International Journal of Innovative Research and Scientific Studies, 8(12) 2025, pages: 289-297

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Classification of organometallic reactions using machine learning

Walter Bonke Mahlangu[1*], Nomasonto Rapulenyane[1] Taurai Hungwe[2], Somandla Ncube[3]

[1]*Department of Chemistry, Sefako Makgatho Health Sciences University, Medunsa, 0204, South Africa.*
[2]*Department of Computer Science and Information Technology, Sefako Makgatho Health Sciences University, Medunsa, 0204, South Africa.*
[3]*Department of Chemistry, Durban University of Technology, Durban 4000, South Africa.*

Corresponding author: Walter Bonke Mahlangu (*Email: Bonkemahlangu@Gmail.Com*)

Abstract

Classifying organometallic reactions into distinct reaction types is fundamentally important for understanding mechanisms and predicting reactions, for synthesis optimisation. Fundamental to classification of organometal reactions is reaction representation, but current methods often fail to capture organometal chemical transformation adequately. The study has adopted a hybrid fingerprinting approach, whereby new fingerprints were concatenated with permutation important Morgan fingerprints to create 49 to 63 bits fingerprints. The hybrid fingerprints were used to build KMeans clustering, Guassian mixtures, Affinity propagation and Heirarchical clustering models for classification of organometal reactions. The models successfully classified reactions into 6-8 major organometal reaction types. Of note, the fingerprints consistently outperformed Morgan fingerprints across all clustering models with good Davies–Bouldin Index (DBI) ranging from 0.3 to 0.6 and Silhouette score from 0.3 to 0.8. Furthermore, the clustering models were visualized using Principal Component Analysis. Affinity Propagation and KMeans demonstrated superior performance over Hierarchical and GMM algorithms in distinguishing major reaction categories. In contrast, the Hierarchical model excelled at identifying sub-level classifications compared to the other methods. Consequently, Affinity Propagation and KMeans are recommended for broad reaction type classification, while the Hierarchical approach is better suited for resolving detailed subclass distinctions. The observed variability in model performance further highlighted the importance of feature selection and representation in clustering models. Future studies should look into improving the fingerprints to recognize subcategories.

Keywords: Chemical data mining, Computational chemistry, Machine learning classification, Organometallic reactions, Reaction mechanism prediction.

DOI: 10.53894/ijirss.v8i12.11078
Funding: This research was supported by the National Research Foundation (NRF) through a NRF Postgraduate Scholarships fund..
History: Received: 24 October 2025 / **Revised:** 21 November 2025 / **Accepted:** 24 November 2025 / **Published:** 18 December 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

The categorization of organometallic reactions into specific reaction classes is essential in building models for predicting reactivity patterns, guiding synthetic strategies, and designing catalytic processes with enhanced accuracy and efficiency [1, 2]. Organizing reactions into defined classes such as oxidative addition, reductive elimination, migratory insertion, and σ-bond metathesis researchers can train machine learning algorithms with consistent descriptors, facilitating accurate forecasting of chemical behaviour and enabling rational design in organometallic chemistry [3, 4]. Reactions were systematically classified by their functional groups, products formed and by the inventor's name. Functional group based classification often oversimplifies mechanistic diversity and fails to account for the influence of catalysts, ligands, and reaction conditions, particularly in organometallic systems [5]. Product-based schemes tend to emphasize retrosynthetic outcomes while neglecting mechanistic insights and intermediate species, which are crucial for understanding selectivity and yields [6, 7]. Naming reactions after inventors, while historically significant, introduces redundancy, and excludes newer or less prominent transformations [5]. Recent progress in machine learning has enabled the automation of chemical reaction classifications, allowing algorithms to systematically categorize reactions based on learned patterns and features [8-10]. Classifications can be based on two strategies a) model-driven which is based on predefined rules, expert systems, or chemical knowledge (e.g., SMARTS patterns, reaction templates) and b) data-driven which relies on machine learning and statistical models that learn patterns from large datasets without manual rule-setting [3, 11]. Central to all is reaction representations or fingerprinting. The accuracy and efficiency of models in reaction classifications and predictions is greatly dependent on the type of fingerprints used in training the models [3, 9].
The Table 1 summarizes and compares features of the model-driven and data-driven approaches [12].

Table 1.
Comparison of data-driven and model-driven approaches.

Feature	Model-Driven	Data-Driven
Foundation	Expert rules, templates	Statistical patterns from data
Interpretability	High	Moderate to low
Scalability	Limited by rule coverage	High with large datasets
Flexibility	Rigid (rule-bound)	Adaptive (learns new patterns)
Use in Organometallics	Mechanistic modeling, SMARTS filters	Catalyst prediction, yield modeling

The drawback of data-driven approach is interpretation or understanding the reasoning or decision process behind results. Subsequently, for chemistry, where mechanistic understanding is crucial, if the models' decision cannot be explained, then it becomes difficult to verify or reproduce it [7, 13]. The other limitation is that these models depend heavily on the structural connectivity within the extended reaction center, which can cause them to overlook reactions that follow the same mechanism but differ in how their atoms or bonds are arranged [3]. The drawback of model-driven strategy is that it focuses on reaction centers only and that might cause the model to fail to identify related subclasses that share broader chemical features or mechanistic traits [14-16]. The other way of representing chemical reactions is through reaction fingerprinting or molecular descriptors [17]. These descriptors are used to characterize reactants and reactions, for instance [18-21].

- Morgan Fingerprints, generates fingerprints by looking at the bond between atoms and types of atoms within a molecule and then turn the information into binary strings
- MACCS keys, generates bit values by asking questions about the molecule for example, does the molecule contain lanthanides, group 1B and 2B metals among others?
- Differential reaction fingerprints, encodes the structural changes occurring during a chemical reaction

These descriptors have been successfully used in quantitative structure-activity relationship (QSAR) studies to predict toxicity of toxicants, molecular docking and discovery of novel drugs[18, 22-24]. They have also been used in organic chemistry reaction predictions [13, 25, 26]. The challenge arises when applying these models in predicting and classifying organometallic reactions. The approaches fail to capture organometallic reaction specific features such as metal-mediated mechanisms, variable oxidation states and coordination chemistry [27]. To address the limitations of relying solely on either data-driven or model-driven frameworks for organometallic reactions, the research was adopted a hybrid modelling approach. A wide range of chemical reactions can systematically be analysed and classified by integrating SMARTS-derived reaction features with novel organometallic transformation fingerprints, and then applying unsupervised machine learning algorithms. This strategy leverages the rigor and transparency of rule-based classification alongside the adaptability and scalability of data-driven learning, enabling a more nuanced understanding and robust classifications of organometallic reactions

2. Methods

The research used a qualitative study design method to develop machine learning models for classification tasks. Secondary data from peer reviewed databases was utilized in training the model to classify organometallic reactions. The training dataset was derived from ORD with dataset name, ord_dataset-de0979205c84441190feef587fef8d6d. The final data set contained 82 organometallic reactions.

2.1. Fingerprinting

The research methodology employs a computational approach to automatically classify organometallic reaction mechanisms by analysing changes in molecular structure using the RDKit cheminformatics toolkit. The section provides a summary of how some fingerprints were generated.

Atom and Bond Analysis: The methodology begins by defining a comprehensive set of metal atoms based on their atomic numbers. Two core functions are used to quantify atomic connectivity: one that counts only non-hydrogen neighbours and another that counts all neighbours (including hydrogen). A third function is used to identify and map the non-hydrogen ligands attached to metal centers. These functions provide the necessary data to track changes in coordination.

Reaction Pre-processing and Atom Tracking: A given reaction is parsed into its reactant and product components. A key step involves using atom map numbers—unique identifiers assigned to atoms—to reliably track specific metal atoms and their ligands as they transform from reactants to products.

Automated Reaction Classification: The changes in coordination and connectivity are then used as a basis for classifying the reaction into distinct types:

Dissociation/Association: Identified by a decrease or increase of one non-hydrogen neighbour, respectively.

Oxidative Addition/Reductive Elimination: Identified by an increase or decrease of two total neighbors (including hydrogen), respectively.

Transmetallation: Detected by tracking a ligand (identified by its map number) that breaks its bond with one metal center in the reactants and forms a new bond with a different metal center in the products.

2.2. Morgan Fingerprints

Fingerprints are created by analysing the atomic connectivity and element types in a molecule, which are then encoded into binary strings. These strings form a distinctive 2048-bit fingerprint that uniquely represents each molecule. Only Important features were used in training the model. The important features in training the model were identified by using permutation importance. Permutations were performed for each feature in triplicates with a fixed random state for reproducibility. Model generalization was improved by reducing dimensionality using reduced permutation importance, and features were ranked based on their impact on model performance. For a trained model f, the importance of feature j is computed by permuting its values across the dataset and measuring the change in prediction error:

$$\text{Importance}(j) = \mathbb{E}\text{perm}\,[L(f(\text{Xperm}(j)\,, y)\text{-}L(f(\text{X}), y)] \quad (1)$$

Where L is the loss function (mean squared error), X is the original feature matrix and Xperm(j) is a modified version of X where the values in column have been randomly permuted. If permuting feature j causes a large increase in the loss L, then feature j was important for accurate predictions.

The “classifications fingerprint" and Morgan fingerprints were the concatenated to form final 49 - 63-bit fingerprints, which were used in model training.

2.3. Algorithm Training

Classification models were deployed using different featurization methods i.e. Morgan fingerprints (2046 bits) and novel fingerprints (24 bits). Significant Morgan fingerprint features, identified through permutation importance for each model, were selectively retained and concatenated with novel fingerprint components. This process yielded final composite fingerprints ranging from 49 to 63 bits in length. These datasets were then used to train KMeans clustering, Guassian mixtures, Affinity propagation and Heirarchical clustering models.

2.4. Method Validation

The validation metrics applied, namely, Silhouette Score and Davies–Bouldin Index (DBI) served as complementary internal measures to quantitatively assess the clustering outcomes. These metrics were instrumental in evaluating the structural quality and internal coherence of the clusters produced by the models. The Silhouette score quantifies how well each data point fits within its assigned cluster relative to neighbouring clusters. The Silhouette score range from -1 to +1, with higher scores “+1” indicating well-separated, cohesive clusters and lower scores suggesting ambiguous or overlapping assignments. For example, If a particular reaction lies close to the centroid of its assigned cluster and far from the centroids of other clusters, it will have a Silhouette Score close to +1. Conversely, if the reaction sits near the boundary between clusters it might score closer to 0, suggesting ambiguity in its assignment. A negative score would imply misclassification, where the reaction is more similar to a neighbouring cluster than its own.

The Silhouette score $s(i)$ for a data point i is defined as:

$$s(i) = \frac{b(i)-a(i)}{\max\,(a(i),b(i))} \quad (2)$$

Where: $a(i)$ is the average dissimilarity (Euclidean distance) of point i to all other points in the same cluster. $b(i)$ is the minimum average dissimilarity of point i to all points in any other cluster (i.e. the nearest neighbouring cluster)

The overall Silhouette score s for the clustering solution is the mean $s(i)$ over all data points

$$S = \frac{1}{n}\sum_{i=1}^{n} s(i) \quad (3)$$

The Davies–Bouldin Index offers a reciprocal perspective to that of Silhouette score by measuring the average similarity between each cluster and its most similar counterpart. An ideal DBI value is close to zero, reflecting low intra-cluster dispersion (compact clusters) and high inter-cluster separation (distinct boundaries). In contrast, higher DBI values

signal either diffuse clustering or insufficient separation between groups. Together, these metrics provide a robust framework for assessing clustering performance across algorithms with varying assumptions and geometries, including KMeans, Gaussian Mixtures, Affinity Propagation, and Hierarchical clustering.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max j \neq i \left(\frac{S_i + S_j}{M_{ij}}\right) \quad (4)$$

Where: k is the number of clusters, S_i is the average distance between each point in cluster i and the centroid of cluster i (intra-cluster dispersion). M_{ij} is the distance between the centroids i and j (inter-cluster separation)

3. Results and Discussion

Machine learning results were based on classification algorithms. A series of experiments were conducted to determine the most effective clustering methods. In the first set up, Morgan fingerprints were tested for reaction type classifications for example, metal-mediated mechanisms, variable oxidation states and coordination compounds. The efficiency of clustering methods were evaluated using Silhouette score and DBI as shown in Table 2.

Table 2.
Clustering method validation.

Model	Morgan fingerprint		Novel fingerprints	
	Silhouette Score	Davies-Bouldin Index	Silhouette Score	Davies-Bouldin Index
KMeans clustering	0.233	0.763	0.863	0.302
Guassian Mixture	0.198	0.840	0.401	0.573
Affinity Propagation	0.484	0.722	0.765	0.366
Hierarchical Clustering	0.219	NA	0.356	NA

The Affinity propagation model demonstrated relatively superior performance with a DBI score of 0.722 and Silhouette Score of 0.484 using Morgan fingerprints, which is significantly higher than the 0.198 achieved by the Gaussian Mixture Model, indicating that the clusters formed by the Affinity Propagation method are relatively more cohesive (points are tighter) and better separated. Conversely, the Gaussian Mixture Model's Silhouette Score near 0.2 suggests a less effective partitioning where many data points are near cluster boundaries or fall into overlapping regions. Regarding the DBI, the Gaussian Mixture Model yielded a value of 0.840; since lower values (closer to 0) are optimal for the DBI, this suggests moderate separation and compactness, though the metric was Not Applicable for the Hierarchical Clustering model. KMeans was not effective in producing well separated clusters, this was shown by a high DBI score of 0.763 and a low Silhouette Score of 0.233.

The KMeans Clustering model significantly outperformed the alternative models for novel fingerprint, achieving the highest Silhouette Score of 0.863, which indicates exceptional cluster cohesion and separation. This superior performance was further confirmed by the lowest DBI of 0.302. The Affinity propagation model also demonstrated strong performance with a Silhouette Score of 0.765 and a DBI of 0.366. In contrast, the Gaussian Mixture Model (Silhouette=0.401, DBI=0.573) and Hierarchical Clustering (Silhouette=0.356) showed non-effective clustering, suggesting greater cluster overlap or dispersion. Based on the consistent results from both metrics, the K-Means Clustering model was chosen as the optimal solution for the final segmentation of the data. The novel fingerprints consistently outperformed Morgan fingerprints across all clustering models, particularly KMeans and Affinity Propagation. These results suggest that the novel descriptors capture structural or reactivity features more effectively, leading to more coherent clustering in chemical space. However, it is important to acknowledge limitation of the developed fingerprints organometallic reactions. These have reduced capacity to encode catalytic activity in organometallic systems. They fall short in representing dynamic catalytic behaviour, such as turnover frequency, substrate activation, or transition state stabilization because catalytic performance is influenced by static molecular features and reaction conditions, reaction mechanisms (pathways) and kinetics [27, 28] that are not readily captured by current fingerprint design.[12,28,29]. Nonetheless, these descriptors excel at capturing structural motifs and reaction patterns.

3.1. Principal Component Analysis Visualization of Clustering Models

Principal Component Analysis (PCA) was employed to project the high-dimensional data into two dimensions in order to visually and quantitatively assess the clustering performance of KMeans applied to the novel fingerprint representation. The resulting scatter plot reveals six well-separated clusters, each marked by distinct colours and clearly defined centroids (Figure 1). This visual coherence aligns strongly with the quantitative metrics: the Silhouette Score of 0.863 indicates excellent intra-cluster cohesion and inter-cluster separation, while the DBI of 0.302 confirms minimal overlap and compact cluster formation. The tight grouping of data points around their respective centroids in the PCA plot which reinforces the interpretation that the novel fingerprints capture chemical reactions features. This provides a meaningful distinction that is effectively more than traditional Morgan fingerprints.

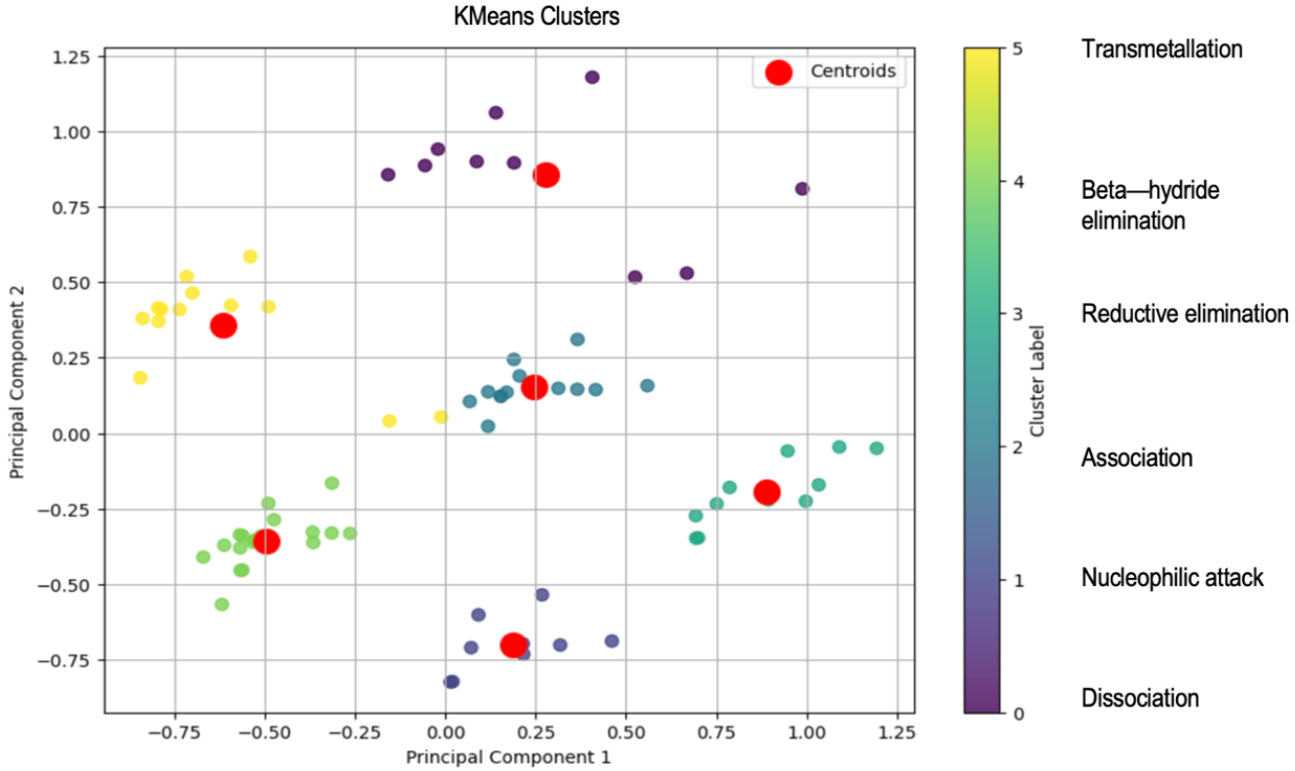

Figure 1.
KMeans clustering PCA using hybrid of novel fingerprints and Morgan fingerprints.

Figure 2, GMM Clusters (PCA) shows the scatter plot showcasing six clusters, each assigned a unique colour and labelled "0 – Reductive elimination", "1 – Not clearly defined", "2 – beta hydride elimination", "3 – Not clearly defined", "4 – Transmetallation and Association", "5 – Not clearly defined". The clustering results revealed a mixed outcome in terms of mechanistic clarity. Among the six identified clusters, only two clusters (Reductive Elimination and β-Hydride Elimination) exhibited well defined mechanistic identities. These clusters likely benefited from distinct structural or fingerprint features that aligned strongly with reaction patterns, enabling the model to consistently group them with high confidence. In contrast, clusters 1, 3, and 5 remained not clearly defined. The improved performance of the novel fingerprints across clustering models, particularly in KMeans and Affinity Propagation highlights their ability to capture chemically relevant features. However, the comparatively low mechanistic clarity observed in the GMM clusters suggests that the issue may lie not in the descriptors themselves, but in the clustering algorithm's assumptions and structure. GMM relies on probabilistic distributions and soft assignments, which can blur boundaries between mechanistically distinct reactions, especially when feature separation is subtle or nonlinear. The ambiguity is reflected in the presence of multiple "Not clearly defined" clusters, despite the fingerprints' demonstrated capacity to distinguish those clusters in other models. Cluster 4, labeled as Transmetallation and Association, appears partially resolved but still encompass mechanistic heterogeneity. These findings also underscore the importance of aligning clustering algorithms with the chemical structure of the data.

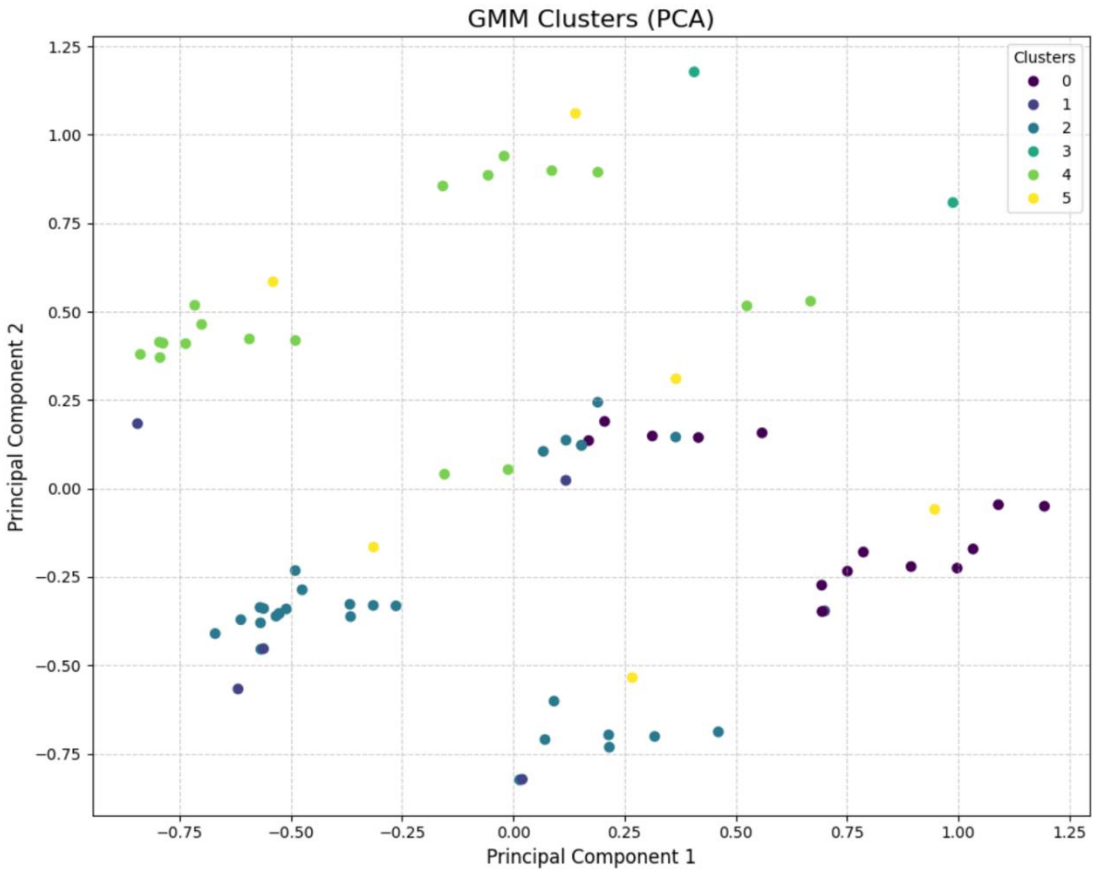

Figure 2.
GMM clustering PCA using hybrid of novel fingerprints and Morgan fingerprints.

The Affinity propagation clustering model Figure 3 exhibited robust performance in characterising the dataset's underlying structure. The dataset was partitioned into eight distinct clusters, which were subsequently examined through manual analysis to assign contextually meaningful labels i.e. “0 – Association”, “1 – beta hydride elimination”, “2 – Reductive elimination”, “3 – Dissociation”, “4 – ligand coordination”, “5 – Nucleophillic attack”, “6 – Transmetallation”, “7 - Intramolecular β-Hydride Elimination”. Quantitatively, it achieved a Silhouette Score of 0.765, indicating well-defined and cohesive clusters with substantial separation between them. Complementing this, the DBI was calculated at 0.366, further affirming the model’s effectiveness by reflecting low intra-cluster dispersion relative to inter-cluster distances. These metrics collectively suggest that KMeans and Affinity Propagation were the two best algorithms in categorising the dataset into major reaction classes, outperforming GMM and Hierarchical models in all metrices. The datapoints were close to their respective centroids, except in dissociation reactions, whereby the points were relatively scattered away from the centroids. The scatteredness in the cluster can be linked to the models’ inability to recognize reaction subcategories. Recognizing subcategories within organometallic dissociation reactions such as neutral ligand loss, anionic dissociation, or photochemical ejection can potentially enhances the effectiveness of ML clustering. These mechanistic distinctions introduce chemically meaningful variation into the feature space, allowing clustering algorithms to group reactions more coherently. This can improve the resolution of unsupervised models and enables dimensionality reduction techniques like PCA or Uniform Manifold Approximation and Projection (UMAP) to preserve mechanistic axes, resulting in interpretable visualizations that align with chemical intuition. Subcategory aware clustering can also strengthens modelling tasks. When clusters reflect subtle mechanistic patterns, they can serve as high-quality labels for supervised learning, improving the accuracy of predictions related to activation energies, selectivity, or reaction outcomes. Furthermore, these refined clusters can be mapped onto reaction networks, supporting retrosynthetic analysis and catalyst design [3, 6].

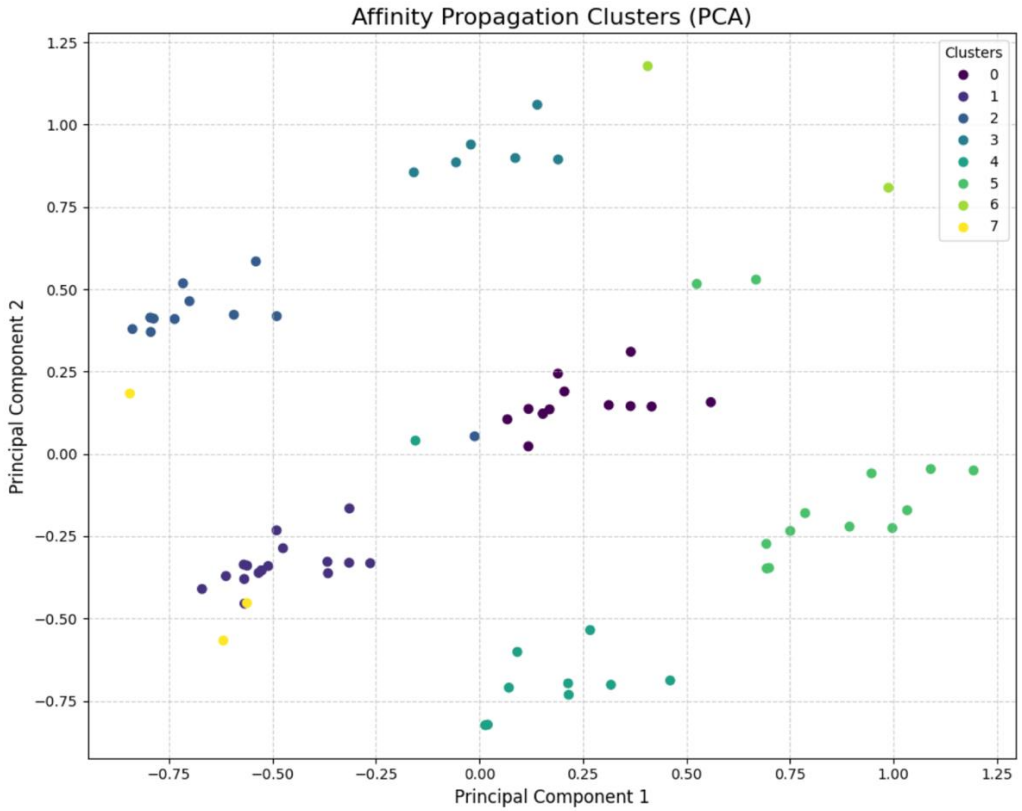

Figure 3.
Affinity Propagation clustering PCA using hybrid of novel fingerprints and Morgan fingerprints.

The hierarchical clustering dendrogram Figure 4 provides a visual representation of the structural similarity among chemical reactions, as encoded by their molecular fingerprints. Using the Jaccard distance metric, reactions are grouped based on shared substructural features, enabling the identification of clusters with high internal consistency. The x-axis, denoting individual chemical reactions, reveals how these transformations relate to one another in terms of fingerprint-derived descriptors. The vertical height of each branch reflects the dissimilarity between reaction clusters, with shorter branches indicating greater similarity. For example, Reactions 48 and 55 are very similar, sharing a reductive elimination process that causes the metal's formal oxidation state (OS) to drop by 2 units. The analysis groups these two reactions into a larger class that also contains reaction 23, despite reaction 23 only having an OS drop of 1 unit. Therefore, the clustering suggests a hierarchical relationship: all three are in the same overall class, but the difference in the OS change (-2 vs-1) defines their distinct subclasses. Hierarchical clustering frameworks benefit especially from subcategory recognition, enabling multi-level classification that mirrors how chemists conceptualize reactivity. [3,6]

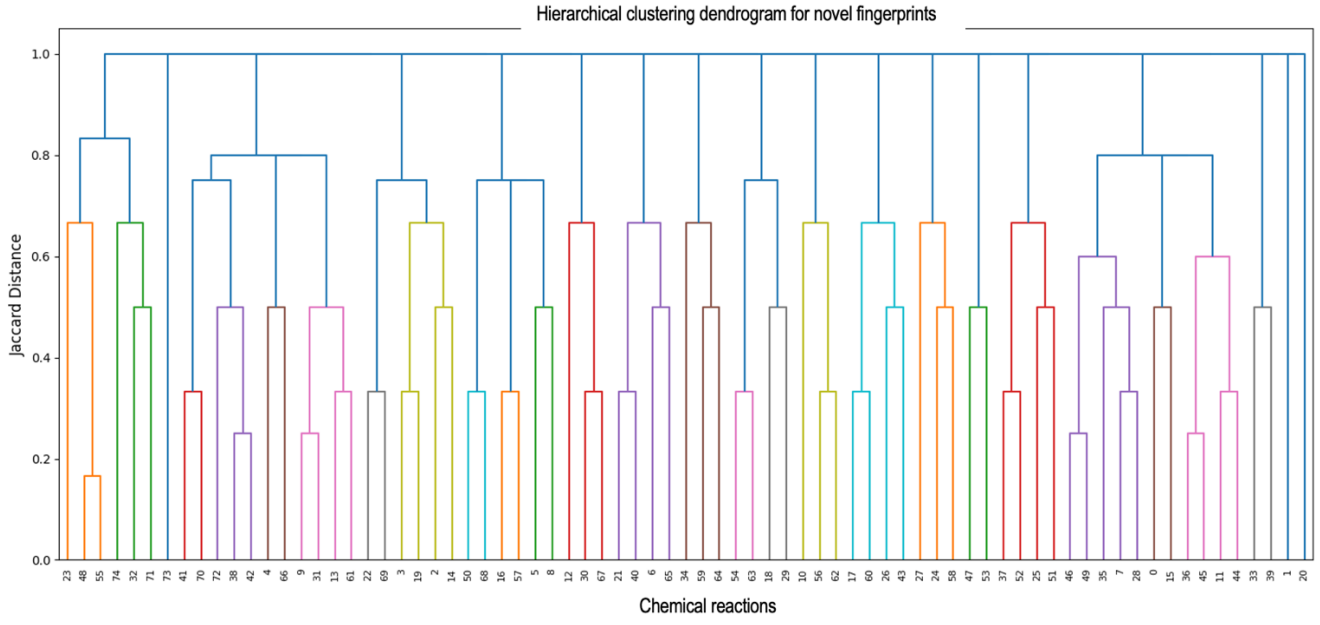

Figure 4.
Hierarchical clustering dendrogram for novel fingerprints.

4. Conclusion

New organometallic reaction fingerprints were successfully developed. The novel fingerprints demonstrated better performance than Morgan fingerprints in all clustering analyses with good DBI ranging from 0.3 to 0.6 and Silhouette score from 0.3 to 0.8. Notably, the PCA plots aligned with the metric scores. and reaction data was classified into 6 – 8 distinctive major reaction types namely, association, beta hydride elimination, reductive elimination, dissociation, ligand coordination, Nucleophillic attack, Transmetallation, intramolecular β-hydride elimination. Affinity propagation and KMeans outperformed Hierarchial and GMM algorithms in classifying major reaction types. Hierarchical model performed better in recognising sub level classifications compared to Affinity propagation, KMeans and GMM models. Therefore, when classifying major reaction types Affinity propagation and KMeans should be used, and for recognising sub level reactions classes Hierarchical model should be used. The observed variability in model performance further highlighted the importance of feature and model selection in unsupervised reaction modelling. Future studies should look into improving the fingerprints to recognize reaction subcategories and organometallic catalytic activity.

References

[1] S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich, and C. Corminboeuf, "Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts," *Chemical Science*, vol. 12, no. 20, pp. 6879-6889, 2021.

[2] J. Burés and I. Larrosa, "Organic reaction mechanism classification using machine learning," *Nature*, vol. 613, no. 7945, pp. 689-695, 2023. https://doi.org/10.1038/s41586-022-05639-4

[3] N. Schneider, D. M. Lowe, R. A. Sayle, and G. A. Landrum, "Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity," *Journal of Chemical Information and Modeling*, vol. 55, no. 1, pp. 39-53, 2015. https://doi.org/10.1021/ci5006614

[4] H. Kraut, J. Eiblmaier, G. Grethe, P. Löw, H. Matuszczyk, and H. Saller, "Algorithm for reaction classification," *Journal of Chemical Information and Modeling*, vol. 53, no. 11, pp. 2884-2895, 2013. https://doi.org/10.1021/ci400442f

[5] G. M. Ghiandoni *et al.*, "Development and application of a data-driven reaction classification model: Comparison of an electronic lab notebook and medicinal chemistry literature," *Journal of Chemical Information and Modeling*, vol. 59, no. 10, pp. 4167-4187, 2019. https://doi.org/10.1021/acs.jcim.9b00537

[6] M. H. Lin, Z. Tu, and C. W. Coley, "Improving the performance of models for one-step retrosynthesis through re-ranking," *Journal of Cheminformatics*, vol. 14, no. 1, pp. 1-13, 2022. https://doi.org/10.1186/s13321-022-00594-8

[7] Z. Zhong *et al.*, "Recent advances in deep learning for retrosynthesis," *WIREs Computational Molecular Science*, vol. 13, no. 1, p. e1694, 2023. https://doi.org/10.1002/wcms.1694

[8] M. Meuwly, "Machine learning for chemical reactions," *Chemical Reviews*, vol. 121, no. 16, pp. 10218–10239, 2021. https://doi.org/10.1021/acs.chemrev.1c00033

[9] D. Probst, P. Schwaller, and J. L. Reymond, "Reaction classification and yield prediction using the differential reaction fingerprint DRFP," *Digital Discovery*, vol. 1, no. 2, pp. 91–97, 2022. https://doi.org/10.1039/D1DD00006C

[10] C. Kieffer, A. J. Genot, Y. Rondelez, and G. Gines, "Molecular computation for molecular classification," *Advanced Biology*, vol. 7, no. 3, p. e2200203, 2023. https://doi.org/10.1002/adbi.202200203

[11] L. Chen, "Reaction classification and knowledge acquisition," *Handbook of Chemoinformatics: From Data to Knowledge*, vol. 4, pp. 348-390, 2003. https://doi.org/10.1002/9783527618279.ch12

[12] F. Van den Bergh, M. Van Wyk, B. Van Wyk, and G. Udahemuka, "A comparison of data-driven and model-driven approaches to brightness temperature diurnal cycle interpolation," *SAIEE Africa Research Journal*, vol. 98, no. 3, pp. 81-86, 2021. https://doi.org/10.23919/SAIEE.2007.9488131

[13] K. Jorner, A. Tomberg, C. Bauer, C. Sköld, and P.-O. Norrby, "Organic reactivity from mechanism to machine learning," *Nature Reviews Chemistry*, vol. 5, no. 4, pp. 240-255, 2021. https://doi.org/10.1038/s41570-021-00260-x

[14] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A survey on active deep learning: From model driven to data driven," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1-34, 2022. https://doi.org/10.1145/3510414

[15] X.-B. Jin, R. J. Robert Jeremiah, T.-L. Su, Y.-T. Bai, and J.-L. Kong, "The new trend of state estimation: From model-driven to hybrid-driven methods," *Sensors*, vol. 21, no. 6, p. 2085, 2021. https://doi.org/10.3390/s21062085

[16] Z. Xu and J. Sun, "Model-driven deep-learning," *National Science Review*, vol. 5, no. 1, pp. 22-24, 2018. https://doi.org/10.1093/nsr/nwx099

[17] J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik, "Neural networks for the prediction of organic chemistry reactions," *ACS Central Science*, vol. 2, no. 10, pp. 725-732, 2016. https://doi.org/10.1021/acscentsci.6b00219

[18] J. W. Campbell and K. D. Vogiatzis, "Augmenting MACCS keys with persistent homology fingerprints for protein–ligand binding classificationa," *Journal of Chemical Information and Modeling*, vol. 65, no. 15, pp. 8113-8126, 2025. https://doi.org/10.1021/acs.jcim.5c00934

[19] E. Fernández-de Gortari, C. R. García-Jacas, K. Martinez-Mayorga, and J. L. Medina-Franco, "Database fingerprint (DFP): An approach to represent molecular databases," *Journal of Cheminformatics*, vol. 9, no. 1, p. 9, 2017. https://doi.org/10.1186/s13321-017-0195-1

[20] S. Zhong and X. Guan, "Count-based morgan fingerprint: A more efficient and interpretable molecular representation in developing machine learning-based predictive regression models for water contaminants' activities and properties," *Environmental Science & Technology*, vol. 57, no. 46, pp. 18193-18202, 2023.

[21] L. O. Aman, "Prediction of binding affinity for erbb inhibitors using deep neural network model with morgan fingerprints as features," *arXiv preprint arXiv:2501.05607*, 2025. https://doi.org/10.48550/arXiv.2501.05607

[22] M. Li *et al.*, "Mechanism and kinetics of ClO-mediated degradation of aromatic compounds in aqueous solution: DFT and QSAR studies," *Chemical Engineering Journal*, vol. 412, p. 128728, 2021. https://doi.org/10.1016/j.cej.2021.128728

[23] M. Er-rajy *et al.*, "3D-QSAR studies, molecular docking, molecular dynamic simulation, and admet proprieties of novel pteridinone derivatives as plk1 inhibitors for the treatment of prostate cancer," *Life*, vol. 13, no. 1, p. 127, 2023. https://doi.org/10.3390/life13010127

[24] T. Huang, G. Sun, L. Zhao, N. Zhang, R. Zhong, and Y. Peng, "Quantitative structure-activity relationship (QSAR) studies on the toxic effects of nitroaromatic compounds (NACs): A systematic review," *International Journal of Molecular Sciences*, vol. 22, no. 16, p. 8557, 2021. https://doi.org/10.3390/ijms22168557

[25] P. Schwaller, A. C. Vaucher, T. Laino, and J.-L. Reymond, "Prediction of chemical reaction yields using deep learning," *Machine Learning: Science and Technology*, vol. 2, no. 1, p. 015016, 2021.

[26] A. M. Zuranski, J. I. Martinez Alvarado, B. J. Shields, and A. G. Doyle, "Predicting reaction yields via supervised learning," *Accounts of Chemical Research*, vol. 54, no. 8, pp. 1856-1865, 2021. https://doi.org/10.1021/acs.accounts.0c00770

[27] A. C. Albéniz, "Experimental studies of reaction mechanisms in organometallic chemistry and catalysis," *ACS Publications*, vol. 42, no. 4, pp. 285-287, 2023. https://doi.org/10.1021/acs.organomet.3c00072

[28] W. Yang, T. T. Fidelis, and W.-H. Sun, "Machine learning in catalysis, from proposal to practicing," *ACS Omega*, vol. 5, no. 1, pp. 83-88, 2019. https://doi.org/10.1021/acsomega.9b03673