Science Advances
AAAS

# Supplementary Note 3
(Details of Dimensional Reduction Analysis)
for

## Leveraging Data Mining, Active Learning, and Domain Adaptation in a Multi-Stage, Machine Learning-Driven Approach for the Efficient Discovery of Advanced Acidic Oxygen Evolution Electrocatalysts

Rui Ding *et al.*

*Corresponding author. Email: junhongchen@uchicago.edu, chenyuxin@uchicago.edu

## Supplementary Note Discussion SND 3-1
## Dimensional Reduction Analysis on the Domain Knowledge Datasets

PCA(*1*) is a powerful dimensionality reduction technique that has been widely adopted across various domains, from data science and machine learning to computer vision and bioinformatics. At its core, PCA aims to transform high-dimensional data into a lower-dimensional space while preserving the most important information and patterns within the data. The underlying principle of PCA is to identify the directions of maximum variance within the data, known as the principal components. These principal components are essentially the eigenvectors of the covariance matrix of the data, sorted in descending order of their corresponding eigenvalues. By projecting the data onto the subspace spanned by the top principal components, PCA can effectively capture the most significant sources of variation, effectively reducing the dimensionality of the data without compromising the essential characteristics. One of the key advantages of PCA is its ability to remove noise and redundant features from the data, making it a crucial preprocessing step for many machine learning algorithms. By reducing the dimensionality of the data, PCA can also improve the computational efficiency of downstream analyses, as well as facilitate better data visualization and interpretation. Additionally, PCA can be used to identify the most influential features or variables within a dataset, which can provide valuable insights for domain experts and decision-makers.

In contrast, t-SNE(*2*) is a nonlinear dimensionality reduction technique that excels at visualizing high-dimensional data in a low-dimensional space, typically two or three dimensions. Unlike PCA, which focuses on preserving the global structure of the data, t-SNE is primarily concerned with preserving the local structure, ensuring that similar data points in the high-dimensional space are represented by nearby points in the low-dimensional embedding. The t-SNE algorithm works by first computing the pairwise similarity between data points in the high-dimensional space, using a Gaussian kernel to model the probability distribution. It then seeks to find a low-dimensional representation of the data where the pairwise similarities are preserved as accurately as possible, using a student's t-distribution to model the similarity in the low-dimensional space. By minimizing the Kullback-Leibler divergence between these two probability distributions, t-SNE is able to generate a faithful low-dimensional projection of the data, often revealing intricate structures and patterns that might not be easily discernible in the original high-dimensional space. The power of t-SNE lies in its ability to capture both global and local structures within the data, making it particularly useful for exploratory data analysis, cluster identification, and data visualization. It has been widely adopted in various fields, including bioinformatics, image processing, and natural language processing, where the visualization of high-dimensional data can provide valuable insights and aid in the understanding of complex relationships within the data.

Overall, PCA and t-SNE are complementary dimensionality reduction techniques that play crucial roles in the field of data science and machine learning, each with its own unique strengths and applications. By leveraging these powerful tools, researchers and practitioners can gain a deeper understanding of their data, uncover hidden patterns, and ultimately make more informed decisions.

Building upon our preliminary unsupervised data mining results, we employed these techniques to further examine our dataset. This examination revealed a significant overlap between positive and negative samples across various thresholds identified in our analysis by dimensional reduction of our domain knowledge dataset from high dimension to two-dimensional space as shown in **Fig. SN 3-1-Fig. SN 3-2**. Such an overlap is indicative of the complexity and nuanced nature of our dataset, suggesting that simpler ML methods might fall short in providing accurate and detailed predictions. This is particularly evident in the Silhouette Coefficients and Davies-Bouldin Index values obtained from our analysis (**Table SN 3-1**). The Silhouette Coefficients were supposed to be values close to +1, but were found notably low by the PCA and t-SNE results, not exceeding +0.2 and occasionally entering negative territory, indicating poorly defined clusters with significant overlap. Meanwhile, the Davies-Bouldin Index values were notably high, as they were supposed to be values less than +1 and closer to 0. This has further confirmed the lack of distinct, well-separated clusters in our data. These metrics underscore the challenges in analyzing and interpreting our dataset, which is characterized by high-dimensional complexity and indistinct separability of samples. Therefore, these findings underscore the necessity of employing more sophisticated ML strategies in our subsequent analysis. Advanced ML approaches are essential to navigate the intricacies of our multifaceted dataset effectively, enabling us to discern the underlying patterns and relationships with greater precision and nuance.

The scripts of this part are stored in the GitHub repository in the directory “/PCA and t-SNE Dimensional Reduction Representation”.

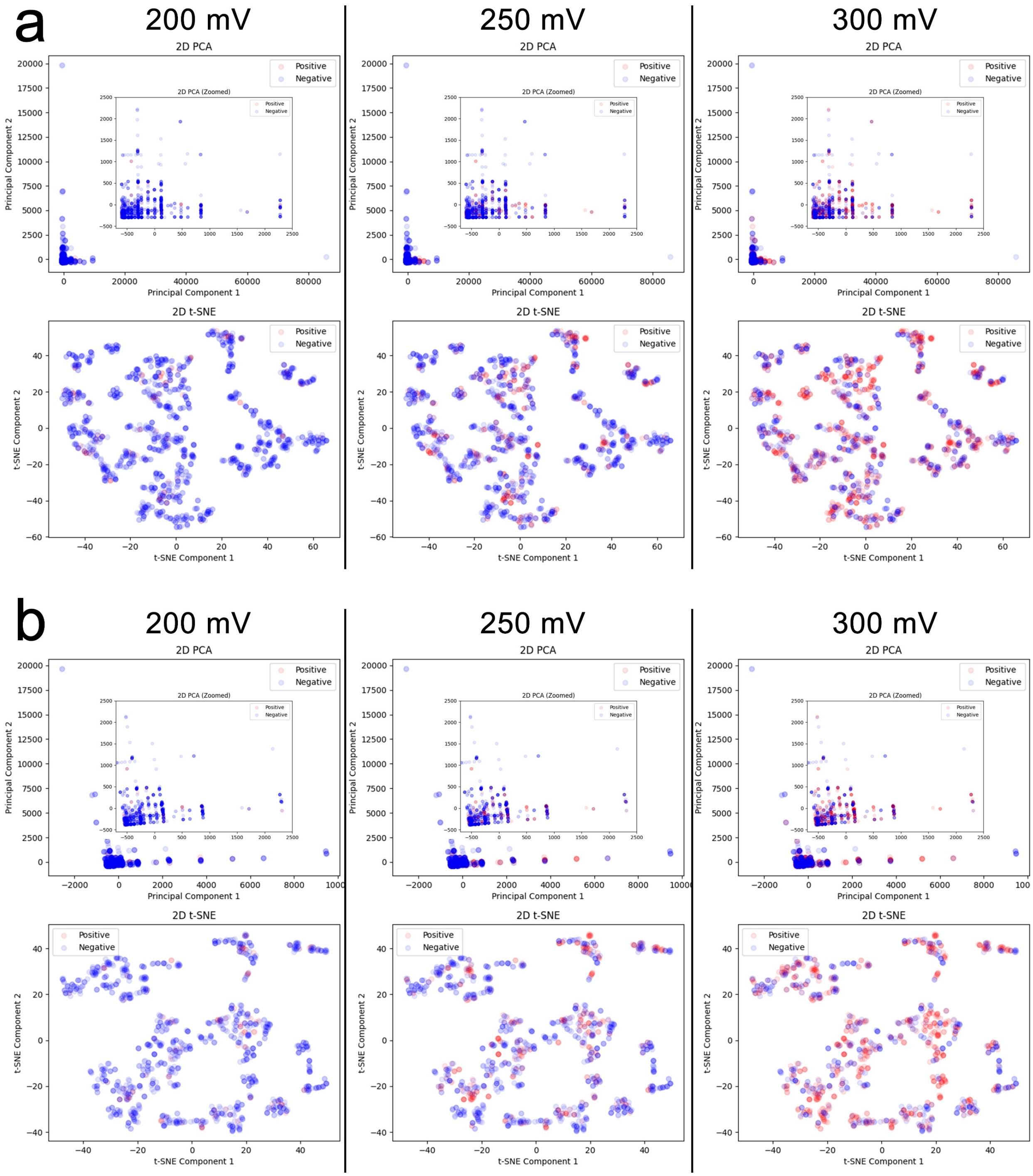

**Fig. SN 3-1** The two-dimensional distribution maps after PCA and t-SNE dimensional reduction transformation of the (a) full and (b) high-quality domain knowledge dataset, under different thresholds previously mentioned for judging the activity of OER electrocatalysts.

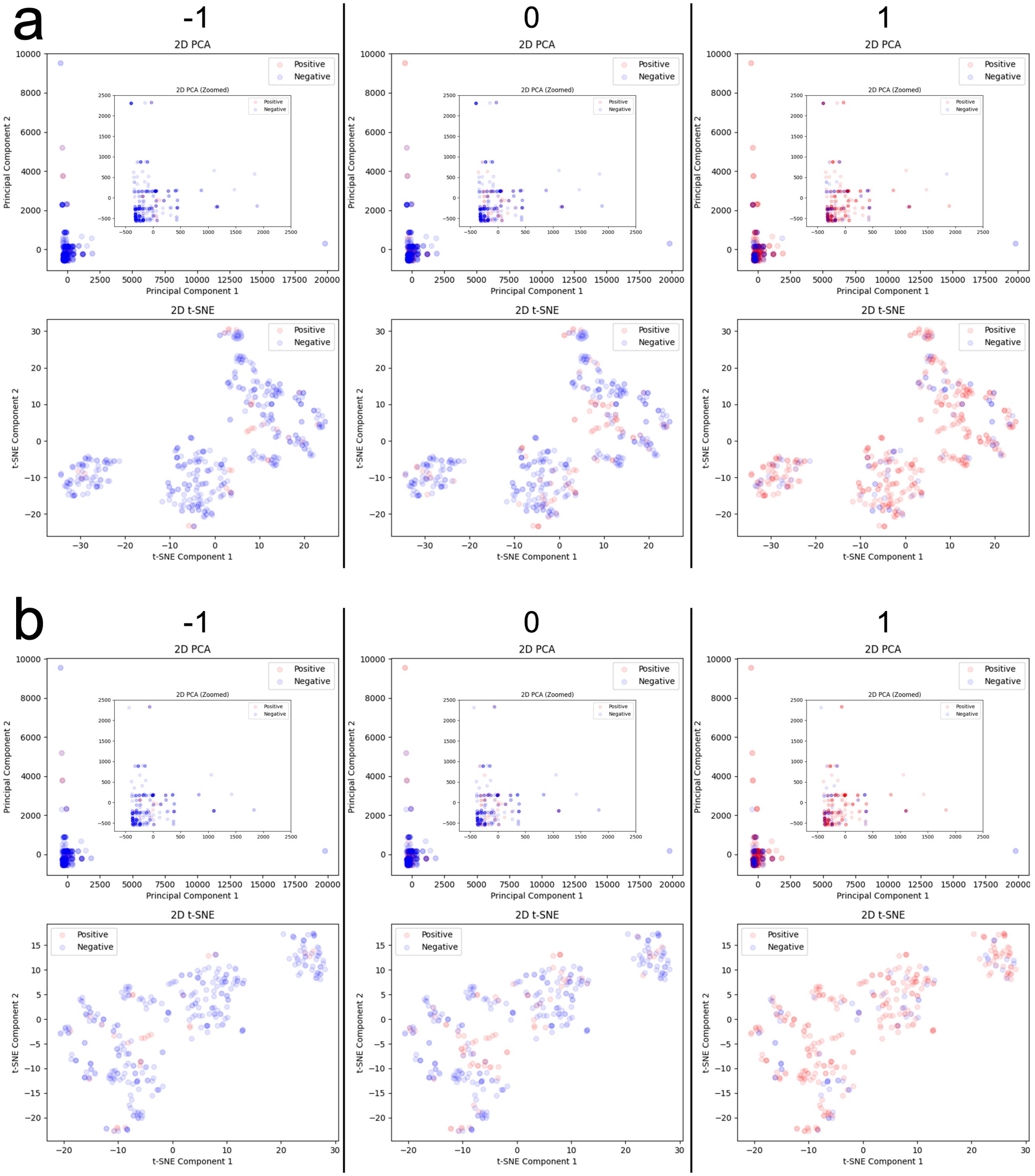

**Fig. SN 3-2** The two-dimensional distribution maps after PCA and t-SNE dimensional reduction transformation of the (a) full and (b) high-quality domain knowledge dataset, under different thresholds previously mentioned for judging the stability of OER electrocatalysts.

**Table SN 3-1** Summary of the Silhouette Coefficient and Davies-Bouldin Index under different thresholds by PCA and t-SNE corresponding to **Fig. SN 3-1-Fig. SN 3-2**.

| | **Activity Full Dataset (mV)** | | | **Activity High-Quality Dataset (mV)** | | | **Stability Full Dataset (mV h$^{-1}$)** | | | **Stability High-Quality Dataset (mV h$^{-1}$)** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Thresholds | 200 | 250 | 300 | 200 | 250 | 300 | -1 | 0 | 1 | -1 | 0 | 1 |
| PCA Silhouette Coefficient | 0.025 | 0 | 0.02 | -0.02 | 0.01 | 0.04 | 0.2 | 0.09 | 0.18 | 0.2 | 0.11 | 0.2 |
| PCA Davies-Bouldin Index | 8.986 | 6.73 | 10.8 | 9.1 | 5.14 | 6.89 | 4.6 | 5.11 | 4 | 3.7 | 3.79 | 3.3 |
| t-SNE Silhouette Coefficient | 0.007 | 0 | 0 | -0 | 0 | 0.01 | -0 | -0 | 0.01 | -0 | -0 | 0 |
| t-SNE Davies-Bouldin Index | 7.836 | 12.9 | 30.9 | 9 | 10.5 | 61 | 6.4 | 9.71 | 21.2 | 5.7 | 7.68 | 16 |

## Supplementary Note Discussion SND 3-2

## t-SNE Analysis of Evolvement of the Active Learning Process

### Methodology

In this study, we also employed t-SNE as a dimensionality reduction technique to visualize and analyze the evolution of the active learning process across multiple experimental iterations. This method enabled us to project high-dimensional encoding inputs (originally 54-dimensional vectors) into a two-dimensional planar space, allowing for an intuitive assessment of sample distributions, clustering patterns, and the effectiveness of the ML-guided exploration and exploitation strategies.

The input data for t-SNE analysis consisted of encoded feature vectors derived from synthesis recipes and process parameters, namely the same that would be used in the ML committees in the active learning process for experimentation guidance. Each feature vector included elemental properties (e.g., atomic number, electronegativity, and oxidation state), mixing proportions, hydrothermal conditions, and post-processing steps, like we have stated in **Supplementary Note 1**. These features were compiled using a custom encoding function and organized into individual CSV files for each experimental batch. Features were standardized using StandardScaler to ensure that all input dimensions contributed equally to the t-SNE embedding, mitigating the influence of varying scales across features.

The t-SNE algorithm relies on several hyperparameters that significantly impact the quality of the resulting embeddings. These parameters were carefully selected based on an iterative optimization process aimed at maximizing the separation of successful and failed samples, as well as enhancing clustering quality for exploitation and exploration categories. We set the perplexity to 75, a value that balances local and global structure preservation in the data. Higher perplexity values were explored but led to over-smoothing, reducing the distinctness of clusters, while lower values fragmented the embedding space, causing poor continuity. The learning rate was set to 5000. This high value was selected to ensure that the embedding converged efficiently, especially given the relatively large dataset size and the high variance inherent in the active learning iterations. A lower learning rate resulted in slow convergence and suboptimal embeddings. The t-SNE algorithm was configured with 4250 iterations to allow sufficient optimization, particularly for later batches where the ML-guided process had introduced more defined clustering patterns. Early exaggeration was set to 42, which amplified the separation of clusters in the initial phase of optimization, ensuring that distinct groups were identifiable before local optimization. The choice of the manhattan metric over the default euclidean was based on its robustness in handling high-dimensional spaces with sparse, categorical, or ordinal data. This metric better captured the nuanced relationships between encoding vectors representing distinct synthesis recipes. Finally, all the 64 (trials per batch) *5 (batches)=320 data points were using the same t-SNE dimensionality reduction settings for fair comparison.

t-SNE was selected as the primary visualization tool for its ability to preserve both local and global structures in high-dimensional datasets. Unlike other dimensionality reduction methods, such as PCA, t-SNE effectively captures non-linear relationships, making it well-suited for complex datasets like ours, where sample distributions evolve iteratively through active learning. The vivid visualizations provided clear insights into how the ML-guided process navigated the design space, avoiding failure zones and optimizing for target performance metrics. This methodological framework, including parameter tuning and embedding quality evaluation, ensures robustness and reproducibility in visualizing the evolution of the active learning process.

The script is available under the repository directory “Machine Learning Databases and Script/AL_failure_tsne”. And the full record of the 5 batches of experimentation could be obtained from repository directory “Experimental Records and Raw Data/ AL Full Experimental Records.xlsx”.

## Supplementary Note 3 References:

1. S. Wold, K. Esbensen, P. Geladi, PRINCIPAL COMPONENT ANALYSIS. *Chemometrics and Intelligent Laboratory Systems* **2**, 37-52 (1987).
2. L. van der Maaten, G. Hinton, Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).