Encoding Strategies for Voltammetry Data and their Machine Learning Applications in Self-Driving Laboratories

By

Rainey Fu

Supervisor: Alán Aspuru-Guzik
Advisor: Sergio Pablo-García Carrillo
April 2024

B.A.Sc. Thesis

Division of Engineering Science
UNIVERSITY OF TORONTO

Abstract

Self-driving laboratories (SDLs) represent a cutting-edge concept in scientific research and experimentation. SDLs utilize automated instruments, recommendation algorithms, and an orchestration device to conduct experiments and analyze data without human intervention. Among the array of experiments conducted by SDLs, cyclic voltammetry (CV) and differential pulse voltammetry (DPV) are prominent, offering insights into electrochemical processes. However, efficiently extracting crucial information, such as overall shape and peaks, from CV and DPV data remains a challenge. This thesis presents a novel encoding technique tailored for CV and DPV data, aimed at enhancing SDLs' understanding of chemical environments. With this encoding method, SDLs can discern intricate patterns and relationships within the data more effectively. Experiments consisting of various machine learning tasks, such as clustering, classification, denoising, and synthetic data generation, that an SDL may encounter showed great results. Beyond SDLs, the utility of this encoding technique extends to any 2-dimensional data. Its versatility opens avenues for broader scientific and industrial applications, empowering researchers and practitioners to glean valuable insights from complex datasets. As SDLs continue to evolve, incorporating innovative methodologies such as this encoding technique promises to accelerate scientific discovery and advance technological frontiers.

Acknowledgment

I am deeply thankful to my advisor Sergio Pablo-García Carrillo for his invaluable guidance, without which this project would not have been achievable. I also extend my gratitude to Prof. Alán Aspuru-Guzik for granting me the opportunity to collaborate with the Matter Lab research group at the University of Toronto.

Additionally, I express my heartfelt appreciation to Lianjiang Fu, Yuchuan Hu, Cindy Fu, and Christina Men for their unwavering support throughout this journey. Their encouragement has played a significant role in shaping who I am today.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Amidst urgent global challenges like climate change, energy sustainability, and healthcare crises, there is a growing need for efficient solutions to address the demands of a growing population and increasing resource demands. Accelerating advancements in materials, technology, and scientific knowledge offer potential avenues for tackling these challenges. However, conventional research methods, marked by gradual progress and limited efficiency, may fall short of meeting the urgency posed by these issues. Self-driving laboratories (SDLs), which integrate laboratory automation and data-driven decision-making, emerge as promising tools to expedite and streamline the exploration of solutions while presenting several advantages over traditional scientific approaches [1]. Developing a fully autonomous self-driving laboratory is a complex endeavor that combines various research disciplines. Machine learning and modeling techniques are utilized to forecast materials properties and propose new experiments. SDLs typically use Bayesian optimization to guide their decision-making algorithm. An example of this is Atlas, a brain for SDLs that has been used to identify the voltage peak in CV experiments to optimize the oxidation potential of a set of metal complexes [2]. Concurrently, robotics, computer vision, and automated characterization methods are employed to conduct experiments and analyze outcomes. Central to the design of autonomous labs is the integration of these disparate technologies into a cohesive platform, facilitating seamless interaction between experiments and computational modeling [3].

SDLs can conduct experiments autonomously, performing tasks with increased speed and precision compared to manual processes. Moreover, they utilize data-driven algorithms to navigate through experimental spaces, enabling efficient exploration based on feedback from existing data, a process known as "closed-loop" experimentation. Additionally, SDLs address issues such as reproducibility challenges and the underrepresentation of negative results in scientific literature by promoting the digitization of research processes. Through automated systems, experimental protocols are meticulously documented, enhancing repeatability and reproducibility. Furthermore, digitization facilitates comprehensive data recording and sharing, emphasizing the importance of negative or null results, thus providing a more accurate depiction of scientific endeavors. The wealth of high-quality data generated by autonomous experimentation serves as a valuable resource for the development of artificial intelligence (AI) in materials science and chemistry. By improving machine learning (ML) and deep learning (DL) models, this data enhances the decision-making capabilities of SDLs, furthering their effectiveness in optimizing materials or processes and facilitating novel discoveries.

SDLs in chemistry and materials science are characterized by two critical dimensions: software autonomy and hardware autonomy. Regarding software autonomy, which governs experiment selection, SDLs are categorized into three types: (1) single iterations of automated experimentation with data-driven methods for selecting subsequent experiments, (2) multiple iterations within closed-loop systems where experimental results inform subsequent rounds of automated experiments, and (3) generative approaches involving multiple iterations of closed-loop optimization within

algorithmically generated search or chemical spaces. By automating high-throughput experimentation and streamlining experiment planning and execution, SDLs possess the potential to substantially accelerate research in chemistry and materials discovery. SDLs have played a pivotal role and made noteworthy advancements in various fields including drug discovery, genomics, chemistry, and materials science [1].

Chapter 2

Background and Motivation

2.1 Electrochemistry

SDLs are commonly used to perform electrochemistry experiments. Electrochemistry is the branch of chemistry that investigates electron mobility and the behavior of charged compounds in the presence of an electric field, ultimately giving rise to the phenomenon known as electricity. The flow of electrons occurs not only through the transfer from one chemical species to another in what is called an oxidation-reduction reaction but can also be transferred through positive charges or charged species. When a substance loses an electron, its oxidation state increases, indicating oxidation. When a substance acquires an electron, its oxidation state decreases, indicating reduction. For example, consider the following redox reaction which has oxidation and reduction components:

$$H_2 + F_2 \rightarrow 2HF \tag{2.1}$$

Oxidation:

$$H_2 \rightarrow 2H^+ + 2e^- \tag{2.2}$$

Reduction:

$$F_2 + 2e^- \rightarrow 2F^- \tag{2.3}$$

A redox reaction is balanced when the number of electrons gained by the oxidant is equal to the number of electrons lost by the reductant. Similar to any balanced chemical equation, the entire process is electrically neutral, meaning that the net charge remains consistent on both sides of the equation. With redox reactions, it is possible to physically separate the oxidation and reduction half-reactions in space, provided there exists a complete circuit using an external electrical link, such as a wire, connecting the two halves. As the reaction progresses, electrons migrate from the reductant to the oxidant through this electrical connection, generating an electric current.

Devices that use redox reactions to generate electricity or use electricity to drive non-spontaneous redox reactions are called electrochemical cells. This device effectively transforms chemical energy into electrical energy or vice-versa. In an electrochemical cell, reduction and oxidation reactions take place at the electrodes. The electrode where reduction occurs is termed the cathode, while oxidation occurs at the anode. An electrode serves as a stable electrical conductor, facilitating the flow of electrical current within non-metallic solids, liquids, gases, plasmas, or even vacuums. While electrodes often exhibit high electrical conductivity, they are not limited to metals [4].

Electrode potential is the voltage of an electrochemical cell composed of a refer-

ence electrode and another electrode to be characterized [5]. Figure 2.1 shows a

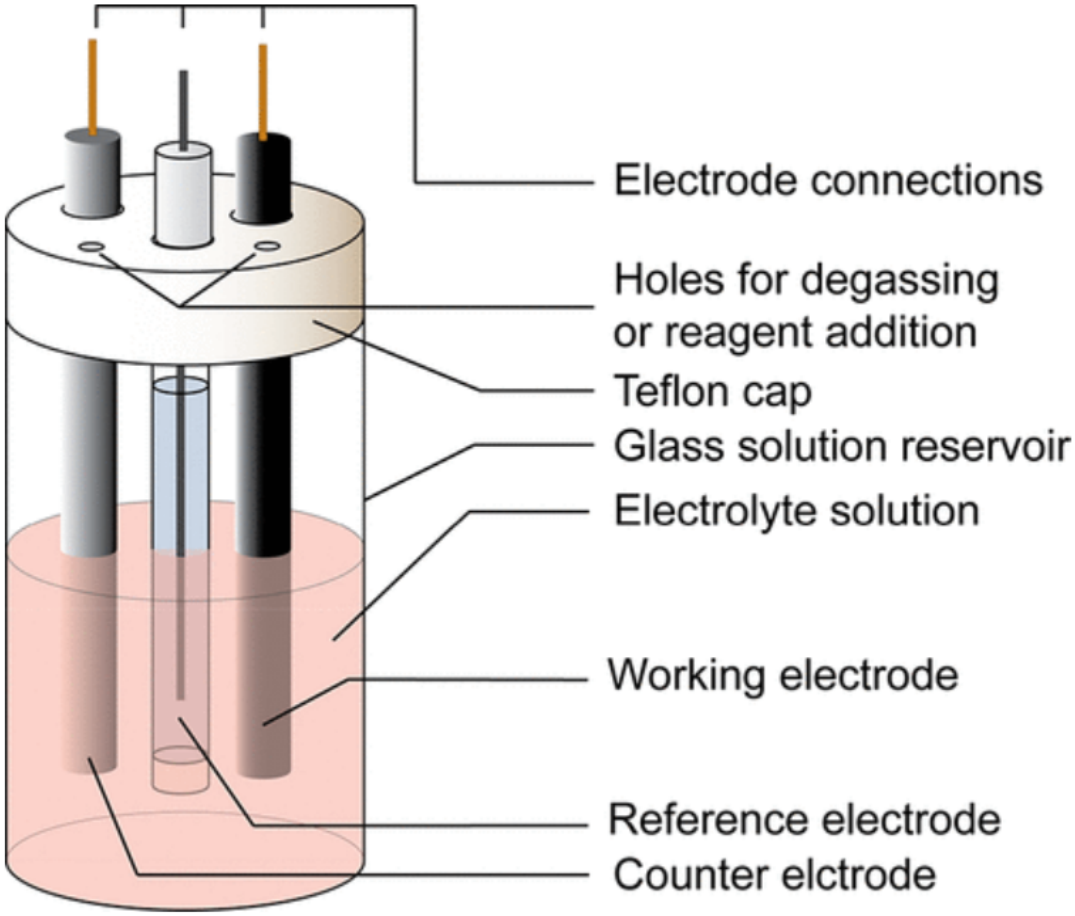


Figure 2.1: Schematic of Electrochemical Cell [6]

three-electrode setup common for electrochemical experiments like cyclic voltammetry. During the flow of current between the working and counter electrodes, the reference electrode is used to precisely measure the applied potential in relation to a stable reference reaction. A potentiostat, as shown in Figure 2.2, is an analytical instrument designed to control the potential of the working electrode within a multi-electrode cell [7]. The potentiostat contains various internal circuits tailored to fulfil

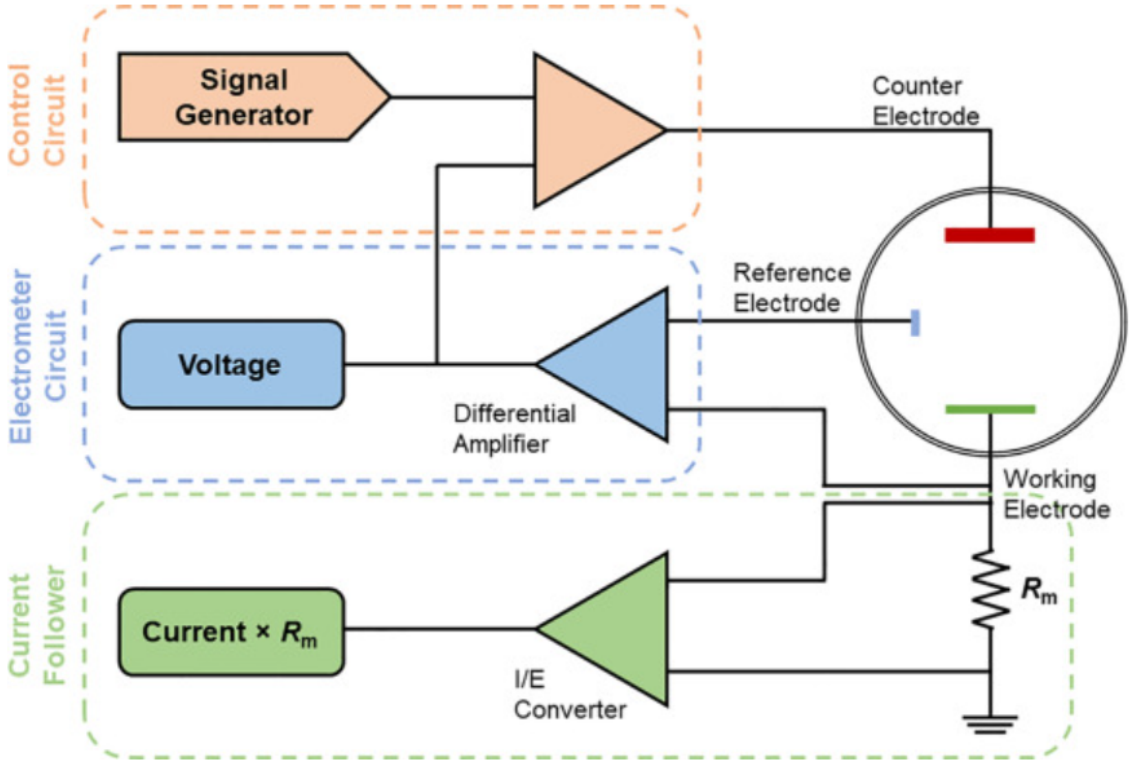


Figure 2.2: Potentiostat Circuit Diagram

this role, facilitating the generation and measurement of potentials and currents. External wires within a cell cable establish connections between the potentiostat circuit and the electrodes within the electrochemical cell. In a three-electrode configuration, the cell cable links the working, counter, and reference electrodes on one terminal and the potentiostat cell cable connector on the opposite end. The potentiostat's internal circuitry governs the applied signal.

The working electrode performs the electrochemical event of interest. Since the event occurs at the electrode's surface, it is crucial that the electrode surface is extremely clean and that the surface area is well-defined. The working electrodes should be

immediately polished after use to ensure there are no surface contaminants that inhibit electron transfer. Even a few hours of air exposure will degrade the electrode surface. Detecting when surface contamination begins to affect data quality is one of the questions this work addresses.

Potentiostats are commonly provided by commercial vendors and are typically governed by proprietary software, employ graphical user interfaces (GUI), and produce already curated data. These devices are commonly used to perform electroanalytical experiments like cyclic voltammetry and differential pulse voltammetry. Commercial potentiostats can vary in their design, but a typical potentiostat is shown in Figure 2.2 and consists of three component circuits: a control circuit, an electrometer, and a current follower [8]. The electrometer circuit utilizes a differential amplifier to measure the difference in potential between the working and reference electrode. Subsequently, the measured potential feeds into the control circuit, which administers a current through the counter electrode, altering the relative potential of the working electrode to align with the user-defined parameters. A single generator ensures this potential adheres to a predefined periodic waveform. The current flowing through the working electrode is then assessed by a current follower circuit, commonly in the form of a current-to-voltage converter. This circuit measures the drop in potential across a grounded resistor, allowing the current to be determined using Ohm's law.

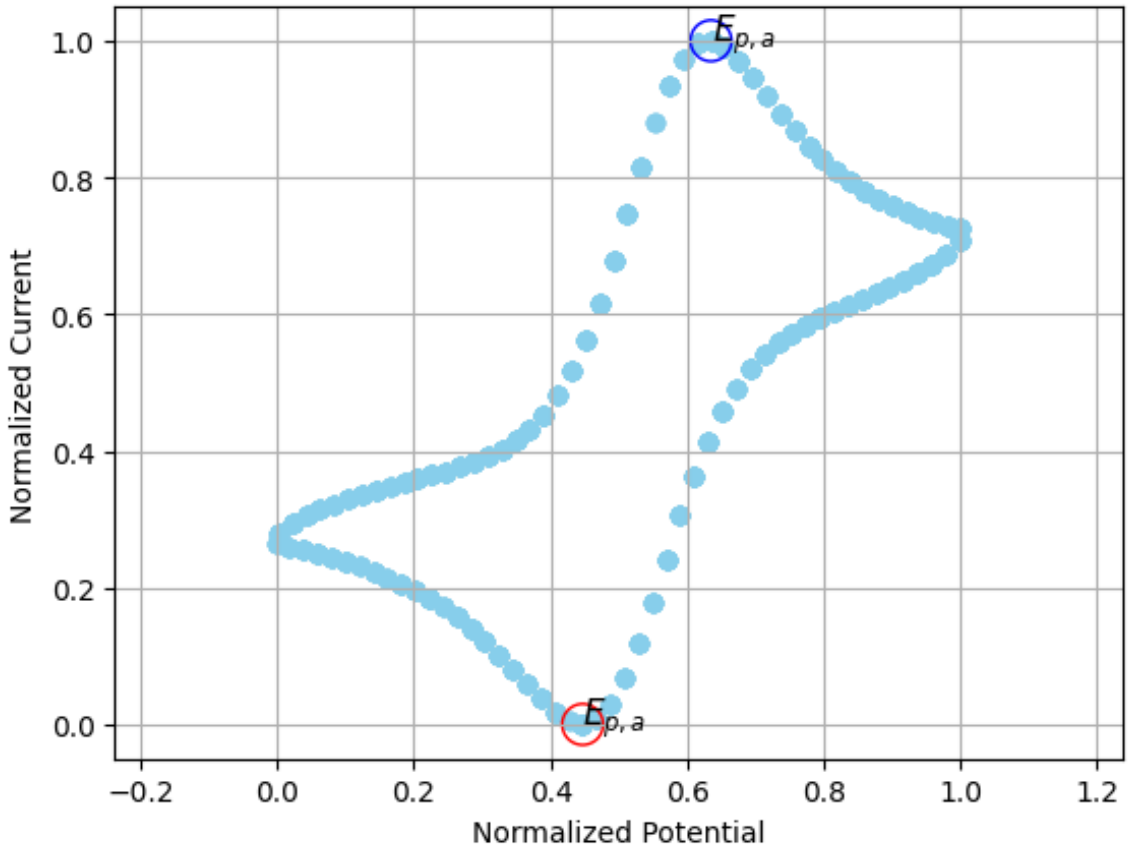


Figure 2.3: Cyclic Voltammogram

2.2 Cyclic Voltammetry

Cyclic voltammetry (CV) is a common electrochemical characterization that extracts important reduction and oxidation information about molecules [9]. Typically, the working electrode potential increases linearly with time. After a predetermined limit is reached, the potential decreases to return to the starting voltage. These cycles can be repeated as many times as needed to bolster confidence in the obtained data. The rate of voltage change over time is known as the experiment's scan rate (Voltage/-

Time) and affects how many data points are gathered throughout the experiment [10].

CV serves as a valuable tool for studying qualitative information about electrochemical processes across diverse conditions. It enables the examination of intermediates in oxidation-reduction reactions and the assessment of reaction reversibility. Other use cases include the determination of electron stoichiometry, analyte diffusion coefficients, and formal reduction potentials, aiding in identification processes [11]. Additionally, in reversible, Nernstian systems, the proportional relationship between concentration and current allows for the determination of unknown solution concentrations via the construction of calibration curves correlating current and concentration [12].

In a typical cyclic voltammogram shown in Figure 2.3, peaks represent electrochemical processes occurring at the electrode surface. The anodic peak ($E_{p,a}$) is observed during the scan where oxidation of the electroactive species occurs at the electrode and corresponds to the potential at which oxidation is most favourable. The current increases as the potential applied to the electrode becomes more positive, reaching a maximum at the peak potential. The cathodic peak is observed during the reverse scan where reduction of the electroactive species occurs at the working electrode and corresponds to the potential at which reduction is most favorable. The current increases as the potential becomes more negative, reaching a maximum at the peak potential [13]. Typically, researchers are especially interested in these peaks.

2.3 Differential Pulse Voltammetry

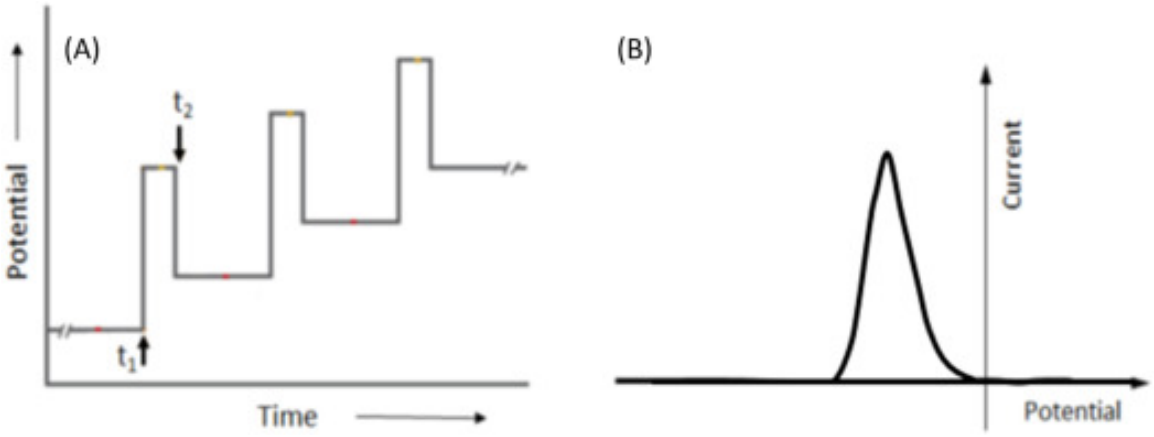


Figure 2.4: DPV Voltammogram

Differential Pulse Voltammetry (DPV) is a more sophisticated electrochemical measurement technique where a series of increasing pulses are applied across the electrodes in an electrochemical cell [14]. The current I_1 is measured right before applying the pulse at time t_1 and I_2 is measured again at the end at time t_2. The difference in current $(\Delta I_2 - I_1)$ is plotted against the potential and results in a peak-like shape. This method helps reduce the impact of charging current by sampling the current just before the potential change. DPV is well suited for measurements with extremely low concentrations of chemicals. This is because the effect of the charging current can be minimized to achieve high sensitivity, and only the faradaic current, the electric current generated by the redox of a chemical at an electrode, is extracted, so electrode reactions can be measured precisely. Furthermore, DPV serves as a versatile tool for the qualitative characterization of chemical compounds and their electrochemical properties. By analyzing the shape, position, and area of the peaks in the DPV curve, researchers can glean insights into the nature of the electroactive species

present, their concentration, kinetics of electron transfer processes, and other relevant electrochemical parameters. This capability makes DPV invaluable in various fields such as analytical chemistry, environmental monitoring, and pharmaceutical research, where understanding the behaviour of chemical compounds at the molecular level is crucial.

Chapter 3
Clustering

3.1 Introduction

Given the capabilities and limitations of SDLs, there exists a need for a quick and accurate characterization of the produced electrical compounds. Clustering experimental results becomes crucial for several reasons. Clustering identifies patterns and similarities among experimental results. This aids in the discovery of underlying trends or relationships between different compounds or experimental conditions. It also facilitates quality control by pinpointing outliers or anomalies in experimental data, ensuring the reliability of data produced by SDLs. Moreover, clustering allows researchers to optimize processes by providing insights into the effects of various parameters on the formation of electrical compounds. This optimization can significantly enhance the efficiency of the voltammetry process in SDLs. Additionally, by classifying different types of electrical compounds based on their properties or characteristics, clustering supports classification and prediction tasks, enabling researchers to predict the behavior of new compounds or classify unknown compounds based on their similarities to known clusters. Finally, clustering provides a structured way to organize and interpret large volumes of experimental data, facilitating decision-making processes related to the selection of compounds for further analysis or the design of future experiments. In essence, clustering experimental results in the context of SDLs used for voltammetry is indispensable for gaining insights, ensuring data

quality, optimizing processes, classifying compounds, and facilitating decision-making processes.

3.2 Data Collection

To analyze how data gathered from SDLs can be clustered, data was gathered through a cost-effective autonomous electrochemistry experimentation that operates through an iterative workflow [15]. The workflow was used to synthesize and characterize 10 distinct metals and 10 distinct ligands, with specific details available in Appendix A.1 and Appendix A.2, resulting in 100 unique complexes. Each complex was synthesized using a metal/ligand concentration ratio of 1:7 to ensure complete complexation. The synthesis process employed 1.0 M NaCl in water as the electrolyte/solvent, and a buffer solution consisting of a 1:1 ratio of HOAc/NaOAc. Following synthesis, comprehensive characterizations were conducted using cyclic voltammetry (CV) and differential pulse voltammetry (DPV) techniques. The experimentation was done using a low-cost electrochemistry platform designed as an alternative to commercial options. The length of these samples may vary due to differing scan rates. Higher scan rates lead to more data points being collected during the experiment and can provide finer resolution of the electrochemical processes occurring. Additionally, it's worth noting that these samples may be duplicated as CV and DPV analyses can be conducted multiple times on the same sample. Importantly, the workflow is adaptable, with the potential to encompass a broader range of parameters, including additional ligands, varying metal/ligand ratios, mixed ligands, different buffer pH levels, and

reaction times. The accumulation of data points is ongoing, contributing to the continuous expansion and refinement of our understanding. The final dataset consists of 800 CV data points and 200 DPV data points. The dataset used in this work can be found in the article preprint [15].

3.3 Curse of Dimensionality

The curse of dimensionality refers to the phenomena that cause various challenges and complications when analyzing data in high-dimensional spaces. As the number of features in a dataset increases, the amount of data needed to generalize accurately grows exponentially. As the number of dimensions increases, the data becomes increasingly sparse. This makes tasks like clustering and classification more challenging. In higher dimensions, the difference between distances between data points starts to become negligible, making measurements like Euclidean distance negligible. As such, algorithms that rely on distance measurements will experience a drop in performance. Furthermore, more dimensions will require more computational resources and time to process the data. It is good practice to aim to have the data in as low-dimension as possible provided relevant information is maintained.

3.4 K-Means

K-Means clustering is an unsupervised machine learning algorithm aimed to divide a set of data points into clusters such that the data points within each cluster are similar

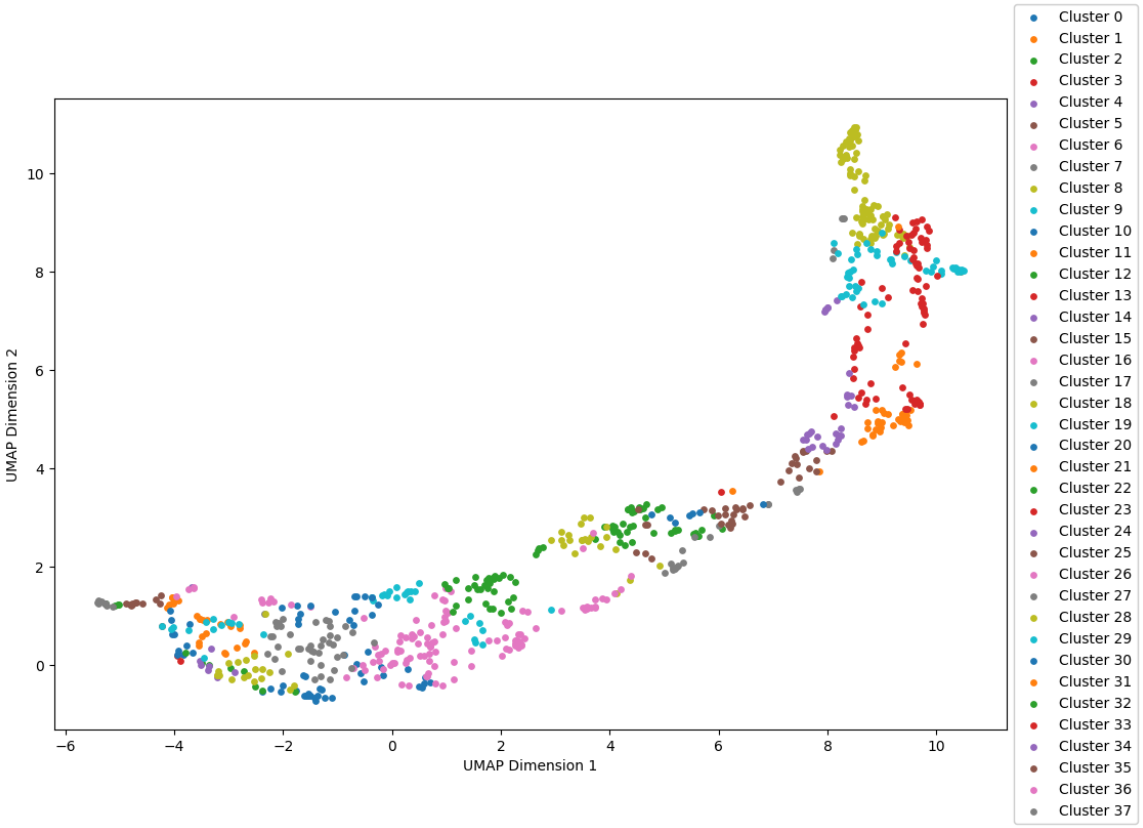


Figure 3.1: CV K-Means Clustering Visualized

and different from the data points in other clusters [16]. The K-Means clustering result for CV data can be seen in Figure 3.1. The algorithm is explained below with K representing the desired number of clusters:

1. Initially, K points are selected randomly as the cluster centroids
2. Each data point is assigned to the closest mean, quantified by the Euclidean distance.
3. Each cluster centroid is updated to reflect the average of data points currently assigned to that cluster

4. This process is repeated for a specified number of iterations

One of the questions that needs to be answered is the choice of K. This means finding a balance between the number of clusters represented by K and the average variance of the clusters while minimizing both. There is no approach for determining K that works better than all others in all cases. For this problem of clustering CV and DPV data, a combination of the Elbow Method and Silhouette method is used. The Elbow Method is done by plotting the within-cluster sum of squares (WCSS) for a range of K and choosing the value K where adding more clusters does not significantly decrease the WCSS. While the Elbow Method can easily eliminate many values of K, it also has drawbacks regarding the shape of the WSCC curve. Determining the exact location of the "elbow" can be subjective and depends on the analyst's interpretation. Different individuals may identify different elbows, leading to inconsistency in results. In cases where the relationship between the number of clusters and WCSS is not distinctly elbow-shaped, the Elbow Method may not provide clear guidance for choosing the appropriate number of clusters. The Silhouette Method addresses some of these drawbacks by providing a more quantitative measure of cluster quality. Instead of relying on subjective interpretation, the Silhouette Method calculates the silhouette coefficient for each data point, which quantifies how similar an object is to its own cluster compared to other clusters. This provides a more objective measure of cluster cohesion and separation. The process for selecting K for this work includes determining a set of candidate K values using the Elbow Method by eliminating obviously suboptimal values and then using the Silhouette method to find optimal K among the potential candidates.

3.5 Density-Based Spatial Clustering of Applications with Noise

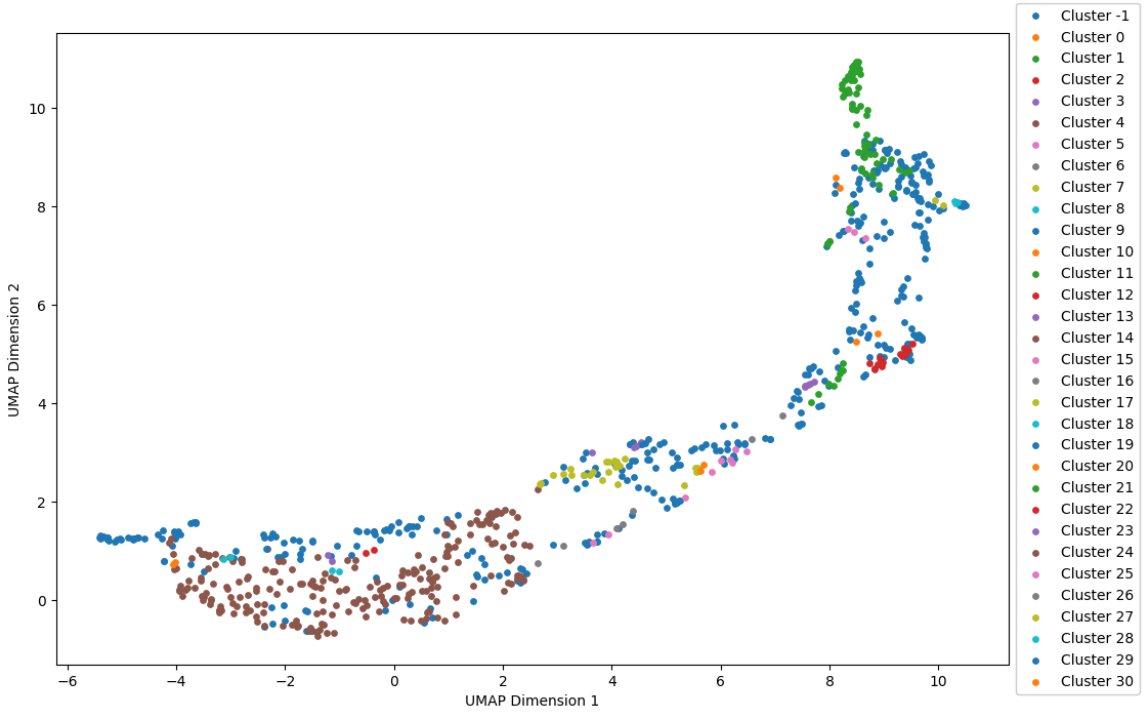


Figure 3.2: CV DBSCAN Clustering Visualized

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is another clustering algorithm that works by partitioning the data into dense regions of points that are separated by less dense areas [17]. It defines clusters as areas of the dataset where there are many points close to each other, while the points that are far from any cluster are considered outliers or noise. In DBSCAN, eps (ϵ) represents the maximum distance between two points for them to be considered neighbours, and minimum samples is the number of points required for a point to be considered a core point. Points that have fewer than minimum samples points are labelled as noise.

The key differentiator for DBSCAN is that the number of clusters does not need to be determined beforehand.

3.6 t-Distributed Stochastic Neighbor Embedding

Dimensionality techniques like t-Distributed Stochastic Neighbor Embedding (t-SNE) are used for visualizing high-dimensional data in a low-dimensional space [18]. This visualization can aid in the clustering process by providing insights into the underlying structure of the data and help in understanding the results of the clustering algorithm. The first step of the algorithm is to create a probability distribution that represents the similarity between neighbors. The similarity between the two data points is represented by their Euclidean distance. For each data point, it is placed in the middle of the Gaussian curve and the rest of the data is placed along the curve. This is represented by the following equation where $j \neq i$ and $p_{i|i} = 1$:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)} \tag{3.1}$$

"The similarity of datapoint x_j to datapoint x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i" [18]. The last variable that has not been discussed yet is sigma. This variable is not chosen directly, but rather by choosing a value for perplexity. Perplexity is defined as:

$$Perp(p) \coloneqq 2^{-\sum_x p(x) log_2(p(x))} \tag{3.2}$$

Perplexity represents the density of data and how many neighbors the central point should have with higher values relating to higher variance. After choosing the perplexity value, the corresponding sigma values are found using binary search. Next, the similarities between data points for low-dimensional representational will also need to be found to ensure that similar data are close together after projection.

3.7 UMAP

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualization similarly to t-SNE [19]. It achieves this by leveraging concepts from algebraic topology and Riemannian geometry. Here's a simplified breakdown of how UMAP works:

1. Constructing a Topological Representation: UMAP starts by creating a fuzzy topological representation of the data. This involves building a simplicial complex, which is a way to represent topological spaces using simple geometric shapes called simplices. The algorithm constructs these simplices based on the proximity of data points to each other.

2. Optimizing Low-Dimensional Representation: Once the topological representation is established, UMAP then optimizes a low-dimensional representation of the data to match this topological structure as closely as possible. It does this by minimizing a measure called cross-entropy, which quantifies the difference between the fuzzy topological structures of the high-dimensional and

low-dimensional data.

3. Efficient Computations: UMAP employs several strategies to make computations efficient. It focuses on computing only the nearest neighbours of each point and uses algorithms like Nearest-Neighbor-Descent for this purpose. Additionally, it utilizes stochastic gradient descent for optimization and smooth approximations of the membership strength function to ensure differentiability.

4. Preserving Topological Structure: The goal of UMAP is to ensure that the low-dimensional representation maintains the essential topological properties of the original data. It achieves this by balancing attractive forces that pull similar points together and repulsive forces that push dissimilar points apart, based on the weights of edges in the topological representation.

3.8 Ramer–Douglas–Peucker Algorithm

The Ramer–Douglas–Peucker (RDP) algorithm is employed to reduce the number of points in a curve approximated by a series of points. It operates by conceptualizing a line between the initial and terminal points within a point set defining the curve. Subsequently, it identifies the point furthest from this line among the intermediary points. If this point, termed the "outlier point", and consequently all intervening points, lie within a specified distance 'epsilon' from the line, they are removed. Conversely, if the outlier point surpasses the epsilon threshold, the curve is segmented into two parts: from the initial point to the outlier point, inclusive and the outlier point and the remaining points. The algorithm is then recursively applied to both re-

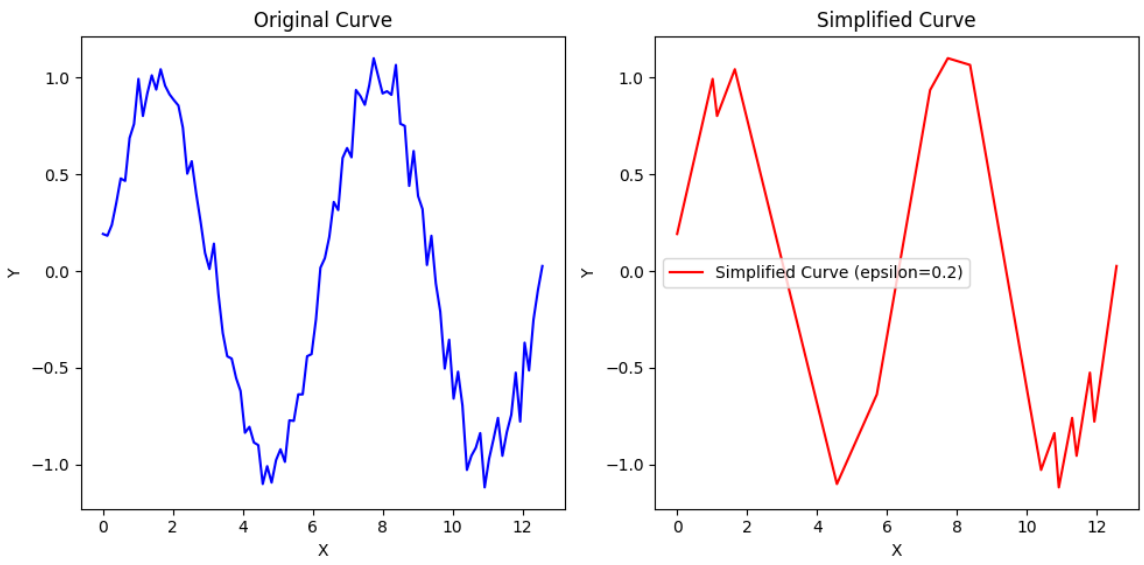


Figure 3.3: RDP Algorithm

sulting segments, and the reduced forms of the curve are reassembled. Since CV and DPV results can be represented as a curve, RDP can be used to remove unnecessary points while maintaining the overall shape of the voltammogram. This will reduce the dimensionality and improve data analysis results.

3.9 Data Preparation and Encoding

Many parameters can be set during CV and DPV analysis, affecting the outcomes of the characterization. Particularly, the experiment's scan rate affects the sampling frequency and the number of points collected within a certain time interval, leading to a variable number of point densities depending on the analyzed compound. Heterogeneity among samples becomes challenging for many ML algorithms, as they often require input data to be the same shape. Similarly, the potential limit at which the

potential begins to return to its initial point will affect the overall shape of the cyclic voltammogram. To handle this, the following steps are used to prepare the data:

1. Split experiment cycles into separate data points
2. Normalize values to fit between [0, 1]
3. Reduce points using the Ramer-Douglas-Peucker algorithm
4. Duplicate data points until the total length reaches the longest cycle's length
5. Order data points based on angular position relative to the center

Due to the curse of dimensionality, the RDP algorithm is used to reduce the number of dimensions. Since the RDP algorithm takes only a variable ϵ, the final length after reduction will be different for each set of data. To ensure the data has the same length as the longest data after RDP reduction, data points are randomly selected and duplicated. Finally, the data is ordered based on its angular position relative to the center for consistency. Plots of the raw data and processed data can be seen in Figure 3.4 with the color of the scatter plot representing the order in which the points appear. The starting point and end point varies across different voltammetry experiments. As such, it is important to order the data points so that comparisons can be informative. A significant reduction in dimensionality by 1/3 can also be seen in the plots. Despite this, the important characteristics of the voltammogram such as the overall shape and peaks are maintained.

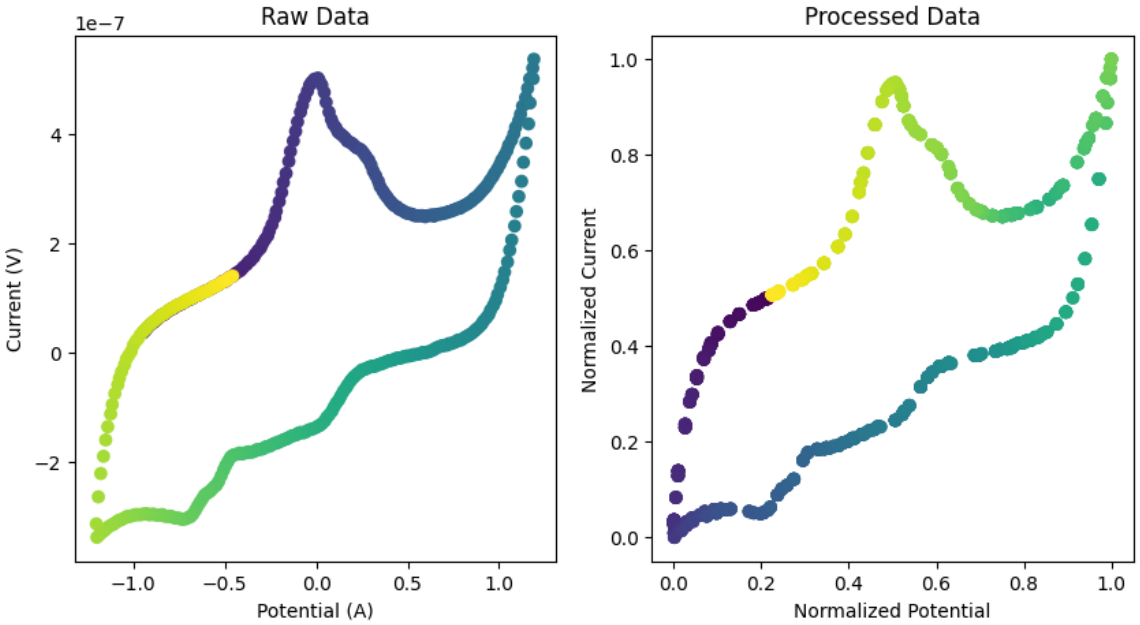


Figure 3.4: Raw Data and Processed Data

3.10 Results and Discussion

To categorize the experimental data, the K-Means clustering algorithm was used post-processing the entire set of experimental voltammetry. With K-Means, a value of K will need to be selected. This is done using the elbow method. Figure 3.5 shows the results of the elbow method applied to the dataset. It can be seen that this methodology identifies multiple potential candidates for K-values, necessitating a more comprehensive analysis to select the most appropriate option. To aid the decision-making process, the Silouhette method is used to analyze promising values [20]. A cluster with a value of 1 means points are perfectly assigned in a cluster and clusters are easily distinguishable, 0 means clusters are overlapping, and -1 means points are assigned to the wrong cluster. The K value should be chosen based on

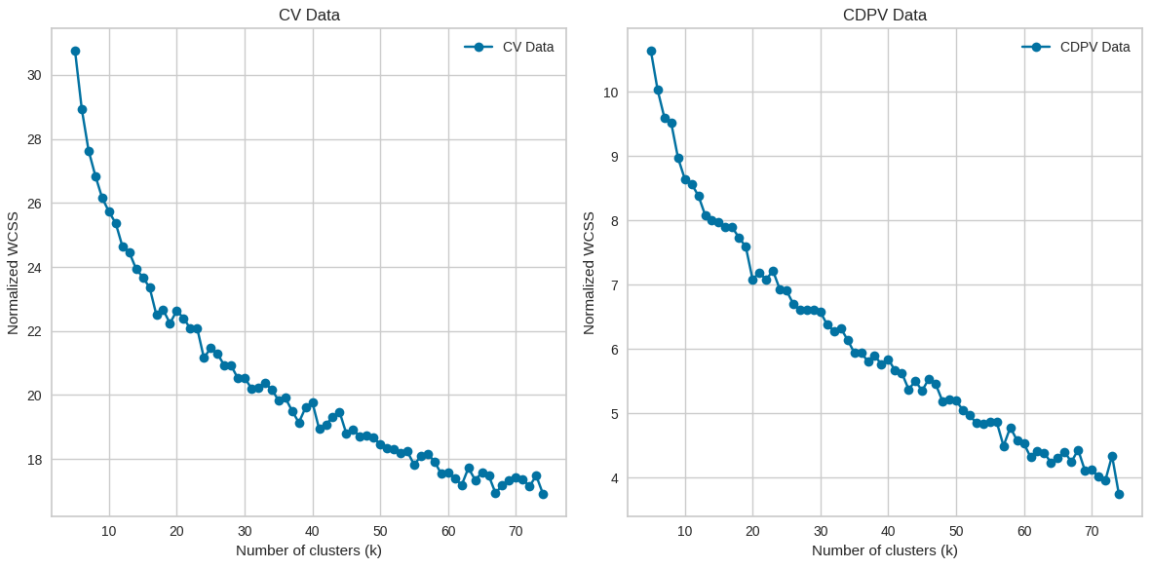


Figure 3.5: K-Means Elbow Method

which value produces the most clusters with Silhouette scores greater than the average score of the dataset, represented by the red-dotted line seen in Figure 3.6 and Figure 3.7. Furthermore, there should not be wide fluctuations in the size of the clusters. The width of the clusters represents the number of data points belonging to the cluster. In Figure 3.6, which showcases the application of the Silhouette method for CV cross-validation, K = 38 yields the highest number of clusters with a score surpassing the mean of the dataset. This configuration not only reduces the number of clusters scoring below zero but also minimizes the variance in cluster sizes. Similarly, in Figure 3.7 showcasing the Silhouette method for DPV, K = 42 results in the best quality of clusters. A subset of the cluster results is available in the appendix. Despite having 100 different combinations of metals and ligands, using a relatively small K value still shows promising results, as the data points within each cluster have similar overall shapes, which is crucial for compound identification.

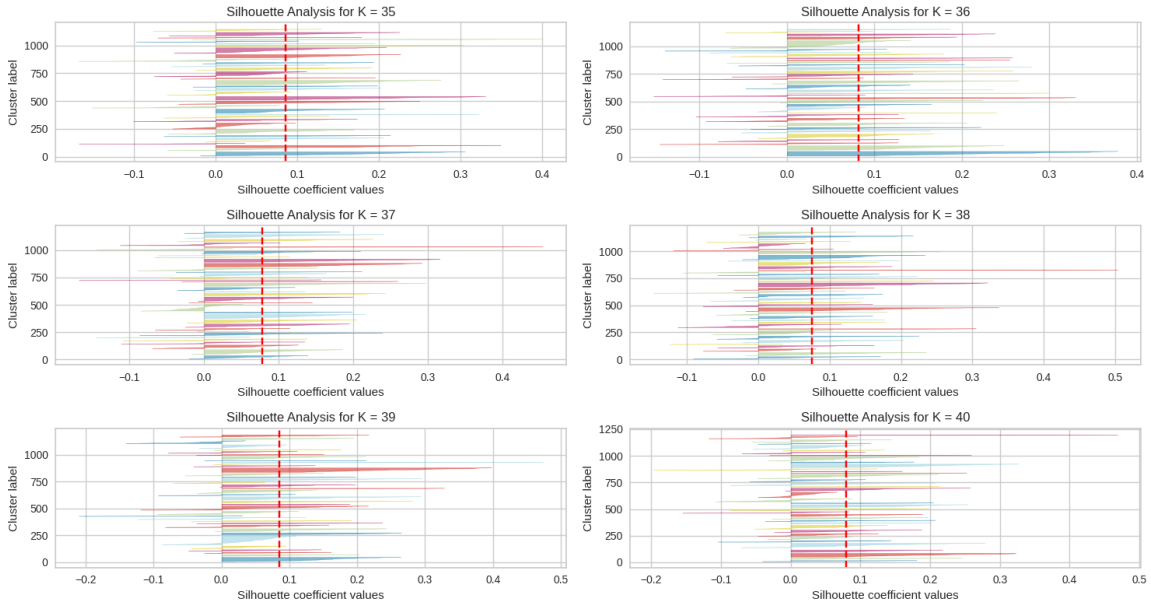


Figure 3.6: CV Silhouette Method

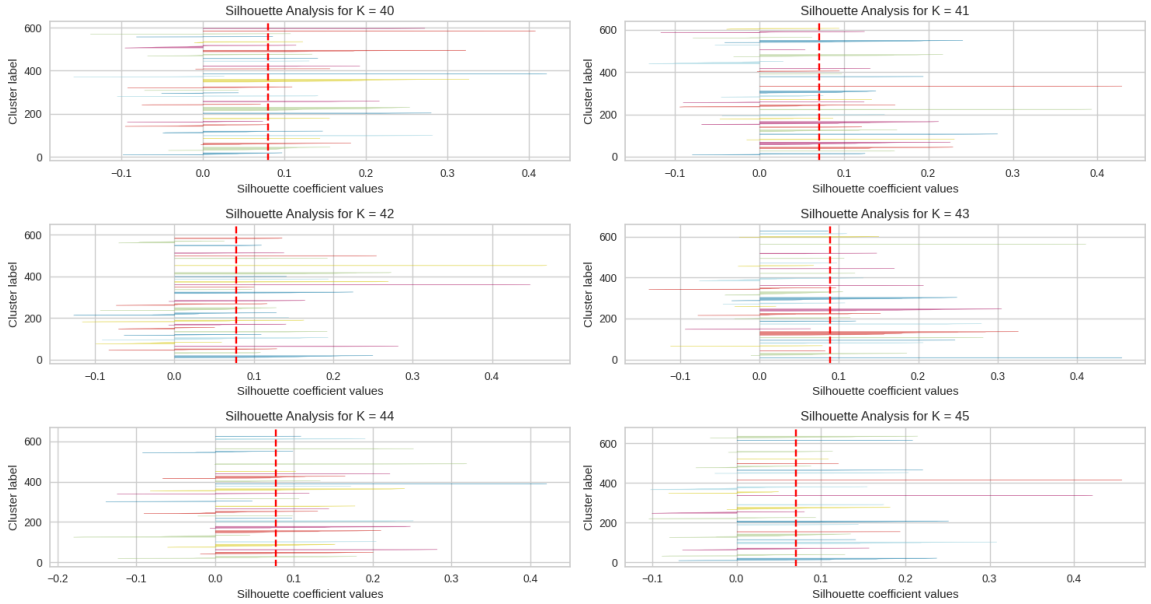


Figure 3.7: DPV Silhouette Method

DBSCAN, as an alternative clustering method, demonstrated significant promise in identifying anomalous data points. With the appropriate parameters, DBSCAN effi-

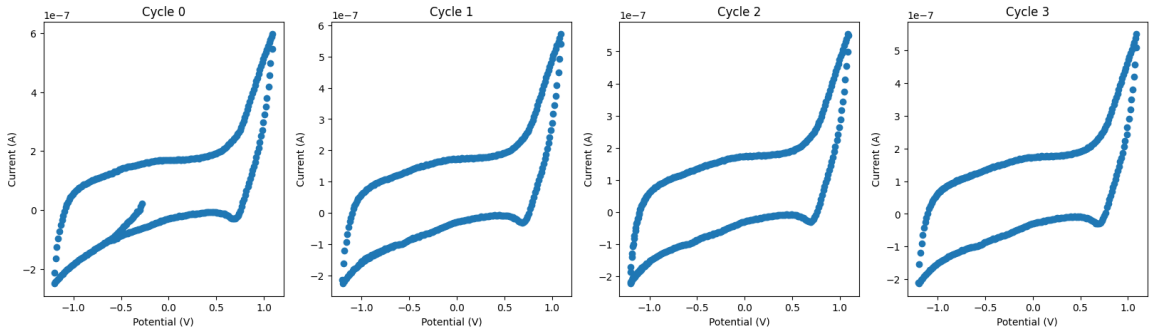


Figure 3.8: DBSCAN Clusters

ciently grouped cycles originating from the same experiment. As depicted in Figure 3.8, DBSCAN defines a cluster comprising cycles solely from a single experiment. This capability could be seamlessly incorporated into SDLs as an error validation mechanism. Any cycle not assigned to the same cluster as others from the identical experiment could trigger an error notification, prompting intervention and investigation.

To further demonstrate the efficacy of the encoding, t-SNE and UMAP projections are created to visualize the data in 2-D and show how the shapes, metals, and ligands are distributed. Interactive plots made with Bokeh are available on Github. As seen in Figure 3.9 and Figure 3.10, t-SNE emphasizes local structure and tends to agglomerate similar data points into tight clusters. As a result, t-SNE plots often show clearer separation between clusters but may not preserve the global structure as effectively. t-SNE primarily preserves local neighborhoods, which leads to tighter clusters of similar points. However, it may not always capture the global structure accurately, especially for complex datasets. t-SNE embeddings can vary significantly with different random initializations and parameter choices, making it less stable and potentially

more sensitive to noise in the data. UMAP tends to focus more on preserving global structure and maintaining relative distance between clusters. Therefore, clusters in the UMAP plot are usually well-separated and evenly distributed. UMAP tries to preserve local and global neighborhoods, resulting in more evenly spaced clusters and better representation of both local and global structures. UMAP embeddings are generally more stable across different runs and parameter settings compared to t-SNE. Figures 3.11 and 3.12 illustrate these characteristics, especially when contrasted with the projections generated by t-SNE.

Utilizing machine learning techniques to classify voltammetry data based on the overall shape presents numerous benefits over merely employing a script to identify the number of peaks, as has traditionally been done. Machine learning models can be trained to recognize patterns and variations regarding the overall shape and number of peaks. They can adapt to experimental conditions, electrode materials, and analytes without needing manual adjustment of parameters. Voltammetry data can often be noisy, especially at low concentrations. ML models can be trained to distinguish true peaks from noise more effectively than simple peak-finding algorithms. Voltammograms can vary in characteristics due to factors such as electrode deterioration, surface roughness, and solution composition. ML models can learn to handle this variability and provide more reliable peak classification across different experimental conditions. Additionally, ML models can learn when the electrode deteriorates and automatically polish it. ML models can automatically extract relevant features from voltammogram data such as peak heights, peak widths, peak potential, and overall shape. This allows for more comprehensive analysis beyond locating peaks. Once

trained, ML models can be integrated into larger data analysis pipelines to classify cyclic voltammetry data rapidly and efficiently, potentially saving time and effort compared to manual analysis or parameter tuning for peak-finding algorithms. These automated analysis techniques can be integrated into an SDL, updating them with newly generated data to increase their accuracy as more experiments are performed.

In summary, clustering techniques play a crucial role in analyzing and interpreting experimental voltammetry results obtained from SDLs. By organizing data into meaningful clusters, clustering techniques like K-Means and DBSCAN and dimensionality reduction techniques like t-SNE and UMAP uncover patterns, similarities, and trends that enhance our understanding of the electrochemical compounds of interest. The choice of appropriate clustering algorithms and parameter selection methods, such as the Elbow Method and Silhouette Method, have been discussed to ensure meaningful and reliable clustering results. The results obtained from clustering algorithms and dimensionality reduction techniques have provided valuable insights into the underlying structure of the experimental data, facilitating compound identification, error detection, and decision-making processes in SDLs.

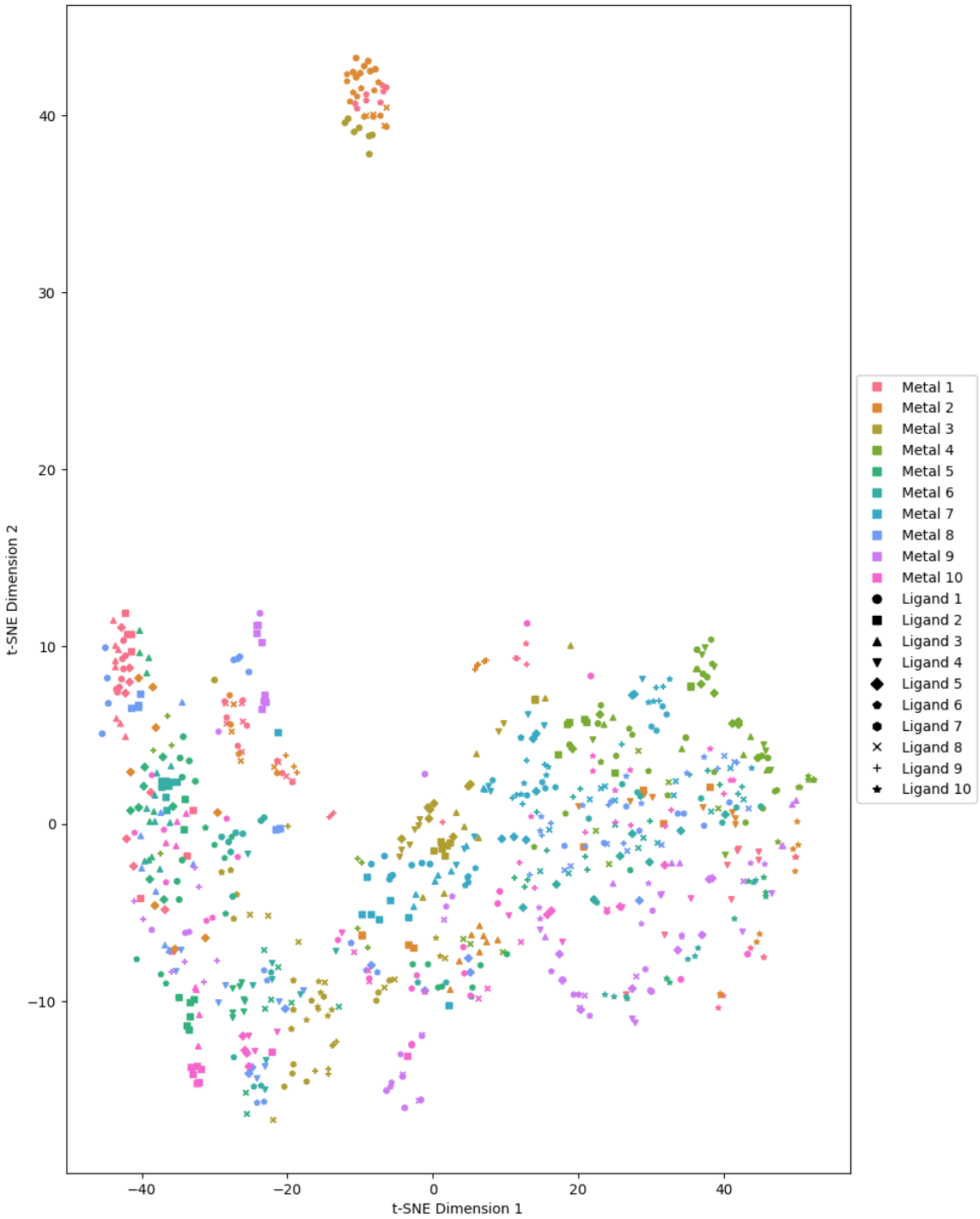


Figure 3.9: Cyclic Voltammetry t-SNE Projection

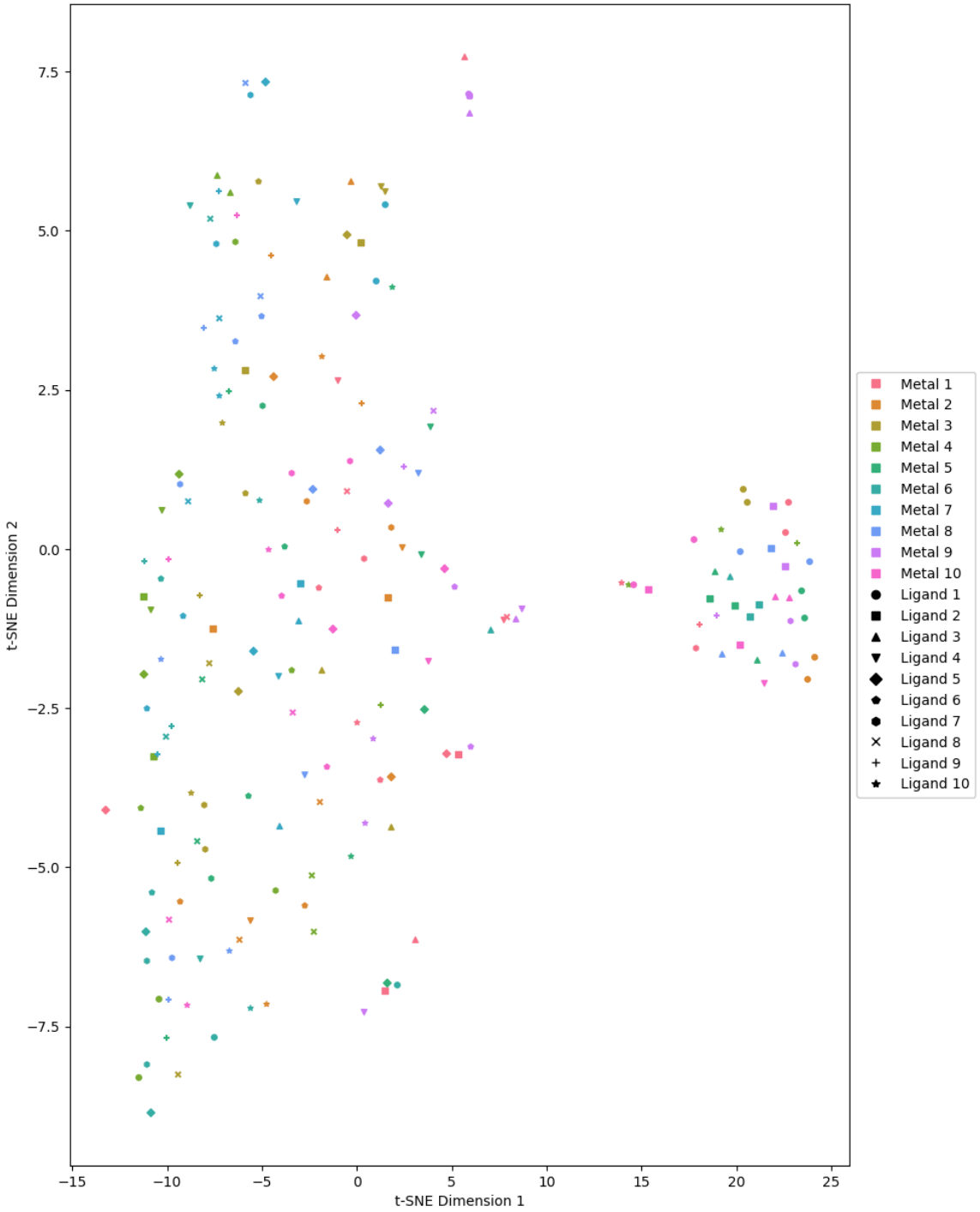


Figure 3.10: Differential Pulse Voltammetry t-SNE Projection

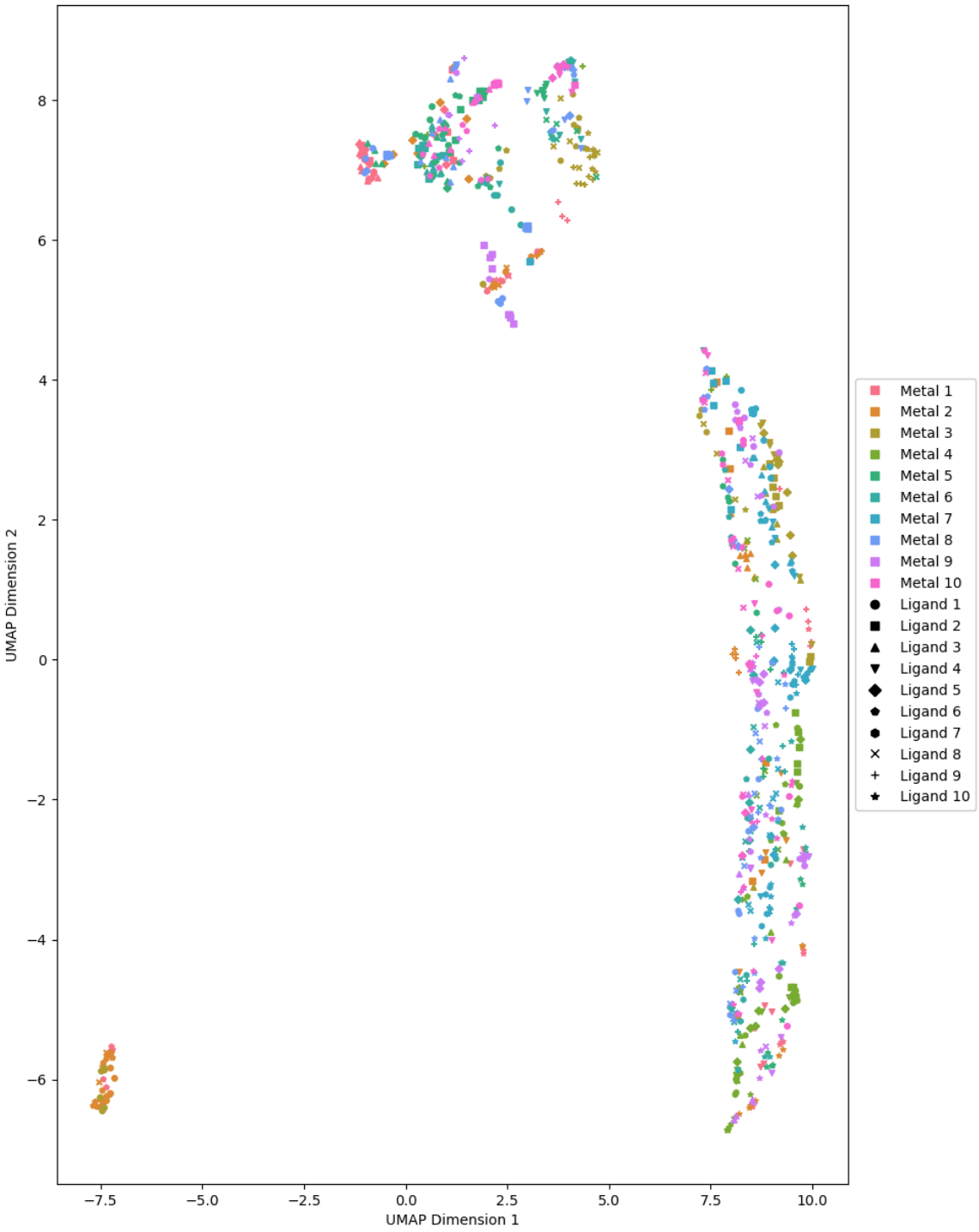


Figure 3.11: Cyclic Voltammetry UMAP Projection

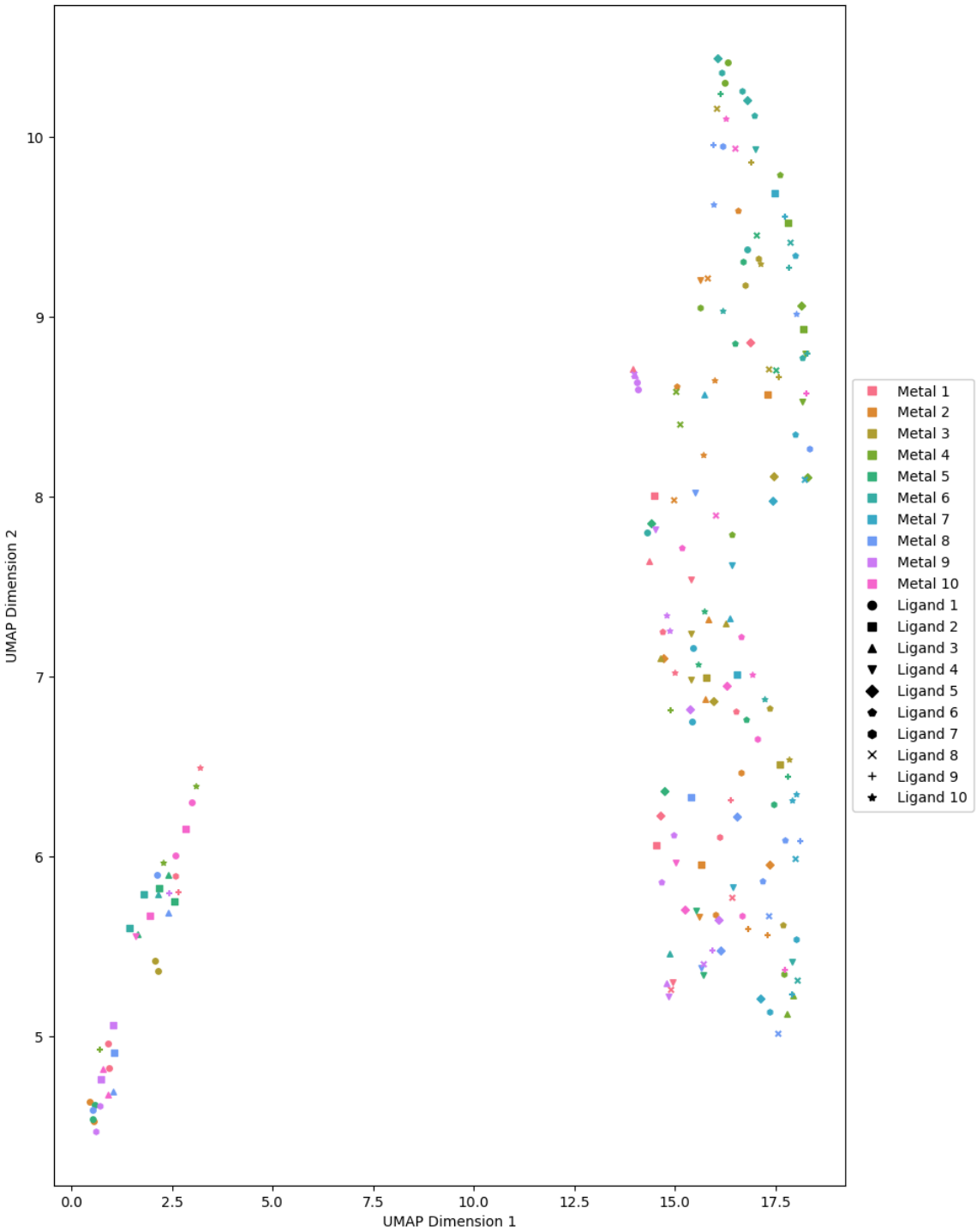


Figure 3.12: Differential Pulse Voltammetry UMAP Projection

Chapter 4

Classification

4.1 Introduction

To further demonstrate the feasibility of using this encoding technique for various machine-learning tasks, a classifier is trained to predict what metals and ligands were used to generate each voltammogram. It is important to note that the dataset used is relatively small for a deep-learning task. For training, the dataset was split with 80% for training, 10% for validation, and 10% for testing. An important insight to consider is the similarity between voltammetry data and images. After all, each point has a potential and current value, which is similar to an image's RGB values. The main difference is that an image is 2-dimensional while voltammetry data is 1-dimensional. Many previous works have used convolutional neural networks for classification tasks [21]. Using this as inspiration, the proposed model architecture for voltammetry data classification uses 1-dimensional convolutional layers.

4.2 Variational Autoencoders

Since one of the major challenges faced is the dataset size, one method to address this is to create synthetic data. A variational autoencoder (VAE) is similar to the autoencoder neural network architecture shown in Figure 4.1, with the main difference

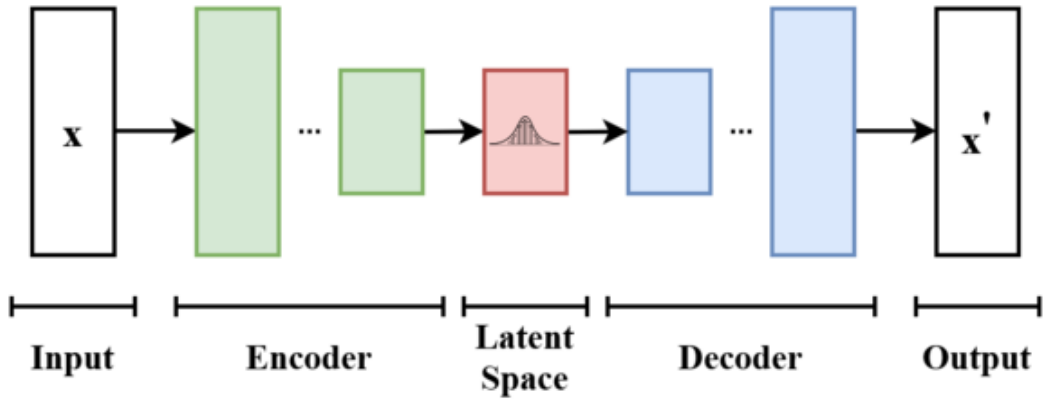


Figure 4.1: Autoencoder Diagram

being that VAEs connects the encoder to its decoder through a probabilistic latent space that corresponds to the parameters of a variational distribution [22]. The encoder maps each point from the dataset into a distribution within the latent space rather than a single point in that space. The distribution is typically Gaussian with a mean and a variance. Once the VAE is trained, different points can be sampled from the learned latent space distribution. These samples represent different configurations of the input data in the latent space. The sampled points from the latent space are then fed into the decoder network, which generates reconstructions of the input data corresponding to those points. By sampling multiple points from the latent space and decoding them, a diverse set of synthetic data samples that resemble the original data distribution is generated. The variability in the latent space allows for the generation of novel and diverse data samples that capture the underlying characteristics of the training data.

4.3 Conditional Variational Autoencoders

While traditional VAEs learn a latent space for the dataset, conditional variational autoencoders (CVAEs) expand this concept by introducing conditional dependencies between the input data and the latent variables. In the context of generating synthetic data, CVAEs offer a more controlled approach by allowing the generation process to be conditioned on additional information, such as class labels or other attributes associated with the data. By conditioning the generation process on known attributes or labels, CVAEs can generate synthetic data samples that not only capture the underlying data distribution but also adhere to specific conditions or constraints defined by the conditioning variables. This enables the targeted generation of synthetic data for different classes or categories, even in the absence of labeled data. In this case, the metal and ligand are encoded using one-hot encoding and passed to the decoder to generate data belonging to the same class.

4.4 Classifier Model Architecture

The classifier architecture can be seen in Figure 4.2. The model consists of several convolutional layers followed by max-pooling layers to encode the data and reduce dimensions. All layers except for the output layer use the ReLU activation function. The output layer is a dense layer with 10 units and a softmax activation function. The Adam optimizer and categorical cross-entropy loss are used to train the model. Additionally, the model uses L2 regularization and early stopping to prevent overfit-

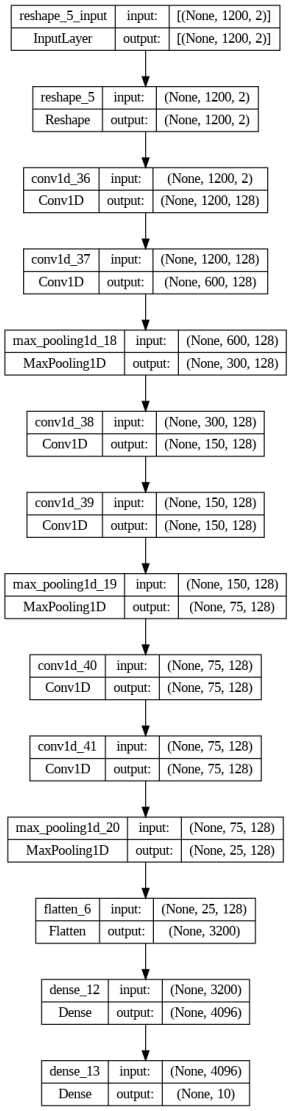


Figure 4.2: Classification Model Architecture

ting and ensure smooth convergence. The Glorot uniform initializer is used for weight initialization to facilitate better gradient flow and prevent exploding gradients.

4.5 Results and Discussion

Model	Accuracy (%)
CV Ligands	75.13%
CV Metals	79.24%
DPV Ligands	30.00%
DPV Metals	25.00%

Table 4.1: Classification Results

Separate classifiers were trained each with a unique task of classifying CV ligands, CV metals, DPV ligands, and DPV metals. The accuracy of the classifiers can be seen in Table 4.1 and the results were much better for the CV data compared to the DPV data. This difference can likely be attributed to the size of the datasets. After incorporating synthetic data generated with the CVAE into the training process, there was a significant improvement in accuracy for classifying CV data as seen in Table 4.2. However, the DPV ligands classifier actually saw a decrease in performance. Again, this is likely due to the size of the dataset. In utilizing VAEs to generate

Model	Accuracy (%)
CV Ligands	77.86%
CV Metals	85.00%
DPV Ligands	25.00%
DPV Metals	25.00%

Table 4.2: Classification Accuracy with Synthetic Data

synthetic data, several key considerations impact classifier performance, especially when dealing with small datasets. Firstly, the quality and diversity of the original data influence the effectiveness of the synthetic data produced by VAEs. With limited variation or complexity in a small dataset, the VAE might struggle to capture the true underlying data distribution accurately, potentially resulting in synthetic data that fails to fully represent the characteristics of the real data. This mismatch can detrimentally affect classifier performance. Additionally, the risk of overfitting is heightened in small datasets, where the classifier may excessively specialize in training data patterns that do not generalize well. Introducing synthetic data from a VAE can compound this issue if the VAE itself overfits the small dataset, producing synthetic data overly similar to the training data, which provides minimal additional information for the classifier and can lead to decreased performance on unseen data. VAEs implicitly learn the probability distribution of the input data. However, if the distribution of the real data is significantly different from the distribution learned by the VAE due to the small dataset size, the synthetic data generated by the VAE may not accurately represent the true data distribution. This distribution mismatch can confuse the classifier, as it may encounter data points in the synthetic dataset that deviate from the real data distribution, leading to suboptimal performance. Table 4.4 provides insights into the precision, recall, and F1-score of each metal type classification, along with the number of instances (support) for each metal type. Precision indicates the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positives that were correctly identified. F1-score, the harmonic mean of precision and recall, provides a balanced measure between the two. Overall, the classifier model achieved an accuracy of 85%, indicating

	Precision	Recall	F1-Score	Support
Metal 1	0.88	0.88	0.88	8
Metal 2	0.80	1.00	0.89	8
Metal 3	1.00	1.00	1.00	4
Metal 4	1.00	0.83	0.91	12
Metal 5	1.00	0.71	0.83	7
Metal 6	0.88	0.78	0.82	9
Metal 7	0.82	0.90	0.86	10
Metal 8	0.50	0.40	0.44	5
Metal 9	0.78	1.00	0.88	7
Metal 10	0.82	0.90	0.86	10
Accuracy			0.85	80
Macro Avg	0.85	0.84	0.84	80
Weighted Avg	0.86	0.85	0.85	80

Table 4.3: CV Metals Classification Report

its effectiveness in classifying different metal types. However, it is important to note variations in performance across metal types. For instance, Metal 3 achieved perfect precision, recall, and F1-score, suggesting the model's excellent ability to classify this particular metal type accurately. On the other hand, Metal 8 exhibited lower precision and recall scores, indicating potential challenges in accurately distinguishing this metal type from others. In terms of macro-average and weighted-average metrics, both hover around 0.85, indicating a reasonably balanced performance across all metal types. These metrics consider the average performance across all classes, with macro-average treating all classes equally, while weighted-average considers the

contribution of each class based on its support. Table 4.4 shows the classification

	Precision	Recall	F1-Score	Support
Ligand 1	0.88	0.78	0.82	9
Ligand 2	0.88	0.88	0.88	8
Ligand 3	0.75	0.86	0.80	7
Ligand 4	0.45	0.71	0.56	7
Ligand 5	0.78	0.70	0.74	10
Ligand 6	1.00	0.86	0.92	7
Ligand 7	0.71	0.50	0.59	10
Ligand 8	0.67	0.89	0.76	9
Ligand 9	1.00	0.67	0.80	3
Ligand 10	1.00	0.90	0.95	10
Accuracy			0.78	80
MacroAvg	0.81	0.77	0.78	80
WeightedAvg	0.80	0.78	0.78	80

Table 4.4: CV Ligands Classification Report

report for classifying ligands. The classifier achieved an accuracy of 78% overall, indicating its capability to classify different metal types to some extent. However, upon closer examination, there are notable variations in performance across metal types. For instance, Metal 6 demonstrates excellent precision, recall, and F1-score, suggesting the model's proficiency in accurately classifying this metal type. Conversely, Metal 4 exhibits lower precision, recall, and F1-score, indicating challenges in effectively distinguishing this metal type from others. Table 4.5 and Table 4.6 show the classification reports for DPV ligands and metals. However, it is difficult to

	Precision	Recall	F1-Score	Support
Ligand 1	1.00	1.00	1.00	1
Ligand 2	0.00	0.00	0.00	2
Ligand 3	0.33	0.25	0.29	4
Ligand 4	0.33	0.33	0.33	3
Ligand 5	0.50	0.50	0.50	4
Ligand 6	0.00	0.00	0.00	1
Ligand 7	0.00	0.00	0.00	1
Ligand 8	0.00	0.00	0.00	0
Ligand 9	0.00	0.00	0.00	1
Ligand 10	1.00	0.33	0.50	3
Accuracy			0.30	20
MacroAvg	0.32	0.24	0.26	20
WeightedAvg	0.42	0.30	0.33	20

Table 4.5: DPV Ligands Classification Report

draw definitive conclusions from this data due to the small sample size. To further assess the performance of these classification models, receiving operating characteristic (ROC) curves and area under the ROC curve (AUC) values can be used to gain valuable insights into the discrimination capabilities and robustness of the models when distinguishing between different metals and ligands. ROC curves and AUC values help assess the robustness of the classification model by showing how well it performs across different thresholds and levels of noise. A smooth ROC curve with a high AUC suggests that the model can reliably discriminate between different metals and ligands even in the presence of noise or variability. Furthermore, there may be a need

	Precision	Recall	F1-Score	Support
Ligand 1	0.33	1.00	0.50	2
Ligand 2	0.00	0.00	0.00	1
Ligand 3	0.25	0.50	0.33	2
Ligand 4	0.00	0.00	0.00	3
Ligand 5	0.00	0.00	0.00	1
Ligand 6	0.50	0.25	0.33	4
Ligand 7	0.00	0.00	0.00	2
Ligand 8	1.00	0.50	0.67	2
Ligand 9	0.00	0.00	0.00	3
Ligand 10	0.00	0.00	0.00	0
Accuracy			0.30	20
MacroAvg	0.23	0.25	0.20	20
WeightedAvg	0.26	0.25	0.22	20

Table 4.6: DPV Metals Classification Report

to choose a classification threshold that balances sensitivity and specificity according to specific requirements or constraints. ROC curves provide a visual aid for selecting an appropriate threshold based on the desired trade-off between true positives and false positives. For example, when integrating with an SDL, minimizing false positives (misclassification of metals or ligands) might be prioritized over maximizing true positives. In Figure 4.3 and Figure 4.4 the ROC curves show good results for both metals and ligands. The area under the ROC curve (AUC) calculation summarized the ROC curve analysis into a scalar value, which ranges between 0 and 1. The closer the AUC score to value 1, the better the application's overall performance. In Figure

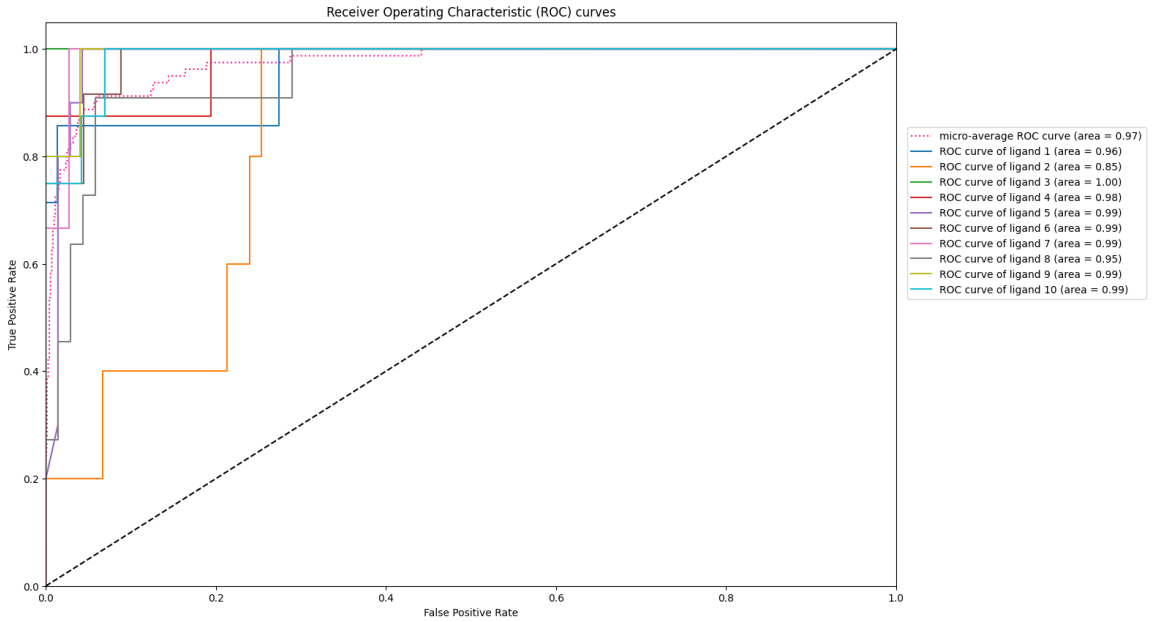


Figure 4.3: CV Ligand ROC Curves

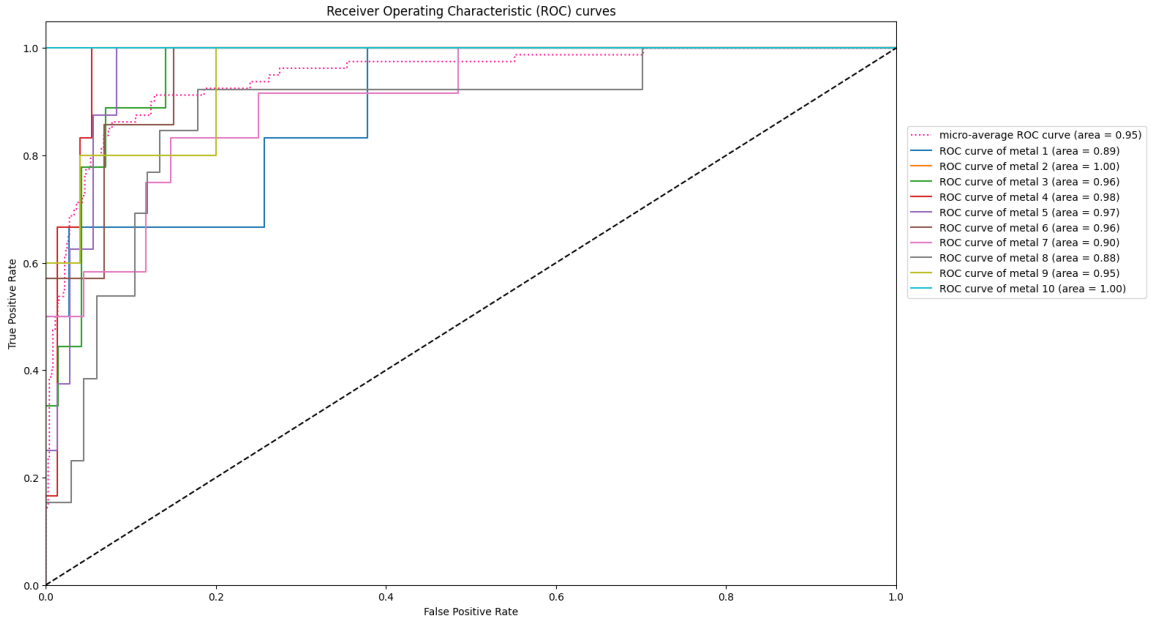


Figure 4.4: CV Metal ROC Curves

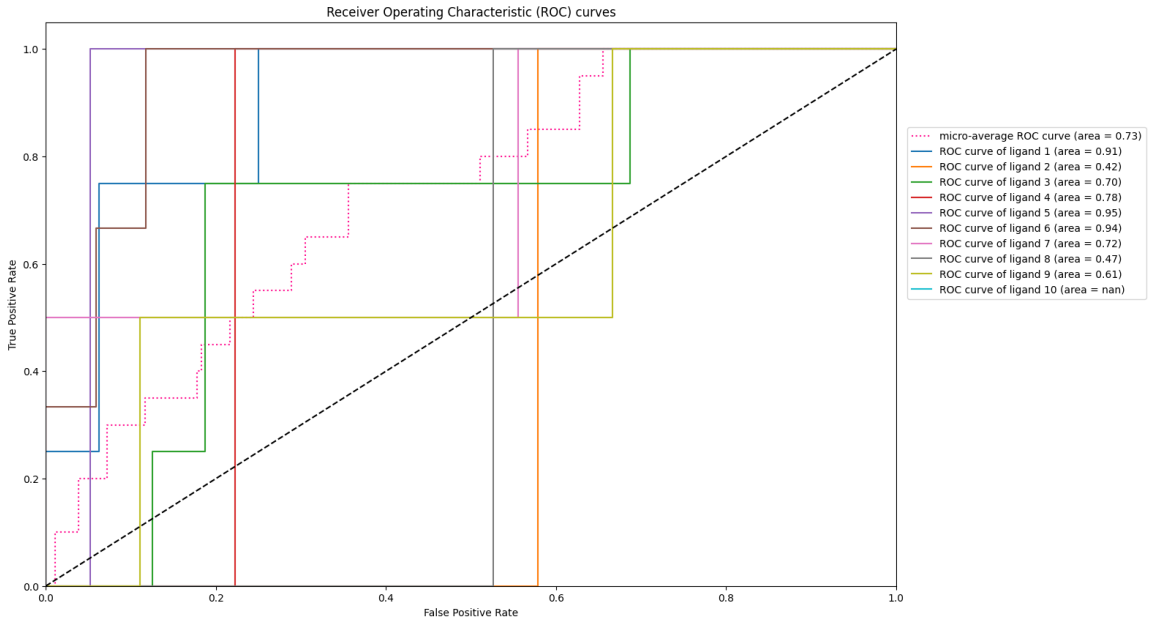


Figure 4.5: DPV Ligand ROC Curves

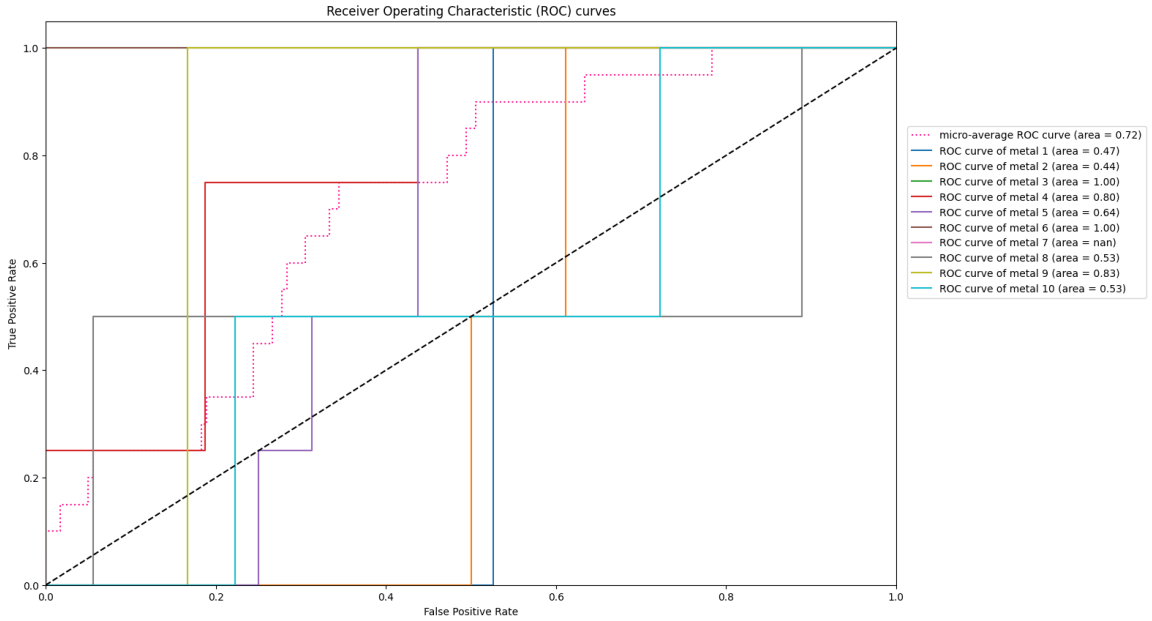


Figure 4.6: DPV Metal ROC Curves

4.5 and Figure 4.6, the ROC curves show that the classifier outperforms a random classifier by having an AUC value above 0.5. The data itself may be causing issues

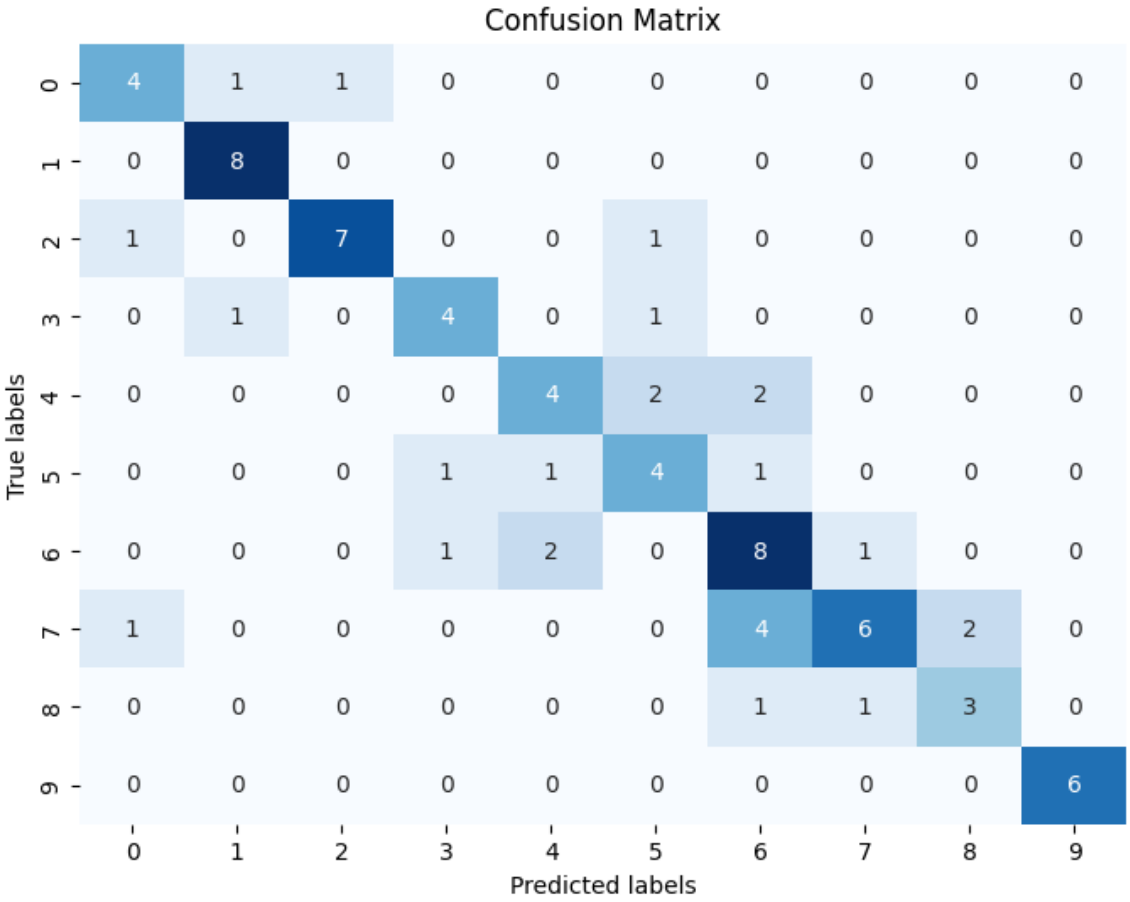


Figure 4.7: CV Ligand Confusion Matrix

for classification as some ligands and metals may be more difficult to distinguish than others. To investigate this, the confusion matrices are provided. From the confusion matrix for ligands 4.7, ligand 7 was often misclassified as metal 6. However, this misclassification is understandable. As seen in Figure 4.8, the voltammograms for ligand 6 and ligand 7 are quite similar. From the confusion matrix for metals 4.9, metal 1 was difficult to recognize with many metals being misclassified as metal 1. From

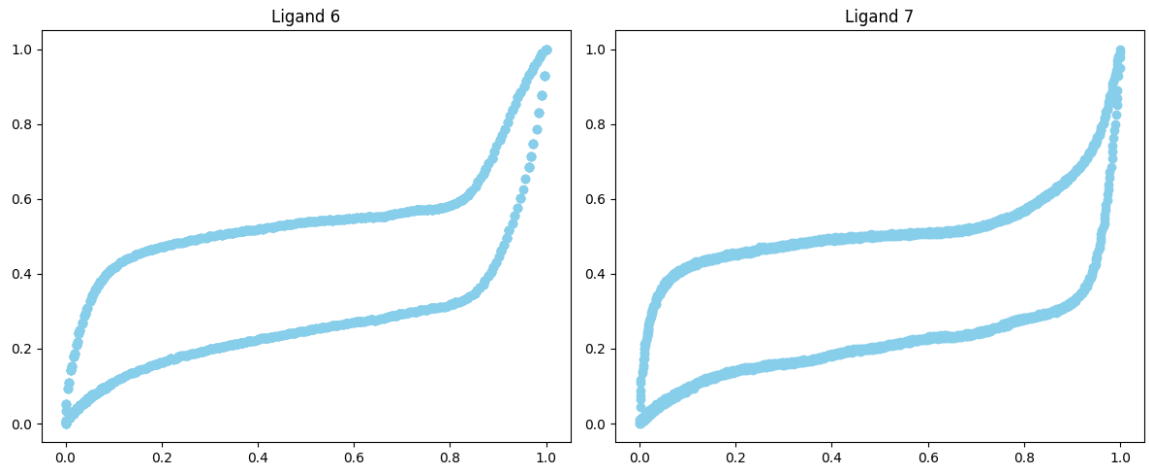


Figure 4.8: Ligand 6 and Ligand 7 Voltammogram Comparison

the DPV confusion matrices seen in Figure 4.11 and Figure 4.10, it is hard to draw any definitive conclusions due to the dataset size. A major challenge in supervised learning is providing good examples during training. However, despite using a small dataset, these results are promising.

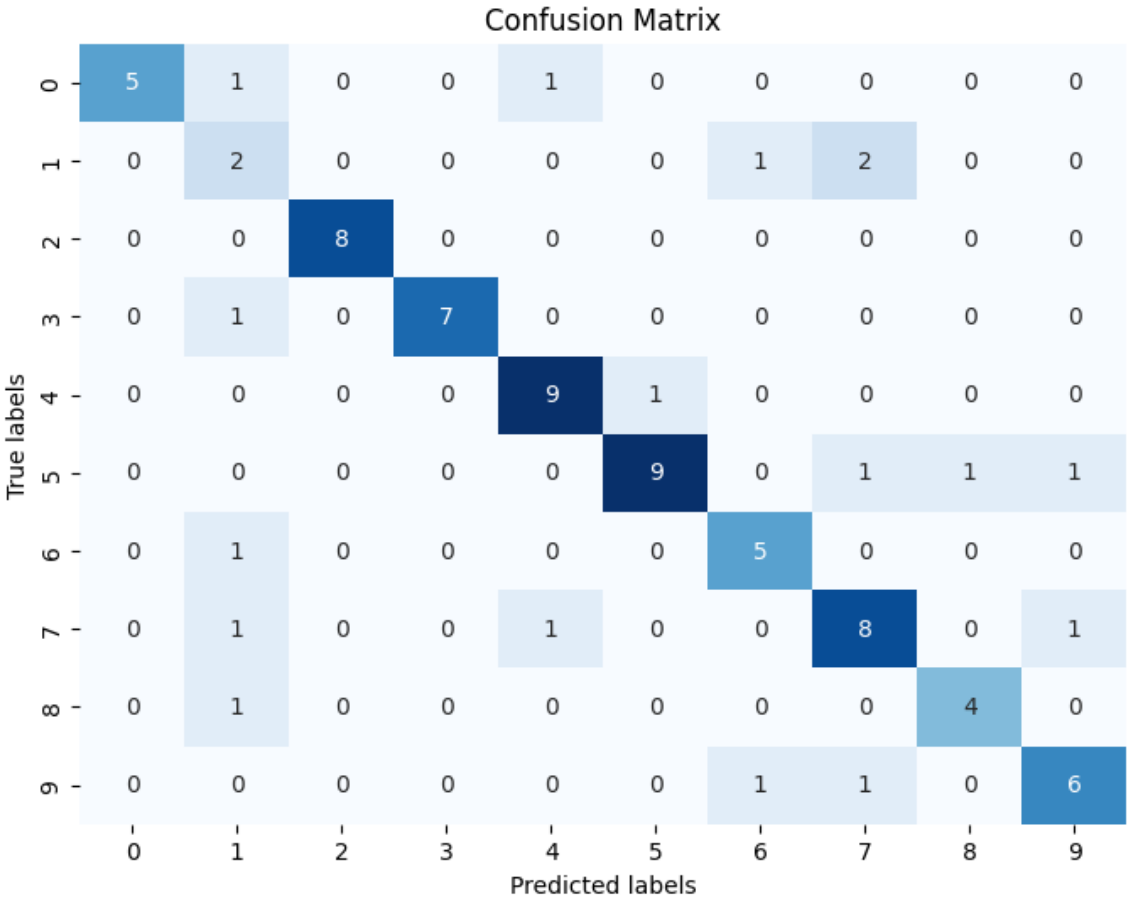


Figure 4.9: CV Metal Confusion Matrix

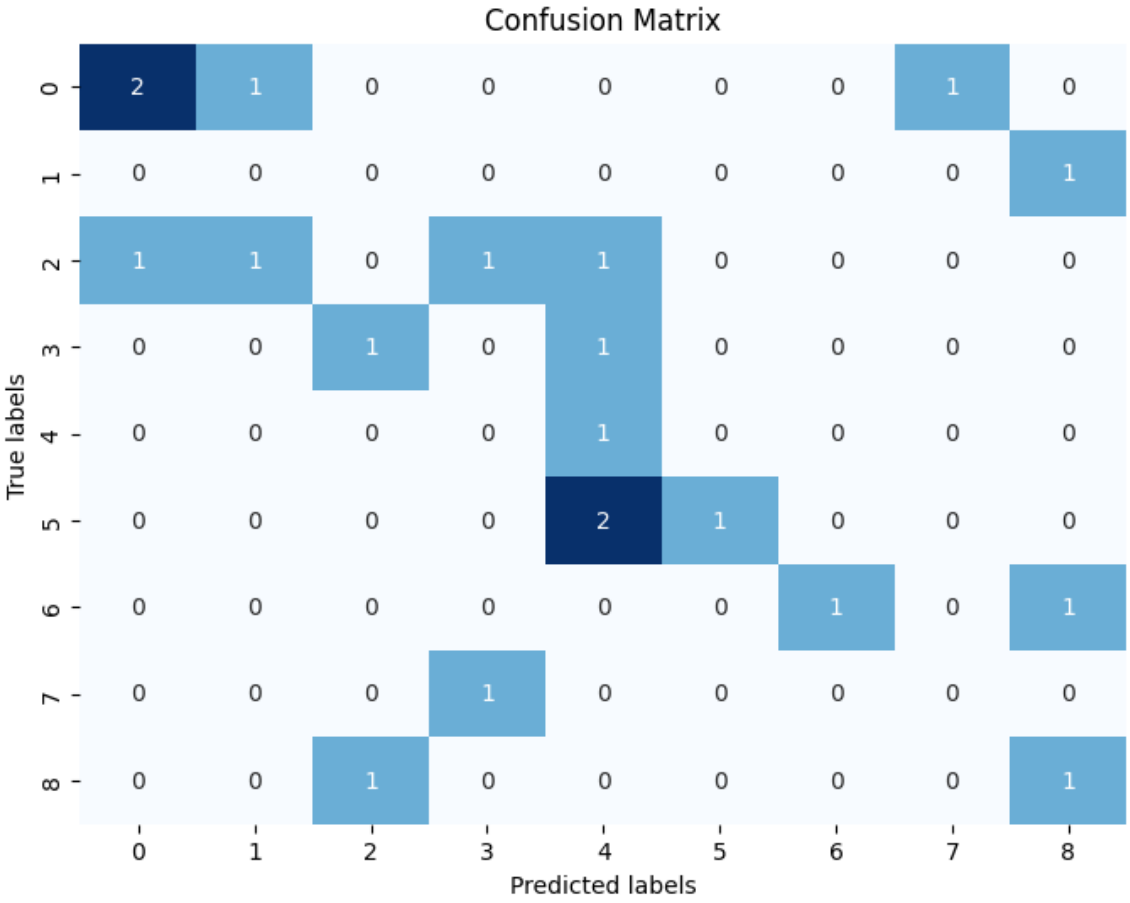


Figure 4.10: DPV Ligand Confusion Matrix

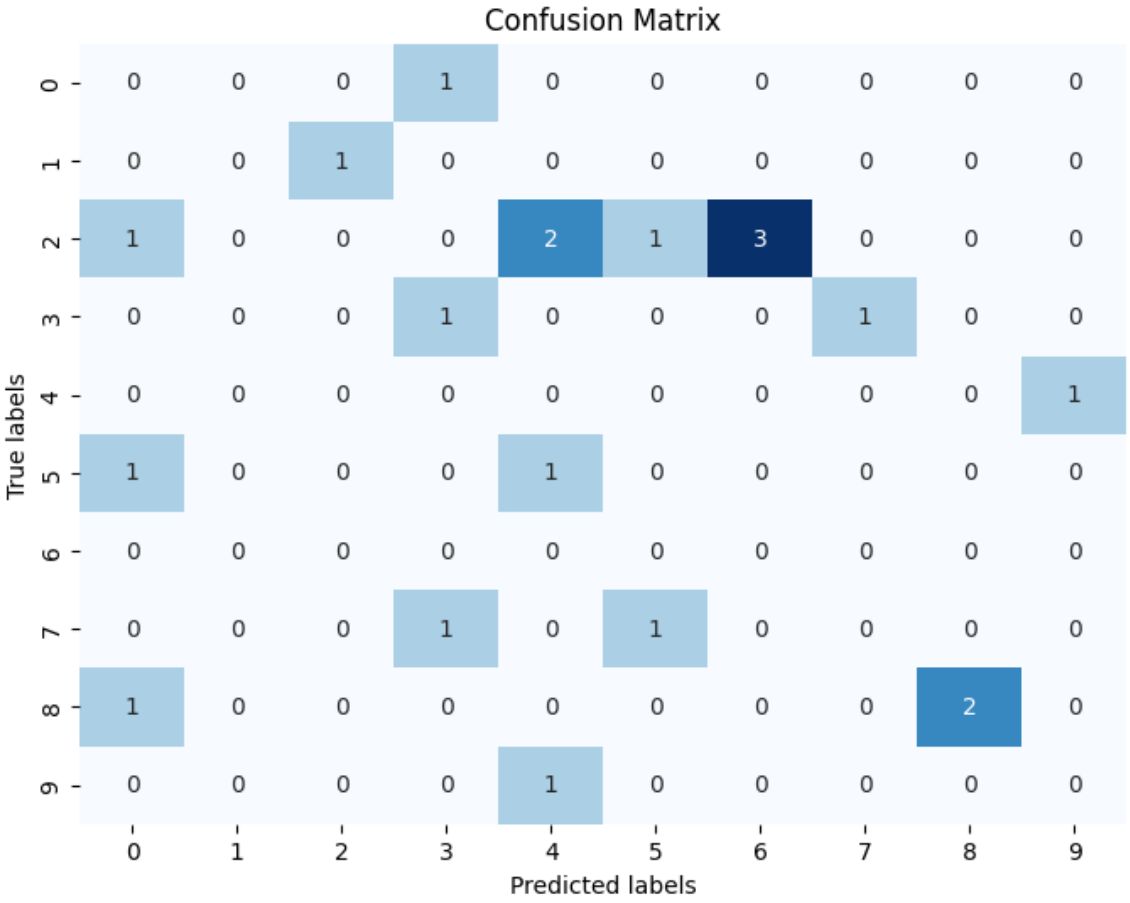


Figure 4.11: DPV Metal Confusion Matrix

Chapter 5

Denoising

5.1 Introduction

As the previously discussed dataset was generated using a low-cost potentiostat that lacks the accuracy of commercial options [15], we attempt to improve the quality of the data obtained by this potentiostat by applying ML to denoise the raw data with the commercial potentiostat data as a reference.

5.2 Autoencoder

As previously shown in Figure 4.1, an autoencoder is a neural network used to learn an efficient low-dimensional encoding of data. An autoencoder consists of an encoder and a decoder. The encoder transforms the input data into an encoded representation, and the decoder attempts to recreate the data from the encoded representation. Since the goal is to try and improve the data quality, the commercial potentiostat data is used for the decoder instead. This way, the low-cost potentiostat data is used to create an encoded representation, and an equivalent commercial potentiostat data is decoded. The main problem to solve is how to pair results from the two potentiostats. While the metal and ligand used for each experiment are recorded, numerous other variables can influence the data. Therefore, the challenge revolves around accurately

aligning the outcomes generated by the low-cost potentiostat with their counterparts from the commercial one. This alignment is crucial for ensuring the reliability and validity of the encoded representations created by the autoencoder. Without precise pairing, the encoded representations may not accurately capture the underlying patterns in the data, leading to suboptimal performance of the autoencoder. To address this challenge, the clustering technique previously described can be employed. By leveraging the recorded information on metals and ligands, combined with other relevant variables, similar experimental results can be grouped. This clustering process helps in identifying pairs of results that share comparable characteristics, despite potential variations introduced by the different potentiostats.

5.3 Results and Discussion

In Figure 5.1, both the As seen in Figure 5.1, both the input and output are similar in overall shape. However, the output contains a much more defined duck-shaped voltammogram, which is typically expected. The results show promising outcomes and indicate that an autoencoder can effectively transform data from the low-cost potentiostat to resemble data from the commercial potentiostat. By leveraging the capacity of deep neural networks to learn complex patterns and relationships within the data, it becomes feasible to enhance the quality of measurements obtained from low-cost instruments, thereby expanding their utility in research and industrial applications. However, despite the promising results, several drawbacks and considerations must be acknowledged. Firstly, the effectiveness of the transformation heavily

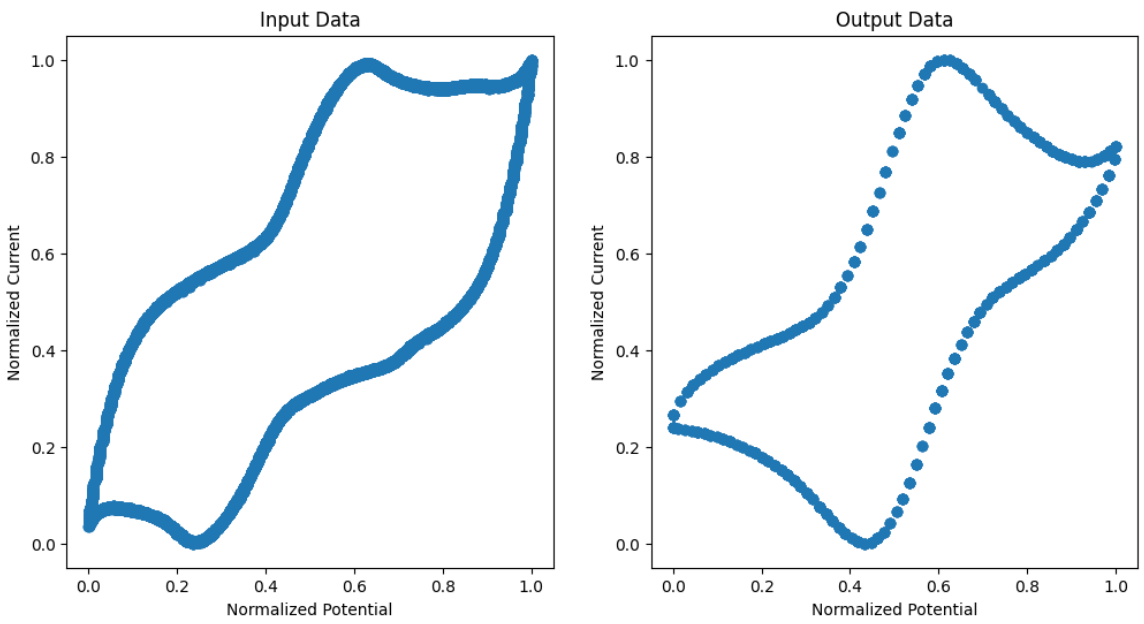


Figure 5.1: AutoEncoder Results

relies on the quality and diversity of the training data. Insufficient or biased training samples may lead to suboptimal performance and generalization issues, especially when dealing with complex electrochemical processes or diverse experimental conditions. While the autoencoder can effectively capture and replicate the dominant features present in the data, it may struggle with preserving subtle nuances or domain-specific characteristics inherent to the commercial potentiostat. Variations in hardware specifics, measurement protocols, or environmental factors could introduce discrepancies between the transformed and reference datasets. In conclusion, while autoencoders offer a promising avenue for enhancing the capabilities of low-cost potentiostats, their deployment must be accompanied by rigorous validation and consideration of the aforementioned limitations. Future research could focus on optimizing the autoencoder architecture, exploring alternative deep-learning techniques,

and investigating strategies for addressing data heterogeneity to further improve the robustness and versatility of the proposed approach.

Chapter 6

Conclusion

In summary, the novel technique introduced for encoding CV and DPV data represents a pivotal advancement in the realm of SDLs. Shown through accurately segmenting voltammograms based on their characteristics and demonstrated efficacy across various machine learning tasks, including clustering, classification, denoising, and synthetic data generation, this technique signifies a significant step towards autonomous SDLs. Furthermore, its versatility extends beyond SDLs and can be applied to any 2-dimensional data. Moving forward, further exploration into alternative curve simplification algorithms and integration of the encoding technique into operational SDL frameworks stand as promising avenues for future research and development.

Bibliographic references

1. Tom, G. *et al.* Self-driving laboratories for chemistry and materials science. doi:10.26434/chemrxiv-2024-rj946 (2024).

2. Hickman, R. *et al.* Atlas: a brain for self-driving laboratories. doi:10.26434/chemrxiv-2023-8nrxx (2023).

3. Strieth-Kalthoff, F. *et al.* Delocalized, asynchronous, closed-loop discovery of organic laser emitters. doi:10.26434/chemrxiv-2023-wqp0d (2023).

4. Faraday, M. S. (B. *On electro-chemical decomposition* 1970.

5. Electrode potential. doi:doi:10.1351/goldbook.E01956 (2019).

6. Elgrishi, N., Rountree, K. J., McCarthy, B. D., Rountree, E. S., Eisenhart, T. T. & Dempsey, J. L. A practical beginner's guide to cyclic voltammetry. *Journal of chemical education* **95,** 197–206. doi:10.1021/acs.jchemed.7b00361 (2018).

7. *Handbook of electrochemistry* (ed Zoski, C. G.) (Elsevier Science, London, England, 2006).

8. Wain, A. J. & Dickinson, E. J. in *Nanoscale electrochemistry* (eds Wain, A. J. & Dickinson, E. J.) 1–48 (Elsevier, 2021). doi:https://doi.org/10.1016/B978-0-12-820055-1.00008-3.

9. Nicholson, R. S. & Shain, I. Theory of stationary electrode polarography. single scan and cyclic methods applied to reversible, irreversible, and kinetic systems. *Analytical chemistry* **36,** 706–723. doi:10.1021/ac60210a007. eprint: https://doi.org/10.1021/ac60210a007 (1964).

10. Heinze, J. Cyclic voltammetry—"electrochemical spectroscopy". new analytical methods (25). *Angewandte chemie international edition in english* **23,** 831–847. doi:https://doi.org/10.1002/anie.198408313. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.198408313 (1984).

11. Nicholson, R. S. & Shain, I. Theory of stationary electrode polarography. single scan and cyclic methods applied to reversible, irreversible, and kinetic systems. *Analytical chemistry* **36,** 706–723. doi:10.1021/ac60210a007 (1964).

12. Libretexts. *Cyclic voltammetry* 2023.

13. Grimshaw, J. in *Electrochemical reactions and mechanisms in organic chemistry* (ed Grimshaw, J.) 1–26 (Elsevier Science B.V., Amsterdam, 2000). doi:https://doi.org/10.1016/B978-044472007-8/50001-X.

14. *Electroanalytical methods* 1st ed. en (ed Scholz, F.) (Springer, Berlin, Germany, 2005).

15. Pablo-García, S. *et al.* An affordable platform for automated synthesis and electrochemical characterization. doi:10.26434/chemrxiv-2024-cwnwc (2024).

16. MacQueen, J. B. *Some methods for classification and analysis of multivariate observations* in *Proc. of the fifth berkeley symposium on mathematical statistics and probability* (eds Cam, L. M. L. & Neyman, J.) **1** (University of California Press, 1967), 281–297.

17. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A density-based algorithm for discovering clusters in large spatial databases with noise* in *Knowledge discovery and data mining* (1996).

18. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9,** 2579–2605 (2008).

19. McInnes, L., Healy, J. & Melville, J. *Umap: uniform manifold approximation and projection for dimension reduction* 2018. doi:10.48550/ARXIV.1802.03426.

20. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20,** 53–65. doi:https://doi.org/10.1016/0377-0427(87)90125-7 (1987).

21. Sharma, N., Jain, V. & Mishra, A. An analysis of convolutional neural networks for image classification. *Procedia computer science* **132.** International Conference on Computational Intelligence and Data Science, 377–384. doi:https://doi.org/10.1016/j.procs.2018.05.198 (2018).

22. Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, E. A. & Lima Netto, S. in *Variational methods for machine learning with applications to deep networks* 111–149 (Springer International Publishing, 2021). doi:10.1007/978-3-030-70679-1_5.

23. Dirac, P. A. M. *The principles of quantum mechanics* (Clarendon Press, 1981).

24. Knuth, D. *Knuth: computers and typesetting* https://www-cs-faculty.stanford.edu/~knuth/abcde.html.

25. Knuth, D. E. in. Chap. 1.2 (Addison-Wesley, 1973).

26. J., A. & Faulkner, L. R. *Student solutions manual to accompany electrochemical methods: fundamentals and applicaitons, 2e* en (John Wiley & Sons, Nashville, TN, 2002).

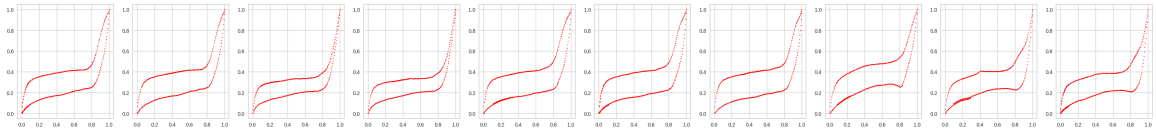
27. Yoshikawa, N., Akkoc, G. D., Pablo-García, S., Cao, Y., Hao, H. & Aspuru-Guzik, A. Does one need to polish electrodes in an eight pattern? automation provides the answer. doi:10.26434/chemrxiv-2024-ttxnr (2024).

28. Kramer, M. Autoassociative neural networks. *Computers amp; chemical engineering* **16,** 313–328. doi:10.1016/0098-1354(92)80051-a (1992).

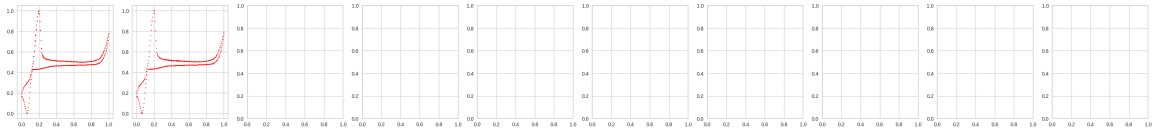
Appendix A

CV K-Means Cluster Results

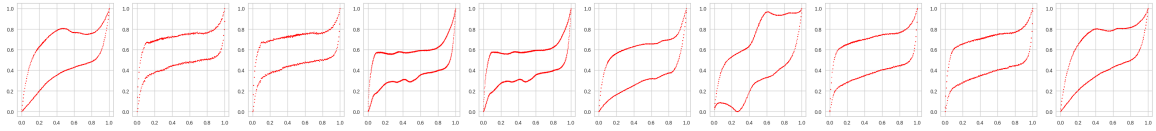
Cluster 1:



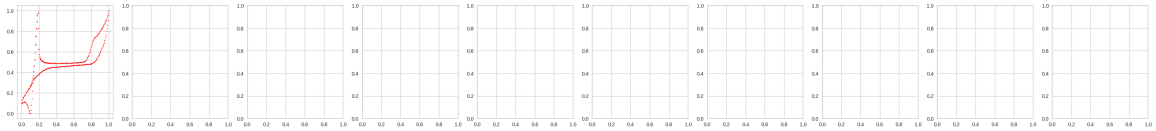
Cluster 2:



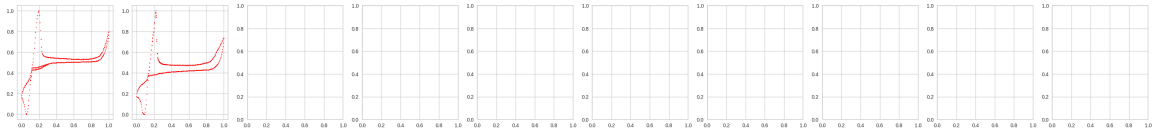
Cluster 3:



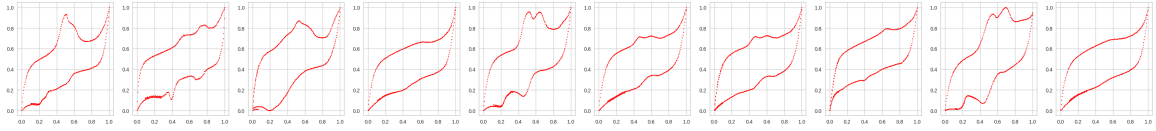
Cluster 4:



Cluster 5:

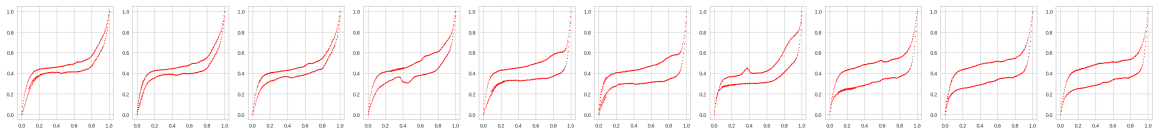


Cluster 6:

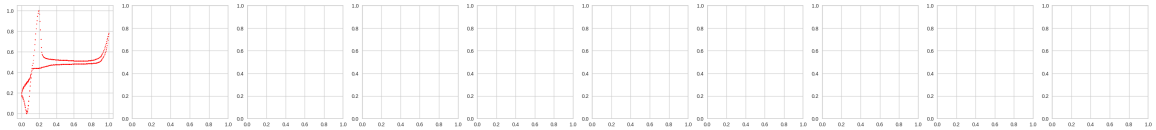


Cluster 7:

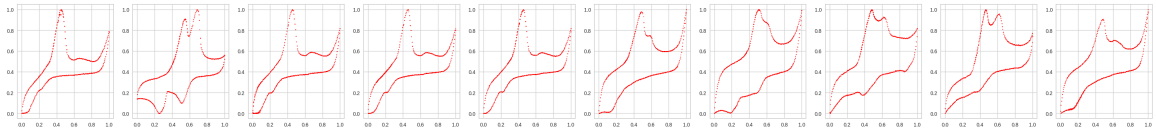
Appendix A. CV K-Means Cluster Results



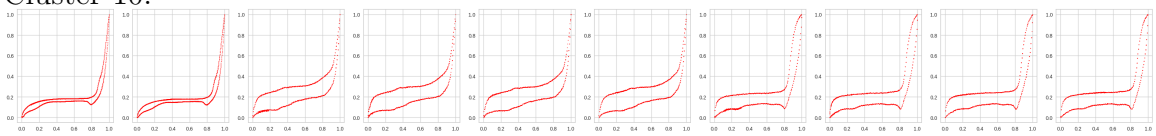
Cluster 8:



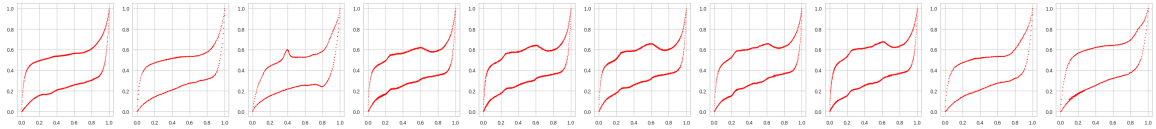
Cluster 9:



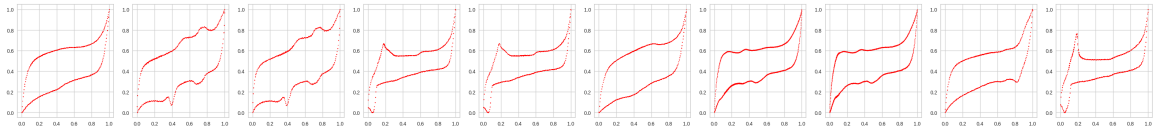
Cluster 10:



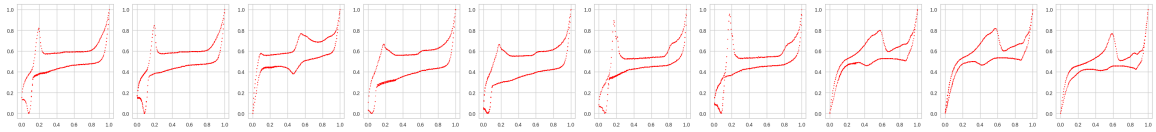
Cluster 11:



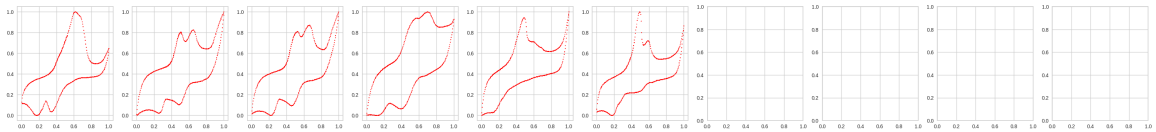
Cluster 12:



Cluster 13:

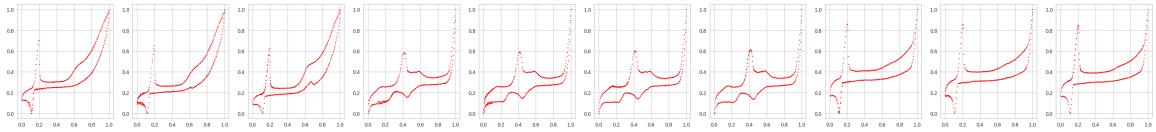


Cluster 14:

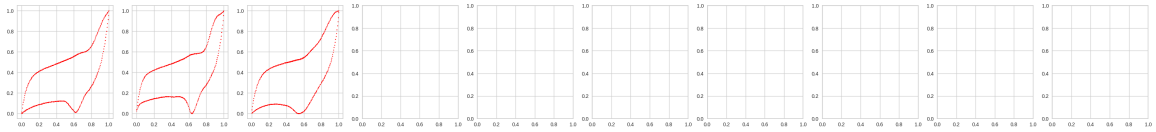


Appendix A. CV K-Means Cluster Results

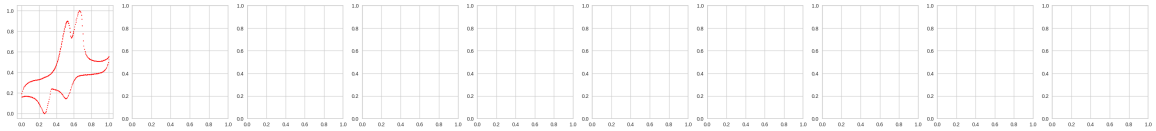
Cluster 15:



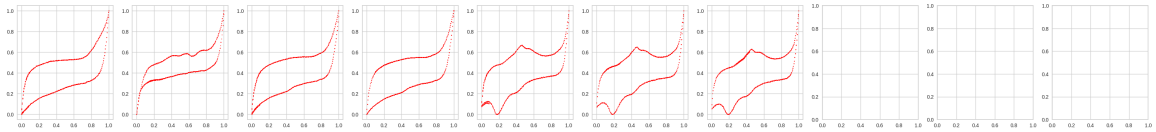
Cluster 16:



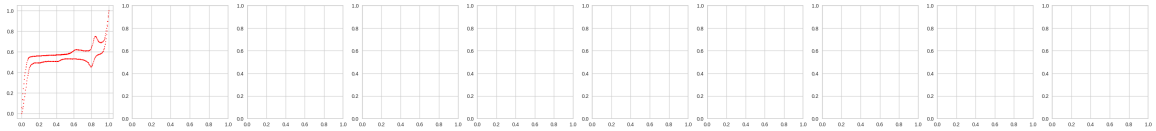
Cluster 17:



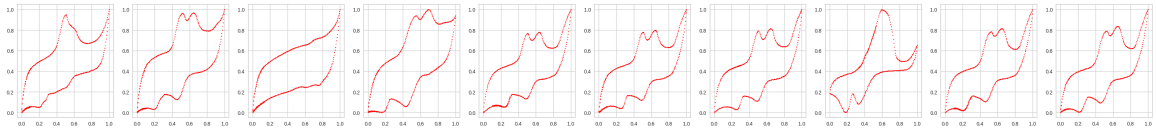
Cluster 18:



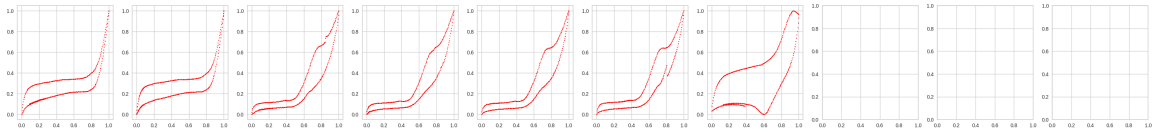
Cluster 19:



Cluster 20:

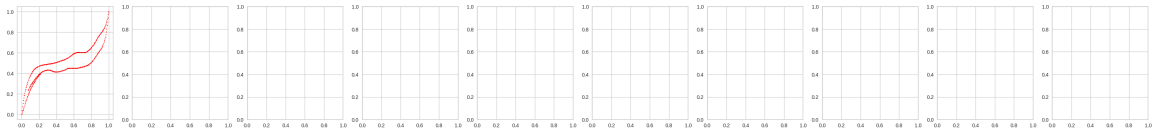


Cluster 21:

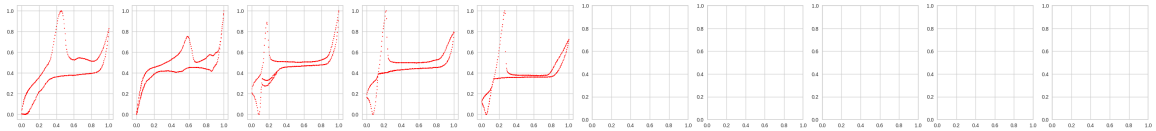


Cluster 22:

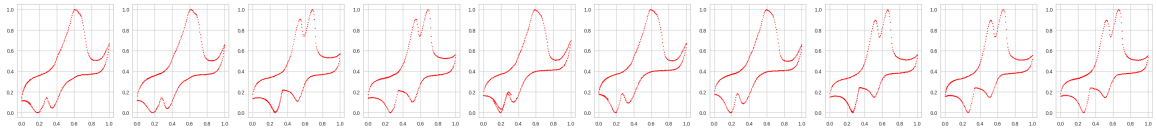
Appendix A. CV K-Means Cluster Results



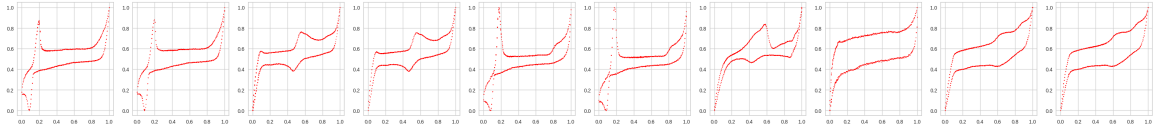
Cluster 23:



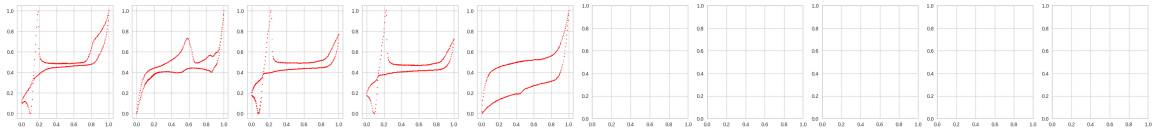
Cluster 24:



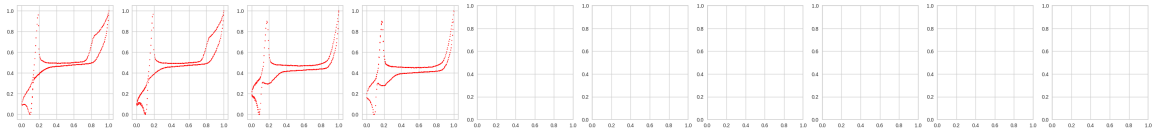
Cluster 25:



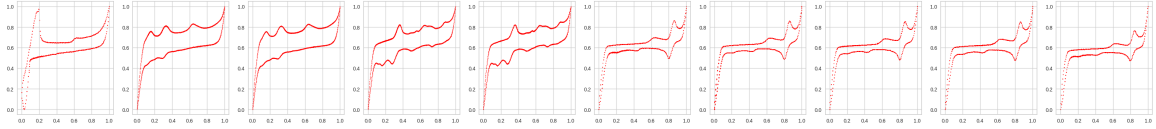
Cluster 26:



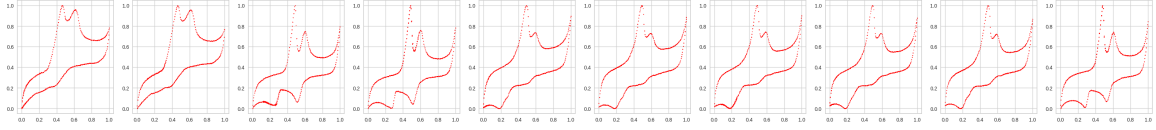
Cluster 27:



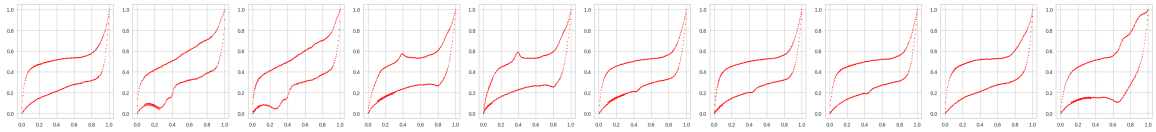
Cluster 28:



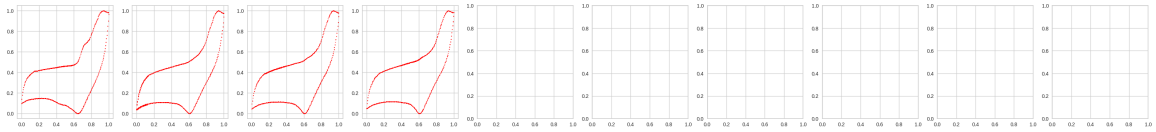
Cluster 29:



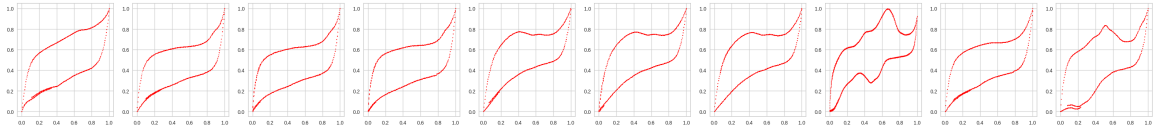
Cluster 30:



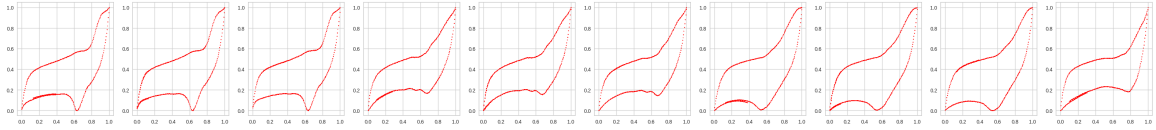
Cluster 31:



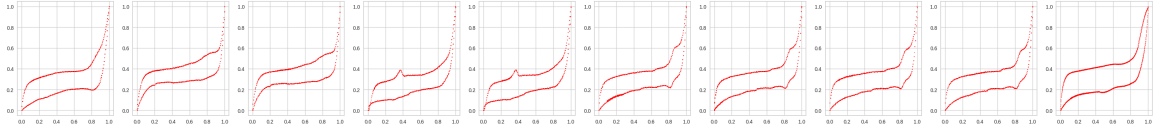
Cluster 32:



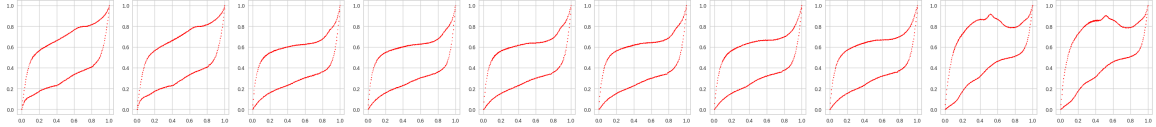
Cluster 33:



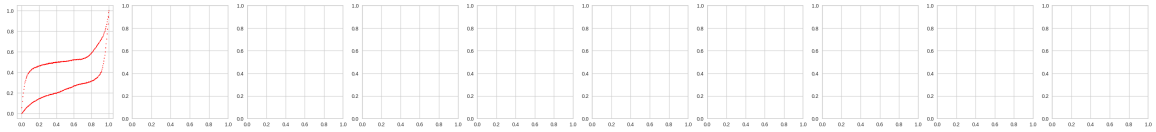
Cluster 34:



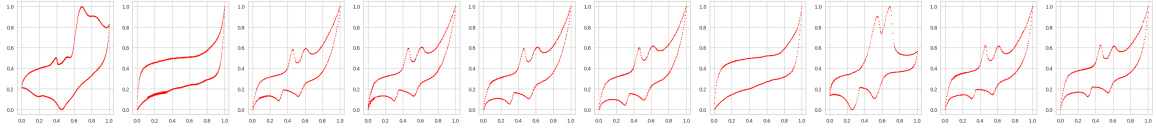
Cluster 35:



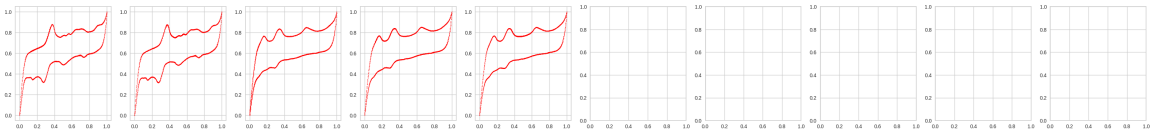
Cluster 36:



Cluster 37:



Cluster 38:



A.1 Metals and Ligands

Entry	Metal	Form	CAS
1	V(IV)	$VOSO_4xH_2O$	123334-20-3
2	Cr(III)	$CrK(SO_4)_212H_2O$	7788-99-0
3	Mn(II)	$MnSO_4H_2O$	10034-96-5
4	Fe(II)	$FeSO_47H_2O$	7782-63-0
5	Co(II)	$CoSO_47H_2O$	10026-24-1
6	Ni(II)	$NiSO_46H_2O$	10101-97-0
7	Cu(II)	$CuSO_45H_2O$	7758-99-8
8	Zn(II)	$ZnSO_47H_2O$	7446-20-0
9	Cd(II)	$CdSO_48/3H_2O$	7790-84-3
10	Pd(II)	Na_2PdCl_4	13820-53-6

Table A.1: Table of Metals

Entry	Ligand	SMILES	CAS
1	ammonia	N	1336-21-6
2	hydrazine	NN	7803-57-8
3	ethylenediamine	NCCN	107-15-3
4	ethanolamine	NCCO	141-43-5
5	diethanolamine	OCCNCCO	111-42-4
6	triethanolamine	OCCN(CCO)CCO	102-71-6
7	piperidine	N1CCCCC1	110-89-4
8	morpholine	N1CCOCC1	110-91-8
9	pyridine	n1ccccc1	110-86-1
10	2,2'-bipyridine (in HCl salt form)	c1ccc(nc1)c2ccccn2	336-18-7

Table A.2: Table of Ligands