ClusterMol Literature Review

Nicholas Limbert
LMBNIC008@myuct.ac.za
University of Cape Town
Cape Town, South Africa

ABSTRACT

Molecular Dynamics (MD) is a computer simulation method for generating and analysing the trajectories of atoms and molecules. This is a numerical method where atoms positions are recalculated over extremely small periods of time to generate an overall motion picture of the molecular system for a longer period of time. These simulations play an important role in determining the characteristics of a molecule, specifically uncovering conformational change in the molecular system. Molecular Dynamic Simulations of molecules produce a large number of data points based on their trajectory/shape over time. Clustering analysis is typically used to determine the dominant conformations of a molecule. Validation is a fundamental part of clustering analysis and it is essential for achieving meaningful results. There are numerous validation indices used in clustering analysis, with no consensus on a standardised method or index to validate a clustering algorithm. This paper outlines and reviews the most common and applicable validation methods available in order to determine which of these methods may be best suited for validating clustering analysis of flexible molecules.

CCS CONCEPTS

• **Applied computing** → *Computational biology*; • **Mathematics of computing** → *Cluster analysis*; • **Theory of computation** → *Unsupervised learning and clustering*;

KEYWORDS

Clustering Analysis, Molecular Dynamic Simulations (MD), Carbohydrate Molecules, Dominant Conformations, Cluster Validity Indices (CVI's), Quantitative Comparisons

1 INTRODUCTION

The data produced by molecular dynamic (MD) simulations is growing rapidly due to the increase in computational power [17], size of molecular systems and length of simulations. These simulations produce a large number of molecular frames (snapshots) over time, with many of these frames representing similar conformations [20]. Molecular Dynamic simulations provide a method to generate the evolution of a molecular system over time [7, 8, 13, 18]. This allows us to simulate complex interactions of molecules and atoms to better understand their various characteristic. Allowing us to develop a better understanding of the various macromolecular structures and variations in the molecular conformations of the molecule [14]. Clustering analysis is typically used to group similar molecular conformations into partitions for analysis.

Cluster analysis is a technique for grouping a set of data objects into clusters. A cluster is defined as a collection of data objects where objects within the same cluster are similar to one another while dissimilar to data objects in another cluster. Clustering analysis does not require prior knowledge. Prior knowledge refers to all additional data that is available about the data set excluding the data set itself. In terms of clustering analysis prior knowledge would refer to the possible underlying patterns already known, distributions of data set, the correct number of clusters within the data set and correctly label data. Labelled data objects have additional meaningful information tagged. MD data is not tagged as it would take a large amount of time and resources to label a dataset of this size. MD simulations do not produce any additional prior knowledge that would be relevant, therefore clustering analysis is the primary method for analysing the data as it does not require prior knowledge. Clustering makes use of a similarity metric [1] between the objects in the data set. This metric is used to form clusters of similar objects and differentiate dissimilar objects in other clusters.

Many clustering algorithms focus on optimising some characteristic such as the similarity measure between objects or the validity measure of the data [16, 17]. The validity measure is commonly referred to as the Cluster Validation Index or CVI. This index measures and provides insight into the quality of a clusters generated by a clustering method. A clustering method that generates good quality clusters usually has high intra-cluster similarity and low inter-cluster similarity. Intra-cluster similarity refers to the cohesion or compactness of a cluster. This indicates how well a cluster is formed and how similar the objects within a specific cluster are. While inter-cluster similarity is used to determine how dissimilar different clusters are from one another. This is also referred to as cluster separation. In many instances both of these two measures are combined by sum or ratio to generate a Cluster Validation Index (CVI). A good CVI is either indicated by a high or low numerical value depending on the methods used to determine the intra and inter cluster metrics.

Validation techniques and indices fall into the following categories - internal, external and relative validation [2]. External techniques focus on validating the generated clusters against that of the predefined correct clusters, however, this data is usually not available. Internal validation techniques make use of data that is only available from the clusters generated, no external data is used. Finally, relative validation is methods that compares different clustering methods against one another as well as compare methods with different parameter inputs. Relative validation methods make use of internal CVI's to compare partitions from the same dataset but generated with different input parameters. Relative validation plays an important role in determining which clustering method generates the best quality clusters as well as the optimal cluster count.

Clustering analysis can lead to an incorrectly partitioned data set. This is due to the fact that initially, we do not know the correct or actual number of clusters in the data set. Clustering analysis techniques have some of the following shortcomings: influence of outliers, large cluster formations, defining of similarity metric, linkage factors, algorithms constraints and prior knowledge. Outliers can often influence the number of clusters, developing small clusters of individual objects. However, if the outlier is included in the cluster the intra-cluster variation is increased - we aim for low intra-cluster variation. Large cluster formations are possible when objects are on the borders of two or more clusters, thereafter the clusters are joined to reduce intra-cluster variation and create a single homogeneous cluster. Selecting a similarity metric is an important step to enable the algorithms to differentiate the data correctly. If a correct similarity metric is not chosen the algorithm will not be able to effectively cluster the data. Linkage factors can lead to a variety of different cluster formations in Hierarchical clustering. Clustering is seen as an optimisation problem and therefore may not produce the global optimal solution. This gives rise to multiple different methods for clustering and consequently has given rise to numerous different cluster validation indices. There is no formal definition for cluster validation nor a consensus on how to compare the performance of different clustering algorithms.

This paper reviews the literature on cluster validation techniques and indices. Firstly, we summarise the various validation techniques currently available. Secondly, we outline validation techniques that are currently implemented and used in MD simulations. Thereafter, we provide an in-depth analysis of the validation indices with regards to their quantitative measurement and limitations. Finally, we propose which indices would be best suited for validation measure of Carbohydrate MD data.

2 CLUSTERING METHODS

2.1 Clustering Categories

Clustering algorithms differ in their approach to form clusters. Most algorithms behave based on input parameters, geometry or density distributions [12]. The underlying geometry or patterns in the data set may influence the shapes of cluster formations. This will largely depend on the type of clustering method implemented and how the method aggregates the clusters. The data set may have an underlying statistical density distribution that will also affect the formation of clusters. Many of these algorithms fall into one of the following four categories [2].

(1) Centroid-based clustering (Partitioning-based clustering). Each cluster is represented by a single vector. A data objects inclusion in the cluster is based on its proximity (usually distance) to this vector.
(2) Hierarchical clustering (Linkage-based clustering). Clusters are formed based on each object's proximity to one another. There is a connectivity link between the objects in the cluster to minimise some distance measurement.
(3) Density-based clustering. Clusters are defined as high-density areas and outliers or low-density areas are not included in the clusters.
(4) Distribution-based clustering. Clusters are defined by statistical distribution models. Clusters are formed as distribution areas around a centroid. Each data point may have a probability that it belongs to a specific cluster. If the data is artificially generated we may know the distribution model in advance. MD simulations produce data that we would expect to have an underlying Boltzmann distribution [20].

2.2 Parameter Selection and Validation

The input parameters depend largely on the data set [15]. It is critical that optimal parameters are chosen for a particular clustering algorithm. Determining the optimal parameter is usually done based on some validation measure, therefore the choice of validation index is just as critical in performing reliable clustering analysis. Reliable clustering analysis results in good quality clusters which have characteristics of high intra-cluster similarity and low inter-cluster similarity. The main parameter needed for most clustering algorithms is k, the number of clusters in the data set. This is usually not known, and much research has gone into trying to determine the optimal number of clusters in a dataset [6]. We initially select a value for k and then compute a selected CVI, this is then repeated with different values of k until we find a possible optimal value for the number of clusters in the data set. We compare the values of the CVI for each cluster count. Another parameter needed in distribution-based clustering is the assumption of the underlying distribution [17]. This generates clusters that are based on a distribution model where each point has a probability that it belongs to a cluster. Other parameters that algorithms may require in clustering are the diameter of clusters, cluster size, cluster density, cut-off values and linkage function. The diameter of clusters is a distance cut-off function where no data points outside this distance will be considered for the cluster. This distance cut-off is typically associated with centroid-based clustering where the distance is measured from an internal vector within a cluster. Certain algorithms may also have cluster size limitation parameters to stop the formation of extremely large clusters. Hierarchical clustering makes use of cluster size limitations and linkage functions. Linkage functions are used to determine the distance between sets of observations. A data object is linked to a cluster based on this linkage function. Cluster density defines what qualifies as high-density areas. Many of these parameters refer to a cut-off or stoping values, these are parameters that will stop the algorithm once it satisfies a set of conditions or criteria.

2.3 Clustering Molecular Dynamic Simulations

Clustering analysis requires a similarity metric for the comparison of molecular structures. A clustering algorithm is then used to group structures according to their similarity values. The root-mean-square deviation (RMSD) is used as the similarity metric for clustering based on molecular conformation [1, 15, 20, 21]. In some cases both molecular orientation and conformation [1] are used as the similarity metric, however, this has only been implemented for protein absorption where orientation is an important factor.

3 CLUSTER ANALYSIS VALIDATION

Large data sets are time-consuming to analyse and clustering analysis provides an approach to make sense of the data. Clustering analysis of MD simulation data allows us to understand how a molecule may move between multiple different conformations without having to visually analyse the entire data set. Many validation techniques use graphical summaries [19, 21] to interpret the results and validity of the clustering analysis. Dendrograms illustrate a tree-like structure of the data over time. They depict how the clusters are formed through merging similar clusters in hierarchical clustering. This allows the viewer to understand the formation of structures and to choose a point in the algorithm they deem to have the correct clustered the data. The 2D RMSD plot is often used to illustrate how many unique conformations the MD has sampled and hence the initial number of clusters. The plot maps the RMSD of structures in all the trajectory frames to each other. However, as the dimensions of the data increase, it becomes increasingly harder to produce and interpret graphical summaries [11, 12]. Non-visual metrics such as numerical indicators and ranking measurements can help determine the validity of an algorithm without the need for graphical comparisons [12]. These are typically referred to as Cluster Validity Indices. A few notable CVI's include the Davies-Bouldin index, Calinski-Harabasz index, Dunn index, S-Dbw index and Silhouette index. These indices along with others will be outlined and discussed later in the paper.

There are three types of validation; internal, external and relative validation. External validation is primarily used to compare two sets of clustered data. Comparing a correctly clustered set of data with a data set clustered by an alternative algorithm.

Relative cluster validity compares the behaviour of a clustering algorithm with varying parameter input. In order to find the optimal input parameters and hence optimal partitions of the data. Monte Carlo Simulations have also shown promising results with relative valuation to determine the "stopping" or "cut-off" value for the number of clusters [6]. It is important to note that relative validation should only be used for comparing the same algorithm.

Finally, internal validation validates only the data available in the data set and clusters. Many of the internal validation indices focus on compactness and separation. Compactness is the internal metric for variation within a cluster. Separation is how well dissimilar clusters are separated. Various measurements and methods are used to determine compactness and separation. There is usually a tradeoff between the two measurements as some CVI's will prioritise compactness over separation and visa-versa. As there are multiple different methods for determining the compactness of a cluster as well as the separation between clusters this gives rise to a large number of internal validation indices available [3]. Cluster compactness can be determined by the sum of distances between a centroid, cluster diameter and variance within a cluster to name a few. Cluster separation is often calculated by the distance between centroids, closest points between clusters or the inter-cluster variance. The indices should be carefully selected based on data set and the type of clusters formations desired. Overall the S-Dbw index [12] is one of the most widely implemented internal valuation indices and includes both intra-cluster density (compactness) and inter-cluster variance (separation). The S-Dbw index combines both intra-cluster density and inter-cluster variance to get a numerical value which we aim to minimise. The specifications of the S-Dbw index will be outlined later in the paper.

The main issue with clustering validations is the wide variety of different metrics available and how to select a metric suitable for the dataset. This is due to the fact that a validation metric cannot measure all aspects of an implementation "correctness [15]." The correctness of a clustering algorithm is defined by how well it can generate clusters that are similar to the natural partitions of the data. This would include clusters with high intra-class similarity and low inter-class similarity. The CVI's provide different measurements of the validity of the clusters in terms of intra-cluster density, inter-cluster variance and in some cases error measurements against correct clusters. Intra-cluster density measurements try to determine how well a cluster is structured. Should a data object be included in the cluster or will it increase the intra-cluster variation? We only include data objects that try to minimise this intra-cluster variation. Inter-cluster variation determines the separation between distinct clusters. This measurement is used to determine whether clusters should be combined if similar. How similar a neighbouring cluster is to another. Where correctly clustered data is available for comparison we can use error measurements to asses the similarity of two clusterings.

4 VALIDATION TECHNIQUES FOR CLUSTERING MOLECULAR DYNAMIC SIMULATIONS

Table 1 indicates some of the most widely used internal validation indices (CVI's). The selection of these 15 CVI's was based on numerous papers where these indices performed well against others while also providing a range of different measurements of cluster compactness and separation. Formal mathematical definitions and implementations can be found in the attached reference papers. Arbelaitz et al [3] provide an extensive comparison between 30 different validation indices. They developed criteria to compare these indices against one another. As mentioned previously, choosing a CVI is an important aspect of clustering analysis with a specific data set and these indices should be trialled with the data as there are no clear advantages of a specific index. The Silhouette index, S-Dbw index and score function performed well in the majority of tests [15].

As mentioned previously two measures make up a cluster validity index. Firstly, cluster cohesion or compactness is a measure of how well a cluster is formed. This looks at all the objects within a cluster to see how similar they are. This is also referred to as intra-cluster similarity and we aim to maximise the intra-cluster similarity. Secondly, cluster separation or inter-cluster similarity is how similar clusters are to others. Ideally, we want distant dissimilar clusters with a low inter-cluster measure over the whole data set. These two measurements are either summed or there is a ratio implemented between the two. A ratio is used when the CVI favours one measure over the other. Once the two measures have been combined, we aim to either maximise or minimise the CVI.

C-index computes the sum of distances T over all pairs of objects from the same cluster. The C-index is calculated by taking the T minus Tmin over Tmax minus Tmin. Tmax is defined as the largest distance between all pairs while Tmin is defined as the smallest distance between all pairs. A good C-index is represented by a low value.

Calinski-Harabasz index is a ratio type index. A good index is represented by high values. The compactness of a cluster is determined by the sum of distances between all objects in the cluster to the local centroid while the cluster separation is calculated by the distances between the local centroids and a globally specified centroid.

COP index another ratio type index that has low values for good cluster formations. The cluster compactness is measured by the distance from all objects to the centroid while the separation makes use of furthest neighbour distance. This is the distance from the two furthest objects between two clusters.

CS index is a ratio-type index that has low values for good cluster formations. Cluster compactness is estimated by the cluster diameters while separation is estimated by the nearest neighbour distance. This is the distance from the two closest points between two clusters.

Davies-Bouldin index is summation based index where low values represent a good cluster. The cluster compactness is calculated as the distance from all the points in a cluster to the centroid. The cluster separation is then determined by the distance between all centroids from the clusters.

The Dunn index is used to determine the quality of a specific cluster. The cluster seperation is determined by the nearest neighbour distance between two clusters. The compactness is then calculated by the maximum cluster diameter. The two are combined using a type of ratio and a good Dunn index is represented by high values.

The Gamma index is an adaption of the of Goodman and Kruskal's Gamma statistic. It use the amount of times objects not in the same cluster have larger separation than those that are in the same cluster. The full implementation is outlined in [4].

The SV-index and OS index are two recently proposed indices [22]. SV-index makes us of nearest neighbour distance for separation and the distance from all border points to a centroid for cluster compactness. The OS index has a complex calculation and is outlined in [22]. Both these indices are ratio based and good clusters are represented by high values.

The Pseudo F-statistic and SSR/SST Ratio are both regression-based validation methods. SSR/SST Ratio is also known as the coefficient of determination. The measures the variance explained by the data. A low value illustrates that clustering is likely poor. While the Pseudo F-Statistic is a ratio based index that makes use of the intra-cluster variance or cluster compactness over the total variance of all objects. High Pseudo F-statistic and SSR/SST represent good clustering.

S-Dbw index is a ratio type index with a low value for good cluster implementation. Cluster compactness is estimated by the average scattering within a cluster while separation is determined by the average number of points between clusters. The calculation is outlined in [12].

The score function is a summation type index where a high value represents good clustering. The score function calculates the compactness of the cluster as all objects within a cluster to the local centroid while separation is calculated as the distance from the local centroids to the globally defined centroid.

Silhouette index is summation type index with a high value for good cluster implementation. The cluster compactness is determined by the sum of the distances between all the points in the same cluster. The separation is estimated on the nearest neighbour distance, this is the distance between the two closest points of different clusters.

Shao et al. [20] use several validation metrics such as the pseudo-F statistic, Davies-Bouldin index (DBI) [9], SSR/SST ratio and the "critical distance" in the clustering of MD simulations. Each metric is used for different cluster characteristics. The DBI is used to determine the compactness and separation of all the clusters. The pseudo-F statistic uses the ratio of inter and intra-cluster variance again to determined the overall compactness and separation of the clusters. SSR/SST checks whether adding a cluster will add new information. High SSR/SST values are seen as better and one could compare a range of different cluster counts to determine the ideal cluster count. The critical distance index is used to determine the ideal cluster count. It is shown that low DBI values and high pseudo-F values indicate good partition while the SSR/SST ratio and the "critical distance" are used to determine ideal cluster counts. The validation indices behaved as expected, with high pseudo-F statistic and low DBI values for cluster formation. Including a constant SSR/SST ratio when the ideal cluster count is reached. Together these metrics provided a compressive validation of the algorithms implemented, however, none individually could be used as a metric.

Abramyan et al. [1] use three internal cluster validation techniques; Calinski Harabasz (CH), Devies-Bouldin (DB), and Silhouette (S) indices when clustering of MD simulations. These simulations focus on clustering protein adsorption MD simulation data. The CH index asses a ratio of inter and intra-cluster variation. The DB index makes use of a similar internal cluster metric, however, the distance between clusters is used for separation measurement. While high values of the silhouette index are achieved by proximity within a cluster, each cluster contributes to the overall value for the index. All of the three indices were implemented and evaluated to determine the effectiveness of the algorithms.

Melvin et al. [17] implement a wide variety of different algorithms and make use of the Silhouette index as their basis of comparison. The Silhouette provides a way to determine how similar an object is to its cluster, the compactness compared to other clusters and the separation. The primary use of Silhouette index is for determining the cut-off value for the number of clusters, however, Melvin et al. use to compare different algorithms. Additional metrics are available however were not implemented in the literature.

The main reason for such a variety in clustering validation indices is that no single measure can capture all different aspects of the clustering problem [5, 10]. This being the correct number of clusters and the correct natural partitions in the dataset. Ideally, multiple Cluster Validity Indices (CVI's) results should be interpreted when validating results. Several CVI values that indicate good clustering results provide more significance to the results, therefore the use of multiple CVI's is recommended. S-Dbw index, Silhouette index and Davies-Bouldin index are the most widely used CVI's in the literature and provide consistent results with a variety of different datasets. We expect them to have similar results when validating MD simulation clusterings.

There is limited research into the use of CVI's in their application of clustering MD simulations, specifically clustering Carbohydrate molecules. Most literature available focuses on implementing several algorithms and do not focus on CVI compassions or implementations. The use of a single CVI is common amongst MD simulation clustering however it is shown that, no single CVI can be recommended. CVI's are often overlooked in the clustering analysis process. Many packages make use of regularly available CVI's and do no investigate the use of other CVI's. As some algorithms aim to optimise the CVI it would be irrational to not carefully select a CVI for the given data set.

5 DISCUSSION

Correctly clustered MD data is rarely available to do baseline comparative tests against different clustering algorithms (external analysis). We are therefore unable to make use of external validation methods to determine the effectiveness of a clustering algorithm. It is evident that much of the literature focuses on the implementation of internal metrics that focus on the data available from the cluster formations. Since we usually only have access to the MD simulation data set, internal validation methods would be the primary source of validations.

Due to the flexibility of carbohydrate molecules, validations indices that can detect both compactness and separation would be most suitable. Allowing clusters of similar confirmations to be of any size while also having distinct dissimilar clusters. A validation index must also be able to integrate with an objective function of an algorithm should it require one. Many of these indices may be adapted to validation clustered MD simulation data.

The literature provides a good guideline to determine possible CVI's that will be effective in MD clustering. They also outline the implementations and calculations of the most promising CVI's. One of the main limitations of these CVI's is their lack of use in MD simulation clustering. No single or group of cluster validation indices has not been identified to provide comprehensive validation of an algorithm.

6 CONCLUSIONS

Choosing a CVI is an important aspect of cluster analysis and careful consideration should be done when selecting a CVI. This should be based on the data set, the algorithm implemented and characteristics of the clusters desired.

We aim to implement a range of different clustering algorithms for Molecular Dynamic simulations of Carbohydrate molecules. The evidence is not clear on which validation indices would perform best. Therefore, a possible direction for this research should include the implementation of multiple different CVI's, allowing us to determine the most promising CVI's for this specific clustering analysis. By finding CVI's that perform well we can then compare a range of different clustering algorithms.

There are two areas within the topic that need further research. Firstly, what type of clustering characteristics are best suited for Carbohydrate molecules. This will determine whether the clustering indices implemented should look for a larger number of clusters with minimal internal variance or a smaller number of clusters with greater internal variance. How dissimilar should cluster formations in Carbohydrate molecules be? Secondly, how can current implementations of MD clustering be adapted in order to include these validation measures?

Finally, we aim to provide motivation for the use of these validation indices in Carbohydrate Molecular Dynamic Simulations.

REFERENCES

[1] Tigran M. Abramyan, James A. Snyder, Aby A. Thyparambil, Steven J. Stuart, and Robert A. Latour. 2016. Cluster analysis of molecular simulation trajectories for systems where both conformation and orientation of the sampled states are important. *Journal of Computational Chemistry* 37, 21 (2016), 1973–1982. https://doi.org/10.1002/jcc.24416

[2] Charu C. Aggarwal and Chandan K. (1980-) Reddy. 2014. *Data clustering: algorithms and applications.* CRC Press/Taylor and Francis Group.

[3] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesus M. Perez, and Inigo Perona. 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition* 46, 1 (2013), 243–256. https://doi.org/10.1016/j.patcog.2012.07.021

[4] Frank B Baker and Lawrence J Hubert. 1975. Measuring the Power of Hierarchical Cluster Analysis. *J. Amer. Statist. Assoc.* 70, 349 (1975), 31–38. http://www.tandfonline.com/doi/abs/10.1080/01621459.1975.10480256

[5] J.C. Bezdek and N.R. Pal. 1998. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 28, 3 (1998), 301–315.

[6] Derrick Boone. 2013. Determination of the Number of Clusters in a Data Set: A Stopping Rule x Clustering Algorithm Comparison. *International Journal of Strategic Decision Sciences* 2 (10 2013), 1–13. https://doi.org/10.4018/jsds.2011100101

[7] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. 2009. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* 30, 10 (2009), 1545–1614.

[8] David A. Case, Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. 2005. The Amber biomolecular simulation programs. *Journal of Computational Chemistry* 26, 16 (2005), 1668–1688.

[9] David L Davies and Donald W Bouldin. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1, 2 (1979), 224–227.

[10] Vladimir Estivill-Castro. 2002. Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter* 4, 1 (2002), 65–75.

[11] M. Feher and J. M. Schmidt. 2001. Metric and multidimensional scaling: efficient tools for clustering molecular conformations. *Journal of chemical information and computer sciences* 41, 2 (2001), 346.

[12] M. Halkidi and M. Vazirgiannis. 2001. Clustering validity assessment: finding the optimal partitioning of a data set. (2001), 187–194. https://doi.org/10.1109/ICDM.2001.989517

[13] Berk Hess, Carsten Kutzner, David van Der Spoel, and Erik Lindahl. 2008. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of chemical theory and computation* 4, 3 (2008), 435–447.

[14] Adam Hospital, Josep Ramon Goni, Modesto Orozco, and Josep L. Gelpi. 2015. Molecular dynamics simulations: advances and applications.(Report). 8, 1 (2015), 37–47.

Table 1: Validation Indices

Validation Index	Cluster Cohesion (Compactness)	Cluster Separation	Ratio or Summation	Good Partition Indicator	Comments
C-index	Normalised metric	n/a	Summation	Low	
Calinski-Harabasz index	Distance from the points in a cluster to the clusters centroid.	Distance from the cluster centroid to global centroid.	Ratio	High	
COP index	Distance from the points in a cluster to its centroid	Furthest neighbour distance	Ratio	Low	
CS index	Cluster diameters	Nearest neighbour distance	Ratio	Low	
Davies-Bouldin index	Distance from the points in a cluster to the clusters centroid	Distance between cluster centroids	Summation	Low	
Dunn index	Maximum cluster diameter	Nearest neighbour distance	Ratio	High	Many different variations and implementation.
Gamma index	n/a	n/a	Summation	Low	Adaption of Goodman and Kruskal's Gamma index. Deals with the strength of association between two variables.
OS index	Distance from the border points in a cluster to its centroid	A more complex mathematical definition is defined	Ratio	High	SV-index with a more complex seperation function.
Pseudo F-Statistic	Intra-cluster variance	Inter-cluster variance	Ratio	High	The pseudo-F statistic is a ratio of the between-cluster variation to the within-cluster variation.
S-Dbw index	Average scattering within a cluster	Average number of points between the clusters	Ratio	Low	Specific formal definitions for cluster compactness and separation.
Score function	Distance from the points in a cluster to its centroid	Distance from the cluster centroid to global centroid	Summation	High	
Silhouette index	Distance between all the points in the same cluster	Nearest neighbour distance	Summation	High	
SSR/SST(explained variation/total variation)	n/a	n/a	Ratio	High	Also know as the coefficient of determination. Measures the variance explained by the data. (Allowing us to determine "cut-off values")
SV-Index	Distance from the border points in a cluster to its centroid	Nearest neighbour distance	Ratio	High	

[15] Pablo Jaskowiak, Davoud Moulavi, Antonio Furtado, Ricardo Campello, Arthur Zimek, and Jorg Sander. 2016. On strategies for building effective ensembles of relative clustering validity criteria. *Knowledge and Information Systems* 47, 2 (2016), 329–354. https://doi.org/10.1007/s10115-015-0851-6

[16] Mary E. Karpen, Douglas J. Tobias, and III Brooks, Charles L. 1993. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. (polypeptide Tyr-Pro-Gly-Asp-Val). *Biochemistry* 32, 2 (1993), 412–420.

[17] Ryan Melvin, Ryan Godwin, Jiajie Xiao, William Thompson, Kenneth Berenhaut, and Freddie Salsbury. 2016. Uncovering Large-Scale Conformational Change in Molecular Dynamics without Prior Knowledge. *Journal of Chemical Theory and Computation* 12, 12 (2016), 6130–6146. https://doi.org/10.1021/acs.jctc.6b00757

[18] Mark T. Nelson, William Humphrey, Attila Gursoy, Andrew Dalke, Laxmikant V. Kale, Robert D. Skeel, and Klaus Schulten. 1996. NAMD: A parallel, object-oriented molecular dynamics program. (Jan. 1996).

[19] Peter J. Rousssew. 1988. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *International Abstracts in Operations Research* 39, 1 (1988). https://doi.org/10.1057/iaor.1988.183

[20] Jy Shao, Sw Tanner, N. Thompson, and Te Cheatham. 2007. Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *Journal Of Chemical Theory And Computation* 3, 6 (2007), 2312–2334. https://doi.org/10.1021/ct700119m

[21] Thibault Tubiana, Jean-Charles Carvaillo, Yves Boulard, and StÃĭphane Bressanelli. 2018. TTClust: A Versatile Molecular Simulation Trajectory Clustering Program with Graphical Summaries. *Journal of Chemical Information and Modeling* 58, 11 (2018), 2178–2182. https://doi.org/10.1021/acs.jcim.8b00512

[22] Kr Zalik and B Zalik. 2011. Validity index for clusters of different sizes and densities. *Pattern Recognition Letters* 32, 2 (2011), 221–234.