Accurate Scoring of Drug Conformations at the Extreme Scale

Boyu Zhang,* Trilce Estrada,† Pietro Cicotti,‡ Pavan Balaji,§ Michela Taufer*
*University of Delaware {bzhang, taufer}@udel.edu
†University of New Mexico {estrada}@cs.unm.edu
‡San Diego Supercomputer Center {pcicotti}@sdsc.edu
§Argonne National Laboratory {balaji}@anl.gov

Abstract—**We present a scalable method to extensively search for and accurately select pharmaceutical drug candidates in large spaces of drug conformations computationally generated and stored across the nodes of a large distributed system. For each ligand conformation in the dataset, our method first extracts relevant geometrical properties and transforms the properties into a single metadata point in the three-dimensional space. Then, it performs an octree-based clustering on the metadata to search for predominant clusters. Our method avoids the need to move ligand conformations among nodes because it extracts relevant data properties locally and concurrently. By doing so, we can perform accurate and scalable distributed clustering analysis on large distributed datasets. We scale the analysis of our pharmaceutical datasets a factor of 400X higher in performance and 500X larger in size than ever before. We also show that our clustering achieves higher accuracy compared with that of traditional clustering methods and conformational scoring based on minimum energy.**

I. Problem overview

The design of new pharmaceutical drugs relies on finding small molecules, called ligands, that dock into proteins and play an essential role in turning protein functions on or off. Studying protein-ligand interactions in the wet lab is extremely expensive and time demanding, especially for high-throughput experimental structure determination by X-ray crystallography and nuclear magnetic resonance spectroscopy. Computer simulations are used to accelerate this process and to reduce costs. The computational search for candidate drugs (i.e., ligands that dock well in a protein) is a search under uncertainty in a very large space of potential docking conformations of a given ligand; this space is shaped by the protein, the ligand, the computational methods, and the degrees of freedom to be explored [?]. Cutting-edge distributed systems, such as cloud infrastructures and high-end clusters, provide scientists with an efficient and scalable way to perform computationally expensive protein-ligand docking simulations at a rate never seen before. At the same time, this capability leads to large datasets of resulting ligand conformations that are distributed across the nodes of the system, resulting in new challenges for scientists who have to analyze the data and select the more promising conformations for experiments in the wet lab.

When analyzing large ligand datasets, we can rely on two orthogonal methods: the highly scalable but inaccurate scoring based on the conformations' energy and the poorly scalable but highly accurate comparison of the conformations' geometry. The scoring based on the conformations' energy reduces the number of candidates from hundreds of thousands up to 10 to 100 conformations based on their energy [?]; however, the method leaves scientists with the tedious task of subjectively selecting a possible near-native ligand manually by using visualization tools such as VMD [?] or Chimera [?]. Several studies have shown the inaccuracy of such energy scoring [?] [?]. On the other hand, traditional scoring based on the geometry of conformations relies on directly comparing the root-mean-square deviations (RMSDs) of the conformations' geometries and then clustering the conformations based on the RMSD (i.e., calculating the inner variance of a cluster by looking at the conformations' RMSDs), thus requiring the communication of ligands from several nodes to one. Previous studies have shown the accuracy but lack of scalability of these methods when the entire distributed dataset is analyzed, and a gain in scalability but a potential loss in accuracy when a subset of data is sampled and analyzed [?].

Our work is motivated by the need to find a transformative approach at the intersection of the energy-based scalability and geometry-based accuracy. To this end, we transform the clustering problem into a search for densities that capture the geometry of all the ligand conformations concurrently. The search does not require substantial communication as do more traditional geometry-based analyses. Specifically, our method accurately maps similar ligand conformations (each conformation of n atoms in the 3D space) to metadata points (each point of three coordinates, x, y, and z) in a 3D space in close proximity concurrently. Thus subspaces of the metadata space with higher point concentrations (or densities) can be associated with the most frequently found ligand conformations in a docking simulation that naturally converges toward conformations of interest for scientists. Each node performs a local 3D clustering on its own metadata points to search for dense clusters and to exchange densities (or aggregates) with other compute nodes, allowing each node to obtain a global convergence view for subsets of the original dataset. The scalability of 400X in performance and 500X in data size is achieved because no communication of ligand conformations or metadata is performed; the accuracy is preserved because geometries are accurately mapped into metadata.

II. Distributed octree-based clustering

The overall method first extracts the relevant geometrical properties of each ligand conformation and represents the properties as three-dimensional points (i.e., metadata) and then performs an octree-based clustering to search for densest metadata subspaces.

A. *Capturing relevant geometrical properties*

Docking simulations generate hundreds of thousands of independent ligand conformations docked in the pocket of a protein. We concurrently extract the geometrical shape (property) of the ligand conformations in parallel across the nodes of the distributed system. To this end, we perform a space reduction by mapping the atom coordinates of each ligand conformation to a single metadata point of three coordinates in the 3D space. Our space reduction has the desired property of projecting ligands with a similar geometry closer into the newly defined 3D space of metadata points. Thus the cluster with the highest density of mapped metadata points can be associated with those ligand conformations of interest for the pharmaceutical search that occur over and over from the independently executed docking simulations.

We consider two variations of the mapping algorithm; both are based on projections and linear interpolations. The variations share the backbone reduction technique but differ in terms of the final metadata representation. In both variations, given a ligand with p atomic coordinates (x_i, y_i, z_i, with i from 1 to p), we perform a projection of the coordinates in the three planes (x, y), (y, z), and (z, x). Each projection results in a set of 2D points on the associated 2D plane. For each projection, we compute the best-fit linear regression line over the projected points and compute the three slopes of the three lines.

In the first variation, we use the three slopes as the coordinates of the 3D point to encode the conformational geometry of its corresponding ligand. We call this variation “3D mapping.” Figure 1 shows an example of metadata generated from multiple conformations of the ligand 1hbv when docked in the HIV protease as part of the Docking@Home project [**?**].

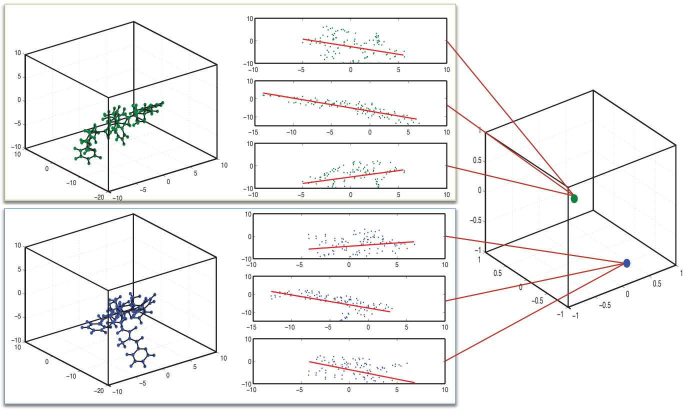


Fig. 1. Capturing relevant geometrical properties by using projection and linear interpolation for the 3D mapping variation.

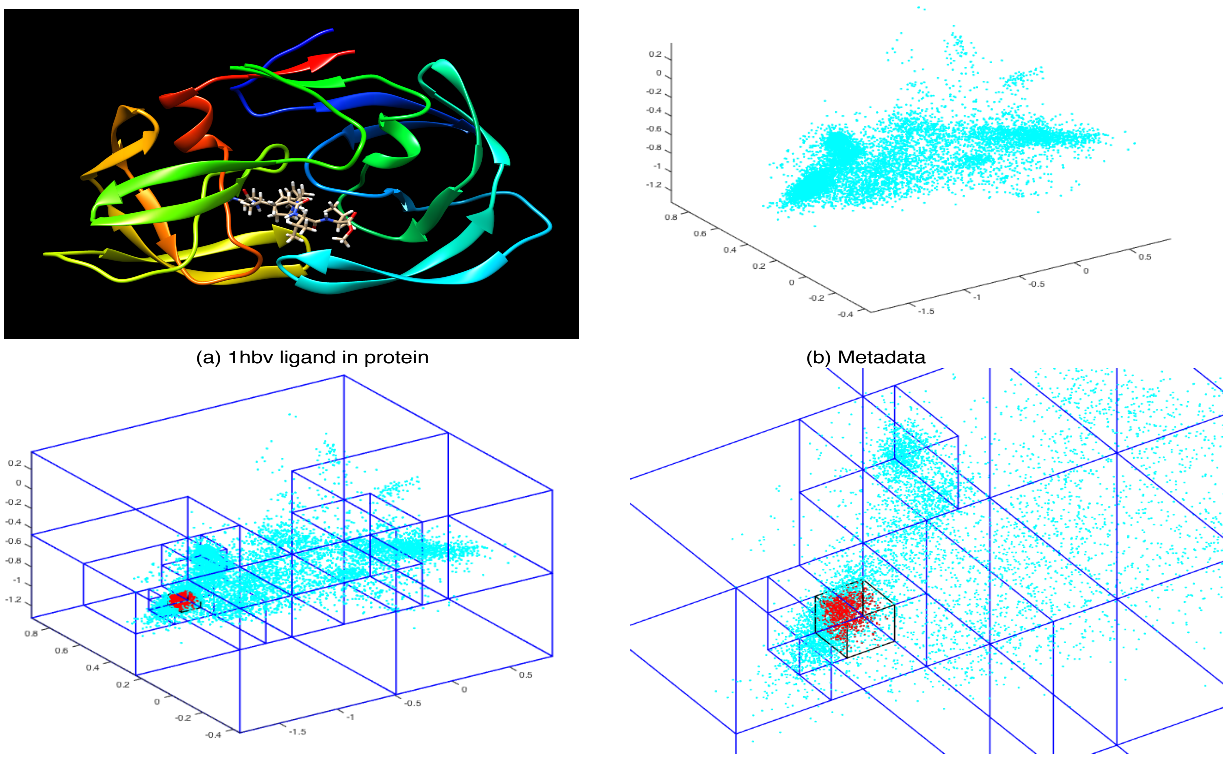
In the second variation, we use the logarithmic values of the three slopes to encode the 3D point representing the conformation geometry. If the slope is negative, we use the negative logarithm of the absolute value. Contrary to the “3D mapping” variation, this variation better captures the geometrical properties of conformations that are in an almost-vertical position inside the protein pocket. In the 3D mapping variation, when ligand conformations are in this position, the three resulting slopes are large. When the shape rotates or changes slightly, the resulting slopes change significantly. This approach may result in conformations with similar shape and in an almost-vertical position to be unmapped into a dense metadata subspace. When using the logarithm of the slopes as metadata, however, we decrease the changes in the metadata coordinates and thus increase the chance for ligand conformations with similar shape to form a dense-enough subspace. We call this variation “3Dlog mapping.”

B. *Searching for densest metadata subspaces*

By dealing with property-encoding metadata (i.e., three-dimensional points for the 3D and 3Dlog mappings) rather than raw atom coordinates, we implicitly transform the analysis problem from a clustering or classification problem into a search of the smaller subspaces in the newly defined metadata space (i.e., an octant for the 3D and 3Dlog mappings) with high property aggregates. We search for these subspaces concurrently across the nodes by building on each node an octree (i.e., by recursively partition the 3D space of metadata on the node into fixed-sized octants, each of which forms the tree nodes).

We count the aggregates of close property-encoding points on each node in a distributed way. Each node has a partial view of the entire datasets; it counts the scalar property aggregates (SPAs) representing the locally stored metadata densities and identifies the densest octants in its local octree (by “densest” we mean the deepest nodes with a minimum number of metadata items or aggregates). Only at this point does each node shuffle the densities (or aggregates) with the other compute nodes. After shuffling the aggregates, each compute node sums the aggregates for a subspace of interest assigned to the node in order to obtain the global cluster densities for the subspace, while searching for convergence. The search algorithm is depicted in Figure 2 for a simplified case that maps the conformations into a 2D space. In the example Node 0 is in charge of the two subspaces on the left, and Node 1 is in charge of the two subspaces on the right.

Figure 3 shows an example of a dataset of 1hbv ligand conformations when docked in the HIV protease. Figure 3(a) shows one ligand conformation in the docking site of the HIV protease. Figure 3(b) is the result of the 3D mapping, after the mapping of 10,000 1hbv ligand confirmations into metadata points has been performed. The compute nodes build an octree by assigning octkeys to the points. Figure 3(c) shows an example of the generated octree for the 3D mapping points in Figure 3(b). Figure 3(d) shows the deepest and densest



(a) 1hbv ligand in protein

(b) Metadata

(c) Octree

(d) Densest octant

Fig. 3. Example of a metadata space of 3D points generated from a dataset of ligand conformations and its octree built to identify the densest octant.

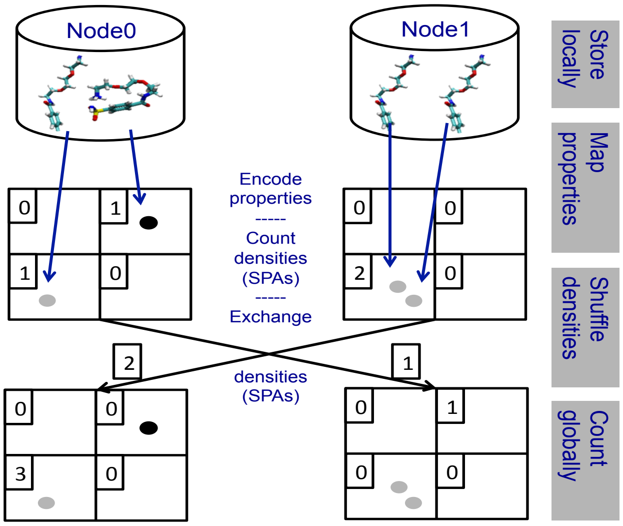


Fig. 2. Examples of exchange scalar property aggregations in our search for a simplified 2D space.

octant (points in red) that is identified by our tree search when looking for the deepest tree node with at least 500 points).

C. Algorithmic implementation

Algorithmically, metadata points belongs to a specific tree node based on its key. The point's key is generated as follows. We initially determine the edge size (i.e., 3D resolution) of the 3D space containing all the projected conformations. We divide the initial space into eight subspaces of the same size, half the original edge size. Every subspace is given a unique identifier ranging from 0 to 7 for the 3D space, based on its position in the 3D space. The key of each point is extended by attaching the subspace identifier to the point's key by padding the left side with the identifier. This process is recursively repeated an arbitrary number of times on each subspace to produce a complete key for each metadata point, that is, key[1...Nkey], where Nkey is the number of digits selected to represent each point. As observed in [?], [?], Nkey can be empirically defined. A value key of 15 digits is sufficient to capture diverse geometries in the dataset of ligand conformations considered in this work.

Each compute node explores its 3D space (e.g., the octree in Figure 3(c)), moving up or down along the tree branches depending on whether a "dense enough" tree node is found. The exploration is performed as a binary search along the tree levels. For example, if the tree is built with each point in the metadata space containing a key of 15 digits, then the tree has up to 15 levels; our search starts at level 8, and we reach a solution (i.e., the deepest tree node with a defined minimum number of points) by exploring up to 4 levels of the tree. Specifically, we start from level 8 and branch to either level 12 or level 4 depending on whether any node is found at level 8 with at least a given number of points or not. Once at the new level, the same criterion is applied to decide whether to move up or down within either the lower half the tree (moved down in the previous iteration) or the upper half (moved up). During the search, compute nodes exchange scalar property aggregates. As shown in Figures 2, each compute node transforms its locally stored ligand conformations into a local 3D space, as described in the steps involving capture of the relevant properties, and exchanges only partial knowledge on its metadata with the other compute nodes. Since compute nodes work on disjointed sets of ligand conformations and metadata, they can map ligand conformations into metadata concurrently and count aggregates locally in advance to perform a global summation.

The MapReduce paradigm naturally accommodates the capturing of properties from local data and the iterative search for densities in its Map and Reduce functions, respectively. Thus, we integrated our method into the MapReduce-MPI framework rather than implementing a new MPI-based framework from scratch [?].

III. Accuracy study

Any scalability method for the analysis of pharmaceutical data that is not accurate is a waste of resources and time. Therefore, before addressing the scalability study, we compare the accuracy of the three methods: (1) our clustering method based on the mapping of conformations into metadata and the octree-based search; (2) previous work based on the direct comparison of ligand conformations in terms of their root-mean-square-deviation (RMSD) and their probabilistic hierarchical clustering [?]; and (3) the naïve selection approach based on only the lowest conformation energy [?].

Datasets: We consider 23 protein-ligand complexes for HIV protease (an aspartic acid protease protein), 21 protein-ligand complexes for trypsin (a serine protease protein), and 12 protein-ligand complexes for P38alpha kinase (a serine/threonine kinase protein) sampled with Docking@Home project over five years [?]. HIV protease (HIV PR) is a protein in the HIV virus that is essential for its replication in human cells. Several protease inhibitors (e.g., saquinavir, ritonavir, indinavir, and nelfinavir) are available for treating HIV infection [?]. Trypsin is a protease that breaks down other proteins in the digestive system. Recent studies suggest that inhibitors of trypsin can have potential application in breast cancer treatment. Here we simulate the docking of several drugs that, when docking in the protease, can act as inhibitors by deactivating the trypsin-like protease and are, therefore, potential agents capable of stopping the spread of breast cancer [?]. P38alpha is the most flexible protein among the three proteins considered. It is involved in the regulation of cellular stress responses as well as the control of proliferation and survival of many cell types. Several promising compounds that inhibit P38alpha are being investigated as potential therapies for arthritic and inflammatory diseases [?] and are part of our study. Figure 4 shows the three proteins. For each protein-ligand complex, we use the ligand conformations sampled with Docking@Home and randomly distributed them across the nodes of our distributed systems. On average, each complex dataset contains around 210,000 ligand conformations.

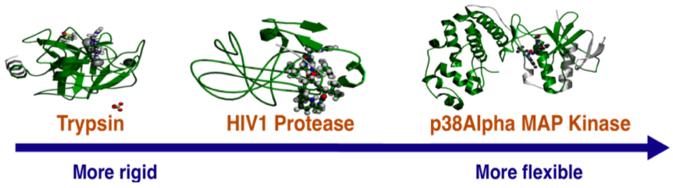


Fig. 4. Three proteins whose results from the Docking@Home datasets are analyzed for this accuracy study.

Accuracy analysis: For each protein, its accuracy is the number of complexes with captured near-native conformations observed in nature over the total number of complexes for that protein. Note that a near-native conformation has an RMSD from the experimentally observed conformation that is smaller than or equal to two angstroms.

For our clustering, we apply the two mapping methods described in Section II (i.e., 3D and 3Dlog mappings) to the data and reduce the ligand geometries into metadata points. The minimum number of metadata points in the tree nodes selected by the octree-based search is set to 0.5% of the number of ligands conformations in each protein-ligand complex dataset. For example, when the dataset contains 100,000 ligand conformations, the density threshold for this dataset is 500 points. When the dataset contains 500,000 ligand conformations, the density threshold for this dataset is 2,500 points. We capture a near-native conformation if the arithmetic median of the conformations associated with the metadata point in the selected tree node is below or equal to 2 Å. The use of the median is preferred as the accuracy metric over the mean because it is less affected by extreme values, although the majority of our overall results are not very sensitive to whether the median or mean is used for selection.

For the probabilistic hierarchical clustering, the distance metric used to cluster each ligand is the RMSD of its atom coordinates versus all the other ligands already in the cluster. If a simulation converges, the largest cluster with lower internal variance is likely the cluster that contains more near-native conformations. We capture a near-native conformation if the centroid of the selected cluster is a near-native conformation [?].

For the energy-based approach, we consider 100 D@H conformations selected based on their lowest energy versus the same crystal structure, which we denote as the naïve approach. Here, we identify the near-native conformation if the arithmetic median of the lowest energy conformations is below or equal to 2 Å.

For the three methods, we perform the clustering and selection of the near-native candidates without using any information about the crystal structures available for the complexes. The crystal structures play an important role only in the validation phase when, for each complex, we calculate the RMSD of the clustering candidate with respect to its crystal structure. Table I summarizes the accuracy for our method (octree-based clustering) with density threshold equal to 0.5%

of the dataset, for the probabilistic hierarchical clustering, and for the energy-based approach.

TABLE I

Comparison of the number of hits for different scoring approaches: our octree-based clustering; a probabilistic hierarchical clustering; and an energy-based scoring method.

Protein	Octree-based clustering		Hierarchical	Min. Energy
	3D	3Dlog		
HIV	**23/23(100%)**	20/23(87.0%)	20/23(87.0%)	8/23(34.8%)
Trypsin	15/21(71.4%)	**17/21(81.0%)**	16/21(76.2%)	5/21(23.8%)
P38alpha	**10/12(83.3%)**	8/12(66.7%)	6/12(50.0%)	1/12(0.8%)
All	**48/56(85.7%)**	45/56(80.4%)	42/56(75.0%)	14/56(25.0%)

When comparing our method with the other two methods, we observe that for all three proteins (i.e., HIV protease, trypsin, and P38alpha), we always get better accuracy. In particular, when mapping the ligand conformation into metadata using the two different mapping techniques, we observe that the 3D mapping has higher accuracy than 3Dlog mapping when considering the HIV and P38alpha proteins. In particular, for the HIV protease, the 3D mapping method captures 23 of the 23 near-native conformations (100%). For the P38alpha protease, the 3D mapping method captures 10 of the 12 near-native conformations (83.3%). In contrast, for the trypsin protein, the 3Dlog mapping has higher accuracy than 3D mapping does. In particular, the 3Dlog mapping method captures 17 of the 21 near-native conformations (81.0%). The reason is that the ligands docked in the HIV and p38alpha proteins' pockets are long and with a high degree of freedom, whereas the ligands docked in trypsin's pockets are relatively small and rigid (with very low degrees of freedom). In this case, the 3Dlog mapping method achieves better accuracy than the 3D method achieves. The reduced flexibility of the conformations explains why the 3Dlog mapping works well for the associated clustering with trypsin but not for the other two proteins. Specifically, when a small and rigid ligand conformation is in a near-vertical position in a pocket, its slope is very large. If the conformation position slightly changes, the slope also changes significantly, because of the projections. In the case of trypsin, some conformations may have similar shapes and be in near-vertical positions; and their slopes may differ to the extent such that the mapping may not result in a dense-enough subspace containing the metadata. By taking the log of the slopes, we reduce the slope differences when dealing with vertical ligand conformations.

IV. Scalability Study

For our scalability study, since the energy-based scoring is highly inaccurate, we exclude it from our set of analysis methods considered. We compare the performance (speedup) and data scalability of our clustering method based on the mapping of conformations into metadata and the octree-based search with the traditional clustering method based on direct comparisons of ligand conformations and probabilistic hierarchical clustering of their RMSD [**?**].

Platforms: The hierarchical clustering is executed on a dedicated cluster at the University of Delaware that is composed of 8 dual quad-core compute nodes (64 cores), each with two Intel Xeon 2.50 GHz quad-core processors and 48 GB RAM. The nodes are connected by high-speed DDR InfiniBand. Our distributed octree-based clustering is executed on up to 256 compute nodes of Fusion, a 320-node computing cluster at the Laboratory Computing Resource Center at Argonne National Laboratory (ANL). Each of Fusion's compute nodes contains two Nehalem 2.6 GHz dual-socket, quad-core Pentium Xeon processors and 36 GB of RAM. The nodes are connected by InfiniBand QDR at 4 GB/s per link.

Datasets: We use ligand conformations sampled with Docking@Home for the *1dif* ligand-HIV protease complex. The size of the datasets ranges from 512 MB to 2 TB (i.e., 200K conformations and 800 million conformations, respectively). The conformations are distributed across the nodes of the two platforms in a load-balancing way.

Scalability analysis: Figure 5 shows the execution times of both the octree and the hierarchical clustering methods. The time reported for the hierarchical clustering is the execution time on 8 nodes of the dedicated cluster at the University of Delaware and includes both communication time, in which the distributed dataset is sent to a centralized node through InfiniBand, and analysis times, in which the hierarchical clustering performs comparisons of the conformations' geometries. Data ranges from 0.5 GB (200K ligand conformations) to 4 GB (1.6 million ligand conformations). The time reported for the octree-based clustering is the total time for the MapReduce-MPI key steps, including the map, shuffle, and reduce phases on the shared cluster at ANL. The number of nodes ranges from 8 to 256, and the data ranges from 4 GB (1.6 million ligand conformations) to 2 TB (800 million ligand conformations).

The hierarchical clustering is not able to scale to more than a dataset of 4 GB (1.6 million ligand conformations) and 8 nodes because it needs to move the whole distributed dataset onto one node and perform all-to-all comparisons among the conformation records on that node. The type of comparison (i.e., the computation of the root-mean-square-deviations among all the conformations) quickly fills the node's memory. A comparison of the times for the hierarchical clustering for the 4 GB dataset versus the octree-based clustering for the same dataset reveals a performance speedup of 400X for our analysis method. Specifically, the same type of analysis on the identical dataset is performed in 1999 seconds by the hierarchical clustering and in 5 seconds by our octree-based clustering.

The study of the weak scalability on the ANL Fusion cluster when the number of conformations increases from 25 million ligands (64GB) to 800 million ligands (2TB) and the number of nodes increases from 8 nodes (64 cores) to 256 nodes (2,048 cores) reveals that our analysis of the pharmaceutical dataset of interest scales up to 500X in data size. Specifically, the hierarchical clustering can deal with up to 4 GB of data without encountering substantial slowdown due to memory swap, whereas the octree-based clustering can deal with up to 2 TB of data without encountering a major slowdown due

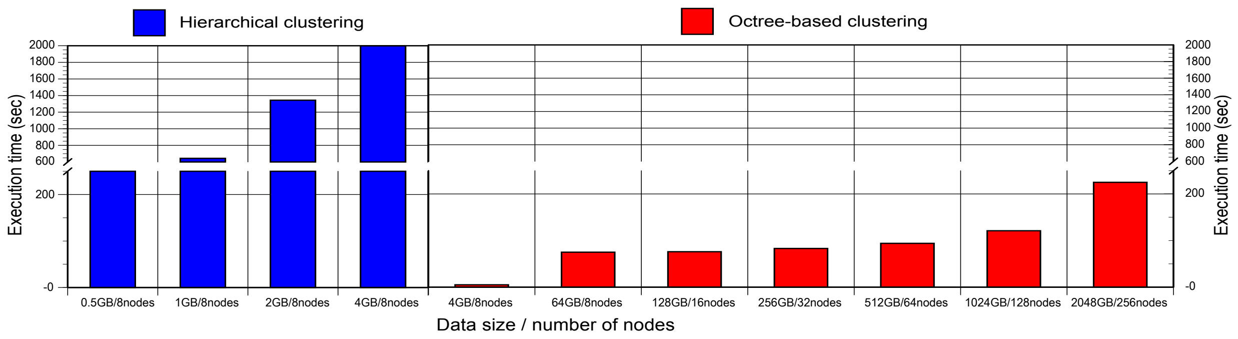


Fig. 5. Performance comparison of our distributed clustering vs. the hierarchical clustering when the size of data increases.

to the iterative exchange of densities in the MapReduce-MPI shuffling.

We note that we achieve the outstanding scalability in size and time while preserving and increasing the accuracy of our analysis over the entire dataset of generated conformations. In other words, our analysis considers all the generated conformations and performs a single pass on all the simulation results across all the node of the distributed system. In contrast, sample-based methods perform clustering analysis of the ligand conformations only on a subset of the conformations and thus too often trade off accuracy for scalability [**?**], [**?**]. Since these sample-based methods deal with reduced-size datasets and lower levels of accuracy, they are not considered in our comparison.

V. Conclusion and future work

This paper presents a scalable and accurate octree-based clustering method for selecting drug candidates in large distributed protein-ligand docking datasets. The scalability is achieved by applying metadata mapping locally on each compute node and by avoiding substantial data communication of ligand conformations or metadata. The accuracy is achieved by accurately capturing the geometries of ligand conformations using metadata. We measure both the scalability and accuracy of a real dataset ranging from 512 MB to 2 TB. Our method based on metadata mapping and octree-based clustering is approximately 400X faster and can analyze approximately 500X larger datasets compared with a traditional hierarchical clustering based on direct comparisons of ligand conformations. Accuracy results on 56 ligands docking in 3 proteins (i.e., HIV, trypsin, and P38alpha) show that our method can achieve 100%, 81.0%, and 83.3% clustering accuracy, respectively, whereas the hierarchical-based clustering achieves 87.0%, 76.2%, and 50.0% clustering accuracy and the energy-based scoring achieve only 34.8%, 23.8%, and 0.8% accuracy.

Acknowledgment

This material is based upon work supported in part by NSF CCF1318445 and by the U.S. Department of Energy, Office of Science, contract DE-AC02-06CH11357. The research used the Argonne LCRC "Fusion" cluster. The authors thank the Docking@Home volunteers.

References

[1] A. N. Jain, "Bias, reporting, and sharing: Computational evaluations of docking methods," *J. Comp.-Aided Mol. Desi.*, vol. 22, pp. 201–212, 2008.

[2] O. Rahaman, T. Estrada, D. Doren, M. Taufer, C. L. Brooks III, and R. S. Armen, "Evaluation of several two-step scoring functions based on linear interaction energy, effective ligand size, and empirical pair potentials for prediction of protein-ligand binding geometry and free energy," *J. Chem. Inf. Mod.*, vol. 51, no. 9, pp. 2047–2065, 2011.

[3] W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics," *J. Mol. Graph.*, vol. 14, pp. 33–38, 1996.

[4] S. Ceri and R. Manthey, "Chimera: A model and language for active DOOD systems," in *Proc. of the 2nd East-West Database Workshop, Workshops in Computing*, 1994.

[5] D. Wei, H. Zheng, N. Su, M. Deng, and L. Lai, "Binding energy landscape analysis helps to discriminate true hits from high-scoring decoys in virtual screening," *J. Chem. Inf. Model.*, vol. 50, no. 10, pp. 1855–64, 2010.

[6] T. Estrada, R. S. Armen, and M. Taufer, "Automatic selection of near-native protein-ligand conformations using a hierarchical clustering and volunteer computing," in *Proc. of the First ACM International Conference on Bioinformatics and Computational Biology*, 2010.

[7] Docking@home. [Online]. Available: http://docking.cis.udel.edu

[8] T. Estrada, B. Zhang, P. Cicotti, R. S. Armen, and M. Taufer, "Reengineering high-throughput molecular datasets for scalable clustering using MapReduce," in *Proc. of the 14th IEEE International Conference on High Performance Computing and Communications*, 2012.

[9] B. Zhang, T. Estrada, P. Cicotti, and M. Taufer, "On efficiently capturing scientific properties in distributed big data without moving the data - a case study in distributed structural biology using MapReduce," in *Proc. of the 16th IEEE International Conferences on Computational Science and Engineering*, 2013.

[10] K. Backbro, S. Lowgren, K. Osterlund, J. Atepo, T. Unge, J. Hulten, N. M. Bonham, W. Schaal, A. Karlen, and A. Hallberg, "Unexpected binding mode of a cyclic sulfamide HIV-1 protease inhibitor," *J. Med. Chem.*, vol. 40, pp. 898–902, 1997.

[11] F. Dullweber, M. T. Stubbs, D. Musil, J. Sturzebecher, and G. Klebe, "Factorising ligand affinity: A combined thermodynamic and crystallographic study of trypsin and thrombin inhibition," *J. Mol. Bio.*, vol. 313, no. 3, pp. 593–614, 2001.

[12] Z. Wang, B. J. Canagarajah, J. C. Boehm, S. Kassisa, M. H. Cobb, P. R. Young, S. Abdel-Meguid, J. L. Adams, and E. J. Goldsmith, "Structural basis of inhibitor selectivity in MAP kinases," *Struct.*, vol. 6, pp. 1117–1128, 1998.

[13] S. Lorenzen and Y. Zhang, "Identification of near-native structures by clustering protein docking conformations," *Proteins: Struct., Funct., Bioinf.*, vol. 68, pp. 187–194, 2007.

[14] D. Kozakov, K. H. Clodfelter, S. Vajda, and C. J. Camacho, "Optimal clustering for detecting near-native conformations in protein docking," *Biophys. J.*, vol. 89, no. 2, pp. 867–875, 2005.