HDBSCAN clustering for identification of states with lignin dimers bound to β-cyclodextrin from molecular dynamics simulations

Brian Novak

Table of contents

Abstract

Cyclodextrins might be useful for separating lignin compounds due to their unique structural features. Unbiased molecular dynamics simulations of three lignin dimers or dimer derivatives were conducted to investigate dimer-cyclodextrin bound states using three carefully chosen yet relatively simple collective variables (CVs) and HDBSCAN clustering. All configuration types with lignin dimers bound to the cyclodextrin center which were visually observed in the trajectories were separated into distinct clusters or groups of cluster. That is in constrast with the original analysis which used many CVs, PCA to reduce dimensionality, and DBSCAN for clustering where it was not possible to separate some of the configuration types into distinct clusters. For the one dimer where the configuration types were correctly separated into distinct clusters, the proportions of configurations of each type were very similar to the proportions computed in this work. Compared to the original analysis, use of only three simple CVs improved interpretability and use of HDBSCAN made separation of configuration types into different clusters easier.

Introduction

Lignin, one of the most abundant biopolymers on Earth, occurs mainly in plant cell walls.[1] Lignin is biodegradable and biocompatible[2] while still being relatively stable which makes it attractive for research in pharmacological and biomedical applications. It consists of three types of monomers, which link with each other in various ways, producing a complex, branched structure. To extract specific compounds from lignin, it is necessary to first break down the lignin and then isolate the desired compounds from a mixture containing numerous compounds.

Cyclodextrins are a promising class of molecules for the separation of lignin compounds. Cyclodextrins are cyclic, cone-shaped oligosaccharides featuring an internal hydrophobic cavity capable of encapsulating guest molecules and a hydrophilic exterior which ensures water solubility and enhances the stability of guest-cyclodextrin complexes compared to the unbound guest and cyclodextrin.[3,4] Cyclodextrins are employed as selective adsorbents in various fields, including agriculture, food[4], pharmaceuticals[4], and biotechnology.

In our previous research[5], we investigated the interactions between β-cyclodextrin and lignin dimer derivatives in aqueous solution using a combination of experimental techniques, molecular dynamics simulations with GROMACS 2018.3[6,7], and docking with AutoDock Vina[8–10]. The chemical structures are illustrated in Figure 1. A summary of the publication[11] is available. During the molecular dynamics simulations, we

observed multiple types of dimer-cyclodextrin bound states. We estimated the proportions of those states in unbiased simulations using the following procedure:

1. Computed a large number of collective variables including angles between the lignin dimer and β-cyclodextrin principal axes, distances between atoms in the β-cyclodextrin molecule to atoms in the lignin dimer molecule, and lignin dimer dihedral angles
2. Applied principal component analysis (PCA)[12,13] to reduce the number of dimensions to two
3. Clustered the trajectory configurations using Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[14,15] to attempt to separate the different states into distinct clusters
4. Counted the number of points in each cluster corresponding to bound states and computed proportions of each one

The procedure details can be found in the Supporting Information for the previous study[5] on pages 9-18. The primary challenge was encountered in step 3, where the states were not distinctly separated. Trial and error was necessary to select DBSCAN parameters that successfully differentiated the clusters.

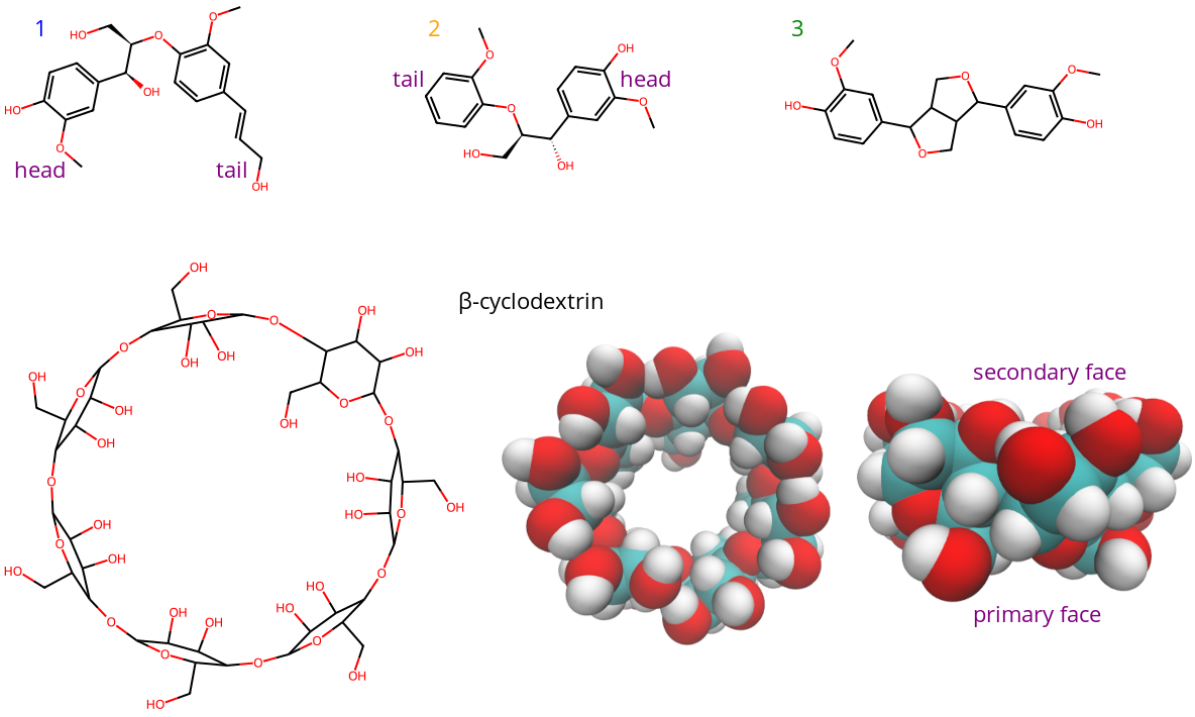

Figure 1: Structures of lignin dimer derivatives (1[16], 2[17], 3[18]), and β-cyclodextrin[19]. The 3D representations in the lower right show β-cyclodextrin from top and side views, with hydrogen atoms in white, carbon atoms in cyan, and oxygen atoms in red. The face of β-cyclodextrin with two hydroxyl (-OH) groups per unit (seen in the top view) is referred to as the secondary face, while the other face is referred to as the primary face. RDKit[20] 2023.03.3 was used for the 2D representations and VMD[21] 1.9.3[22] was used to create the 3D representations.

In this study, three carefully selected collective variables were utilized to distinguish the observed states, which eliminated the need for dimensionality reduction to be able to visualize clusters and improved interpretability. Additionally, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[23–25] was used instead of DBSCAN for clustering which improved the ability to separate states into different clusters. Finding the point where the number of clusters as a function of the `min_cluster_size` = `min_samples` parameter for HDBSCAN started to never increase provided a way to identify a "best" number of clusters (see Determination of `min_cluster_size` for details on the algorithm). Additionally, this eliminated the trial and error that was previously required to separate the clusters using PCA and DBSCAN.[5] The analysis uncovered configuration types with the lignin dimer bound to the cyclodextrin center that were not evident through mere visual inspection of the trajectories. However, these were subsets of the visually observed types, and only one cluster was dominant for each group of clusters corresponding to the visually

observed types. In the previous study[5], some configuration types that were observed visually for dimers 1 and 2 could not be split into separate clusters. Therefore, the current analysis also improved upon the previous study by providing a more complete characterization of the configurations.

Methods

Collective variables

Three collective variables were defined based on important configurations seen in unbiased simulation trajectories. The collective variables were computed using PLUMED[26] 2.6.6[27].

Normal distance

The first collective variable was a signed distance from the cyclodextrin to the lignin dimer along the direction of a vector pointing from the cyclodextrin center through the secondary face:

$$d_n = \frac{\overrightarrow{CL} \cdot \overrightarrow{CS}}{\left|\overrightarrow{CS}\right|} \tag{1}$$

In Equation 1, $\overrightarrow{CL}$ is the vector pointing from the cyclodextrin center of mass (COM) to the lignin COM and $\overrightarrow{CS}$ is the vector pointing from the cyclodextrin COM to the center of the hydroxyl oxygen atoms in the cyclodextrin secondary face.

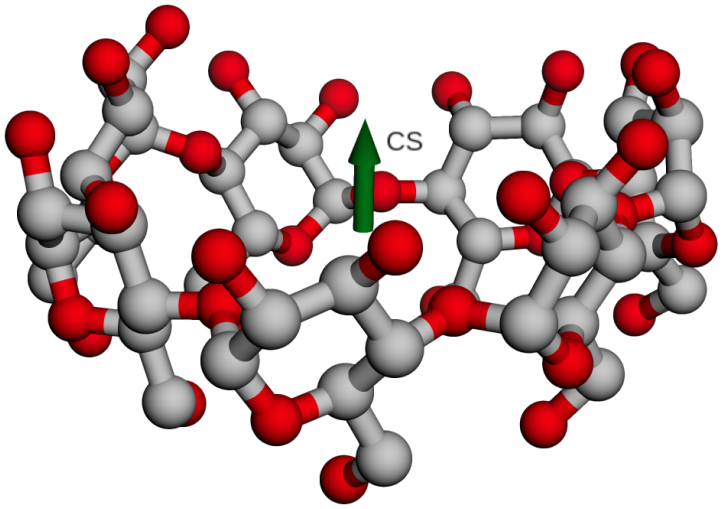

Figure 2: $\overrightarrow{CS}$ represented as a green arrow. In the cyclodextrin structure, gray atoms are carbon, red atoms are oxygen, and hydrogen atoms are not shown.

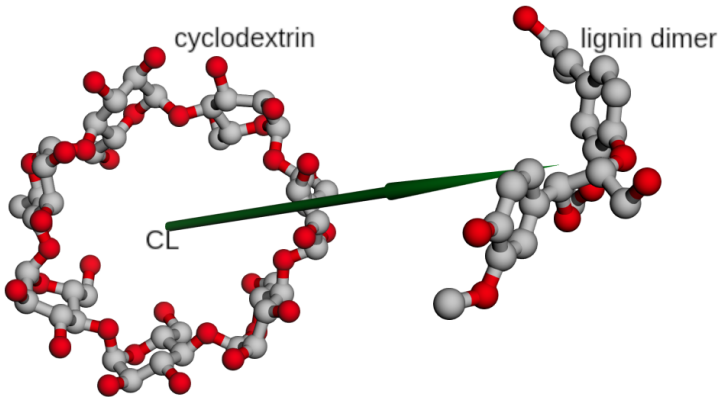

Figure 3: $\overrightarrow{CL}$ represented as a green arrow. In the molecular structures, gray atoms are carbon, red atoms are oxygen, and hydrogen atoms are not shown.

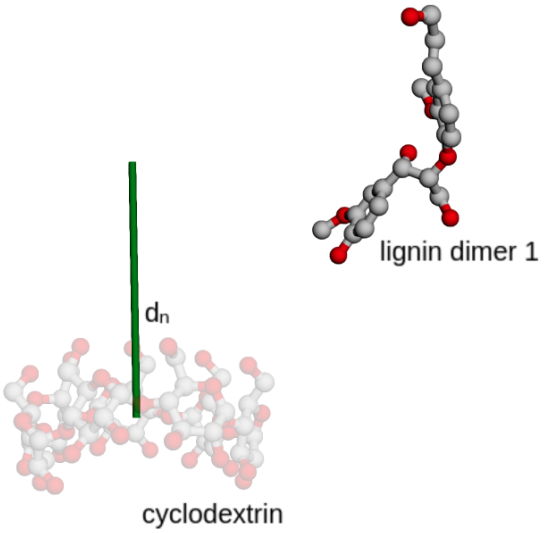

Figure 4: The distance from the cyclodextrin COM to the lignin dimer COM in the direction of $\overrightarrow{CS}$ defined in Equation 1, d_n, represented as a green cylinder. In the molecular structures, gray atoms are carbon, red atoms are oxygen, cyclodextrin is translucent, and hydrogen atoms are not shown.

Tangential distance

The second collective variable was the distance from the cyclodextrin COM to the lignin dimer COM in the direction perpendicular to $\overrightarrow{CS}$. It is the length of the second leg of a right triangle with hypotenuse $\left|\overrightarrow{CL}\right|$ and one leg d_n:

$$d_t = \sqrt{\left|\overrightarrow{CL}\right|^2 - d_n^2} \tag{2}$$

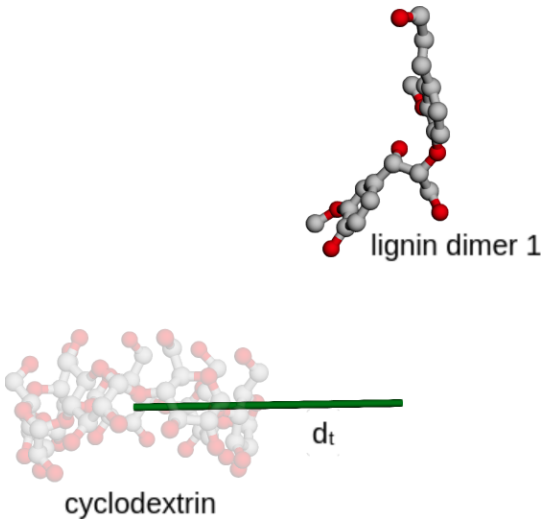

Figure 5: The distance from the cyclodextrin COM to the lignin dimer COM in the direction perpendicular to $\overrightarrow{CS}$ defined in Equation 2, d_t, represented as a green cylinder. In the molecular structures, gray atoms are carbon, red atoms are oxygen, cyclodextrin is translucent, and hydrogen atoms are not shown.

Relative orientation

The third collective variable was the cosine of the angle between the vector from the center of the ring atoms in the lignin head to the center of the ring atoms in the lignin tail ($\overrightarrow{HT}$) and $\overrightarrow{CS}$:

$$\cos\theta = \frac{\overrightarrow{HT} \cdot \overrightarrow{CS}}{\left|\overrightarrow{HT}\right| \left|\overrightarrow{CS}\right|} \tag{3}$$

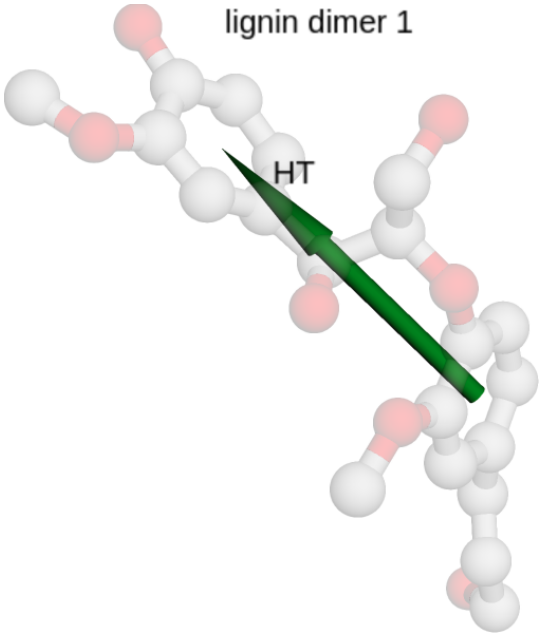

Figure 6: The lignin head to tail vector, $\overrightarrow{HT}$, represented as a green arrow. In the translucent lignin structure, gray atoms are carbon, red atoms are oxygen, and hydrogen atoms are not shown.

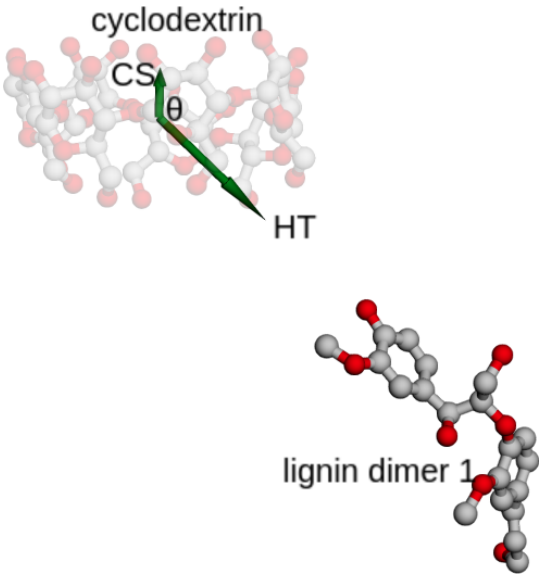

Figure 7: The angle θ between $\overrightarrow{HT}$ and $\overrightarrow{CS}$, represented as a green arrow. The origin of $\overrightarrow{HT}$ is shifted to the cyclodextrin COM. In the molecular structures, gray atoms are carbon, red atoms are oxygen, cyclodextrin is translucent, and hydrogen atoms are not shown.

HDBSCAN clustering

Determination of `min_cluster_size`

The best value of `min_cluster_size` = `min_samples` used in the HDBSCAN algorithm was determined using the following algorithm:

1. Find the last value of `min_cluster_size`, S, where the number of clusters corresponding to $S + 1$ is larger than the number of clusters correponding to S. The number of clusters corresponding to S is C.
2. Choose the smallest value of `min_cluster_size` > S where the number of clusters is equal to C.

Results

Number of clusters

The numbers of clusters as a function of `min_cluster_size` are shown in Figure 8, Figure 9, and Figure 10. The best values for `min_cluster_size` and the corresponding numbers of clusters were determined using the algorithm described in Determination of `min_cluster_size`.

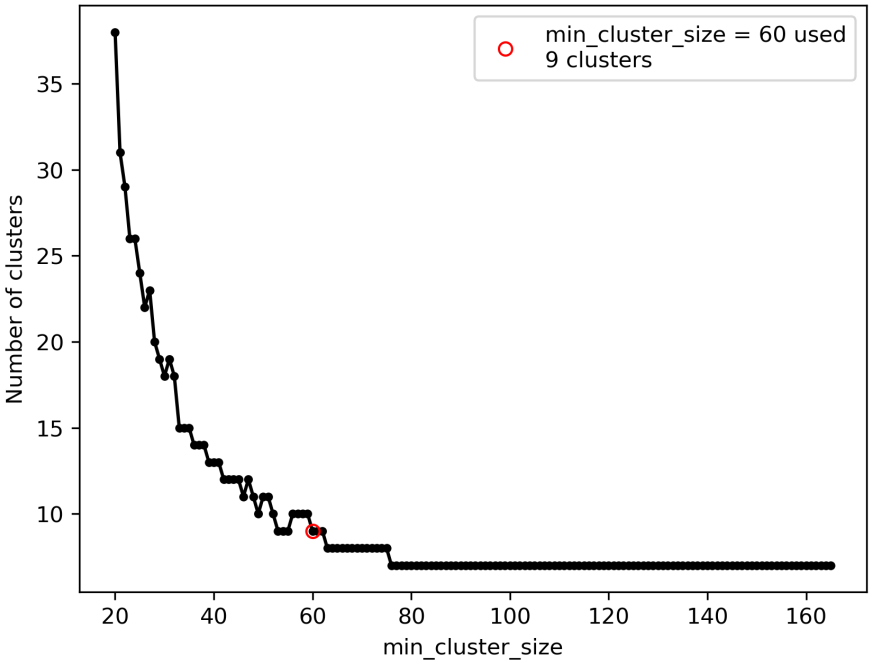

Figure 8: Number of clusters for dimer 1 as a function of `min_cluster_size` for the HDBSCAN algorithm. The best value for `min_cluster_size` and the corresponding number of clusters is circled in red and indicated in the legend.

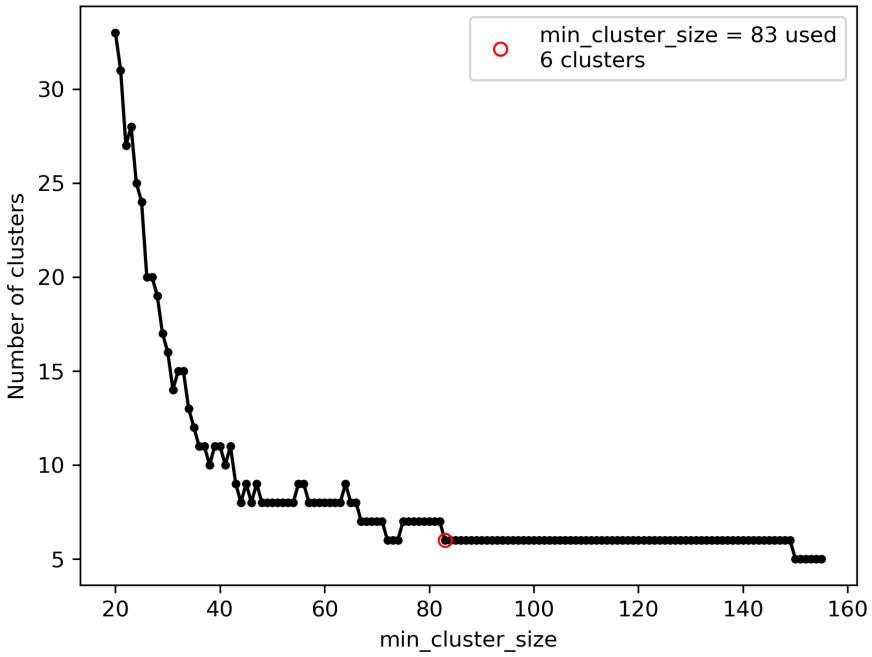

Figure 9: Number of clusters for dimer 2 as a function of `min_cluster_size` for the HDBSCAN algorithm. The best value for `min_cluster_size` and the corresponding number of clusters is circled in red and indicated in the legend.

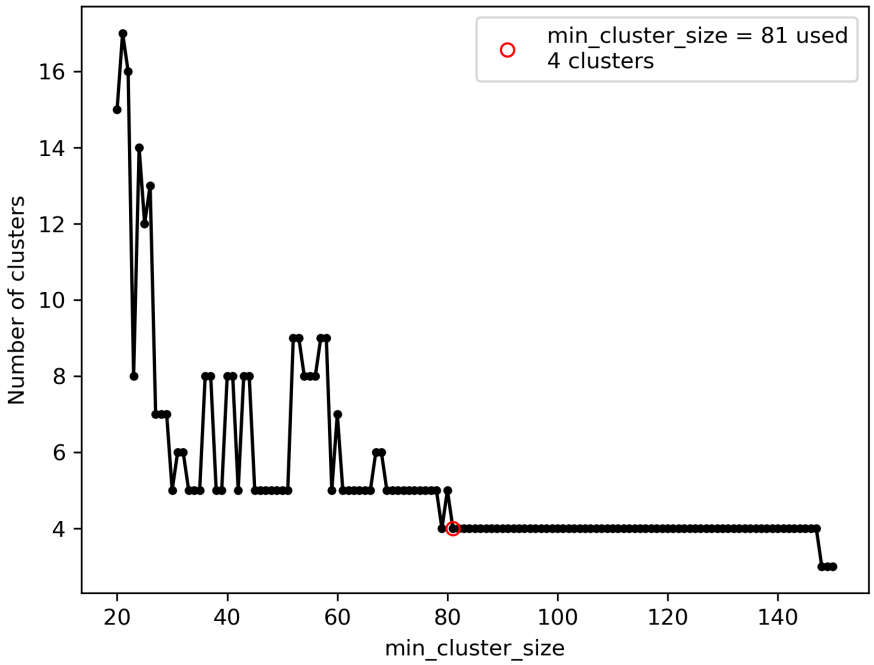

Figure 10: Number of clusters for dimer 3 as a function of `min_cluster_size` for the HDBSCAN algorithm. The best value for `min_cluster_size` and the corresponding number of clusters is circled in red and indicated in the legend.

Cluster scatter plots

The 3D scatter plots of the collective variables showing the HDBSCAN clusters for each dimer are shown in Figure 11, Figure 12, and Figure 13.

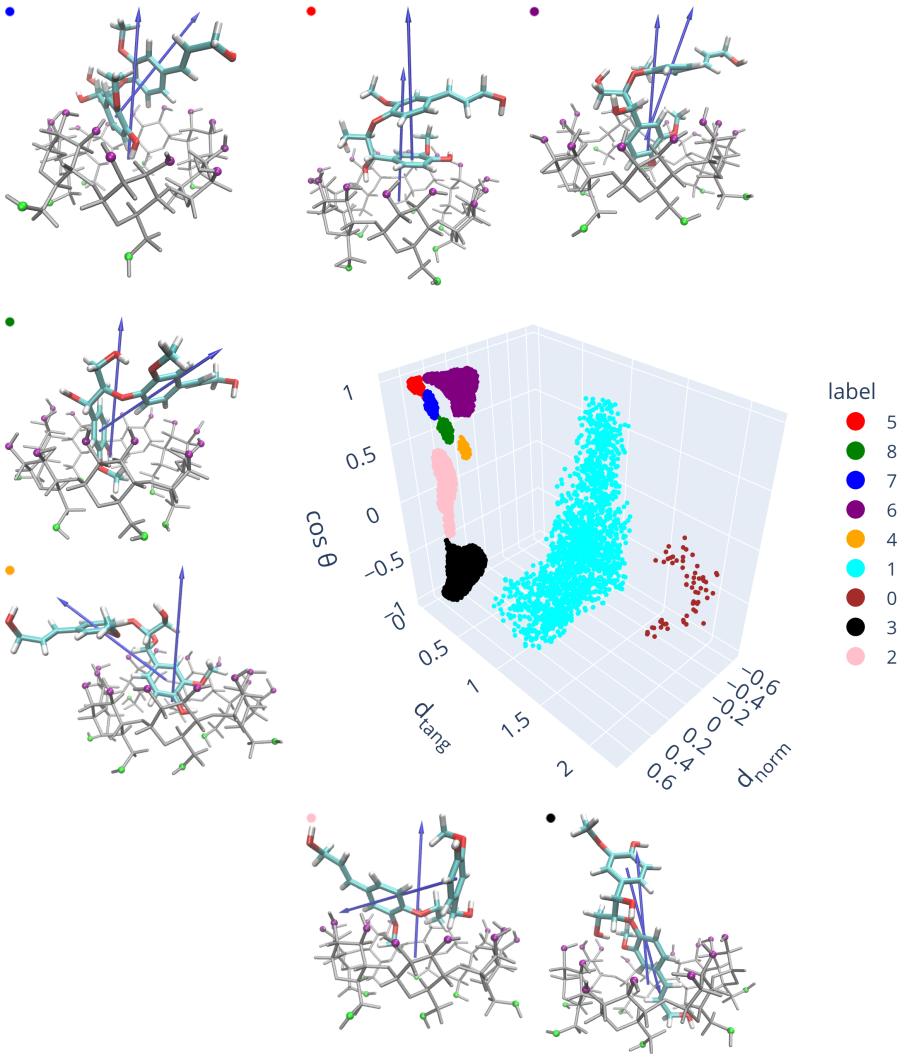

Figure 11: 3D scatter plot of the collective variables showing the HDBSCAN clusters for dimer 1. Representative configurations for each cluster are shown around the scatter plot except for the cyan and brown clusters which include configurations where the lignin dimer interacts (weakly) with the outside of the cyclodextrin ring or where there is no direct contact between the lignin dimer and cyclodextrin at all. The cyclodextrin atoms and bonds are gray except for the primary face oxygen atoms are green and the secondary face oxygen atoms are purple. Atom colors in the lignin dimer molecules: H=white, C=cyan, O=red. The arrows in the configurations point in the directions of $\overrightarrow{CS}$ and $\overrightarrow{HT}$. HDBSCAN outlier points are not plotted.

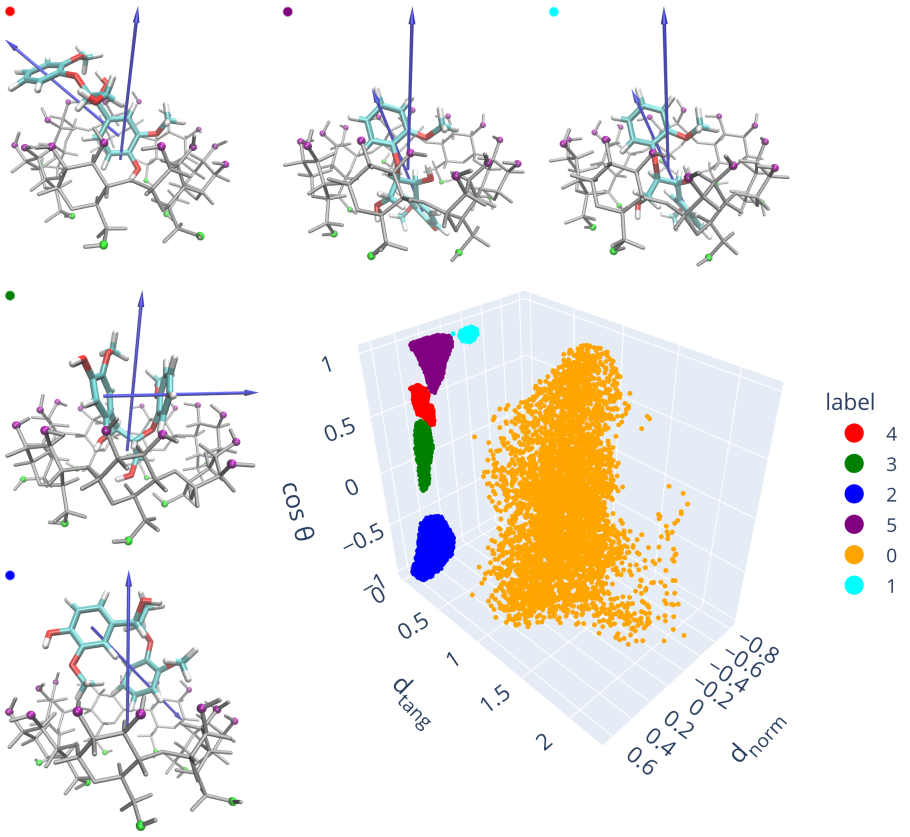

Figure 12: 3D scatter plot of the collective variables showing the HDBSCAN clusters for dimer 2. Representative configurations for each cluster are shown around the scatter plot except for the orange cluster which includes configurations where the lignin dimer interacts (weakly) with the outside of the cyclodextrin ring or where there is no direct contact between the lignin dimer and cyclodextrin at all. The cyclodextrin atoms and bonds are gray except for the primary face oxygen atoms are green and the secondary face oxygen atoms are purple. Atom colors in the lignin dimer molecules: H=white, C=cyan, O=red. The arrows in the configurations point in the directions of $\overrightarrow{CS}$ and $\overrightarrow{HT}$. HDBSCAN outlier points are not plotted.

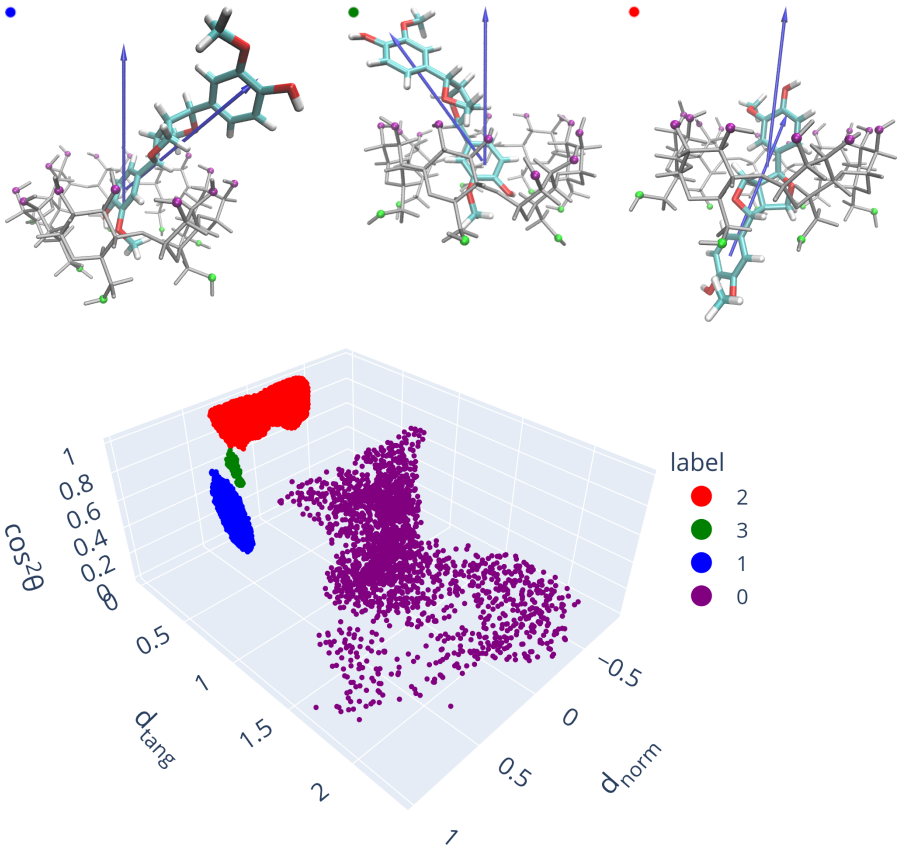

Figure 13: 3D scatter plot of the collective variables showing the HDBSCAN clusters for dimer 3. Representative configurations for each cluster are shown around the scatter plot except for the purple cluster which includes configurations where the lignin dimer interacts (weakly) with the outside of the cyclodextrin ring or where there is no direct contact between the lignin dimer and cyclodextrin at all. The cyclodextrin atoms and bonds are gray except for the primary face oxygen atoms are green and the secondary face oxygen atoms are purple. Atom colors in the lignin dimer molecules: H=white, C=cyan, O=red. The arrows in the configurations point in the directions of $\overrightarrow{CS}$ and $\overrightarrow{HT}$. HDBSCAN outlier points are not plotted.

Definition of cluster groups

For clusters containing configurations with a lignin dimer bound to the center of the cyclodextrin, clusters with similar binding configurations were grouped together. Those groups corresponded to configurations that were visually observed in the trajectories.

The cluster groups are described below in terms of the normal distances from the cyclodextrin COM to the lignin dimer COM, head, and tail. This is probably more intuitive than using a combination of d_n and cos θ. The cyclodextrin center COM to lignin dimer COM normal distance is d_n and the cyclodextrin COM to lignin dimer head and tail normal distances are defined analagous to d_n (Equation 1).

Dimer 1 cluster groups

For dimer 1, the cluster groups were named head-sec, center-sec, and tail-sec. The head-sec group consisted of clusters where the lignin dimer head distance was usually less than the lignin dimer tail and COM distances and the COM of the lignin dimer was closer to the cyclodextrin secondary face than the cyclodextrin primary face (positive distance). The center-sec group consisted of clusters where the lignin dimer COM, head, and tail were about the same distance from the cyclodextrin COM and the COM of the lignin dimer was closer to the cyclodextrin secondary face than the cyclodextrin primary face (positive distance). The tail-sec group consisted of clusters where the lignin dimer tail distance was usually less than the lignin dimer head and COM distances and the COM of the lignin dimer was closer to the cyclodextrin secondary face than the cyclodextrin primary face (positive distance). Kernel density estimate (KDE) plots of the cyclodextrin COM to lignin dimer normal distances (COM, head, tail) for each cluster group are shown in Figure 14, Figure 15, and Figure 16.

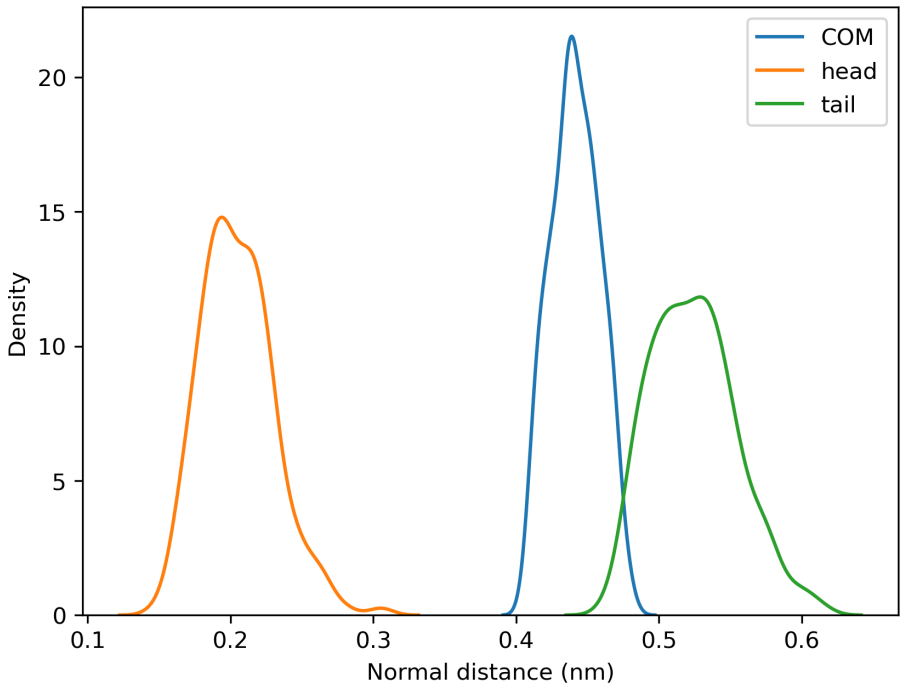

Figure 14: KDE plots (Scott's Rule bandwidth) of the cyclodextrin to lignin dimer normal distances for dimer 1 for the head-sec cluster group. In the legend, COM refers to the cyclodextrin COM to lignin dimer COM distance, head refers to the cyclodextrin COM to lignin dimer head distance, and tail refers to the cyclodextrin COM to lignin dimer tail distance.

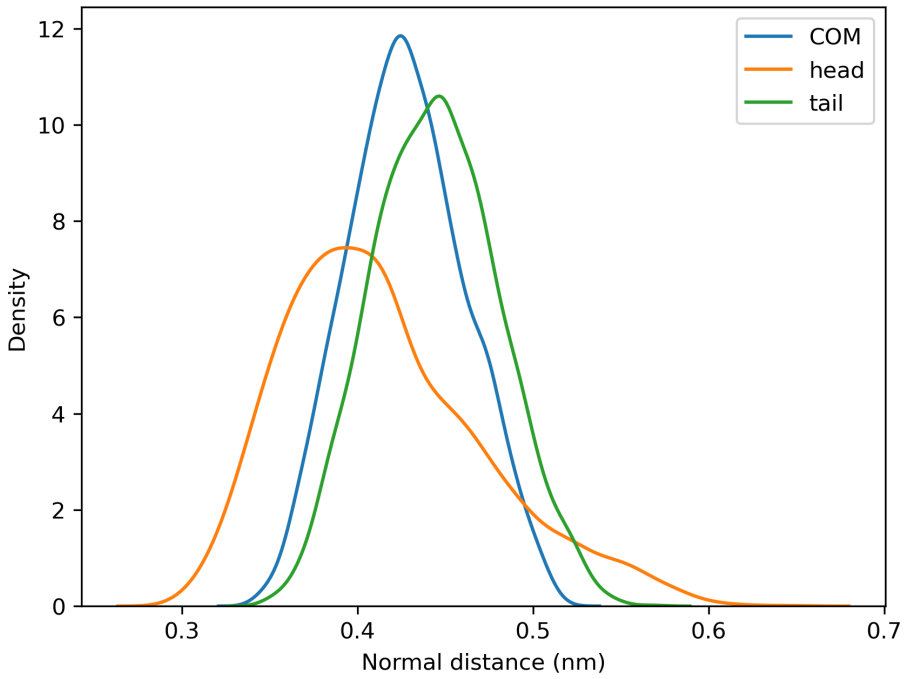

Figure 15: KDE plots (Scott's Rule bandwidth) of the cyclodextrin to lignin dimer normal distances for dimer 1 for the center-sec cluster group. In the legend, COM refers to the cyclodextrin COM to lignin dimer COM distance, head refers to the cyclodextrin COM to lignin dimer head distance, and tail refers to the cyclodextrin COM to lignin dimer tail distance.

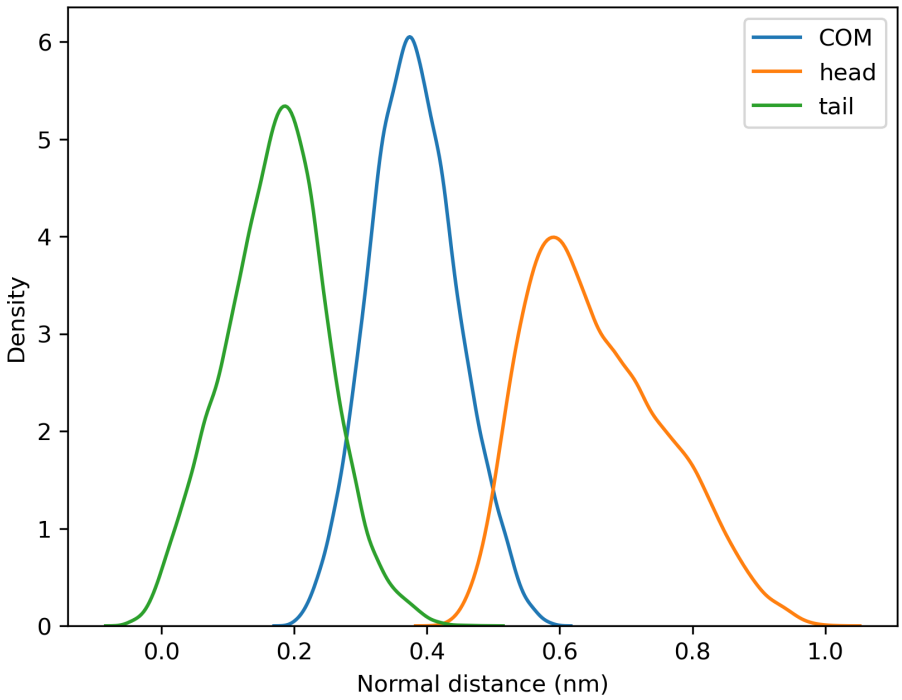

Figure 16: KDE plots (Scott's Rule bandwidth) of the cyclodextrin to lignin dimer normal distances for dimer 1 for the tail-sec cluster group. In the legend, COM refers to the cyclodextrin COM to lignin dimer COM distance, head refers to the cyclodextrin COM to lignin dimer head distance, and tail refers to the cyclodextrin COM to lignin dimer tail distance.

Dimer 2 cluster groups

For dimer 2, the cluster groups were named head-sec, center-sec, tail-sec, and tail-pri. The head-sec, center-sec, tail-sec groups were defined analogously to dimer 1. The tail-pri group consisted of clusters where the lignin dimer tail distance was usually greater than the lignin dimer head and COM distances and the COM of the lignin dimer was closer to the cyclodextrin primary face than the cyclodextrin secondary face (negative distance). KDE plots of the cyclodextrin COM to lignin dimer normal distances (COM, head, tail) for each cluster group are shown in Figure 17, Figure 18, Figure 19, and Figure 20.

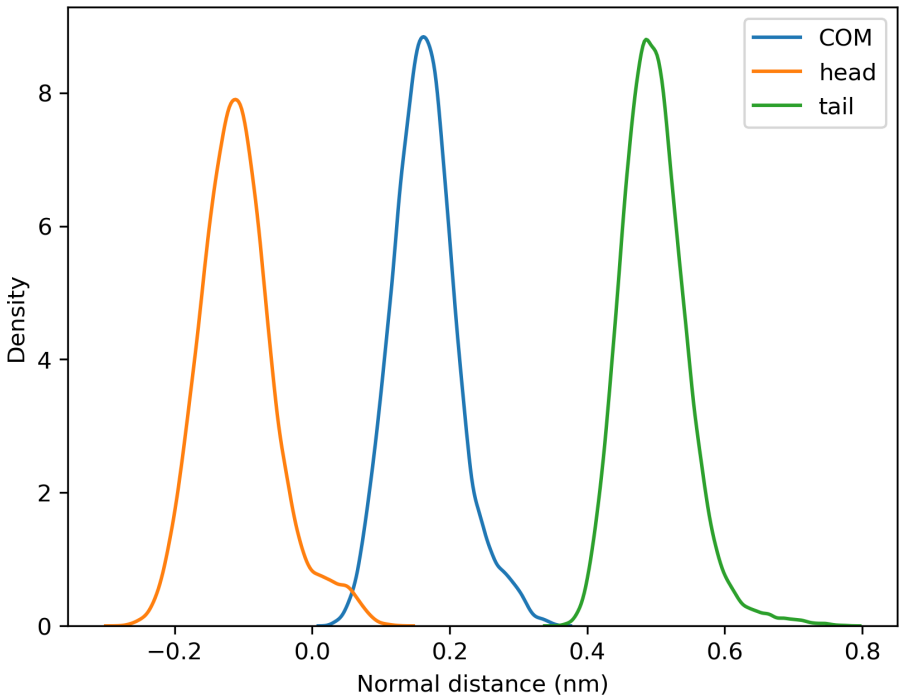

Figure 17: KDE plots (Scott's Rule bandwidth) of the cyclodextrin to lignin dimer normal distances for dimer 2 for the head-sec cluster group. In the legend, COM refers to the cyclodextrin COM to lignin dimer COM distance, head refers to the cyclodextrin COM to lignin dimer head distance, and tail refers to the cyclodextrin COM to lignin dimer tail distance.

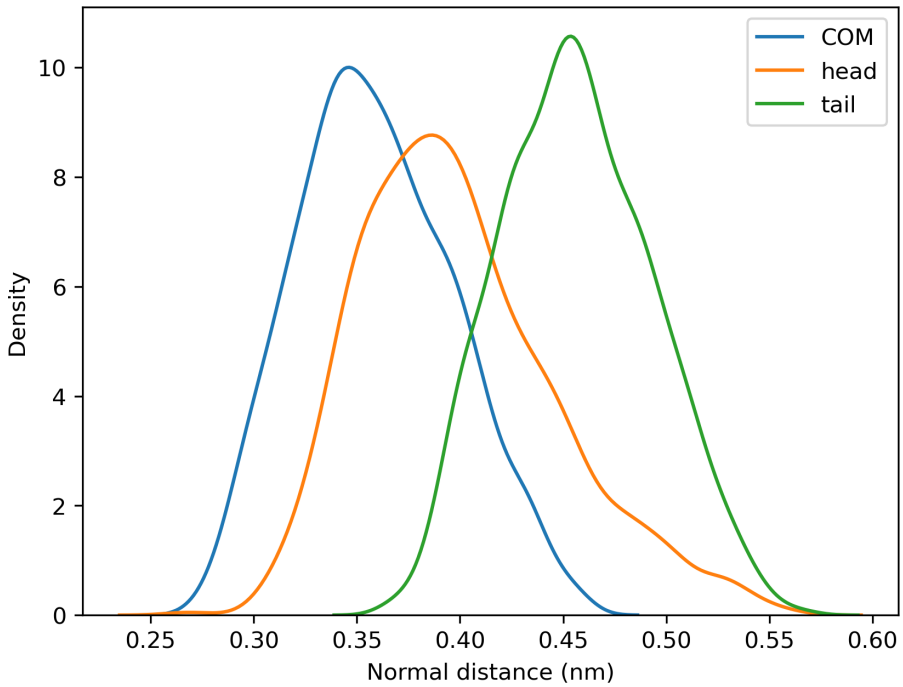

Figure 18: KDE plots (Scott's Rule bandwidth) of the cyclodextrin to lignin dimer normal distances for dimer 2 for the center-sec cluster group. In the legend, COM refers to the cyclodextrin COM to lignin dimer COM distance, head refers to the cyclodextrin COM to lignin dimer head distance, and tail refers to the cyclodextrin COM to lignin dimer tail distance.

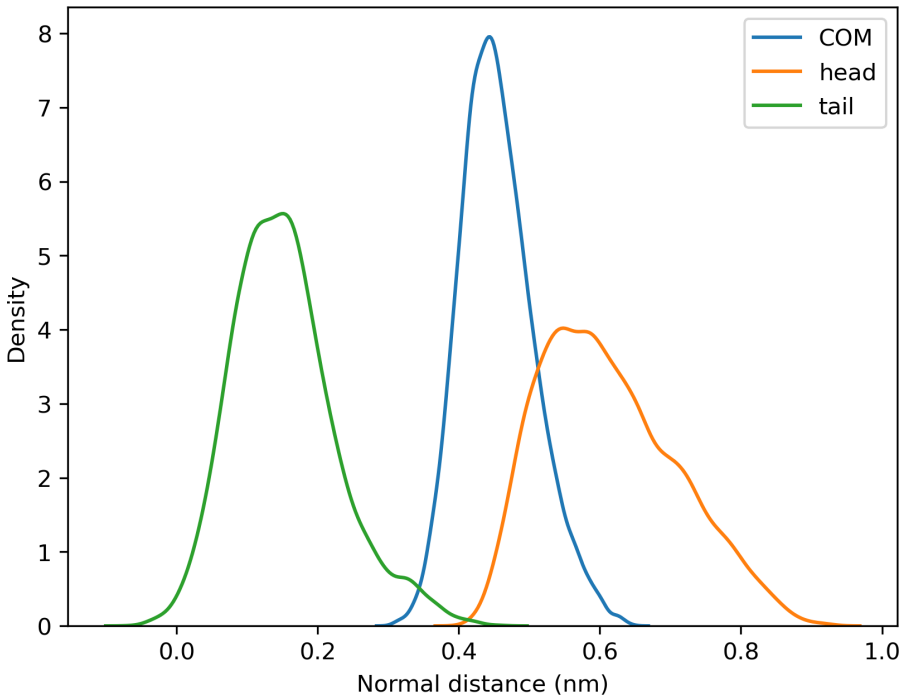

Figure 19: KDE plots (Scott's Rule bandwidth) of the cyclodextrin to lignin dimer normal distances for dimer 2 for the tail-sec cluster group. In the legend, COM refers to the cyclodextrin COM to lignin dimer COM distance, head refers to the cyclodextrin COM to lignin dimer head distance, and tail refers to the cyclodextrin COM to lignin dimer tail distance.

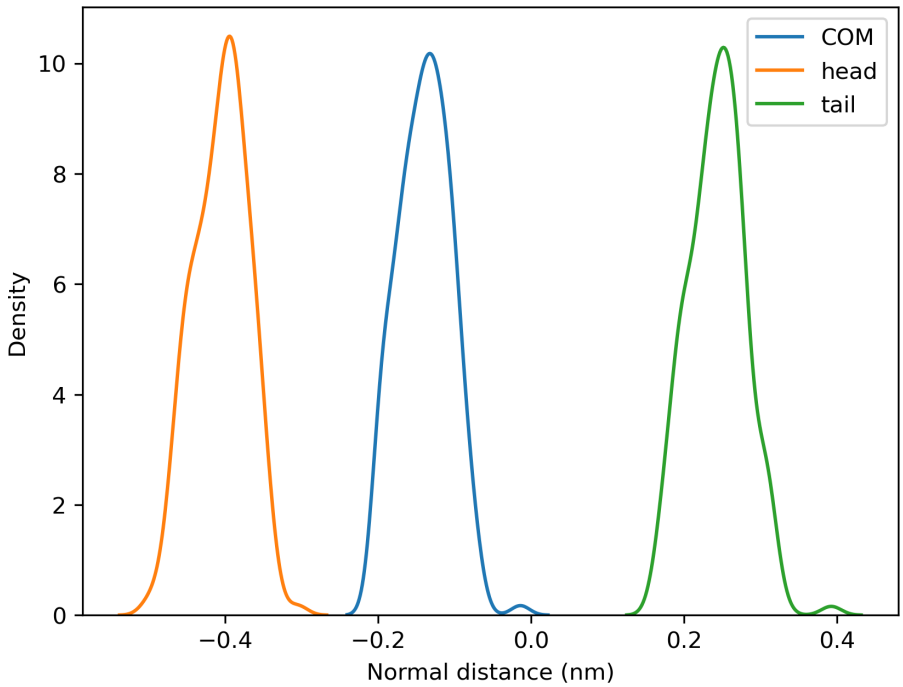

Figure 20: KDE plots (Scott's Rule bandwidth) of the cyclodextrin to lignin dimer normal distances for dimer 2 for the tail-pri cluster group. In the legend, COM refers to the cyclodextrin COM to lignin dimer COM distance, head refers to the cyclodextrin COM to lignin dimer head distance, and tail refers to the cyclodextrin COM to lignin dimer tail distance.

Dimer 3 cluster groups

For dimer 3, the cluster groups were named end-sec and center. The end-sec group consists of clusters where the distance for one of the lignin dimer ends (molecule is symmetric) was usually less than the distance for the other lignin dimer end and the lignin dimer COM, and the COM of the lignin dimer was closer to the cyclodextrin secondary face than the cyclodextrin primary face (positive distance). The center group consists of clusters where the lignin dimer COM was located near the cyclodextrin COM. KDE plots of the cyclodextrin COM to lignin dimer normal distances (COM, ends) for each cluster group are shown in Figure 21 and Figure 22.

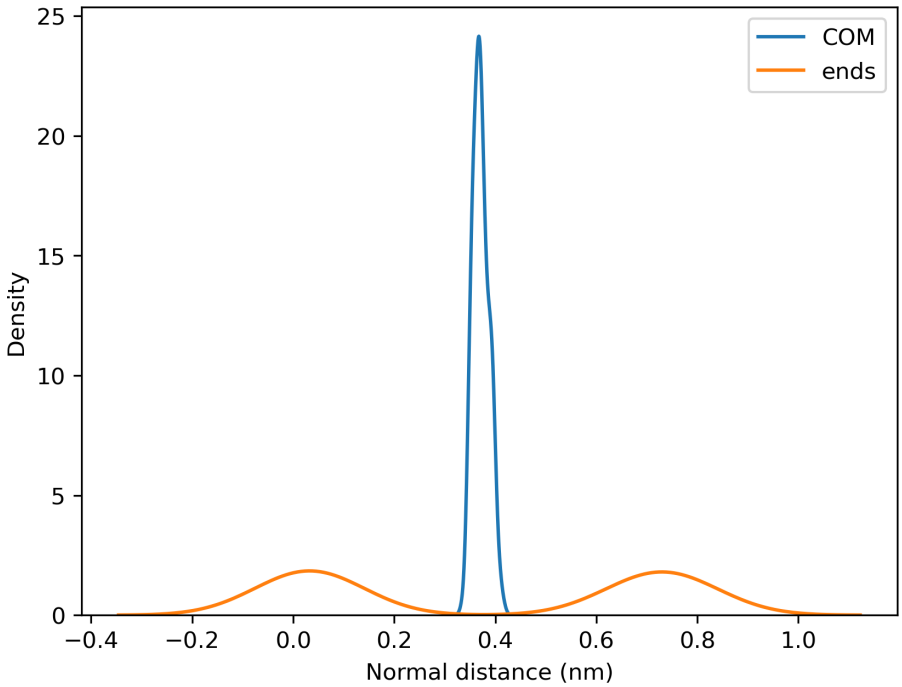

Figure 21: KDE plots (Scott's Rule bandwidth) of the cyclodextrin to lignin dimer normal distances for dimer 3 for the end-sec cluster group. In the legend, COM refers to the cyclodextrin COM to lignin dimer COM distance, and ends refers to the cyclodextrin COM to lignin dimer end distance.

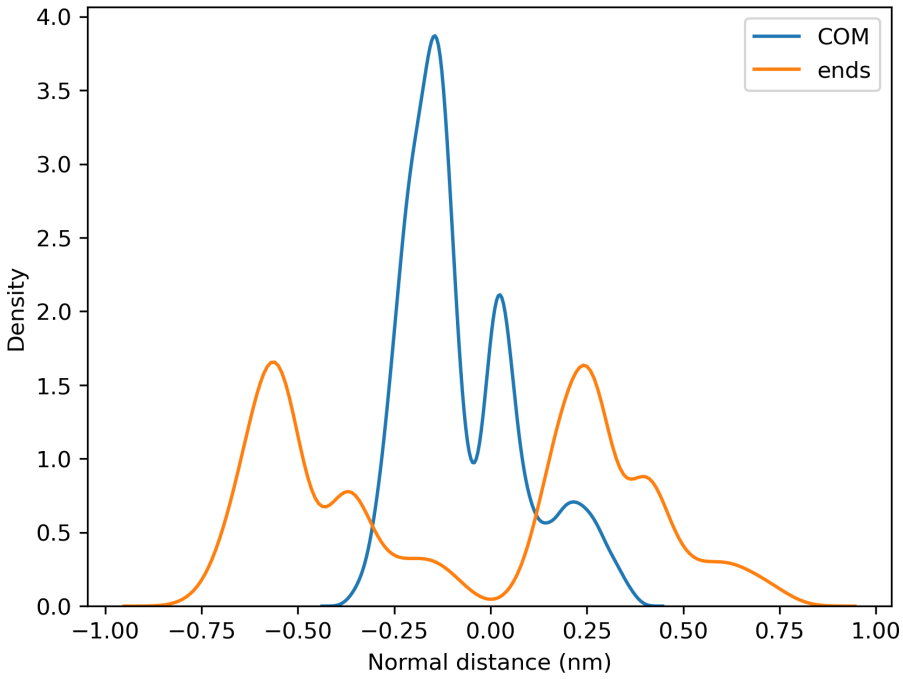

Figure 22: KDE plots (Scott's Rule bandwidth) of the cyclodextrin to lignin dimer normal distances for dimer 3 for the center cluster group. In the legend, COM refers to the cyclodextrin COM to lignin dimer COM distance, and ends refers to the cyclodextrin COM to lignin dimer end distance.

Proportion of configurations in each cluster and cluster group

For clusters and cluster groups which contained configurations with a lignin dimer bound to the center of the cyclodextrin, the proportions of those configurations belonging to each cluster or cluster group were calculated. The proportions are shown in Table 1, Table 2, and Table 3.

Table 1: Proportions of center-bound configurations belonging to each cluster and cluster group for dimer 1. Refer to Figure 11 to see the clusters and configurations corresponding to the label numbers. The head-sec, center-sec, and tail-sec labels refer to cluster groups (see Dimer 1 cluster groups). The clusters in each cluster group are listed in the rows above them up until another cluster group label is encountered or the top of the table.

Label	Fraction
4	0.0020
5	0.0060
6	0.5051
7	0.0054
8	0.0067
head-sec	0.5253
2	0.0571
center-sec	0.0571
3	0.4176
tail-sec	0.4176

There were 3 cluster groups defined for dimer 1: head-sec, center-sec, and tail-sec (see Dimer 1 cluster groups). The head-sec group consisted of 5 clusters. However, most of the configurations belonged to cluster 6. Clusters 4, 5, 7, and 8 accounted for a total of about 2.0% of all configurations, while cluster 6 had about 50.5% of all configurations. The center-sec and tail-sec groups each consisted of a single cluster with about 5.7% and 41.8% of all configurations, respectively.

Table 2: Proportions of center-bound configurations belonging to each cluster and cluster group for dimer 2. Refer to Figure 12 to see the clusters and configurations corresponding to the label numbers. The head-sec, center-sec, tail-sec, and tail-pri labels refer to cluster groups (see Dimer 2 cluster groups). The clusters in each cluster group are listed in the rows above them up until another cluster group label is encountered or the top of the table.

Label	Fraction
4	0.0229
5	0.7153
head-sec	0.7382
3	0.0280
center-sec	0.0280
2	0.2300
tail-sec	0.2300
1	0.0037
tail-pri	0.0037

There were 4 cluster groups defined for dimer 2: head-sec, center-sec, tail-sec, and tail-pri (see Dimer 2 cluster groups). The head-sec group consisted of 2 clusters. However, most of the configurations belonged to cluster 5. Cluster 4 accounted for about 2.3% of all configurations, while cluster 5 had about 71.5% of all configurations. The center-sec, tail-sec, and tail-pri groups each consisted of a single cluster with about 2.8%, 23.0%, and 0.4% of all configurations, respectively.

Table 3: Proportions of center-bound configurations belonging to each cluster and cluster group for dimer 3. Refer to Figure 13 to see the clusters and configurations corresponding to the label numbers. The end-sec and center labels refer to cluster groups (see Dimer 3 cluster groups). The clusters in each cluster group are listed in the rows above them up until another cluster group label is encountered or the top of the table.

Label	Fraction
1	0.0717
3	0.0027
end-sec	0.0744
2	0.9256
center	0.9256

There were 2 cluster groups defined for dimer 3: end-sec and center (see Dimer 3 cluster groups). The end-sec group consisted of 2 clusters. However, most of the configurations belonged to cluster 1. Cluster 3 accounted for about 0.3% of all configurations, while cluster 5 had about 7.2% of all configurations. The center group consisted of a single cluster with about 92.6% of all configurations.

Comparison to previous work

A comparison of the proportions of configurations in each cluster group from this work (3 CVs + HDBSCAN) and the previous work[5] (Many CVs + PCA + DBSCAN) was made. The results of this comparison are illustrated with bar charts in Figure 23, Figure 24, and Figure 25. For dimers 1 and 3, some clusters or groups defined in this work were not able to be separated into separate clusters in the previous analysis. For dimer 2, the cluster groups were consistent with previous work, and the proportions for the cluster groups were similar between this work and the previous work.

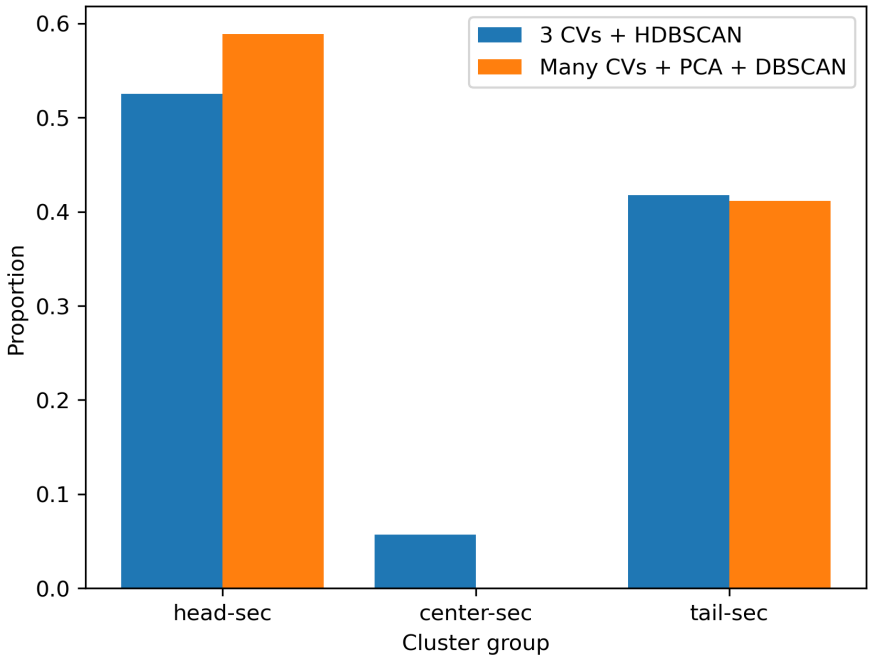

Figure 23: Comparison of the proportions of configurations in each cluster group for dimer 1 between this work (3 CVs + HDBSCAN) and the previous work[5] (Many CVs + PCA + DBSCAN). In the previous work, no cluster corresponding to the center-sec cluster group was found. Based on the difference in proportions for the head-sec and center-sec groups in this work, the center-sec group was counted as part of the head-sec group in the previous work. The tail-sec proportion was similar in this work compared to the previous work.

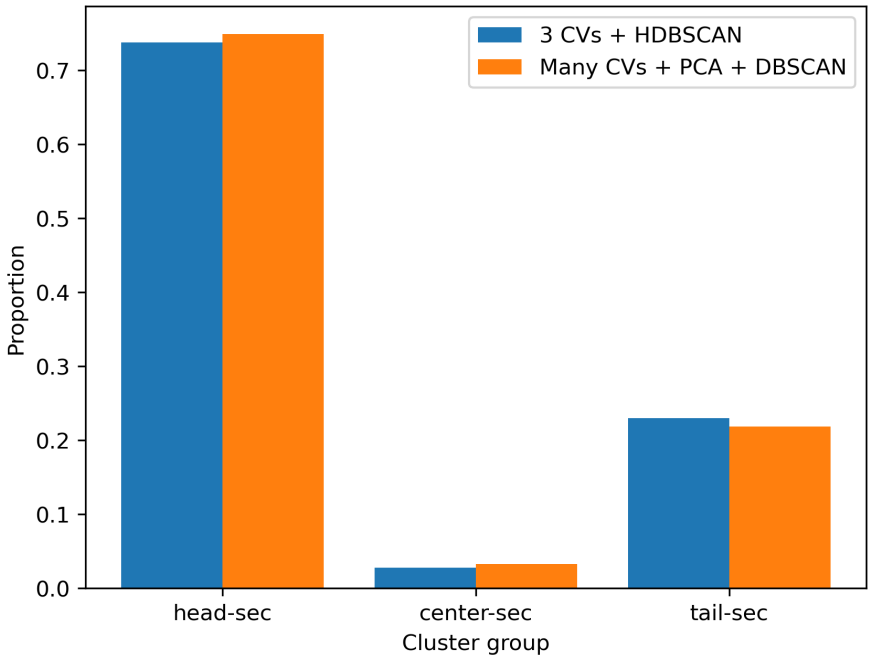

Figure 24: Comparison of the proportions of configurations in each cluster group for dimer 2 between this work (3 CVs + HDBSCAN) and the previous work[5] (Many CVs + PCA + DBSCAN). The cluster groups were the same for this work and the previous work, and the proportions for the cluster groups were similar.

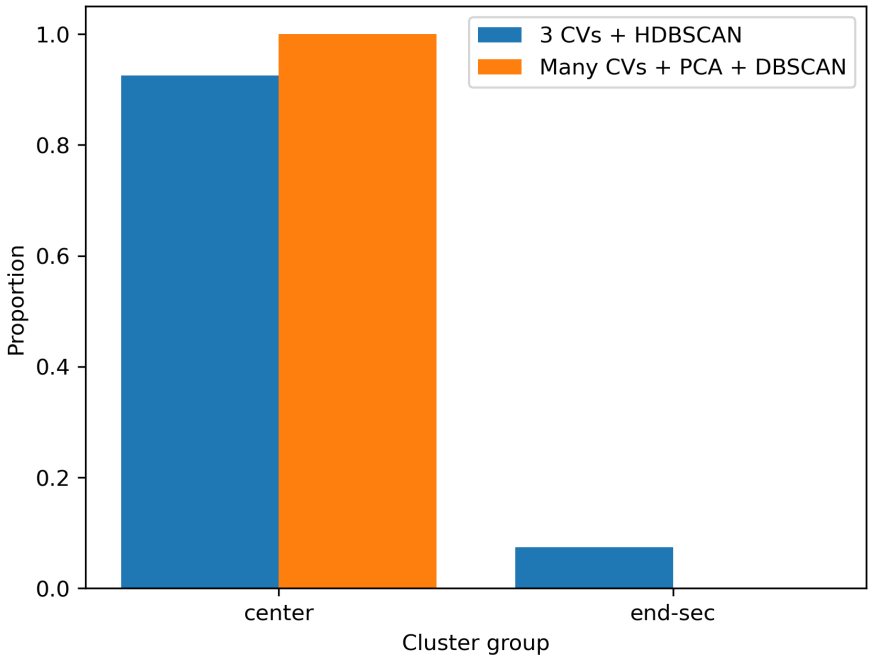

Figure 25: Comparison of the proportions of configurations in each cluster group for dimer 3 between this work (3 CVs + HDBSCAN) and the previous work[5] (Many CVs + PCA + DBSCAN). In the previous work, no cluster corresponding to the end-sec cluster group was found. Only a single cluster corresponding to center bound configurations was found.

Conclusion

Data from previous unbiased molecular dynamics simulations of lignin dimer derivatives and β-cyclodextrin in aqueous solution[5] were reanalyzed to determine the proportions of each type of configuration with the lignin dimer bound to the center of the cyclodextrin. In the previous work, many collective variables (CVs) were used to describe the system, principal component analysis (PCA) was used to reduce the dimensionality of the CVs, and DBSCAN clustering was used to identify the clusters of configurations.

In this work, three CVs were used to describe the system:

1. The cyclodextrin center of mass (COM) to lignin dimer COM in the direction normal to the cyclodextrin secondary face (d_n).
2. The cyclodextrin COM to the lignin dimer COM in the direction tangential to the cyclodextrin secondary face (d_t).

3. The cosine of the angle between the vector from the cyclodextrin COM to the oxygen atoms in the cyclodextrin secondary face and the vector from the lignin head to tail ($\cos\theta$).

Use of these three relatively simple CVs increased the interpretability of the results and did not require dimensionality reduction for cluster visualization.

Instead of DBSCAN, HDBSCAN clustering was used in this work to identify the clusters of configurations which made it easier to separate configuration types into different clusters. HDBSCAN clustering was able to identify the configuration types corresponding to the lignin dimer bound to the center of the cyclodextrin which were visually observed in the trajectories for all three dimers, as well as some subtypes of those main configuration types. In the previous analysis, there were visually observed configuration types which were not separated into their own clusters for two of the dimers.

When there were multiple subtypes for a main configuration type, the subtypes were grouped together since for each main configuration type there was always one dominant subtype. When looking at the proportions of configurations in each cluster group, the proportions were very similar between this work and the previous work for dimer 2 which had all of the main configuration types separated into their own clusters, validating the results of this work.

Although three CVs was few enough to easily visualize the clusters, for computation of binding free energies it would be more efficient to use as few CVs as possible. In the previous work[5], the binding free energies were computed using just the cyclodextrin COM to lignin dimer COM distance which is almost certainly not the best choice for a single CV. It would be useful to attempt to reduce the number of CVs to one or two using dimensionality reduction techniques such as UMAP[28], SGOOP[29], TICA[30], etc. and check if the main configuration types can still be separated into their own clusters.

In addition to using an improved CV(s) for computation of binding free energies, the forcefield parameters for the lignin dimer and cyclodextrin might be improved to better match experimental results[5].

References

(1) Weng, J.-K.; Chapple, C. The Origin and Evolution of Lignin Biosynthesis. *New Phytol.* **2010**, *187* (2), 273–285. https://doi.org/10.1111/j.1469-8137.2010.03327.x.

(2) Wang, H.-M.; Yuan, T.-Q.; Song, G.-Y.; Sun, R.-C. Advanced and Versatile Lignin-Derived Biodegradable Composite Film Materials Toward a Sustainable World. *Green Chem.* **2021**, *23* (11), 3790–3817. https://doi.org/10.1039/D1GC00790D.

(3) Saokham, P.; Muankaew, C.; Jansook, P.; Loftsson, T. Solubility of Cyclodextrins and Drug/Cyclodextrin Complexes. *Molecules* **2018**, *23* (5, 5), 1161. https://doi.org/10.3390/molecules23051161.
(4) Sevim, S.; Sanlier, N. Cyclodextrin as a Singular Oligosaccharide: Recent Advances of Health Benefit and in Food Applications. *J. Food Sci.* **2024**, *89* (12), 8215–8230. https://doi.org/10.1111/1750-3841.17527.
(5) Dean, K. R.; Novak, B.; Moradipour, M.; Tong, X.; Moldovan, D.; Knutson, B. L.; Rankin, S. E.; Lynn, B. C. Complexation of Lignin Dimers with β-Cyclodextrin and Binding Stability Analysis by ESI-MS, Isothermal Titration Calorimetry, and Molecular Dynamics Simulations. *J. Phys. Chem. B* **2022**, *126* (8), 1655–1667. https://doi.org/10.1021/acs.jpcb.1c09190.
(6) Kutzner, C.; Páll, S.; Fechner, M.; Esztermann, A.; de Groot, B. L.; Grubmüller, H. More Bang for Your Buck: Improved Use of GPU Nodes for GROMACS 2018. *J. Comput. Chem.* **2019**, *40* (27), 2418–2431. https://doi.org/10.1002/jcc.26011.
(7) *Welcome to the GROMACS documentation! — GROMACS 2018.3 documentation*. https://manual.gromacs.org/documentation/2018.3/index.html (accessed 2024-11-01).
(8) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **2021**, *61* (8), 3891–3898. https://doi.org/10.1021/acs.jcim.1c00203.
(9) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455–461. https://doi.org/10.1002/jcc.21334.
(10) *AutoDock Vina*. https://vina.scripps.edu/# (accessed 2024-11-02).
(11) *Complexation of Lignin Dimers with β-Cyclodextrin and Binding Stability Analysis by ESI-MS, Isothermal Titration Calorimetry, and Molecular Dynamics Simulations*. https://brian-novak.vercel.app/career/website/lignin/complexation-of-lignin-dimers-with-v-cyclodextrin-and-binding-stability-analysis-by-esi-ms-isothermal-titration-calorimetry-and-molecular-dynamics-simulations/ (accessed 2024-11-02).
(12) *PCA — scikit-learn 1.5.2 documentation*. https://scikit-learn/stable/modules/generated/sklearn.decomposition.PCA.html (accessed 2024-11-05).
(13) Salem, N.; Hussein, S. Data Dimensional Reduction and Principal Components Analysis. *Procedia Comput. Sci.* **2019**, *163*, 292–299. https://doi.org/10.1016/j.procs.2019.12.111.
(14) *DBSCAN — scikit-learn 1.5.2 documentation*. https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.DBSCAN.html (accessed 2024-11-05).
(15) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*; KDD'96; AAAI Press: Portland, Oregon, 1996; pp 226–231.

(16) *(1S,2R)-1-(4-Hydroxy-3-methoxyphenyl)-2-{4-[(1E)-3-hydroxy-1-propen-1-yl]-2-methoxyphenoxy}-1,3-propanediol*. https://www.chemspider.com/Chemical-Structure.10188622.html (accessed 2024-11-25).
(17) *(1S,2R)-1-(4-Hydroxy-3-methoxyphenyl)-2-(2-methoxyphenoxy)-1,3-propanediol*. https://www.chemspider.com/Chemical-Structure.58837283.html (accessed 2024-11-25).
(18) *4,4′-Tetrahydro-1H,3H-furo[3,4-c]furan-1,4-diylbis(2-methoxyphenol)*. https://www.chemspider.com/Chemical-Structure.204822.html (accessed 2024-11-25).
(19) *β-Cyclodextrin*. https://www.chemspider.com/Chemical-Structure.10469496.html (accessed 2024-11-25).
(20) *RDKit*. https://www.rdkit.org/ (accessed 2024-11-02).
(21) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.
(22) *VMD 1.9.3 Documentation*. https://www.ks.uiuc.edu/Research/vmd/current/docs.html (accessed 2024-11-02).
(23) Campello, R. J. G. B.; Moulavi, D.; Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*; Pei, J., Tseng, V. S., Cao, L., Motoda, H., Xu, G., Eds.; Springer: Berlin, Heidelberg, 2013; pp 160–172. https://doi.org/10.1007/978-3-642-37456-2_14.
(24) Campello, R. J. G. B.; Moulavi, D.; Zimek, A.; Sander, J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Trans. Knowl. Discov. Data* **2015**, *10* (1), 5:1–5:51. https://doi.org/10.1145/2733381.
(25) *HDBSCAN — scikit-learn 1.5.2 documentation*. https://scikit-learn/stable/modules/generated/sklearn.cluster.HDBSCAN.html (accessed 2024-11-05).
(26) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604–613. https://doi.org/10.1016/j.cpc.2013.09.018.
(27) *PLUMED: Introduction*. https://www.plumed.org/doc-v2.6/user-doc/html/index.html (accessed 2024-11-02).
(28) McInnes, L.; Healy, J.; Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. https://doi.org/10.48550/arXiv.1802.03426.
(29) Tiwary, P.; Berne, B. J. Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems. *Proc Natl Acad Sci USA* **2016**, *113* (11), 2839–2844. https://doi.org/10.1073/pnas.1600917113.
(30) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139* (1), 015102. https://doi.org/10.1063/1.4811489.