-diagonal blocks represent intercluster similarities.

As demonstrated in the upper panels in Figure 5(a and b) dark blocks occur along the diagonal for both data sets, representing high degrees of similarity between structures within a given cluster, while paler blocks in off-diagonal regions represent greater dissimilarity between structures in different clusters. It is expected that some similarity will be present when comparing structures from any two clusters, as distinct binding modes can share SWIF magnitudes in a nontrivial number of dimensions. This can be physically interpreted as these binding modes

Table 2. Contacts Observed in the Most Populated Binding Modes from 1000 Simulations of the BChE−DIM5 Complex

| 1000 Sims DIM5 | ASN68 | ASP70 | GLN119 | ALA277 | SER287 | TYR332 | SER198 | GLU325 | HIS438 | GLY116 | GLY117 | ALA199 | TRP82 | ALA328 | PHE329 | TRP231 | PRO285 | LEU286 | VAL288 | PHE398 | ILE69 | GLN71 | PHE73 | PRO74 | GLY75 | PHE76 | MET81 | ASN83 | SER79 | TYR114 | GLY115 | PHE118 | THR120 | TYR128 | GLU197 | ASN397 | TRP430 | MET437 | GLY439 | TYR440 | ILE442 | Pop (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mode | PAS | | | | | | CAT | | | OAH | | | CBS | | | ABS | | | | | OML | | | | | | | | Additional Protein Residues (APR) | | | | | | | | | | | | | |
| 0 | | | Ph | | AK1 | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | AK2 | AK1 | AK1 | AK1 | Ph | | | | | | | | | | PO4 | AK1 | PO4 | | PO4 | | AK2 | | | | | 28.7 |
| 1 | | | Ph | | AK1 | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | Ph | AK1 | AK1 | AK1 | Ph | | | | | | | | Ph | | PO4 | AK1 | PO4 | | PO4 | | AK2 | | AK2 | | | 20.2 |
| 2 | | | Ph | | AK1 | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | Ph | AK1 | AK1 | AK1 | Ph | | | | | Ph | AK2 | | Ph | | PO4 | | PO4 | | | | AK2 | AK2 | AK2 | | | 18.7 |
| 3 | | | Ph | | | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | PO4 | AK1 | AK1 | AK1 | Ph | | | | | | | | | | PO4 | AK1 | PO4 | | PO4 | | AK2 | | AK2 | | | 13.2 |
| 4 | | | PO4 | | AK1 | Ph | AK1 | | | AK1 | PO4 | AK1 | Ph | Ph | PO4 | AK1 | Ph | AK1 | PO4 | AK1 | AK2 | | | | | | | | | | | AK1 | PO4 | | | | | | | | | 9.2 |
| 5 | Ph | | PO4 | | AK1 | AK2 | PO4 | | AK1 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | AK2 | AK1 | AK1 | AK1 | Ph | | | | | | | | | | | | PO4 | | | | | | | | | 6.6 |
| 6 | | Ph | AK1 | | AK1 | Ph | AK1 | | PO4 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | PO4 | AK1 | AK1 | AK1 | | | | | | Ph | AK2 | | | | | | PO4 | | | | Ph | | | | | 3.4 |

🟩 Electrostatic 🟥 Hydrogen Bonding 🟨 Charge-Dipole 🟦 π-stacking 🟪 van der Waals 🟦 Non-polar □ Backbone

sharing partial intermolecular contact motifs. For example, the structurally similar DIM5 and DAP4 exhibit sharing of partial contact motifs across the most populated binding modes observed.

As a secondary assessment of clustering validity, the extent to which each cluster centroid represents other cluster members was examined. This provides a sense of the structural deviation within each cluster and is thus a useful measure of how well single-structure visualizations represent each binding mode ensemble. The lower panels in Figure 5(a and b) show RMSD values between centroid SWIF's and the SWIF's for all other cluster members (sorted in decreasing similarity) and illustrate that the obtained centroids are generally very good representations of their cluster members.

It is worth noting, however, that the metrics discussed above are internal criteria[21,57] and many clustering studies employ one or more external criteria, or ground truths, as additional metrics to validate their results.[21,58] With respect to studies employing molecular dynamics simulations, a commonly used metric is the all-atom root-mean-squared deviation of the clustered structures, where RMS differences are calculated between atomic coordinates rather than surface-weighted interaction fingerprint components. Such a comparison here would be impractical for several reasons. Most notably, the values of this purely geometric external criterion would be dominated by the structure of the protein, regardless of inhibitor contact with active site residues, and the comparison would overtly exclude information regarding intermolecular interactions that are central to the recognition complex. Indeed, the SWIF approach employed herein focuses on connectivity, rather than geometry, to best characterize the flexible nature of the complex.

**3.5. Massive Sampling Binding Modes.** To explore these clustering results in the context of their biophysical utility, we employ contact tables, a novel visual format presented in our previous work,[30] to summarize interactions between ligand functional groups and residues in the protein active site. The rows and columns in a contact table represent the identified clusters (binding modes) sorted in descending order by population and amino acids in the protein active site, respectively, with cells in the table listing the strongest type of intermolecular interaction observed for contact between that amino acid and the relevant chemical group in the inhibitor. While softwares such as Ligplot+[49] and PLIP[50] may be employed to identify intermolecular interactions, entries in the contact tables provided below are based solely on a standard physicochemical rule set that deduces the chemical interaction type, and thus its relative strength, based on the chemistry of the inhibitor group and that of the amino acid sidechain with which it interacts, with the strongest interaction taking priority in each cell of each table. Backbone interactions were treated as low priority unless known interactions were present (such as hydrogen bonds involving GLY116 and GLY117).

Table 2 characterizes interactions in the BChE−DIM5 complex, with amino acids grouped according to the subsites color-coded in Figure 1 and additional amino acids of interest shown in gray. For clarity, only interactions that were observed in at least 50 percent of structures in each cluster are shown and population percentages are presented in the final column of the table.

While a definitive biochemical analysis of Table 2 will be more appropriately reported elsewhere alongside analyses of numerous additional BChE−inhibitor complexes, immediately evident in this contact table are numerous interactions that are common to most or all identified binding modes, such as the ubiquitous hydrogen bonding of the DIM5 phosphate group to GLY116 and GLY117; the nonpolar contact between the first alkyl chain and numerous residues in the acyl-binding site (ABS); and π-stacking between the DIM5 phenyl group and TYR332. Consistent interactions such as these represent the similarity observed between clusters due to the previously described sharing of partial contact motifs (off-diagonal blocks in the top panels of Figure 5(a and b).

Additional contacts appear to be present primarily in the most populated of binding modes, such as the strong electrostatic and hydrogen bonding interactions between the DIM5 phosphate group and residues SER198 and HIS438 in the catalytic triad; phenyl contact with GLN119 in the peripheral anionic site; and van der Waals contact between (i) the DIM5 phosphate and GLU197 and (ii) the second DIM5 alkyl chain and TRP430, both involving residues in the additional protein residue (APR) group. These consistent contacts among the most populated DIM5 binding modes clearly explain the similarity observed between highly populated clusters in the upper right quadrant of the PDM (upper panel) of Figure 5b.

Below these modes in Table 2, there is significantly more variation in contacts among less populated binding modes, as represented by the low similarity among these clusters in the lower left quadrant of the PDM (upper panel) in Figure 5b. In addition, the distinct recognition contact motif of binding mode (cluster) 4 in Table 2 was well predicted by cluster-to-cluster comparisons involving this binding mode in the PDM for DIM5 (Figure 5b, upper panel), presenting a self-consistent quality control check on the ability of our protocol to properly distinguish between and separate distinct binding modes.

While contact tables represent a step forward in the succinct description of MR binding modes, a dichotomy results from including some information and excluding other information, even when statistically motivated. In generating Table 2, for example, it was often the case that multiple functional groups competed for entry in the table by being simultaneously in contact with specific amino acids. To simplify this process, we

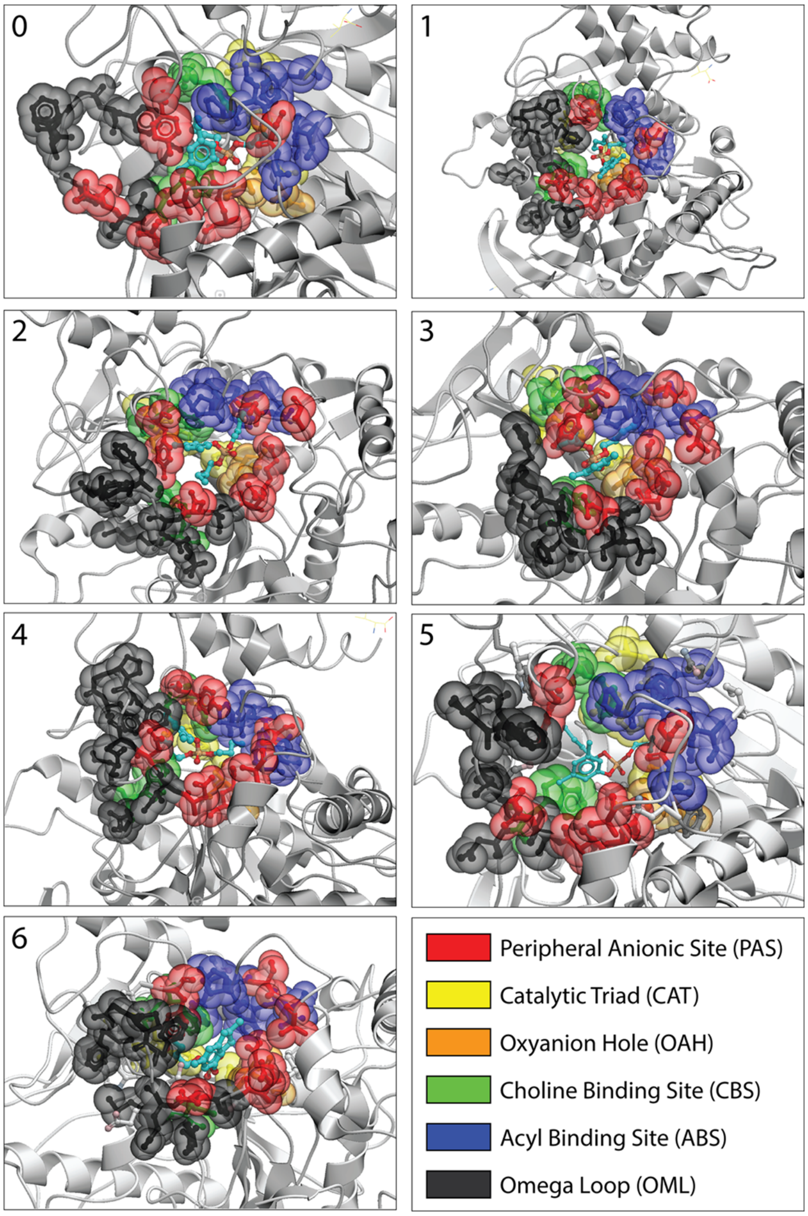

0
1
2
3
4
5
6
🟥 Peripheral Anionic Site (PAS)
🟨 Catalytic Triad (CAT)
🟧 Oxyanion Hole (OAH)
🟩 Choline Binding Site (CBS)
🟦 Acyl Binding Site (ABS)
⬛ Omega Loop (OML)

**Figure 6.** Magnified cluster medoid structures of the BChE−DIM5 complex following the same graphical conventions described in Figure 1, with active site residues colored according to the key. Inhibitor functional groups follow standard chemical color-coding conventions with the phosphate group shown in orange and red and the alkyl and phenyl groups in cyan.

did not account for the varying magnitudes of competing contacts, instead opting to include the functional group whose chemistry would produce the strongest type of interaction with the chemistry of the specific amino acid. Due to this simplification, some resolution is lost when constructing contact tables following this format and certain binding modes may therefore appear more similar than currently allowed for in this representation. We are presently refining our approach to characterizing each tabulated binding mode using a radial distribution approach to quantifying interactions surrounding each functional group within the inhibitor, with the goal of balancing the proximity of nearby amino acids with the strength of their interactions to produce more thorough and well-resolved binding mode characterizations.

Cluster medoid structures (the real data points closest to the absolute average of each cluster) for observed BChE−DIM5 binding modes are visualized in Figure 6 from an external perspective along the approximate direction of the active site gorge cavitation. Most notable is the variation in positioning of the omega loop region of BChE (OML, black), which is known to be highly dynamic[30] and therefore capable of stabilizing different binding modes to varying degrees, as suggested in Table 2.

**3.6. Limited Sampling and Thoroughness.** To ensure the thoroughness of binding mode sampling as a function of data set size, smaller subsets of the 1000-simulation cumulative data sets were randomly selected. Three sets of surface-weighted interaction fingerprints representing 1, 10, and 100 simulations were chosen and each SWIF was then assigned to the cluster it was previously associated with in our analysis of the cumulative data set. The characterization described above was then applied to these three subsets of our data. Table 3 demonstrates the degree to which these smaller subsets of our BChE−DIM5 data sampled the binding modes and contacts present in the

Table 3. Binding Modes and Contacts Observed in 1, 10, and 100 Randomly Selected Simulations of BChE−DIM5

| 1 Sim DIM5 | ASN68 | ASP70 | GLN119 | ALA277 | SER287 | TYR332 | SER198 | GLU325 | HIS438 | GLY116 | GLY117 | ALA199 | TRP82 | ALA328 | PHE329 | TRP231 | PRO285 | LEU286 | VAL288 | PHE398 | ILE69 | GLN71 | PHE73 | PRO74 | GLY75 | PHE76 | MET81 | ASN83 | SER79 | TYR114 | GLY115 | PHE118 | THR120 | TYR128 | GLU197 | ASN397 | TRP430 | MET437 | GLY439 | TYR440 | ILE442 | Pop (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mode | PAS | | | | | | CAT | | | OAH | | | CBS | | | ABS | | | | | OML | | | | | | | | Additional Protein Residues (APR) | | | | | | | | | | | | | |
| 0 | Ph | | PO4 | | AK1 | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | Ph | AK2 | PO4 | AK1 | Ph | AK1 | AK1 | AK1 | Ph | | | | | | AK2 | | | | PO4 | AK1 | PO4 | | PO4 | | | | | AK2 | | 42.6 |
| 1 | Ph | Ph | Ph | | AK1 | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | Ph | AK2 | PO4 | AK1 | Ph | AK1 | AK1 | AK1 | Ph | | | | | | | | | | PO4 | AK1 | PO4 | | PO4 | | | | | AK2 | | 57.4 |
| 2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

| 10 Sims DIM5 | ASN68 | ASP70 | GLN119 | ALA277 | SER287 | TYR332 | SER198 | GLU325 | HIS438 | GLY116 | GLY117 | ALA199 | TRP82 | ALA328 | PHE329 | TRP231 | PRO285 | LEU286 | VAL288 | PHE398 | ILE69 | GLN71 | PHE73 | PRO74 | GLY75 | PHE76 | MET81 | ASN83 | SER79 | TYR114 | GLY115 | PHE118 | THR120 | TYR128 | GLU197 | ASN397 | TRP430 | MET437 | GLY439 | TYR440 | ILE442 | Pop (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mode | PAS | | | | | | CAT | | | OAH | | | CBS | | | ABS | | | | | OML | | | | | | | | Additional Protein Residues (APR) | | | | | | | | | | | | | |
| 0 | | | Ph | | AK1 | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | Ph | AK2 | PO4 | AK1 | AK2 | AK1 | Ph | AK1 | Ph | | | | | AK2 | AK2 | | | | PO4 | AK1 | PO4 | | PO4 | | AK2 | | AK2 | | | 39.6 |
| 1 | | Ph | Ph | | AK1 | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | Ph | AK1 | AK1 | PO4 | Ph | | | | | AK2 | AK2 | | | | PO4 | AK1 | PO4 | | PO4 | | | | AK2 | AK2 | | 26.7 |
| 2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | | Ph | PO4 | | AK1 | Ph | AK1 | | Ph | PO4 | PO4 | AK1 | Ph | Ph | PO4 | AK1 | PO4 | AK1 | PO4 | AK1 | PO4 | Ph | | | | | AK2 | | | | AK2 | AK1 | PO4 | | | | | | | | | 16.6 |
| 5 | Ph | | PO4 | | AK1 | AK2 | AK1 | | AK1 | PO4 | PO4 | | PO4 | | PO4 | AK1 | AK2 | AK1 | PO4 | AK1 | | | AK2 | | | | | | | | | | PO4 | | | AK1 | | | | | | 16.9 |
| 6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

| 100 Sims DIM5 | ASN68 | ASP70 | GLN119 | ALA277 | SER287 | TYR332 | SER198 | GLU325 | HIS438 | GLY116 | GLY117 | ALA199 | TRP82 | ALA328 | PHE329 | TRP231 | PRO285 | LEU286 | VAL288 | PHE398 | ILE69 | GLN71 | PHE73 | PRO74 | GLY75 | PHE76 | MET81 | ASN83 | SER79 | TYR114 | GLY115 | PHE118 | THR120 | TYR128 | GLU197 | ASN397 | TRP430 | MET437 | GLY439 | TYR440 | ILE442 | Pop (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mode | PAS | | | | | | CAT | | | OAH | | | CBS | | | ABS | | | | | OML | | | | | | | | Additional Protein Residues (APR) | | | | | | | | | | | | | |
| 0 | | | Ph | | AK1 | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | AK2 | AK1 | AK1 | AK1 | Ph | | | | | | | | | | PO4 | AK1 | PO4 | | PO4 | | AK2 | | AK2 | | | 26.7 |
| 1 | | Ph | AK1 | | AK1 | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | Ph | AK1 | AK1 | AK1 | Ph | | | | | | | | Ph | | PO4 | AK1 | PO4 | | PO4 | | AK2 | | AK2 | | | 23.5 |
| 2 | | | PO4 | | AK1 | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | Ph | AK1 | AK1 | AK1 | Ph | | | | | Ph | AK2 | | Ph | | PO4 | AK1 | PO4 | | PO4 | | AK2 | | AK2 | | | 17.3 |
| 3 | | Ph | Ph | | | Ph | PO4 | | PO4 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | PO4 | AK1 | AK1 | AK1 | Ph | | | | | | | | | | PO4 | AK1 | PO4 | | PO4 | | AK2 | | AK2 | | | 12.5 |
| 4 | | | PO4 | | AK1 | Ph | AK1 | | | AK1 | PO4 | AK1 | Ph | Ph | PO4 | AK1 | Ph | AK1 | PO4 | AK1 | AK2 | | | | | | | | | | | AK1 | PO4 | | | | | | | | | 7.9 |
| 5 | Ph | | PO4 | | AK1 | AK2 | AK1 | | AK1 | PO4 | PO4 | AK1 | PO4 | AK2 | PO4 | AK1 | AK2 | AK1 | AK1 | AK1 | | | | | | | | | | | | AK1 | PO4 | | | | | | | | | 6.6 |
| 6 | | Ph | AK1 | | | Ph | AK1 | | PO4 | PO4 | PO4 | AK1 | PO4 | Ph | PO4 | AK1 | PO4 | AK1 | AK1 | AK1 | Ph | | | | | Ph | Ph | | | | | | PO4 | | | | Ph | | | | | 5.5 |

🟩 Electrostatic 🟥 Hydrogen Bonding 🟨 Charge-Dipole 🟦 π-stacking 🟪 van der Waals 🟦 Non-polar □ Backbone

cumulative data set (shown in Table 2) and is formatted to highlight the observed differences in sampling of those contacts and binding modes, including rows ordered according to the binding mode populations observed in the cumulative data set and blank rows representing nonsampled clusters.

As expected, smaller sampling sizes detect fewer total binding modes and do not capture the relative populations shown in Table 2 for binding modes that are detected. Moreover, contact motifs for the binding modes that are sampled deviate qualitatively from their massive sampling analogs. While several aspects of the smaller sampling sizes are noteworthy, characterization of these subsets for small numbers of simulations (1 and 10) would depend heavily on the specific randomly selected simulations and we therefore comment only briefly on their thoroughness.

Though the single BChE−DIM5 simulation samples only two of seven observed binding modes, for which relative populations are qualitatively incorrect, it is notable that this random trajectory does sample from the two most populated, and thus thermodynamically important, binding modes while also largely (but incompletely) predicting the contact motifs for these clusters, as presented in Table 2. The single randomly selected BChE−DAP4 simulation fared poorly in comparison, sampling a single binding mode (cluster 2, the third most populated) and showing significant deviation from the cumulative data set in the predicted contact motif for that binding mode. Again, we emphasize that these observations are based on random selection of one simulation from each data set, and the degree of sampling thoroughness will vary greatly between single simulations.

In comparison, the 10-simulation BChE−DIM5 subset samples four of the seven observed binding modes, with relative populations that are a fair approximation of those observed in the cumulative data set and contact motifs for each binding mode that are similar to those observed for, though also showing qualitative deviation from, the cumulative data set. Underscoring the expected variance in sampling thoroughness highlighted above, the randomly selected 10-simulation BChE−DAP4 subset sampled all but one of the modes observed in the cumulative data set while showing poor agreement in relative cluster populations and deviation in contact motifs similar to that observed for BChE−DIM5.

In contrast, the 100-simulation BChE−DIM5 subset shown in Table 3 is much more thorough, sampling from every binding mode observed in the cumulative data set, producing cluster populations that are quantitatively similar to the massive sampling result and contact motifs that closely resemble the 1000-simulation data set, with only a few notable differences. The DAP4 analog, which binds approximately 100 times weaker than DIM5, yields results that qualitatively agree with these observations but also deviate more from the cumulative data set, with poorly predicted relative cluster populations and deviation among contact motifs that is somewhat more pronounced.

## 4. CONCLUSIONS

This study explores a new approach to employ *k*-means clustering and similarly heuristic algorithms in a logical, information-based, and fully reproducible manner that does not require *a priori* knowledge of the number of clusters present ($k$) in one's data set. Our approach employs iterative invocations of *k*-means, followed by minimization of cluster population variance and range, to ascertain $k$ values that generate consistent and well-resolved clusters; assess clustering efficacy through intra- and intercluster similarity matrices; and have been demonstrated to be tractable for massive data sets in tandem with very high-dimensionality *k*-means input (our surface-weighted interaction fingerprints) on a single-user workstation. We expect this protocol to be fully applicable to other high-dimensionality problems of many kinds and easily modified to best fit the problem at hand.

Applied herein to molecular dynamics data and, in particular, to massive all-atom molecular dynamics simulations of butyrylcholinesterase in complex with two previously characterized organophosphate inhibitors, the use of high-dimensionality surface-weighted interaction fingerprints composed of internal coordinates that describe intermolecular interactions, independent of location in Cartesian space, bypasses numerous complications associated with the characterization of molecular recognition complexes. This process, in tandem with our use of surface-weighted interaction fingerprints, illustrates an altogether novel approach to identifying and characterizing molecular recognition complexes that is broadly generalizable. Contact tables provide a novel approach to summarizing the interaction motifs associated with observed binding modes for a given complex and can also be easily modified to fit the specifics of similar studies in myriad chemical contexts.

To assess the impact of sample size on sampling thoroughness, random subsets of 1, 10, and 100 simulations were selected from within our cumulative 1000-simulation data set and the structures from each subset were fit to the clusters observed in the cumulative data set. As expected, very limited data sets consisting of only 1 and 10 simulations fail to sample from all observed clusters, most frequently only approximating the intermolecular contact motifs for the clusters that were sampled, and yield relative weights for those binding modes that are highly inconsistent with the massive sampling result. In contrast, the 100-simulation subset sampled from all binding modes observed in the cumulative data set, yielding strong approximations of the relative populations of those clusters, and provided largely accurate descriptions of the contact motifs associated with each binding mode. For large, flexible complexes that sample numerous binding modes, this observation suggests a lower limit on the sampling needed to obtain the most resolved insight into the interactions between complementary molecular recognition pairs.

## ■ ASSOCIATED CONTENT

### SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b01137.

> Inhibitor charge derivation and charges used; elbow plots for *k*-means trials performed on varying data sets of *n* simulations; secondary population distribution matrices for validation of identification of *k* (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

**Eric J. Sorin** − *Department of Chemistry & Biochemistry, California State University, Long Beach, Long Beach, California 90840, United States;* orcid.org/0000-0003-4081-1142; Phone: 562-985-7537; Email: eric.sorin@csulb.edu

### Authors

**Parker Ladd Bremer** − *Department of Chemistry & Biochemistry, California State University, Long Beach, Long Beach, California 90840, United States*

**Danna De Boer** − *Department of Chemistry & Biochemistry, California State University, Long Beach, Long Beach, California 90840, United States*

**Walter Alvarado** − *Department of Physics & Astronomy, California State University, Long Beach, Long Beach, California 90840, United States*

**Xavier Martinez** − *Department of Computer Engineering & Computer Science, California State University, Long Beach, Long Beach, California 90840, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.9b01137

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors would like to thank the Folding@Home contributors worldwide, who made this study possible, as well as our collaborators in the Nakayama and Schwans Laboratories. P.L.B. was supported by a CSULB Summer Research Assistantship and a Robert B. Henderson Memorial Scholarship. W.A. acknowledges support from the Margaret Heeb and Philip Ord Johnson Scholarship programs. X.M. was supported by an NIH RISE Fellowship (R25GM071538). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## ■ REFERENCES

(1) Gellman, S. H. Introduction: Molecular Recognition. *Chem. Rev.* **1997**, *97*, 1231−1232.

(2) Jeon, J. H.; Kakuta, T.; Tanaka, K.; Chujo, Y. Facile design of organic-inorganic hybrid gels for molecular recognition of nucleoside triphosphates. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 2050−2055.

(3) Hiruta, Y.; Kanazashi, R.; Ayano, E.; Okano, T.; Kanazawa, H. Temperature-responsive molecular recognition chromatography using phenylalanine and tryptophan derived polymer modified silica beads. *Analyst* **2016**, *141*, 910−917.

(4) Zaher, H. S.; Green, R. Fidelity at the molecular level: lessons from protein synthesis. *Cell* **2009**, *136*, 746−762.

(5) Levy, Y.; Onuchic, J. N. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 389−415.

(6) Sorin, E. J.; Nakatani, B. J.; Rhee, Y. M.; Jayachandran, G.; Vishal, V.; Pande, V. S. Does Native State Topology Determine the RNA Folding Mechanism? *J. Mol. Biol.* **2004**, *337*, 789−797.

(7) Haris, P.; Mary, V.; Haridas, M.; Sudarsanakumar, C. Energetics, Thermodynamics, and Molecular Recognition of Piperine with DNA. *J. Chem. Inf. Model.* **2015**, *55*, 2644−2656.

(8) Gleitsman, K. R.; Sengupta, R. N.; Herschlag, D. Slow molecular recognition by RNA. *RNA* **2017**, *23*, 1745−1753.

(9) Nakayama, K.; Schwans, J. P.; Sorin, E. J.; Tran, T.; Gonzalez, J.; Arteaga, E.; McCoy, S.; Alvarado, W. Synthesis, biochemical evaluation, and molecular modeling studies of aryl and arylalkyl di-n-butyl phosphates, effective butyrylcholinesterase inhibitors. *Bioorg. Med. Chem.* **2017**, *25*, 3171−3181.

(10) Vértessy, B. G.; Orosz, F. From "fluctuation fit" to "conformational selection": Evolution, rediscovery, and integration of a concept. *BioEssays* **2011**, *33*, 30−34.

(11) Cremer, P. S.; Flood, A. H.; Gibb, B. C.; Mobley, D. L. Collaborative routes to clarifying the murky waters of aqueous supramolecular chemistry. *Nat. Chem.* **2018**, *10*, 8.

(12) Du, X.; Li, Y.; Xia, Y. L.; Ai, S. M.; Liang, J.; Sang, P.; Ji, X. L.; Liu, S. Q. Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. *Int. J. Mol. Sci.* **2016**, *17*, No. 144.

(13) Sorin, E. J.; Alvarado, W.; Cao, S.; Radcliffe, A.; La, P.; An, Y. Ensemble Molecular Dynamics of a Protein-Ligand Complex: Residual Inhibitor Entropy Enhances Drug Potency in Butyrylcholinesterase. *Bioenergetics* **2017**, *6*, No. 1000145.

(14) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for molecular docking: a review. *Biophys. Rev.* **2017**, *9*, 91−102.

(15) Meng, X. Y.; Zhang, H. X.; Mezei, M.; Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput.-Aided Drug Des.* **2011**, *7*, 146−157.

(16) Ferreira, L. G.; Dos Santos, R. N.; Oliva, G.; Andricopulo, A. D. Molecular docking and structure-based drug design strategies. *Molecules* **2015**, *20*, 13384−13421.

(17) Phillips, J. L.; Colvin, M. E.; Newsam, S. Validating clustering of molecular dynamics simulations using polymer models. *BMC Bioinf.* **2011**, *12*, No. 445.

(18) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* **2007**, *3*, 2312−2334.

(19) De Paris, R.; Quevedo, C. V.; Ruiz, D. D.; Norberto de Souza, O.; Barros, R. C. Clustering molecular dynamics trajectories for optimizing docking experiments. *Comput. Intell. Neurosci.* **2015**, *2015*, No. 916240.

(20) Hartmann, A. K. Practical Introduction to Clustering Data. 2016, arXiv:1602.05124. arXiv.org e-Print archive. https://arxiv.org/abs/1602.05124 (accessed April 13, 2020).

(21) Rodriguez, M. Z.; Comin, C. H.; Casanova, D.; Bruno, O. M.; Amancio, D. R.; Costa, L. d. F.; Rodrigues, F. A. Clustering algorithms: A comparative approach. *PLoS One* **2019**, *14*, No. e0210236.

(22) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Perrot, M.; Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(23) McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Software* **2017**, *2*, No. 205.

(24) Wishart, D. *k*-Means Analysis. *Encyclopedia of Statistics in Behavioral Science*; John Wiley & Sons, Ltd., 2005.

(25) Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881−892.

(26) Fränti, P.; Sieranoja, S. How much can k-means be improved by using better initialization and repeats? *Pattern Recognit.* **2019**, *93*, 95−112.

(27) Liu, Y.; Liu, M.; Wang, X. Towards semantically sensitive text clustering: a feature space modeling technology based on dimension extension. *PLoS One* **2015**, *10*, No. e0117390.

(28) O'Hara, S.; Draper, B. A. Introduction to the Bag of Features Paradigm for Image Classification and Retrieval, 2011. arXiv:1101.3354. arXiv.org e-Print archive. https://arxiv.org/abs/1101.3354.

(29) Law, K. S.; Acey, R. A.; Smith, C. R.; Benton, D. A.; Soroushian, S.; Eckenrod, B.; Stedman, R.; Kantardjieff, K. A.; Nakayama, K. Dialkyl phenyl phosphates as novel selective inhibitors of butyrylcholinesterase. *Biochem. Biophys. Res. Commun.* **2007**, *355*, 371−378.

(30) Alvarado, W.; Bremer, P. L.; Choy, A.; Dinh, H. N.; Eung, A.; Gonzalez, J.; Ly, P.; Tran, T.; Nakayama, K.; Schwans, J. P.; Sorin, E. J. Understanding the enzyme-ligand complex: insights from all-atom simulations of butyrylcholinesterase inhibition. *J. Biomol. Struct. Dyn.* **2020**, 1028.

(31) BIOVIA, D. S. Discovery Studio Modeling Environment. *BIOVIA Workbook*, release 2017; *BIOVIA Pipeline Pilot*, release 2017; BIOVIA Discovery Studio, 2007.

(32) Guex, N.; Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **1997**, *18*, 2714−2723.

(33) Abagyan, R. A.; Totrov, M. M.; Kuznetsov, D. A. ICM: A New Method For Protein Modeling and Design: Applications To Docking and Structure Prediction From The Distorted Native Conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.

(34) An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics* **2005**, *4*, 752−761.

(35) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general AMBER force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(36) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.

(37) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623−1641.

(38) Ellingson, B. A.; Geballe, M. T.; Wlodek, S.; Bayly, C. I.; Skillman, A. G.; Nicholls, A. Efficient calculation of SAMPL4 hydration free energies using OMEGA, SZYBKI, QUACPAC, and Zap TK. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 289−298.

(39) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845−854.

(40) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.

(41) Sorin, E. J.; Pande, V. S. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.* **2005**, *88*, 2472−2493.

(42) Mahoney, M. W.; Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **2000**, *112*, 8910−8922.

(43) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(44) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, No. 014101.

(45) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.

(46) Shen, C.; Liu, H.; Wang, X. W.; Lei, T. L.; Wang, E. C.; Xu, L.; Yu, H. D.; Li, D.; Yao, X. J. Importance of Incorporating Protein Flexibility in Molecule Modeling: A Theoretical Study on Type I-1/2 NIK Inhibitors. *Front. Pharmacol.* **2019**, *10*, No. 345.

(47) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47*, 337−344.

(48) Mordalski, S.; Kosciolek, T.; Kristiansen, K.; Sylte, I.; Bojarski, A. J. Protein binding site analysis by means of structural interaction fingerprint patterns. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 6816−6819.

(49) Laskowski, R. A.; Swindells, M. B. LigPlot+: Multiple Ligand−Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* **2011**, *51*, 2778−2786.

(50) Salentin, S.; Schreiber, S.; Haupt, V. J.; Adasme, M. F.; Schroeder, M. PLIP: fully automated protein−ligand interaction profiler. *Nucleic Acids Res.* **2015**, *43*, W443−W447.

(51) Pham, D.; Dimov, S.; Nguyen, C. Selection of K in K-means clustering. *Proc. Inst. Mech. Eng., Part C* **2005**, *219*, 103−119.

(52) Syakur, M. A.; Khotimah, B. K.; Rochman, E. M. S.; Satoto, B. D. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conf. Ser.: Mater. Sci. Eng.* **2018**, *336*, No. 012017.

(53) Dev, V. R.; Bharathi, G.; Prasad, G. V. S. N. R. V. Prediction of Customer Churn using Fuzzy Balanced Probabilistic C-means Algorithm. *Int. J. Comput. Appl.* **2019**, *178*, 23−30.

(54) Sharma, D.; Thulasiraman, K.; Wu, D.; Jiang, N. A network science-based k-means++ clustering method for power systems network equivalence. *Comput. Soc. Networks* **2019**, *6*, No. 4.

(55) Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc., B* **2001**, *63*, 411−423.

(56) Bicego, M.; Figueiredo, M. A. T. Clustering via binary embedding. *Pattern Recognit.* **2018**, *83*, 52−63.

(57) Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. In *Understanding of Internal Clustering Validation Measures*, IEEE International Conference on Data Mining, 2010; pp 911−916.

(58) Rendón, E.; Abundez, I. M.; Gutierrez, C.; Zagal, S. D.; Arizmendi, A.; Quiroz, E. M.; Arzate, H. E. A comparison of internal and external cluster validation indexes. Proceedings of the 2011 American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications; World Scientific and Engineering Academy and Society (WSEAS): Puerto Morelos, Mexico, 2011, pp 158−163.