Andrew Crow, Andrew Hansell, Miko Miwa, Ella Pickell, Paul Williams
CHBE413 / CHEM452 / CHEM590
December 4th, 2025

# Uncovering Hidden Patterns in Catalyst Heterogeneity Using Dimensionality Reduction

## Abstract

The creation of increasingly complex chemical datasets has driven a need for dimensionality reduction techniques that are able to reveal structure-property relationships in molecular systems. Prior work by the Peters group has demonstrated that kernel principal covariates regression (KPCovR) can identify modes that govern the catalytic behavior but lack a physical interpretation. Therefore in this report, we aimed to replicate and possibly extend that analysis using various unsupervised and supervised dimensionality reduction techniques. Methods such as linear principal component analysis (PCA), kernel PCA, t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) were applied to both the original geometric descriptions and a different set of features centered on the site environment.

We evaluated each method's ability to represent structural variation and if there are any correlations with the catalytic rates that were generated from density functional theory (DFT) trained machine learning (ML) models. Linear PCA identified the significant modes of geometric distortion but did not map clearly onto catalytic activity. Kernel PCA captured some nonlinear separation that was consistent with observations made in previous work, but it still overall had very low interpretability. However, t-SNE and UMAP created low-dimensional plots that had no visual relationship to the catalytic performance despite extensive hyperparameter exploration. Overall, the KPCovR method proved to continue to provide the best relationship to the structure-reactivity relationships as opposed to the other techniques examined. The results highlight the importance of supervised dimensionality reduction and demonstrate that improved featurization is helpful for interpretable chemical dimensionality reduction.

## 1. Introduction and Background

### *1.1 Chemical data analysis*

The expansion of data-driven research in chemistry has created a need for analytical methods capable of handling increasingly large and high dimensional datasets. While certain experimental measurements, like fluorescence intensities or reaction kinetics, can often be analyzed using classical statistical tools such as linear regression, many emerging chemical data problems exceed the capability of traditional techniques. Predicting molecular properties from structural components, relationships among functional groups, or three-dimensional molecular arrangements often requires more sophisticated computational strategies. Machine learning (ML) has therefore become an important tool for identifying structure-property relationships in modern chemical research.

A central challenge in applying ML to chemistry is that chemical inputs are rarely numerical; molecular structures, bonding patterns, and spatial geometry must first be translated into numerical representations. This translation, known as featurization, converts chemical information into vectors. Effective featurization requires careful design to ensure that the numerical vectors retain the chemically relevant information from the original representation [1]. For example, featurizing a molecular structure often requires encoding atomic identities, bond orders, bond lengths, and angular relationships. The choice of features depends strongly on the property or behavior being predicted. Once featurized, these numerical vectors can then be processed by ML algorithms to learn predictive relationships.

ML algorithms generally learn from data using either supervised or unsupervised approaches. In supervised learning, models are trained on labeled data to predict an output variable from a defined set of inputs [2]. This framework encompasses both regression (continuous outputs) and classification (categorical outputs). For example, supervised learning can be used to predict molecular reactivity from structural or compositional descriptors. Supervised learning is effective when specific quantitative or qualitative outputs are known, whereas unsupervised learning is suited to exploring unknown structure, trends, or groupings within unlabeled data. Clustering techniques, for instance, group chemically similar observations and may expose relationships useful for downstream prediction. Molecular clusters may form based on polarity, electronegativity, or structural motifs. Dimensionality reduction is another central unsupervised strategy. These methods compress high-dimensional feature vectors into lower-dimensional representations while retaining the essential structural patterns and variance of the data. Dimensionality reduction is critical in chemical ML because it mitigates overfitting, improves computational efficiency, and often enhances model robustness and precision [3]. Reducing the dimensionality helps maintain a favorable sample-to-feature ratio, reducing the risk of overfitting. Eliminating irrelevant or noisy features also decreases computational cost and can improve predictive performance. Dimensionality reduction strategies include feature-selection methods (where redundant features are removed), linear transformations (where features are linearly transformed into more useful versions), and nonlinear manifold-learning approaches (where the features are projected onto a lower dimension manifold) [4].

A variety of dimensionality reduction algorithms have been developed, each offering distinct advantages and limitations. Principal Component Analysis (PCA) is among the most widely used linear methods and involves determining the principal components of the data based on the highest variance features of the data and creating a new coordinate system based on them [5]. Because PCA is limited to linear structure, Kernel PCA extends the method to nonlinear relationships through kernel mappings. Kernel functions map data into a higher-dimensional space, enabling PCA to capture curved or nonlinear manifolds [4]. More recent nonlinear approaches, such as t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP), provide powerful tools for visualizing complex chemical datasets in two or three dimensions [4]. Because no single method is universally optimal, applying multiple dimensionality-reduction techniques can aid interpretability and validate confidence in the resulting chemical insights.

In this work, we apply several dimensionality reduction techniques to a chemical dataset taken from the literature and evaluate their ability to reveal activity based on molecular structure. The paper this project replicates and expands upon uses Kernel Principle Covariates Regression (KPCovR) for its dimensionality reduction [6,7]. This method is different from the aforementioned ones as it takes into account information from both the features as well as the target data. This method is particularly useful for this application as the validity of the component analysis is dependent on the ability to accurately explain the change in chemical properties shown in corresponding data.

### *1.2 The original study*

In order to understand the relevance and validity of these PCA methods, a summary of the original paper is in order [7]. Industrially, the practice of using single atom metal catalysts on amorphous silica supports is common. These systems, however, are difficult to construct accurate kinetic models for, as each metal atom has a different local environment due to the random nature of the support, leading to differing catalytic properties. How the catalytic properties of a specific site geometry affect the bulk properties is dependent on the probability that the specific site geometry is formed in the metal atom grafting process, as metal atom geometries that are more unlikely to be formed will have less of an impact on the bulk properties and vice versa. Both of these properties are directly related to the geometry of the silanol binding site that the metal binds to (Figure 1). The paper focuses on modeling an ensemble of these different site geometries for the Phillips catalyst, which uses chromium metal atoms dispersed on silica to perform ethylene polymerization. 388 silanol site geometries were produced using a simulated slab of silica, and a portion of these clusters had Density Functional Theory (DFT) calculations performed to calculate catalytic rates and grafting rates. A machine learning model to predict these DFT results was trained on features that specified the position of the fluorine capping atoms, as seen in Figure 2, and the model was used to predict catalytic and grafting rates for the rest of the sites to save computational time. The bulk properties of the ensemble of these sites was calculated with a population balance model and the results were verified via comparison to experimental data.

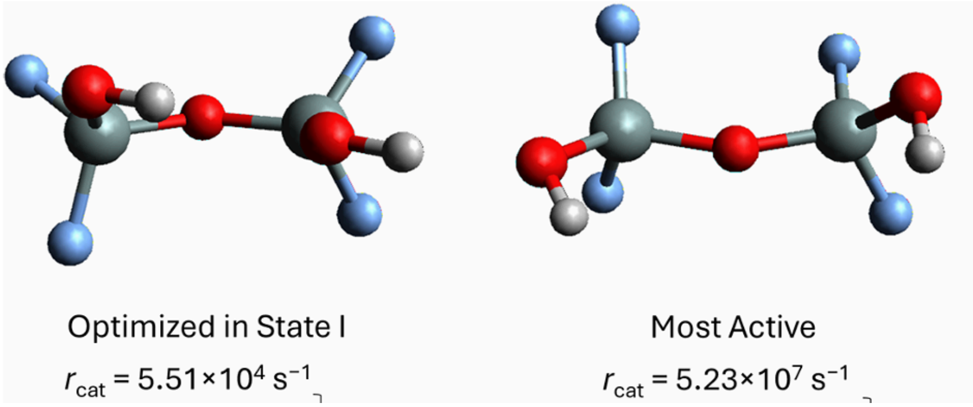
Optimized in State I
$r_{cat}$ = 5.51×10$^4$ s$^{-1}$
Most Active
$r_{cat}$ = 5.23×10$^7$ s$^{-1}$

**Figure 1:** 2 Silanol binding sites that chromium atoms can bind to. Si is shown in grey, O in red, H in white and F atoms that cap the cluster in light blue. Displayed below each molecule is the catalytic rate of the resulting single chromium atom site as calculated by a machine learning algorithm trained on DFT results. The left site is the DFT optimized geometry, and the right site is the geometry with the highest associated reactivity.

The paper then used the six features previously used to predict catalytic and grafting rates in the PCA. These features describe the position of the capping fluorine atoms, which were used to approximate a larger cluster size (Figure 2a). For these and future featurizations, where a pair of symmetrical angles or lengths are seen, e.g. $r_1$ and $r_2$ or $\theta_1$ and $\theta_2$, the corresponding features used in the PCA analysis are the sum and difference between the two measurements as done in the original paper. The original paper used KPCovR analysis to determine the principle components of the system, and came to the conclusion that the reactivity (rcat) of the catalytic site is determined by only two principle components, and thus that there are two modes of geometric change in the site that determine the reactivity of a site. Due to the methods used, however, a physical interpretation of these geometric changes was unable to be gleaned from the PCA. This is the main challenge that this project seeks to address, gaining a more interpretable principle component analysis through the use of more interpretable models, as well as featurizations that capture a more intuitive structure of the atoms that directly form the catalytic site.

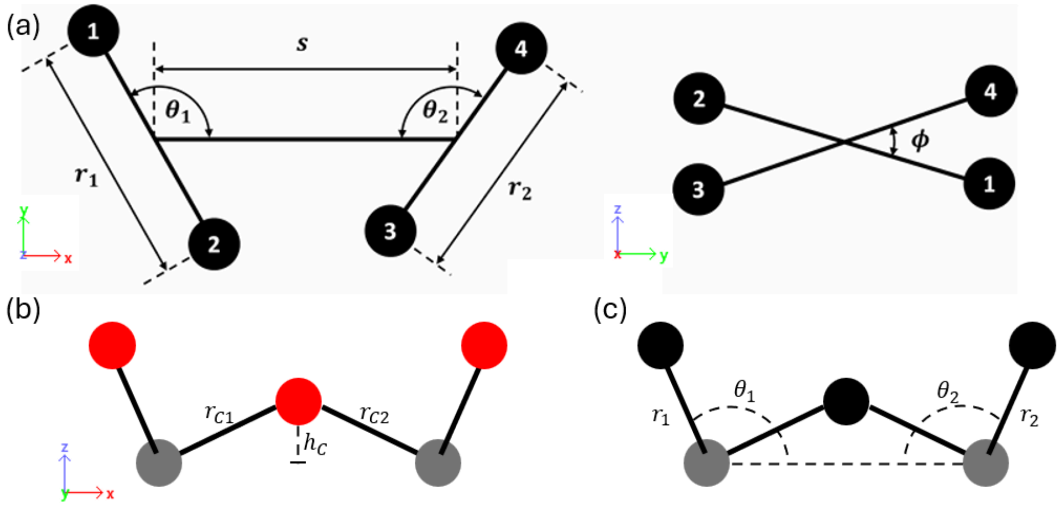

**Figure 2:** (a) Diagram of the features describing the relative position of fluorine capping atoms 1-4. (b) Diagram of the features describing the position of the central oxygen atom at the silanol site created as part of the project. (c) Diagram of the features describing the position of the other atoms (the two fluorine pairs and the one oxygen pair) connected to the two silicon atoms in the site.

### *1.3 Objective*

The objective of using this ensemble of dimensionality reduction methods is twofold. The first is to investigate how well methods that do not take into account target data like KPCovR can explain the calculated catalytic reactivity data. If one or more methods are able to capture the behavior of the reactivity data, then seeing if they agree with each other and the original KPCovR analysis from the paper would help validate the models. The second objective is gaining a better interpretation of the PCA through a more interpretable method, such as linear PCA or kernel PCA. If these methods improve with the second featurization created as part of this project, it would indicate that even further feature engineering could yield more improvements in application to the calculated reactivity data.

## 2. Method

### *2.1 Feaurization*

To simplify the scope of the project, only the catalytic rate was used as a target when evaluating the dimensionality reduction performed. The papers usage of the six features describing the position of the fluorine atoms as the features for PCA, although fully describing the geometry of the site, does not lend itself to interpretability as these capping fluorine atoms are two atoms away from the oxygen atoms pointing out of the silanol site. In order to give the simpler more interpretable dimensionality reduction methods such as linear PCA and kernel PCA a better chance at finding a more direct correlation between geometric features and this catalytic rate, new features were crafted centered around the silicon atoms that fully describe the site's geometry (Figure 2b, Figure 2c). Not shown in the figure are the torsion angles also used as features, not unlike the torsion angle Φ in the original featurization. These torsion angles are the offset of the outward facing left and right oxygen from the plane of the central oxygen and the two silicon atoms in Figure 2b, and the offset of the outward facing atoms (these being either one of the two pairs of fluorine atoms, or the pair of oxygen atoms) relative to each other. A key goal of the featurization created was to allow a simple linear combination of the features to describe the characteristics of the site that the paper qualitatively identified as being correlated to the catalytic rate. These characteristics can be seen in Figure 1, being how far the central oxygen atom sticks out normal to the slab of silica, and how much the OH groups on the silicon atom are sticking away from each other. Within the new featurization, these two hypothesized geometric characteristics that influence reactivity can be easily expressed as the sum of angles $\theta_1$ and $\theta_2$ for the OH groups, and the height of the central oxygen atom in the z direction, $h_c$.

*2.2 PCA and Kernel PCA*

PCA reduces higher dimensionality of structural datasets and isolates variation within a molecular system. When applied to geometric features of a molecule such as the axle widths of the two silanol anchoring groups, the shaft length separating these groups, the angles of these axels (theta 1 and theta 2), and the torsion angle (phi), PCA identifies molecular features that account for the largest variance in site structures. PCA uses a covariance matrix of the standardized descriptors and uses eigenvalues to see how much variance is explained by each principal component [8]. The eigenvalues quantify the magnitude of variation along the corresponding principal component, and the explained variance follows from the ratio of a given eigenvalue to the total variance [9]. In doing so, PCA compresses multidimensional data describing molecular structure into a limited set of linear modes that give an overview of the geometric patterns in the system [8]. In catalytic and surface-chemistry contexts, linear modes often reflect cooperative distortions that influence the local environment of active sites like torsional strain [7]. However, PCA restricts its representation to linear combinations of descriptors, so it cannot effectively describe nonlinear relationships within a catalytic site's local environment.

Kernel PCA addresses this limitation of linear PCA by introducing nonlinearity by directly mapping the molecule's geometric features into a higher dimensionality feature space [10]. Instead of directly using the covariance matrix of the axle widths, shaft length, axle angles, and torsion angle, Kernel PCA constructs a kernel matrix that uses a radial basis function to encode similar features as pairs in an expanded feature space. Taking apart the centered kernel matrix captures the eigenvalues and eigenvectors that define the nonlinear structures (principal components) in the data [11]. Kernel PCA uses a different mathematical framework than linear PCA, but the sizes of its eigenvalues still show the significance of each nonlinear principal component. When these values are normalized, they are analogous to the explained variance in linear PCA. Through this, Kernel PCA visualizes curved or folded structural manifolds that linear PCA fails to detect and visualize. In catalytic systems where axle-angle asymmetry, torsional rotations, and shaft compression interact nonlinearly to shape the molecules local reactive environment, Kernel PCA is a more accurate representation of the molecular geometry while detecting and visualizing the nonlinear structure-function correlations.

### *2.3 t-SNE and UMAP*

t-SNE is a commonly used type of nonlinear dimensionality reduction [12]. Like PCA, t-SNE uses data sets with many features as an input and maps them to a new lower dimensional coordinate system. However, t-SNE does a better job at preserving local neighborhoods of data points.

First, t-SNE takes a point from the data set and computes the similarities to nearby points using a Gaussian probability distribution. The size of the Gaussian probability depends on the "perplexity" which is one of the hyperparameter inputs to the method. Next, the algorithm will assign all the data points to a random distribution of points in a 2-dimensional plane, before iterating through and moving the points based on their higher-dimensional relationships. It adjusts the lower dimensional points using gradient descent. Afterwards, the algorithm computes the proximity probabilities for the low dimensional data using a Student t-distribution. Finally, t-SNE compares the two probabilities and tries to get them to match. Hopefully, clusters start to appear in the lower dimensional data so that some importances can be extracted.

However, t-SNE fails to capture several parts of the higher dimensional space. t-SNE is local neighborhood based and does not include any global structures that could exist. t-SNE also cannot embed new points in the lower dimensional plane without rerunning the algorithm which makes it much slower when compared to other nonlinear dimensionality reduction techniques. Therefore, other methods were investigated as potential comparisons to t-SNE.

In contrast to t-SNE, UMAP is a different nonlinear dimensionality reduction method that will preserve global structures present in the data [13]. UMAP is more commonly used for larger data sets given its faster and more efficient use of data points. First, UMAP will consider the k nearest neighbors for every point in the data set. Then, it will add a weight to each point based on how related each pair of points are. For these processes, UMAP utilizes a local distance function and a smooth exponential kernel. Next, UMAP initializes the points in the lower dimensional space using spectral embedding which is more stable than t-SNE's random or PCA-based initialization. Finally, UMAP manipulates the lower dimensional graph via the edge weights to match what was learned in the higher dimensional space. It does this by minimizing a cross-entropy between the plots using stochastic gradient descent instead of the plain gradient descent that is used in t-SNE. Therefore, the final plot will be displayed that will hopefully have a level of interpretability for useful information to be extracted.

### *2.4 KPCovR*

KPCovR expands the standard PCA by interpolating between the results of Kernel Principal Covariates Analysis and Kernel Ridge Regression (KRR). This interpolation is controlled by the hyperparameter alpha, with an alpha of one corresponding to KPCA and an alpha of zero corresponding to KRR [10]. Kernel Principal Covariates Regression (KPCovR) uses dimensionality reduction with supervised learning to visualize correlations [8]. Whereas PCA and Kernel PCA focus on characterizing geometric variation, KPCovR uses supervised learning to balance structural reconstruction and target feature prediction, such as the catalytic rate [9]. KPCovR does this by using Kernel PCA's reconstruction of the data with the predictive goal of kernel ridge regression, the method creates new variables (principal covariates) that represent both how the structures vary and how they relate to the target feature [8]. By projecting the axle widths, shaft length, axle angles, and torsion angle into a feature space that uses both reconstruction and regression, KPCovR reveals the geometric directions most strongly associated with activity or selectivity. Recent studies of heterogeneous catalytic sites show that this method often uncovers a low-dimensional space where only a few linked structural changes control most of the structure–reactivity behavior. In the ethylene polymerization study by Wimalasiri *et al.*, two principal covariates captured the majority of variation in both grafting propensity and catalytic rate, which emphasizes the value of KPCovR as a tool for mechanistic interpretation in complex structural environment [14].

## 3. Results and Discussion

### *3.1 PCA and Kernel PCA*

The linear PCA analysis of the dataset's significant features, as described by the authors, indicates the primary linear trends that arise from structural variation in the axle widths, the shaft length, the axle angles, and the torsion angle. The first components visualize correlated changes in axle divergence and the overall open conformations of the reaction sites, which define the structural landscape described in the peripheral atom coordinate system of Kim *et. al.* [7].

Kernel PCA applied to the same dataset of significant features shows nonlinear structural variation. The nonlinear mapping contributions from torsional displacement, axle-angle

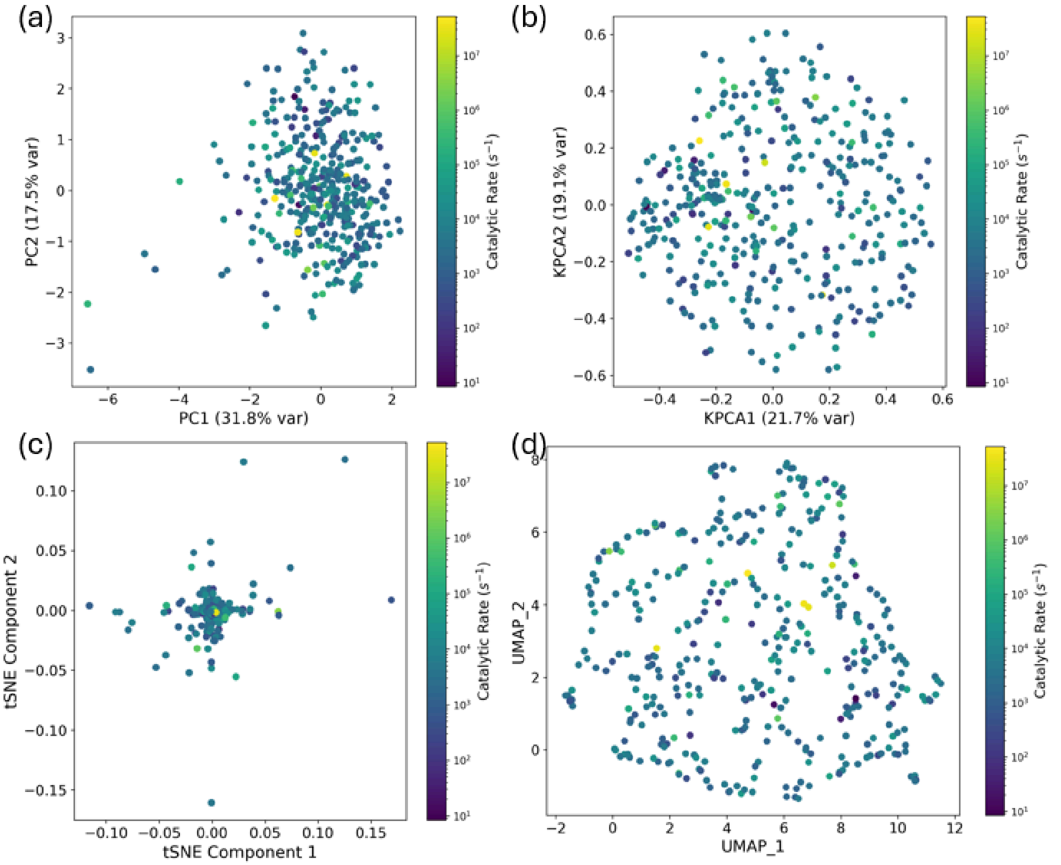

**Figure 3:** Results for the original papers featurization colorized by catalytic rates for (a) linear PCA results, (b) kernel PCA results, (c) t-SNE results, and (d) UMAP results.

asymmetry, and shaft compression, which are molecular features that Kim *et. al.* use as features to describe catalytic activity [7]. The Kernel PCA plots where the most active structures separate to be more clearly defined from the rest, which agrees with the nonlinear way the reaction barrier responds to combined geometric distortions. The pattern of eigenvalues shows that only a few nonlinear features shape the overall structural space, which supports Kim *et. al.*'s conclusion that only a small set of linked distortions largely controls reactivity [7].

Figures 3a-b and 4a-b show how the structural features map onto the catalytic rate across the reduced dimensional spaces. In the linear PCA plots, the distribution of points displays no obvious trend or clustering that aligns with the catalytic rate, and the color gradient forms a largely diffuse pattern across the axes. Kernel PCA provides a slightly more organized separation, regions of higher activity appearing more concentrated, but the overall structure still lacks definitions that would directly explain the reactivity trends. Given the limited interpretability of these visualizations relative to the authors' results, neither feature dataset fully captures the structure–reactivity relationship that the study by Kim *et. al.* emphasizes [7].

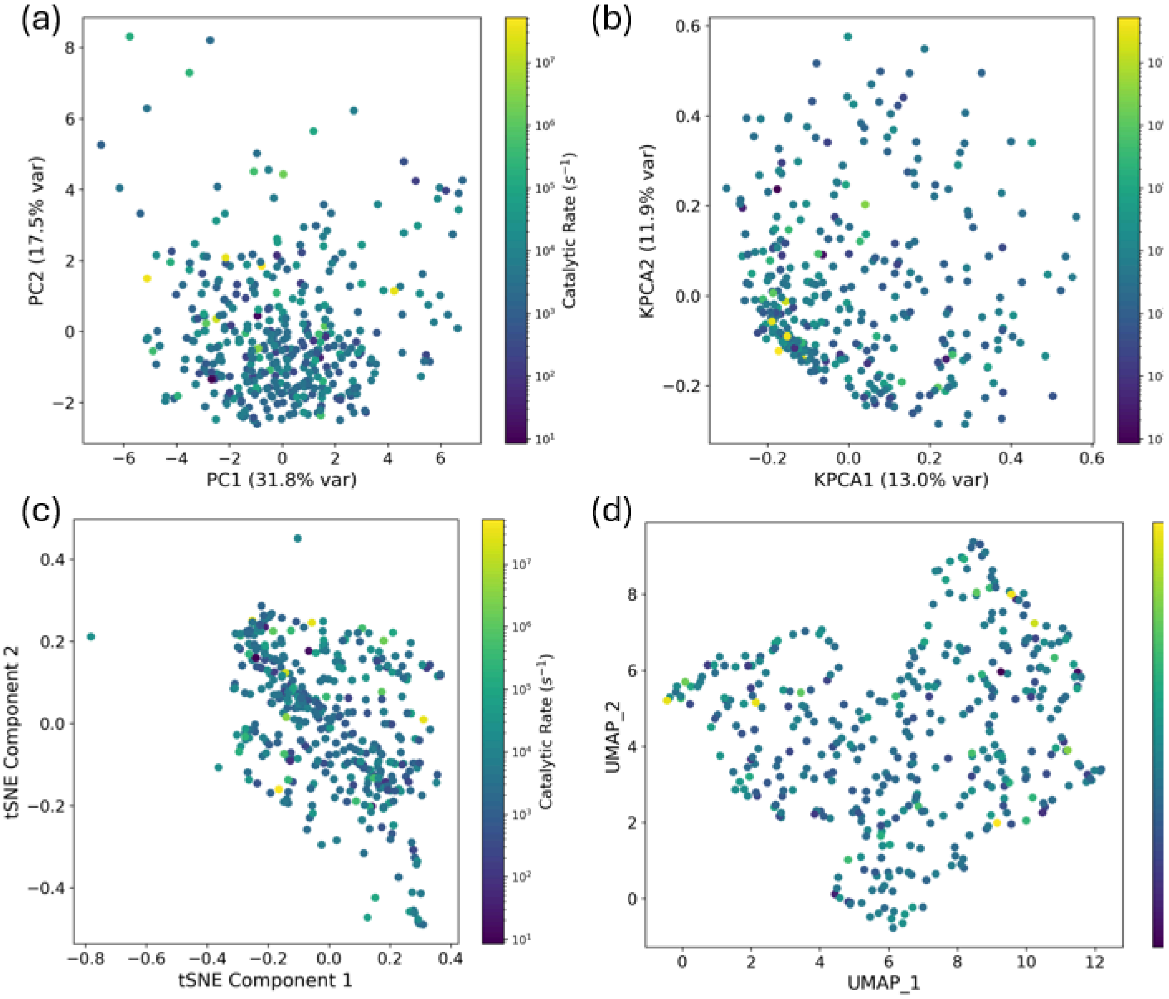

**Figure 4:** Results for this project's featurization colored by catalytic rates for (a) linear PCA results, (b) kernel PCA results, (c) t-SNE PCA results, and (d) UMAP PCA results.

### *3.2 t-SNE and UMAP*

For both t-SNE and UMAP, the same features and targets were used as for the other methods. All the features were appropriately scaled prior to the application of the methods. Several different hyperparameters were manipulated in attempts to generate plots with better interpretabilities. The learning rate, perplexity, max iterations, and initialization method were all changed while investigating t-SNE and the number of neighbors, minimum distance, and metric were all changed for UMAP.

Figures 3c-d and 4c-d show the best plots for t-SNE and UMAP mapped back to the catalytic rate target variable over a logged color scale. Both of the plots generated seem to have no visual trends that could potentially explain the target. Given the low interpretability of these plots, it was determined that neither t-SNE nor UMAP are acceptable alternatives to the techniques investigated by the authors. Some reasons why both t-SNE and UMAP fail to provide some level of correlation to the catalytic rates could be due to the hyperparameters distorting the meaningful relationships, the distances between points are not meaningful, or that the input data simply has no structure. Another possible reason is that our plots are meaningful but not to the targets we are interested in.

### *3.3 KPCovR*

KPCovR uses geometric reconstruction with prediction of catalytic rates, and matches the dimensionality reduction strategy used by Kim *et al* in Fig. 5a [7].

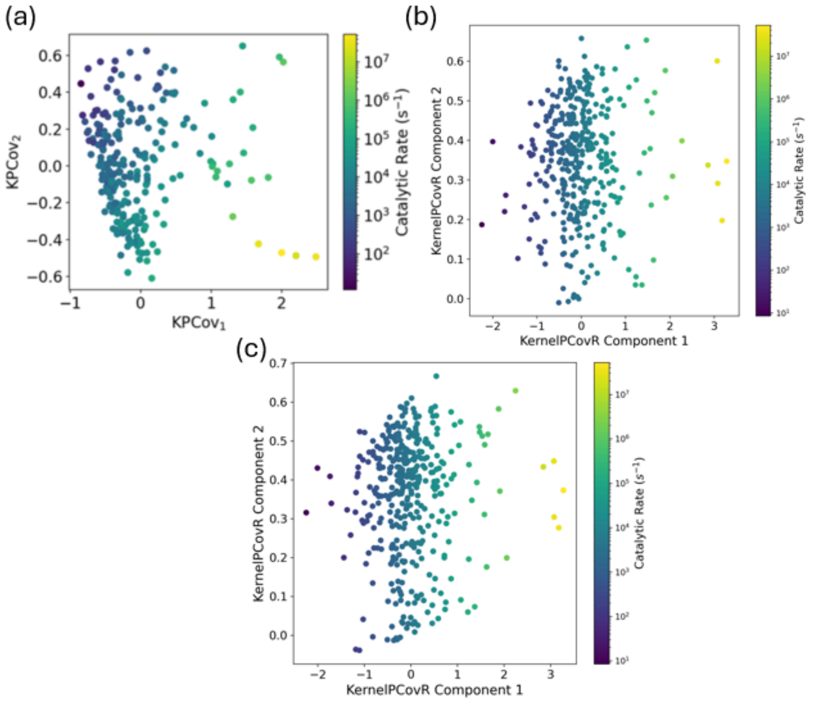

**Figure 5:** Comparison of KPCovR results of (a) the original paper, (b) our attempt using original featurization, (c) our attempt using this project's featurization

The approach identifies structural directions that capture variation in the features while also accounting for changes in catalytic rate. The two-dimensional maps in Figs. 5b-c approximately reproduces the trend reported by Kim *et. al.* in that high-activity sites gather on one side of the covariate space (left), and low-activity sites cluster on the other (right) [7]. KPCovR fills the gap in correlation to the catalytic data left by linear PCA and kernel PCA. Unlike PCA, which isolates only the directions of greatest linear variance, and unlike kernel PCA, which exposes nonlinear structure but remains blind to the target variable, KPCovR aligns the features with the catalytic rate itself. This alignment allows the method to highlight the specific combinations of geometric distortions that drive changes in catalytic activity rather than those that simply account for structural diversity. Together, these changes stabilize or destabilize key states along the reaction pathway, which explains why they dominate the structure reactivity relationship. By isolating these mechanistically meaningful combinations, KPCovR provides a map of how the molecular structural features impact the catalytic rate to reflect the structure reactivity trends noted by Kim *et. al.* [7]. This results still are not able to provide an interpretable geometric interpretation of these two modes of change.

## Conclusion

While utilizing multiple dimensionality reduction techniques is useful for increasing the probability that the data is interpretable, it does not ensure interpretability if none of the methods capture trends in the reactivity. After analyzing the dataset with Linear PCA, Kernel PCA, t-SNE, and UMAP, we were unable to uncover new or consistent structure-reactivity relationships, even when using an expanded and more chemically intuitive featurization. This comparison highlights the importance of thoughtful featurization and supervised dimensionality reduction when seeking interpretable chemical structure-property relationships.

References

[1] Jackson, N. *Molecular Featurization: Machine Learning Meets Molecules*. UIUC CHEM452 – Data Science for Chemistry and Engineering, Champaign, IL, **2025**.

[2] Jackson, N. *ML Overview, Linear Regression*. UIUC CHEM452 – Data Science for Chemistry and Engineering, Champaign, IL, **2025**.

[3] Prince, S. J. D. *Understanding Deep Learning*; MIT Press, **2023**. https://udlbook.github.io/udlbook/.

[4] Jackson, N. *Molecular Featurization: Unsupervised Learning - Dimensionality Reduction*. UIUC CHEM452 – Data Science for Chemistry and Engineering, Champaign, IL, **2025**.

[5] James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. *An Introduction to Statistical Learning with Applications in Python*; Springer, **2023**. https://www.statlearning.com/.

[6] Helfrecht, B. A.; Cersonsky, R. K.; Guillaume Fraux; Ceriotti, M. Structure-Property Maps with Kernel Principal Covariates Regression. *Machine Learning Science and Technology* **2020**, *1* (4), 045021–045021. https://doi.org/10.1088/2632-2153/aba9ef.

[7] Kim, C. A.; Shayesteh Zadeh, A.; Peters, B. Ethylene Polymerization Activity vs Grafting Affinity Trade-off Revealed by Importance Learning Analysis of In Silico Phillips Catalyst. *The Journal of Physical Chemistry C* **2024** *128* (45), 19166-19181 10.1021/acs.jpcc.4c05331

[8] Jolliffe, I. T. *Principal Component Analysis*; Springer: New York, 2002.

[9] Lever, J.; Krzywinski, M.; Altman, N. Points of Significance: Principal Component Analysis. *Nat. Methods* **2017**, *14*, 641–642. https://doi.org/10.1038/nmeth.4346.

[10] Abdi, H.; Williams, L. J. Principal Component Analysis. *Wiley Interdiscip. Rev.: Comput. Stat.* **2010**, *2*, 433–459. https://doi.org/10.1002/wics.101.

[11] Schölkopf, B.; Smola, A.; Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **1998**, *10*, 1299–1319. 10.1162/089976698300017467

[12] van der Maaten, L.J.P.; Hinton, G.E. (Nov **2008**). "Visualizing Data Using t-SNE" (PDF). *Journal of Machine Learning Research*. 9: 2579–2605

[13] McInnes, L.; Healy, J.; Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* **2018**. https://doi.org/10.48550/arXiv.1802.03426.

[14] Wimalasiri, P. N.; Nguyen, N. P.; Senanayake, H. S.; Laird, B. B.; Thompson, W. H. *Amorphous Silica Slab Models with Variable Surface Roughness and Silanol Density for Use in Simulations of Dynamics and Catalysis.* J. Phys. Chem. C 2021, *125*, 23418–23434. https://doi.org/10.1021/acs.jpcc.1c06580