UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

CS/IT Honours
Final Paper 2020

Title: Developing a Framework to Analyse Clustering Algorithms for Molecular Dynamics Trajectories

Author: Robyn McKenzie (MCKROB018)

Project Abbreviation: ClusterMol

Supervisor(s): Associate Professor Michelle Kuttel

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	0
Theoretical Analysis	*0*	*25*	0
Experiment Design and Execution	*0*	*20*	20
System Development and Implementation	*0*	*20*	5
Results, Findings and Conclusions	*10*	*20*	20
Aim Formulation and Background Work	*10*	*15*	15
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	0
Total marks		**80**	**80**

Developing a Framework to Analyse Clustering Algorithms for Molecular Dynamics Trajectories

Robyn McKenzie
mckrob018@myuct.ac.za
University of Cape Town
Cape Town, South Africa

ABSTRACT

Molecular Dynamics (MD) simulations are used to analyse the physical structure and behaviour of molecules. Due to the size of the trajectories that they produce, clustering algorithms are commonly used to reduce the data into a manageable set of partitions which represent the dominant conformations of the molecule. Although these clustering algorithms are often applied to nucleic acids and proteins, which are relatively inflexible molecules, they are less often applied to highly flexible molecules, such as carbohydrates. As clustering algorithms vary in their effectiveness for different types of data, the usefulness of different algorithms for clustering highly flexible molecules must be evaluated. We applied the iMWK-Means and HDBSCAN clustering algorithms to trajectories of carbohydrates with varying levels of flexibility and evaluated the results. While neither algorithm is particularly suited to molecules with extreme flexibility, the findings show that HDBSCAN is able to produce useful clusters from relatively flexible molecules due to its ability to classify frames as noise, while iMWK-Means is particularly suited to identifying subtle details in primarily stable molecules but can also produce useful clusters from molecules with some flexibility.

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

1 INTRODUCTION

Molecular Dynamics (MD) trajectories are powerful tools which simulate the physical behaviour of complex molecules over time. Each trajectory consists of frames which record the position of the molecule at each time step. However, as these trajectories contain upwards of 50 000 frames [13], manual analysis of the data is not feasible. Clustering algorithms are a useful solution to this problem, as they extract the dominant conformations by grouping the simulation frames into clusters of similar frames. This reduces the amount of data to be analysed into a manageable set of groups which, if ideal clustering were to be applied, have maximal dissimilarity between them and maximal similarity within them.

Although trajectories of nucleic acids and proteins are commonly clustered [13, 15], carbohydrates, which are comparatively much more flexible, are not. Different clustering algorithms will produce different results for a single trajectory. This is because they approach clustering differently and are often suited to particular types of data. Clustering algorithms also have different weaknesses, with some only producing clusters of similar shapes, sizes or densities [15], and some having parameters that introduce bias [13] or require tuning to produce optimal results.

iMWK-Means [6] is a k-means variant which addresses some of the weaknesses of traditional k-means by intelligently selecting initial centroids, so that results are deterministic, and preventing bias from being introduced by not requiring a desired cluster count to be supplied [13]. It also can also handle outliers by rescaling the data.

HDBSCAN [3, 4] is a density-based algorithm which implements a partially hierarchical approach. It can disregard noise in the data by classifying it as such and has been recommended for use with intrinsically disordered proteins [13].

In this paper, the iMWK-Means and HDBSCAN algorithms are applied to meningococcal polysaccharrides Y and W, and to the microbial surface polysaccharrides of two Shigella flexneri serotypes. The cluster results are then evaluated to determine if the dominant conformations of the molecules have been captured. As this research is focused on the performance of the clustering algorithms on MD trajectories, the biological properties of the molecules will not be considered. Instead, evaluation will focus on how well-formed the clusters are and how well-defined the differences between them are.

2 BACKGROUND

2.1 Clustering for MD Trajectories

Many factors account for differences between clustering algorithm results. One factor is whether or not the algorithm requires user-specified parameters, such as a desired cluster count. Having to specify parameters can introduce bias [13], and can have a considerable effect on the results [15]. It can also mean that the algorithm will have to be run many times to find an optimal result. Different algorithms also exhibit different sensitivity to outliers and "noisy" data, or may have a tendency to find only find convex clusters or clusters of similar sizes or densities [15]. Ideally, we would want an algorithm that can find concave clusters and clusters of varying sizes and densities.

2.1.1 iMWK-Means. Intelligent Minkowski Weighted K-Means (iMWK-Means) is a k-means variant proposed by de Amorim et al. [6]. It improves upon traditional k-means, which is arguably one of the simplest and most popular clustering algorithms [1, 6]. Traditional k-means begins by randomly selecting a cluster centroid for each cluster, based on the desired cluster count given by the user. Each instance in the data set is then assigned to the cluster whose centroid it is closest to. After each instance is assigned, the cluster centroids are recalculated as an average of the instances assigned to the cluster. The instances are then reassigned, and the process continues iteratively until the cluster assignment results in no change to the cluster centroids, indicating convergence [15]. As the initial centroids are selected randomly, the results are nondeteriministic [15] and k-means must be run multiple times to find optimal clustering.

iMWK-Means eliminates one of the key weaknesses of traditional k-means by not requiring a desired cluster count from the user [13]. Instead, it begins by overestimating the cluster count and then iteratively performing rounds of k-means and rescaling the data by adding weights based on the results. Each instance is rescaled based on the distance from it to the centroid of the cluster that it is in. This results in dense regions of the data having an increased chance of being clustered together in future iterations, while outlier frames that are further from the centroids have less impact on the final clusters. iMWK-Means is also deterministic as it intelligently selects initial centroids [6] rather than randomly seeding them. This means that it does not need to be applied multiple times to find optimal clustering.

2.1.2 HDBSCAN. Hierarchical Density-Based Spatial Clustering of Applications with Noise is an algorithm proposed by Campello et al. [3, 4]. It is a density-based algorithm which identifies clusters of varying shapes by searching for higher density regions in the data separated by lower density regions [13]. It does this by considering the neighbourhoods in the data, which each consist of a core point and the k points which are nearest to it. The distance from a core point to its furthest neighbour is the core distance of the neighbourhood. It then creates a network from the points, where the weight of the edge between any two points is calculated based on the distance between them, and the core distances of the neighbourhoods that each of the points are member of. Edges are then removed from the network, beginning with those with the greatest weights, until removing another edge would completely disconnect two regions of the data. The points with the lowest weighted edges between them are then merged into clusters, as is done in single-linkage hierarchical clustering [15], until every data point is in a single cluster. This merging process results in a hierarchical tree, and the final clusters are produced by cutting the tree based on the minimum cluster size and a stability metric.

HDBSCAN is also able to disregard instances that it identifies as noise by assigning them a cluster label of -1 [13]. Although it has two parameters, the minimum cluster size and minimum neighbourhood size k, primarily referred to as minimum samples in this paper, the option of setting the parameters equal to each other, and to their minimum value, is a suitable default which allows HDBSCAN to be run non-paramterically [3, 13].

2.1.3 Application to MD. Melvin at al.[13] applied iMWK-Means to a series of nucleic acids and preteins. They found iMWK-Means to be ideal for identifying subtle details in stable molecules that HDBSCAN did not detect. HDBSCAN was able to isolate the dominant conformations in relatively stable molecules, and locate stable conformations in unstable or disordered systems. Melvin et al. recommend the algorithms for exploratory clustering based on the fact that both algorithms can be used non-parametrically, and specifically recommend HDBSCAN for use with intrinsically disordered proteins.

2.2 Cluster Validity Indices

To objectively evaluate the quality of a clustering result, Cluster Validity Indices (CVIs) are often used to compare candidate partitions of a data set [13, 15]. A CVI is a mathematical measure which quantifies how well the data in a set has been partitioned. More particularly, this means quantifying how cohesive each cluster is, as well as how well separated the clusters are [2]. However, there are differences in how each CVI evaluates the cluster cohesion, cluster separation and general quality of the clusters. Because of this, the CVIs may disagree on which partition fits the data best and may also be influenced by features of the clustering which do not directly affect their quality, such as the cluster count and comparative size of the clusters [15]. As no single CVI can evaluate all aspects of the clustering, and may not always give a true representation of the cluster quality, it is recommended that multiple CVIs are used together when evaluating cluster results [2, 9].

As CVI results are affected by the data, each CVI should be used as a relative measure, rather than an absolute measure, of cluster quality. CVI results are more useful when comparing different clustering results for a single data set, rather than comparing clustering results across different data sets. For example, CVIs may be used when comparing results over a range of parameter values for a single clustering algorithm, or when comparing results between clustering algorithms [2].

Arbelaitz et al. [2] performed an extensive comparison of different CVIs and their effectiveness with a range of data sets. Their results showed that, among a few others, the Silhouette (S), Davies-Bouldin (DB) and Calinski-Harabasz (CH) indices produced superior results. The DB and CH indices measure cluster cohesion based on the distance from each point in a cluster to its cluster centroid, while the S index uses the distance between all points in a cluster. Cluster separation is based on shortest intercluster distance for the S index, the distance between all cluster centroids for the DB index and the distance from each cluster centroid to the global centroid for the CH index.

Better clustering is indicated by values closer to 1 for the S index (although it will always fall between -1 and 1), by lower values for the DB index and by higher values for the CH index [2, 13].

3 DESIGN AND IMPLEMENTATION

The iMWK-Means and HDBSCAN algorithms were implemented within a Python framework, the purpose of which was to facilitate efficient clustering jobs and be easily extensible so that additional algorithms can be added. The framework processes clustering jobs on MD trajectories and test data, and also includes other helpful functionality for working with clustering algorithms and MD trajectories.

3.1 iMWK-Means and HDBSCAN

The iMWK-Means and HDBSCAN algorithm implementations for MD trajectory data were adapted from a Python library made available by Melvin et al. [13]. This library uses the HDBSCAN implementation by McInnes et al. [12] and code from de Amorim [5], one of the iMWK-Means authors. Slight adaptations to the existing code had to be made so that the algorithms could cluster on either the trajectory or a two-dimensional array. Changes also had to be made to allow the cluster output to be used by other parts of the framework.

The iMWK-Means is non-parametric, so the implementation does not require any user parameters. Although HDBSCAN can be used non-parametrically, its two parameters, *minimum cluster size* and *minimum samples*, do have a drastic affect on results and can be tuned to improve clustering. As expected, *minimum cluster size* provides a lower limit on the number of frames, or instances, per cluster. Slightly less intuitively, *minimum samples* is a measure of how sparse regions of the data can be before they are considered noise. It primarily affects how readily HDBSCAN will label frames as noise. When it is set to 1, the number of frames which are conserved and not labelled as noise will be maximised. As these parameters do have an effect on the cluster results, it will often be necessary to run HDBSCAN with a range of different parameter values in order to find those that are most suited to the data and hence produce the best clustering. The minimum values for minimum cluster size and minimum samples, 2 and 1 respectively, have been suggested as default values for the parameters [3, 13], however, it was found that this tended to produce poor results for our test trajectories.

3.2 Framework Design

The clustering framework, ClusterMol, was developed in Python and is structured as a pipeline to maximise its extensibility. The MDTraj library [11] is used to read in the MD trajectories and make basic changes, such as atom selection, frame selection and downsampling, to them prior to clustering with one of the available algorithms. Alternatively, clustering can also be performed on test data sets from Scikit-learn [14].

After the clustering job is complete, ClusterMol can output time-series plots of frame index against cluster index, frame counts per cluster, CVI results and the largest clusters as Protein Data Bank (.pdb) files.

ClusterMol can be run via command line arguments or it can be given a file containing these arguments. If a file is supplied, it can contain multiple sections of arguments which will each be processed and run in turn. This allows the user to efficiently run many jobs in a row. The arguments include the the choice of algorithm, the source trajectory file or testing data to be used, the required algorithm parameters and which CVIs to compute.

4 EVALUATION

The iMWK-Means and HDBSCAN algorithms were applied to two generated data sets, three testing data sets - Breast Cancer, Iris and Wine, and four MD trajectories. The generated and test data sets were used for the purpose of algorithm validation, as they have a known correct clustering. The MD trajectories were used to evaluate the usefulness of the algorithms for real MD data.

4.1 Algorithm Validation

HDBSCAN and iMWK-Means were tested for validity with two basic data sets - A and B - which were artificially generated with Scikit-learn [14]. Data sets A and B both consist of 100 instances, each with two dimensions or features. In particular, A contains five tight clusters with low intracluster variation and clear separation between clusters, while B contains five clusters with increased overlap and higher intracluster variation. The purpose of these data sets is to test the performance of the algorithms on data which inherently has very clear clusters, as in data set A, as well as on data which has overlapping clusters with less obvious partitions, as in data set B.

HDBSCAN and iMWK-Means were also tested with the Breast Cancer, Iris and Wine data sets, which can be accessed, with their known correct clustering, through Scikit-learn [14]. The Breast Cancer data set contains 569 instances, with 30 features each, of which 212 are correctly classified as malignant, and 357 are correctly classified as benign. The Iris data set contains 150 instances of iris flowers which are each described by four features. The correct clustering categorises the 150 instances into three iris flower varieties of size 50 each. The final two classes overlap significantly and are difficult to distinguish. The Wine data set contains 178 instances describing different wines through 13 features. The correct partitioning consists of three classes with sizes 59, 71 and 48.

The Breast Cancer, Iris and Wine data sets allow the algorithms to be tested on a greater variety of data sets, rather than just A and B. Each of them have more than two features, unlike A and B, and Breast Cancer in particular has a relatively large number of features, which will allow us to see the effect of higher dimensionality on the algorithm results. The data sets also vary in terms of the sizes of the correct clusters. While Iris, A and B have clusters of equal sizes, Breast Cancer and Wine have clusters of different sizes. This allows the effect of this variation to be tested, particularly as some algorithms are biased towards towards clusters of equal sizes.

In attempt to find optimal clustering, most of the data sets were clustered more than once with HDBSCAN while varying the value of *minimum cluster size*. The process of trying different values was less systematic for these validation data sets than for the MD data, which will be described in the following section. Instead, different values for *minimum cluster size* where tried until a result which appeared to come close to the correct clustering and had good CVI results was identified. The selected HDBSCAN results were compared to the iMWK-Means results using percentage accuracy, which is the percentage of instances assigned to the correct cluster, and CVI results. As data sets A and B are two dimensional, their cluster results could also be compared visually on a two-dimensional plane.

4.2 Testing with MD Trajectories

Four carbohydrates molecules were chosen for testing. They have varying levels of flexibility, ranging from relatively stable to extremely flexible, so that the effects of increased flexibility could be evaluated.

The algorithms were first applied to meningococcal polysaccharides Y and W, which will be referred to as MenY and MenW respectively. These molecules are carbohydrates and have similar molecular structure [10]. The MenY and MenW trajectories each consist of a chain of three repeating units (RUs). The trajectories were then downsampled from their original 40021 frames to 4003 frames, by taking every 10th frame. Although MenY and MenW are similar in structure, MenY is expected exhibit a single conformation the majority of the time while MenW experiences more conformational changes and has at least a few dominant conformations [10].

The next trajectories that the algorithms were applied to were the microbial surface polysaccharides of two Shigella flexneri serotypes, specifically serotype Y, simulated with three repeating units and and serotype 6, simulated with six repeating units. These trajectories will be referred to as S.flexneri Y 3RU and S.flexneri 6 6RU respectively. As with the previous two trajectories, each trajectory was downsampled, from 40018 frames to 4002 frames in the case of S.flexneri Y 3RU, and from 36018 frames to 3602 frames in the case of S.flexneri 6 6RU. These molecules are carbohydrates and are expected to contain many small clusters, as S.flexneri Y 3RU is known to be extremely flexible [7] and we would expect the same for S.flexneri 6 6RU.

Prior to clustering, a selection statement, available in Supplementary Material Table S7, was applied to each trajectory. The purpose of the selection statements is to select the relevant, or core, atoms to be clustered on, rather than clustering on all the atoms. This decreases the dimesionality of the clustering job by ignoring irrelevant features. It allows the clustering algorithms to focus on actual conformational changes, rather than being affected by movement of less significant parts of the molecule which would otherwise be detected as additional variation between frames and contribute to noise in the data. The selection statements that were applied to the trajectories excluded all hydrogen atoms and selected only the core structure of the molecule. One, two and four terminal residues from each end of the molecule were also excluded from the selection for S.flexneri Y 3RU, MenY and MenW, and S.flexneri 6 6RU respectively. In addition to the selection statements, the first 802 frames of the Shigella trajectories were ignored as the equilibration phase of the simulation.

As iMWK-Means is non-parametric, it was applied once to each trajectory. HDBSCAN was run multiple times per trajectory while varying the value of *minimum cluster size* in order to find the optimal clustering. For each trajectory, *minimum cluster size* was first set to 2, then to each multiple of 5 between 5 and 100, stopping earlier if the clustering result classified 100% of frames as noise. The value of *minimum samples* was not varied and remained set to 1, as preliminary testing indicated that values any higher than this would result in a large proportion of frames classified as noise. Once the clustering was complete, the results for each trajectory were compared with respect to the CVI scores, by looking for the maximum Silhouette and Calinski-Harabasz scores and the minimum Davies-Bouldin scores. The Visual Molecular Dynamics program (VMD) [8] was then used to visualise the superimposed frames within each cluster so that the definition of the clusters and the distinctness between them could be evaluated. For clarity, only the atoms which were selected for clustering by the selection statement were visualised. Finally, the cluster count and proportion of frames classified as noise was also considered. Ideally the majority of the frames should be put into clusters, rather than being labelled as noise, so that a complete picture of the molecule conformations can be attained. For a similar reason, it is also not helpful if many small clusters are produced, as this provides less insight into the dominant conformations of the molecule.

5 RESULTS AND DISCUSSION

Results from testing with the generated data sets and Breast Cancer, Iris and Wine were evaluated first, followed by the MD trajectory results. Comparisons were made using CVI results and, in the case of the MD data, visualisation of the frames in each cluster.

5.1 Algorithm Validation Results

Figure 1 shows the results of clustering data sets A and B with iMWK-Means and HDBSCAN. HDBSCAN, with a *minimum cluster size* value of 5, correctly partitioned all five clusters in data set A, evidenced by the fact that each cluster, indicated by the different shapes, is filled in with a single and unique colour (Figure 1c). iMWK-Means correctly partitioned three of the five clusters, however, the final two clusters have both been coloured blue, which indicates they were grouped into a single cluster (Figure 1a). The CVI results for data set A, which can be viewed in Tables S1 and S2 in Supplementary Material, support the five cluster split as a better fit for the data than the four cluster split. This is indicated by higher values for S and CH indices and lower values for the DB index when comparing the HDBSCAN results to the iMWK-Means.

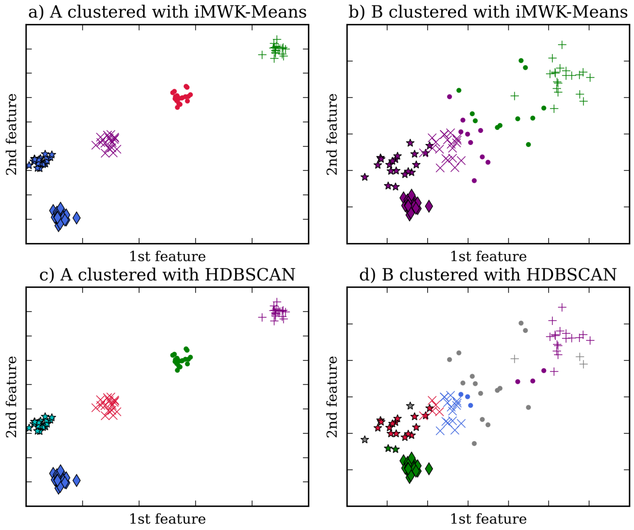

Figure 1: HDBSCAN and iMWK-Means clustering of generated data sets The different shapes indicate the true clusters while the different colours indicate the clusters assigned by the algorithms. In the case of HDBSCAN, grey indicates that the instance was classified as noise. Data set A, in a) and c), contains five clusters with low intracluster variance and high intercluster variance. Data set B, in b) and d), also contains five clusters but with increased intracluster variance and decreased intercluster variance. Data sets A and B both consist of 100 instances, each with two features.

For data set B, iMWK-Means partitioned the data into two clusters (Figure 1b), while HDBSCAN, with a *minimum cluster size* value of 9, came close to isolating four of the clusters, but classified the majority of the most dispersed cluster, which is indicated with the circle markers, as noise (Figure 1d). This result maximises accuracy (the percentage of instances that were assigned to the correct cluster) at 68.0%. Two other tests of HDBSCAN on data set B, with

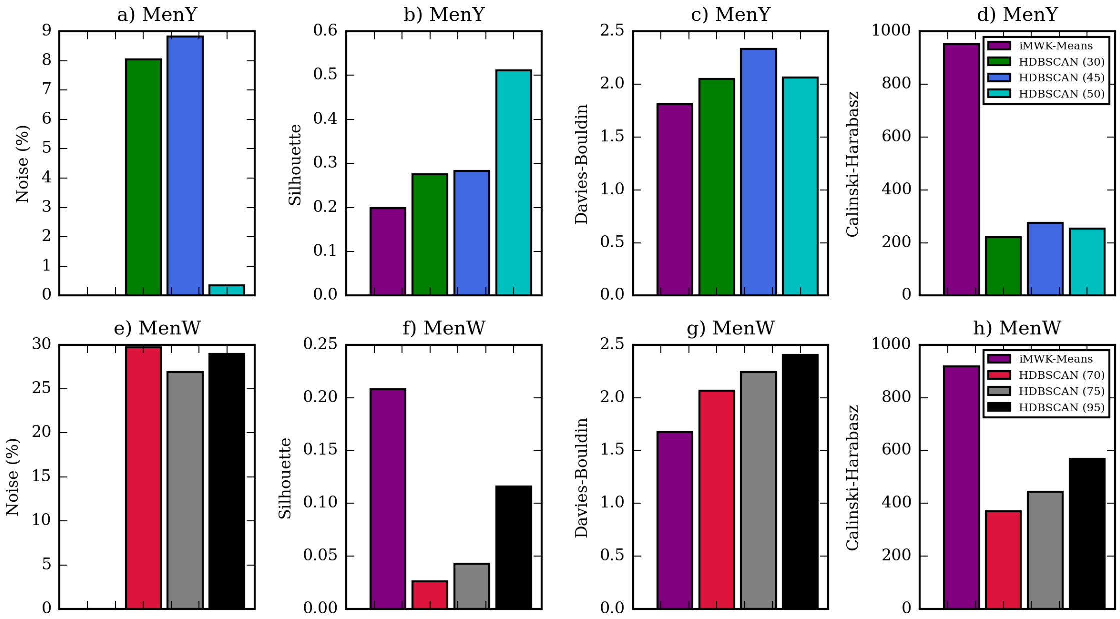

Figure 2: Percentage noise and CVI scores for MenY and MenW cluster results Figures a) to d) compare MenY clustering results (in terms of percentage noise (a), Silhouette index (b), Davies-Bouldin index (c) and Calinski-Harabasz index (d)) from the application of iMWK-Means and three applications of HDBSCAN with *minimum cluster size* values of 30, 45 and 50 - indicated by the number in brackets in the legend in d). Figures e) to h) are similar, but compare the MenW clustering results and use HDBSCAN *minimum cluster size* values of 70, 75 and 95 - indicated by the number in brackets in the legend in h).

minimum cluster size of 6 and 10, can be seen in Table S2. While a *minimum cluster size* value of 9 produced the highest accuracy, a value of 10 labelled the least frames as noise, produced the best CVIs, but had the worst accuracy at 57.0%. Interestingly, the iMWK-Means partitioning of data set B produces the best value of each CVI for the data set, despite arguably producing the least useful clusters.

Although HDBSCAN performed better clustering data sets A and B, iMWK-Means produced better results when clustering the Breast Cancer, Iris and Wine data sets. The iMWK-Means clustered Breast Cancer and Wine with 92.79% and 92.13% accuracy respectively, compared to the HDBSCAN accuracy scores of 69.42% and 60.67% for the same data sets. However, both algorithms partitioned the Iris data set into two clusters, instead of three, and achieved 66.67% accuracy. This data set has relatively low intercluster variation between its last two clusters, making them difficult to differentiate. Once again, the best CVI values do not always indicate the more desirable clustering, with HDBSCAN achieving better CVIs for the Wine data set despite the superior results produced by iMWK-Means. These results can viewed in Tables S1 and S2 in Supplementary Material.

5.2 MD Cluster Results

5.2.1 MenY. Results for each HDBSCAN application to MenY over the range of values for *minimum cluster size* are available in Supplementary Material Table S3. Not all the results for each of the runs are unique, as there are cases where incrementing the value of *minimum cluster size* does not cause any of the clusters to be eliminated because none of their sizes are below the minimum. For example, the HDBSCAN results for *minimum cluster size* values 30 to 40 on MenY are the same.

Out of all the HDBSCAN results, the DB index is minimised with a *minimum cluster size* value of 2. This result classified 42.24% of frames as noise and produced 810 clusters, the largest of which contains only 0.35% of frames in the trajectory. This is ultimately not a useful result, as a collection of many small clusters fails to capture the dominant conformations in the trajectory. Additionally, this minimum DB score coincides with the worst CH and S scores.

The CH and S indices both score their three best, or largest, HDBSCAN results with *minimum cluster size* values of 30, 45 and 50. Values for the DB index are not good for the these results, in fact the largest, and hence worst, DB value is scored for a *minimum cluster size* value of 45. For MenY, the DB index displays a clear preference for results which have larger cluster counts, which is at odds with the results from the S and CH indices. Considering this, we will focus on the HDBSCAN results for *minimum cluster size* values of 30, 45 and 50. Comparisons of these three HDBSCAN results and the iMWK-Means result can be seen in Figure 2a-d.

Of the four clustering jobs compared in Figure 2a-d, HDBSCAN with *minimum cluster size* 50 scores the highest, and best, S index, while iMWK-Means scores the lowest (Figure 2b). However, the CH and DB indices support iMWK-Means as the best result, as it produces a higher CH score (Figure 2d) than any of the HDBSCAN results, by a wide margin, and has the lowest DB score (Figure 2c).

The timeseries for the HDBSCAN results with *minimum cluster size* values of 30, 45 and 50 (Figure 3b-d) show that each time

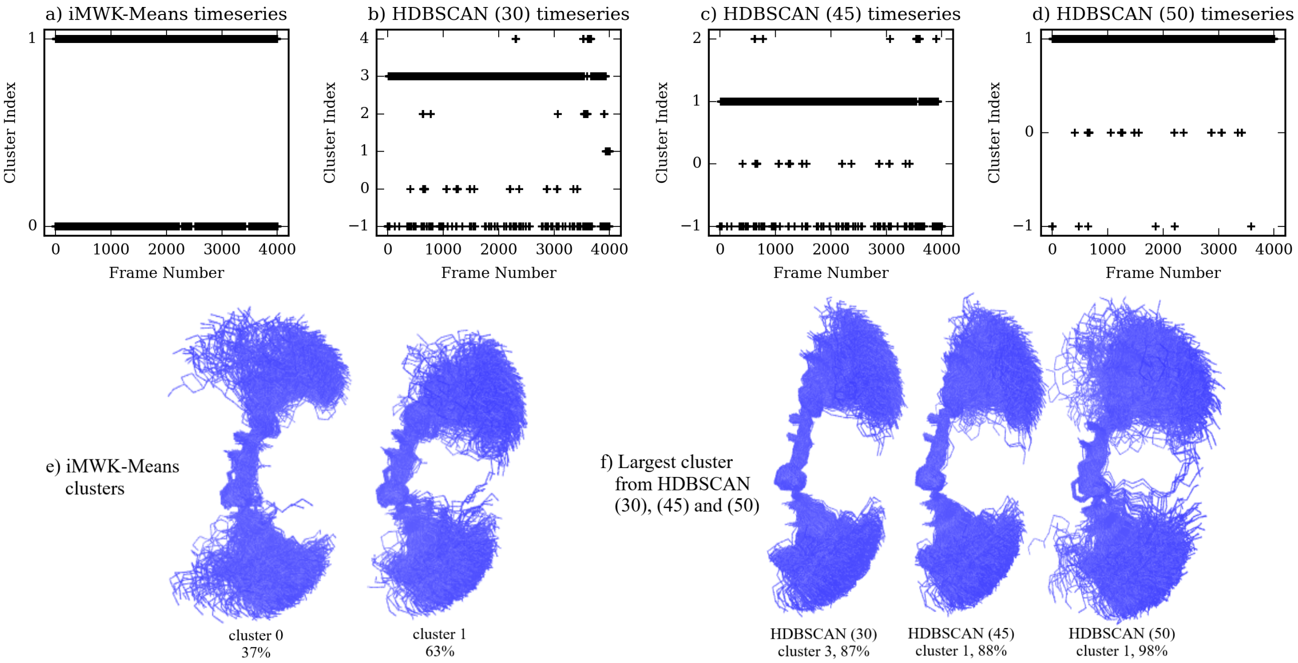

Figure 3: MenY cluster results Figurs a) to d) show the timeseries of the cluster results for iMWK-Means and runs of HDBSCAN with *minimum cluster size* values of 30, 45 and 50 - indicated by the numbers in brackets. The (-1) cluster indicates frames that have been classified as noise. Figure e) shows the two MenY clusters produced by iMWK-Means. The cluster indices are given, as they are labelled in a), as well as the percentage of frames that were assigned to the cluster. Figure f) shows the largest cluster produced by each of the HDBSCAN runs. For all visualisation of the clusters, every 2nd frame in the cluster in superimposed, and, for clarity, the atoms selected for visualisation are the same as those that were clustered on.

HDBSCAN identifies a single large cluster in addition to varying proportions of frames labelled as noise (Figure 2a) and other much smaller clusters. A *minimum cluster size* of 50 labels almost no frames as noise but places 98% of frames into one cluster. A *minimum cluster size* value of 30 or 45 labels just under 10% of frames as noise (Figure 2a) with the most of the remaining frames being assigned to a single cluster.

Visualisations of the largest MenY cluster produced by HDBSCAN for the three values, 30, 45 and 50, of *minimum cluster size* are shown in Figure 3f. The clusters from HDBSCAN with *minimum cluster size* values of 30 and 45, which account for 87% and 88% of frames respectively, exhibit relatively low intracluster variation. This is evident from the fact that the superimposed frames are tightly collected together, with only a few frames straying from the cluster. It is clear that the ends of the molecule are more mobile, based of the fact that the superimposed frames fan out at each end. Contrastingly, the cluster produced by HDBSCAN with a *minimum cluster size* value of 50, which contains 98% of the frames, evidently has higher intracluster variation. The superimposed frames are not as heavily concentrated and more frames are straying from the main part of the cluster. This difference is somewhat expected, and can be attributed to the fact that the *minimum cluster size* 50 cluster contains the majority of the frames that were labelled as noise when *minimum cluster size* was set to 30 or 45.

iMWK-Means splits the MenY trajectory into two large and frequently alternating clusters (Figure 3a). Visualisation of these two clusters (Figure 3e) shows that they each have slightly more intracluster variation than the HDBSCAN *minimum cluster size* 30 and 45 clusters, but not as much as the HDBSCAN *minimum cluster size* 50 cluster. There a some frames in each cluster which stray from the concentrated regions, but, as iMWK-Means does not label and exclude frames it classifies as noise, we should expect this to some extent. The conformations represented by the two clusters are reasonably similar, however, the frames in cluster 1 curve inwards to a greater extent than those in cluster two. This suggests that iMWK-Means has identified a slight difference in conformation that was not identified by HDBSCAN.

5.2.2 MenW. Non-summarised results for the HDBSCAN applications to MenW with varying *minimum cluster size* values are available in Supplementary Material Table S4. Once again, the DB index displayed a strong preference for results with high cluster counts which was contradicted by the S and CH indices. The three best CH index results, indicated by the largest values, are associated with *minimum cluster size* values of 70, 75 and 95. These values also produced S values which were amongst the best. The other acceptable S value, from a *minimum cluster size* value of 85, was disregarded as it coincided with a local minimum of the C index and a local maximum for the DB index. Based on this, we focused on the HDBSCAN results with *minimum cluster size* values of 70, 75 and 95, and these are compared, along with the iMWK-Means result, in Figure 2e-h.

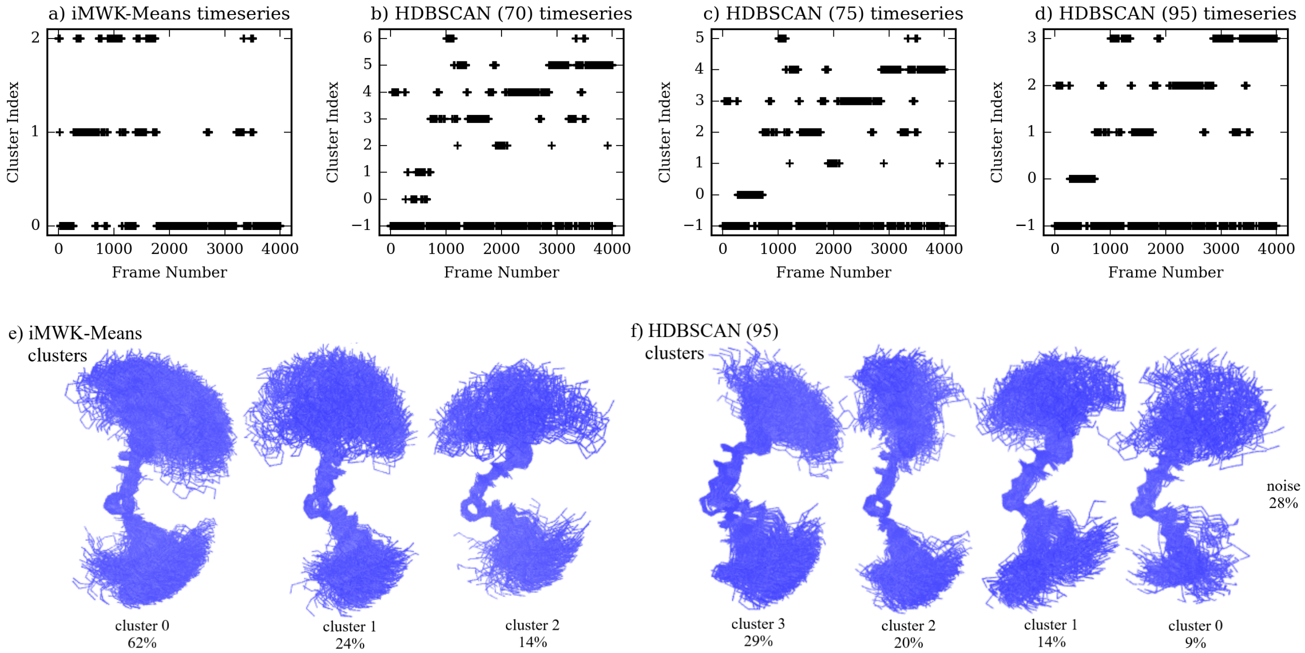

Figure 4: MenW cluster results Figurs a) to d) show the timeseries of the cluster results for iMWK-Means and runs of HDBSCAN with *minimum cluster size* values of 70, 75 and 95 - indicated by the numbers in brackets. The (-1) cluster indicates frames that have been classified as noise. Figure e) shows the three MenW clusters produced by iMWK-Means. The cluster indices are given, as they are labelled in a), as well as the percentage of frames that were assigned to the cluster. Figure f) shows the four clusters produced by HDBSCAN with a *minimum cluster size* value of 95 which corresponds to timeseries d). For all visualisation of the clusters, every 2nd frame in the cluster in superimposed, and, for clarity, the atoms selected for visualisation are the same as those that were clustered on.

iMWK-Means produced three clusters for MenW and is the best result according to all three CVIs, indicated by the the highest S score (Figure 2f), the lowest DB score (Figure 2g) and the highest CH score (Figure 2h). Visualisation of the three clusters shows that represent reasonably different conformations (Figure 4e). Cluster 2 is the most distinct and shows a clear bend in the molecule compared to clusters 0 and 1, which are straighter. The difference between clusters 0 and 1 is less obvious, but it does appear that the bottom half of the molecule tilts upwards to a greater degree in cluster 1 than in cluster 0. Considering the visualised clusters with the timeseries, it seems that the three clusters alternate for approximately the first 1800 frames, after which point cluster 0 is the dominant conformation.

The HDBSCAN result which produced the best combination of CVI values uses a *minimum cluster size* of 95. It maximises, and therefore has the best results for, the S and CH indices (Figure 2f,h) but has the worst DB value (Figure 2g), although this may but due to the smaller cluster count. Visualiation of the four clusters produced by HDBSCAN with *minimum cluster size* 95 shows clearly distinct conformations (Figure 4f). Cluster 2 shows the molecule in its straightest form, while cluster 0 shows the most pronounced curve. Clusters 3 and 1 are the most similar, but cluster 1 appears to be slightly more bent and the top of the molecule tilts to the left rather than the right, as it does in cluster 3. Compared to the iMWK-Means clusters (Figure 4e), the HDBSCAN clusters have less intracluster variation. This is based on the fact that the top and bottom of the clusters fan out to a lesser extent, indicating that the superimposed frames are more similar.

Although the CVI values for the three HDBSCAN results are not as indicative of good clusters as the iMWK-Means CVI results are, this may be partly attributed to the noise cluster. The HDBSCAN results with *minimum cluster size* values of 70, 75 and 95 each classified between 27% and 30% of frames as noise (Figure 2e), which is the minimum for MenW when compared with the other HDBSCAN results which produce acceptable CVI values (Table S4). Due to the fact that HDBSCAN places the frames it classifies as noise into cluster -1, high levels of noise will negatively affect CVI results. When the CVI results are calculated, high levels of variation will be detected in the noise cluster, and this will negate the low levels of variation in the real clusters, resulting in poor CVI values.

Large proprotions of the data being classified as noise is also undesirable as it does not give a complete picture of the molecule conformations within the trajectory, as a substantial amount of the frames have not been assigned to a cluster. Although labelling frames as noise is useful in the case of outliers which would otherwise contribute unnecessarily to intracluster variation, ideally the number of frames labelled as such should be minimised.

5.2.3 *Shigella flexneri.*

The full results of the HDBSCAN applications to S.flexneri Y 3RU and S.flexneri 6 6RU for each *minimum cluster size* value are available in Supplementary Material Tables S5

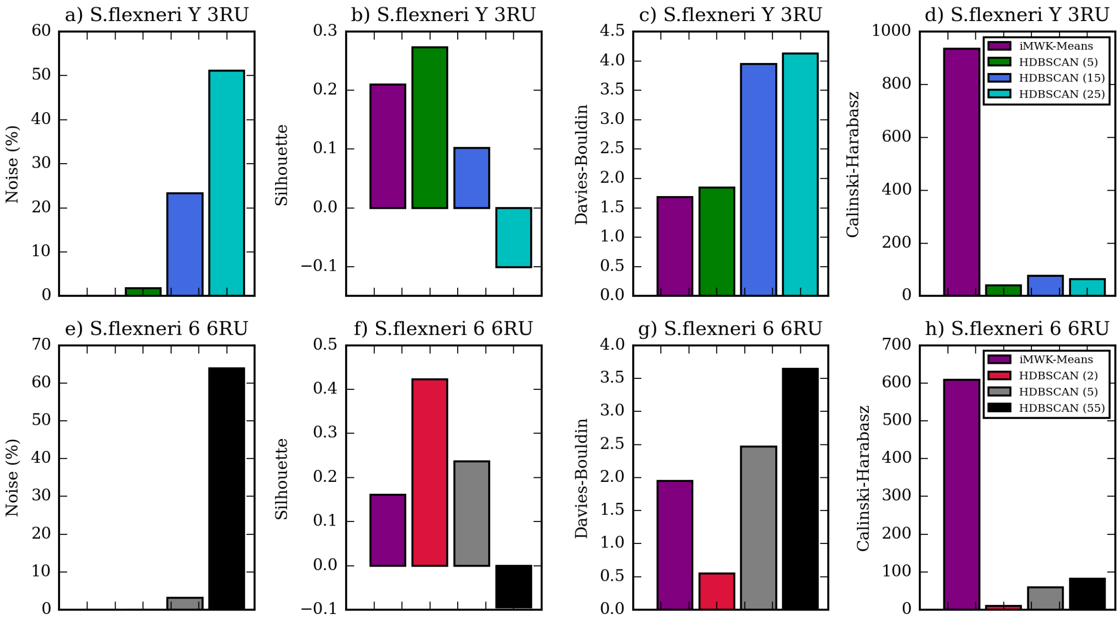

Figure 5: Percentage noise and CVI scores for Shigella cluster results Figures a) to d) compare S.flexneri Y 3RU clustering results (in terms of percentage noise (a), Silhouette index (b), Davies-Bouldin index (c) and Calinski-Harabasz index (d)) from the application of iMWK-Means and three applications of HDBSCAN with *minimum cluster size* values of 5, 15 and 25 - indicated by the number in brackets in the legend in d). Figures e) to h) are similar, but compare the S.flexneri 6 6RU clustering results and use HDBSCAN *minimum cluster size* values of 2, 5 and 55 - indicated by the number in brackets in the legend in h).

and S6. From the *minimum cluster size* values that were used, HDBSCAN does not appear to have produced any useful clustering results for either of the Shigella trajectories. Each HDBSCAN result includes a large cluster of varying size, some small and insignificant clusters and the remaining frames labelled as noise. Three example S.flexneri Y 3RU HDBSCAN timeseries are shown in Figure 6b-d to illustrate this. These three examples, associated with *minimum cluster size* values 5, 15 and 25, were chosen based on their CVI scores which were considered to be the best possible compromises. Similarly, example S.flexneri 6 6RU HDBSCAN timeseries, associated with *minimum cluster size* values 2, 5 and 55, are shown in Figure 7b-d, chosen for the same reason.

For S.flexneri Y 3RU and S.flexneri 6 6RU, iMWK-Means produced two similarly sized clusters which alternate frequently (Figure 6a and Figure 7a). Interestingly, the CH index shows an extreme preference for each of these results, and they score much higher values than any of the HDBSCAN results (Figure 5d,h). For S.flexneri Y 3RU, iMWK-Means also has the best DB score, indicated by the low value on Figure 5c, and the second highest S score (Figure 5b).

Visualisation of the iMWK-Means S.flexneri Y 3RU clusters shows that a lot of variance is contained within each cluster (Figure 6e). Particularly on the lower half of each molecule, the superimposed frames fan out to a large degree. This amount of variance should definitely constitute separate clusters in a ideal clustering result. There does appear to be a difference between these two clusters. The top half of cluster 0 tilts to the right, creating a bend in the molecule, while cluster 1 remains straight. Despite this distinction between the clusters, the level of variance which is not accounted for means that this clustering result has limited use. Based on this, the S.flexneri Y 3RU HDBSCAN clusters are not visualised. It is clear that the trajectory contains a lot of variance between frames, and as the HDBSCAN results each only produce one significant cluster, it is reasonable to assume that the cluster results will not be useful.

In the case of S.flexneri 6 6RU, iMWK-Means has the highest CH score (Figure 5h) and has the second best DB score, indicated by it being the second lowest value (Figure 5g). However, its S score is the second worst, with the only lower score dropping below zero (Figure 5f). Visualisation of the iMWK-Means S.flexneri 6 6RU clusters (Figure 7e) shows that they exhibit even more intracluster variation and no distinction between clusters compared to the S.flexneri Y 3RU clusters (Figure 6e). The bottom half of the superimposed frames fans out extensively, and where for S.flexneri Y 3RU some distinction between clusters was identified, here there appears to be no substantial difference between the two clusters. Clearly the trajectory contains a variety of conformations which have not been identified by iMWK-Means or HDBSCAN, as it never produces more than one significant cluster. As it is evident that a single cluster will have high levels of variation, the HDBSCAN results have not been visualised.

5.3 Discussion

HDBSCAN and iMWK-Means each proved able to produce useful cluster results for the MenY trajectory. HDBSCAN with a *minimum cluster size* value of 30 or 45 produced a only one significant cluster

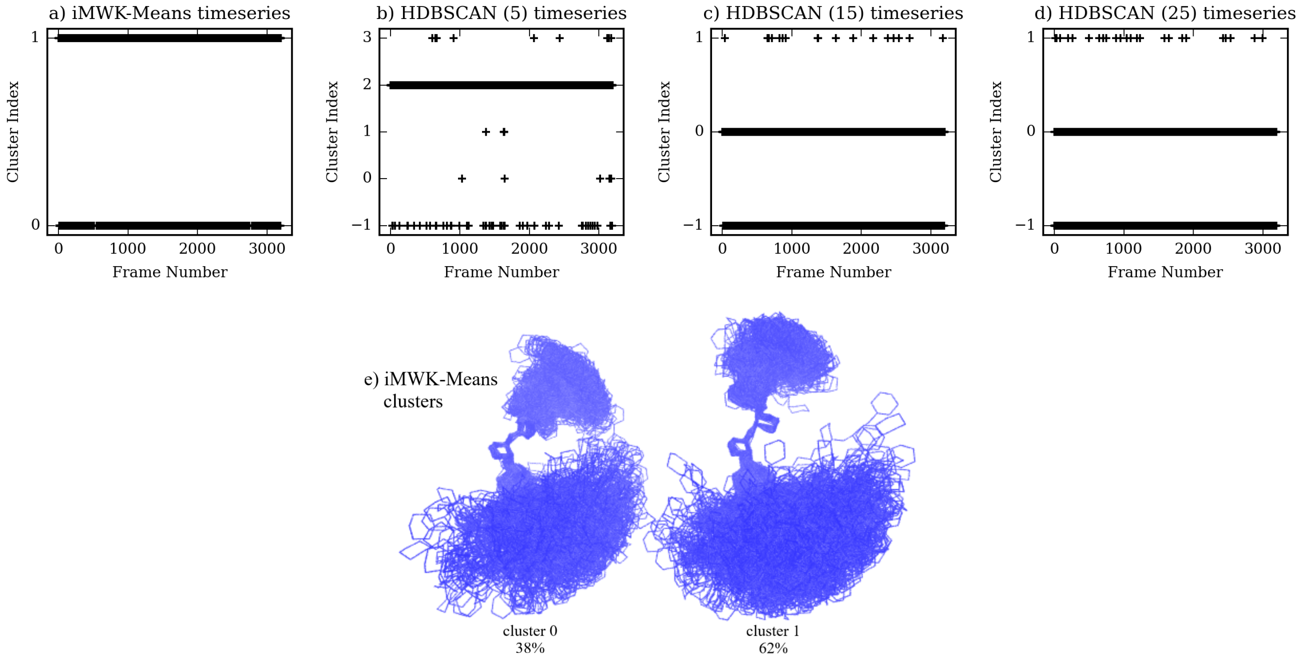

Figure 6: S.flexneri Y 3RU cluster results Figurs a) to d) show the timeseries of the cluster results for iMWK-Means and runs of HDBSCAN with *minimum cluster size* values of 5, 15 and 25 - indicated by the numbers in brackets. The (-1) cluster indicates frames that have been classified as noise. Figure e) shows the two MenY clusters produced by iMWK-Means. The cluster indices are given, as they are labelled in a), as well as the percentage of frames that were assigned to the cluster. For visualisation of the clusters, every 2nd frame in the cluster in superimposed, and, for clarity, the atoms selected for visualisation are the same as those that were clustered on.

accounting for 87-88% of frames while classifying the majority of the remaining frames as noise. The cluster has minimal intracluster variation (Figure 3f), due to the fact that dissimilar or outlier frames could be classified as noise, and clearly indicates that MenY primarily occupies a single conformation, which is consistent with previous analysis of MenY [10]. However, iMWK-Means was able to highlight a subtle difference in conformation by producing two MenY clusters which represented one conformation with slight curvature and another with less (Figure 3e).

HDBSCAN also produced useful clusters from the MenW trajectory, partitioning the data into four distinct conformation with low variance (Figure 4f). Although slightly less than 30% of frames were classified as noise, this is still a useful result. It indicates that there is a lot of variation with the system, while still locating stable conformations through well-formed clusters. Contrastingly, iMWK-Means produced three MenW clusters which were not as well-formed (Figure 4e) as they visually contain a lot more variation, although this is at least partially due to the fact that outliers cannot be excluded as noisy frames. The iMWK-Means result also represents a smaller number of distinct conformations than the HDBSCAN result, as it produced three clusters rather than four.

As expected, based on previous research [10], the MenW cluster results indicate that MenW occupies multiple conformations, one of which bares resemblance to the primary conformation of MenY. Comparing the dominant MenY cluster produced by HDBSCAN (Figure 3f), with a *minimum cluster size* value of 30 or 45, to the MenW clusters, it seems most similar to cluster 2 from the HDBSCAN result (Figure 4f) and cluster 0 from the iMWK-Means result (Figure 4e), as these conformations have the least curvature.

For a relatively stable molecule, such as MenY, it seems that HDBSCAN is able to indicate the stability while disregarding outliers, and iMWK-Means is able to locate detailed differences in conformation. When clustering a molecule such as MenW, which is flexible but contains a number of defined conformations, HDBSCAN will classify a reasonably large number of frames as noise but will still identify stable dominant conformations. This result is consistent with findings from Melvin et al., where the algorithms were applied to nucleic acids and proteins. It seems that HDBSCAN are iMWK-Means are useful as a pair, as iMWK-Means is able to pick out details that HDBSCAN may miss. Additionally, in cases where HDBSCAN classifies many frames as noise, the result can be considered in conjuction with the iMWK-Means clusters, which will not disregard frames as noise, to get a complete picture of the conformations of the molecule.

Although the algorithms achieve useful results with molecules with some flexibility, such as MenW, neither is suited to clustering trajectories of molecules with extreme flexibility, such as S.flexneri Y 3RU and S.flexneri 6 6RU. Besides a slight distinction between the two S.flexneri Y 3RU clusters produced by iMWK-Means (Figure 6e), very little was gained by clustering these trajectories with iMWK-Means and HDBSCAN. It is clear from the visualisation of the clusters, in Figures 6e and 7e, as well as previous research [7], that

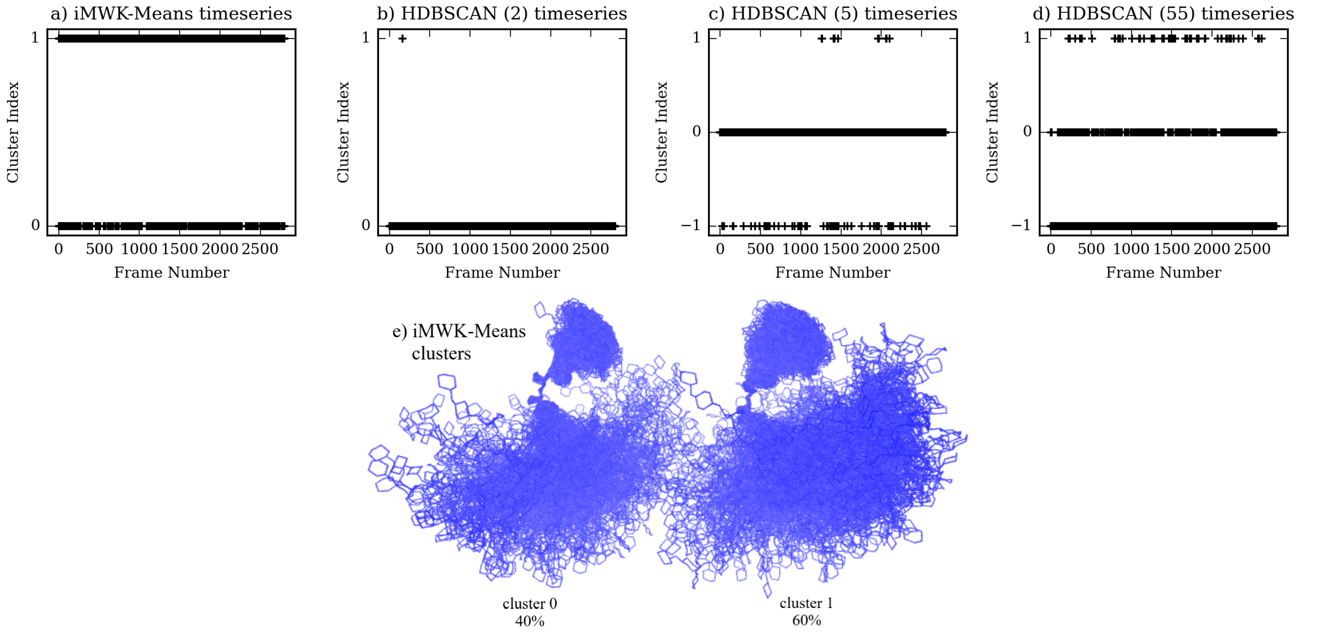

Figure 7: S.flexneri 6 6RU cluster results Figurs a) to d) show the timeseries of the cluster results for iMWK-Means and runs of HDBSCAN with *minimum cluster size* values of 2, 5 and 55 - indicated by the numbers in brackets. The (-1) cluster indicates frames that have been classified as noise. Figure e) shows the two MenY clusters produced by iMWK-Means. The cluster indices are given, as they are labelled in a), as well as the percentage of frames that were assigned to the cluster. For visualisation of the clusters, every 2nd frame in the cluster in superimposed, and, for clarity, the atoms selected for visualisation are the same as those that were clustered on.

there a many dominant conformations within the trajectories than the algorithms we unable to isolate.

6 CONCLUSIONS

The HDBSCAN and iMWK-Means were applied to trajectories of carbohydrates with varying levels of flexibility. Both have been shown to produce useful clusters from trajectories of relatively stable molecules and molecules with some flexibility. For stable molecules, HDBSCAN is able to isolate the dominant conformation while detecting outliers and classifying them as noise. In these cases, iMWK-Means is able to detect finer details in otherwise stable trajectories. HDBSCAN and iMWK-Means can also be used together when clustering more flexible trajectories, as HDBSCAN can isolate the dominant conformations as well-formed clusters but will label a reasonably large proportion of frames as noise. As iMWK-Means does not label frames as noise, it can then be used to account for some of the frames that were disregarded by HDBSCAN. Unfortunately, neither algorithm is suitable for clustering trajectories of molecules with extreme flexibility.

iMWK-Means could be subjected to further testing with a greater variety of carbohydrates molecules, as well as with molecules that have not be downsampled prior to clustering. This would produce additional insights into the suitability of the algorithms for clustering trajectories of flexible molecules.

ACKNOWLEDGMENTS

I would like to thank my project partners, Nicholas Limbert and Wen Kang Lu, for their contributions to the joint framework. I also would like to thank Nicole Richardson, who provided the Shigella flexneri trajectory files. Finally, I would like to thank my supervisor, Associate Professor Michelle Kuttel, for providing valuable guidance throughout the project, as well as the trajectory files for MenW and MenY. This research was funded in part by the National Research Foundation of South Africa.

REFERENCES

[1] Charu C. Aggarwal and Chandan K. Reddy. 2013. *DATA CLUSTERING Algorithms and Applications.* CRC Press/Taylor and Francis Group, University of Minnesota, Minneapolis, Minnesota.

[2] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez, and Iñigo 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

[3] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*, Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 160–172.

[4] Ricardo J. G. B Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. 2015. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10, 1 (2015), 1–51.

[5] Renato Cordeiro de Amorim and Boris Mirkin. 2012. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. *Pattern recognition* 45, 3 (2012), 1061–1075.

[6] Renato Cordeiro de Amorim and Christian Hennig. 2015. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences* 324 (2015), 126–145. https://doi.org/10.1016/j.ins.2015.06.039

[7] Jason Hlozek, Neil Ravenscroft, and Michelle M Kuttel. 2020. Effects of Glucosylation and O-Acetylation on the Conformation of Shigella flexneri Serogroup 2 O-Antigen Vaccine Targets. *The journal of physical chemistry. B* 124, 14 (2020), 2806–2814.

[8] William Humphrey, Andrew Dalke, and Klaus Schulten. 1996. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* 14 (1996), 33–38.

[9] Pablo Jaskowiak, Davoud Moulavi, Antonio Furtado, Ricardo Campello, Arthur Zimek, and Jörg 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

[10] Michelle M Kuttel, Zaheer Timol, and Neil Ravenscroft. 2017. Cross-protection in Neisseria meningitidis serogroups Y and W polysaccharides: A comparative conformational analysis. 446-447 (2017), 40–47.

[11] Robert T Mcgibbon, Kyle A Beauchamp, Matthew P Harrigan, Christoph Klein, Jason M Swails, Carlos X Hernández, Christian R Schwantes, Lee-Ping Wang, Thomas J Lane, and Vijay S Pande. 2015. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical journal* 109, 8 (2015), 1528–1532.

[12] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2, 11 (mar 2017). https://doi.org/10.21105/joss.00205

[13] Ryan L. Melvin, Ryan C. Godwin, Jiajie Xiao, William G. Thompson, Kenneth S. Berenhaut, and Freddie R. 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

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[15] Jianyin Shao, Stephen W. Tanner, Nephi Thompson, and Thomas E. 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

Supplementary Material

A ALGORITHM VALIDATION RESULTS

Table S1: iMWK-Means cluster results for validation data sets

Data set	Actual clusters	Assigned clusters	Accuracy (%)	Silhouette	Davies-Bouldin	Calinski-Harabasz
A	[20, 20, 20, 20, 20]	[20, 20, 20, 40]	80.0	0.744	0.363	714.793
B	[20, 20, 20, 20, 20]	[69, 31]	40.0	0.614	0.531	255.076
Breast Cancer	[212, 357]	[181, 388]	92.79	0.565	0.676	744.595
Iris	[50, 50, 50]	[50, 100]	66.67	0.687	0.383	502.822
Wine	[59, 71, 48]	[67, 57, 54]	92.13	0.182	1.245	150.636

Table S2: HDBSCAN cluster results for validation data sets

Data set	Minimum cluster size	Minimum samples	Actual clusters	Assigned clusters	Accuracy (%)	Silhouette	Davies-Bouldin	Calinski-Harabasz
A	5	1	[20, 20, 20, 20, 20]	[20, 20, 20, 20, 20]	100.0	0.873	0.177	4863.546
B	6	1	[20, 20, 20, 20, 20]	[-19, 20, 22, 18, 13, 8]	65.0	0.363	0.985	103.918
B	9	1	[20, 20, 20, 20, 20]	[-19, 20, 22, 18, 21]	68.0	0.371	1.056	120.233
B	10	1	[20, 20, 20, 20, 20]	[-14, 20, 22, 44]	57.0	0.390	1.021	125.064
Breast Cancer	30	1	[212, 357]	[-63, 38, 468]	69.42	0.541	1.536	563.081
Iris	5	1	[50, 50, 50]	[50, 100]	66.67	0.687	0.383	502.822
Wine	15	1	[59, 71, 48]	[-8, 40, 130]	60.67	0.598	0.686	286.916

B HDBSCAN RESULTS

Table S3: MenY: HDBSCAN cluster results with varied Minimum cluster size

Minimum cluster size	Minimum samples	Cluster count	Largest cluster (%)	Noise (%)	Silhouette	Davies-Bouldin	Calinski-Harabasz
2	1	810	0.35	42.24	-0.186	1.453	5.461
5	1	23	85.49	7.14	-0.026	1.546	61.952
10	1	8	85.49	8.37	0.015	2.002	144.88
15	1	7	85.49	8.34	0.053	2.050	164.252
20	1	6	85.49	8.79	0.069	2.047	187.236
25	1	6	85.49	8.79	0.069	2.047	187.236
30	1	5	86.91	8.04	0.275	2.049	221.314
35	1	5	86.91	8.04	0.275	2.049	221.314
40	1	5	86.91	8.04	0.275	2.049	221.314
45	1	3	88.31	8.82	0.283	2.331	274.272
50	1	2	98.00	0.35	0.511	2.062	253.254
55	1	2	98.00	0.35	0.511	2.062	253.254
60	1	2	98.00	0.35	0.511	2.062	253.254
65	1	2	98.00	0.35	0.511	2.062	253.254
70	1	-	-	100.00	-	-	-

Table S4: MenW: HDBSCAN cluster results with varied Minimum cluster size

Minimum cluster size	Minimum samples	Cluster count	Largest cluster (%)	Noise (%)	Silhouette	Davies-Bouldin	Calinski-Harabasz
2	1	894	0.52	31.98	-0.100	1.267	6.850
5	1	2	97.18	2.40	0.030	3.996	20.826
10	1	38	26.51	36.00	-0.097	1.607	101.062
15	1	27	26.51	38.52	-0.072	1.645	128.682
20	1	21	26.51	39.60	-0.067	1.652	154.301
25	1	14	26.51	32.48	-0.024	1.817	227.096
30	1	14	26.51	32.48	-0.024	1.817	227.096
35	1	12	26.51	32.70	-0.016	1.835	259.043
40	1	12	26.51	32.70	-0.016	1.835	259.043
45	1	12	26.51	32.70	-0.016	1.835	259.043
50	1	12	26.51	32.70	-0.016	1.835	259.043
55	1	10	26.51	29.83	-0.020	1.939	287.250
60	1	9	26.51	29.58	-0.009	2.020	304.916
65	1	8	26.51	28.03	-0.004	2.094	334.179
70	1	7	26.51	29.73	0.026	2.064	370.715
75	1	6	26.51	26.87	0.043	2.241	444.451
80	1	6	26.51	26.87	0.043	2.241	444.451
85	1	2	83.69	7.27	0.111	3.756	254.616
90	1	2	83.69	7.27	0.111	3.756	254.616
95	1	4	28.55	28.93	0.116	2.403	568.230
100	1	4	28.55	28.93	0.116	2.403	568.230

Table S5: S.flexneri Y 3RU: HDBSCAN cluster results with varied Minimum cluster size

Minimum cluster size	Minimum samples	Cluster count	Largest cluster (%)	Noise (%)	Silhouette	Davies-Bouldin	Calinski-Harabasz
2	1	672	0.34	40.67	-0.163	1.465	4.704
5	1	4	97.50	1.81	0.273	1.842	39.693
10	1	3	76.09	22.95	0.013	3.275	56.982
15	1	2	76.09	23.29	0.102	3.945	77.345
20	1	2	76.09	23.29	0.102	3.945	77.345
25	1	2	48.14	51.08	-0.101	4.130	64.005
30	1	3	10.91	84.37	-0.254	3.359	37.760
35	1	4	5.38	88.31	-0.291	3.047	29.188
40	1	4	5.38	88.31	-0.291	3.047	29.188
45	1	3	10.91	84.37	-0.254	3.359	37.760
50	1	3	10.91	84.37	-0.254	3.359	37.760
55	1	3	10.91	84.37	-0.254	3.359	37.760
60	1	3	10.91	84.37	-0.254	3.359	37.760
65	1	2	10.91	86.25	-0.188	3.591	46.187
70	1	2	10.91	86.25	-0.188	3.591	46.187
75	1	2	10.91	86.25	-0.188	3.591	46.187
80	1	2	10.91	86.25	-0.188	3.591	46.187
85	1	2	10.91	86.25	-0.188	3.591	46.187
90	1	2	10.91	86.25	-0.188	3.591	46.187
95	1	-	-	100.00	-	-	-

Table S6: S.flexneri 6 6RU: HDBSCAN cluster results with varied Minimum cluster size

Minimum cluster size	Minimum samples	Cluster count	Largest cluster (%)	Noise (%)	Silhouette	Davies-Bouldin	Calinski-Harabasz
2	1	2	99.93	0.00	0.423	0.547	9.822
5	1	2	96.14	3.22	0.237	2.465	59.032
10	1	2	96.14	3.22	0.237	2.465	59.032
15	1	2	96.14	3.22	0.237	2.465	59.032
20	1	5	31.69	62.63	-0.213	2.495	55.089
25	1	4	31.69	62.63	-0.173	2.744	65.457
30	1	4	31.69	62.63	-0.173	2.744	65.457
35	1	3	33.30	62.09	-0.130	3.183	75.795
40	1	3	33.30	62.09	-0.130	3.183	75.795
45	1	3	33.30	62.09	-0.130	3.183	75.795
50	1	3	33.30	62.09	-0.130	3.183	75.795
55	1	2	33.30	63.88	-0.095	3.643	82.101
60	1	2	33.30	63.88	-0.095	3.643	82.101
65	1	2	33.30	63.88	-0.095	3.643	82.101
70	1	2	33.30	63.88	-0.095	3.643	82.101
75	1	2	33.30	63.88	-0.095	3.643	82.101
80	1	2	5.72	88.92	-0.247	2.789	45.388
85	1	2	5.72	88.92	-0.247	2.789	45.388
90	1	2	5.72	88.92	-0.247	2.789	45.388
95	1	2	5.72	88.92	-0.247	2.789	45.388
100	1	2	5.72	88.92	-0.247	2.789	45.388

C SELECTION STATEMENTS

Table S7: Atom selection statements

Trajectory	Selection statement
MenY, MenW	type != H and (((resname AGL or resname AGA) and not (name O2 or name O3 or name O4)) or (resname ASI and (name O4 or name C2 or name C3 or name C4 or name C5 or name O6))) and not resid 0 1 10 11
S.flexneri Y 3RU	name C1 C2 C3 C4 C5 O2 O3 O5 and not name NH N CT C O SOD and not resid 0 11 and not index 37 58 78 124 145 165 211 223 232
S.flexneri 6 6RU	name C1 C2 C3 C4 C5 O4 O2 O3 O5 and not name NH N CT C O SOD and not resid 0 1 2 3 20 21 22 23 and not index 107 130 147 143 126 193 163 167 216 212 233 229 249 253 279 298 302 319 339 335 315 365 384 388 401 405 425 421 412