UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

MÉTODOS DE CLASIFICACIÓN NO SUPERVISADA PARA TRAZAS DE CONDUCTIVIDAD DE LA FAMILIA DE LAS PORFIRINAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL ELÉCTRICA

CONSTANZA RIVERA KRUMM

PROFESORA GUÍA:
DIANA DULIC

MIEMBROS DE LA COMISIÓN:
MARCOS ORCHARD CONCHA
JORGE SILVA SÁNCHEZ

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERA CIVIL
ELÉCTRICA
POR: **CONSTANZA RIVERA KRUMM**
FECHA: 2021
PROF. GUÍA: DIANA DULIC

MÉTODOS DE CLASIFICACIÓN NO SUPERVISADA PARA TRAZAS DE CONDUCTIVIDAD DE LA FAMILIA DE LAS PORFIRINAS

Utilizar moléculas aisladas como componente electrónico es cada vez más requerido para el desarrollo de la tecnología actual. En este contexto el estudio de la conductividad de moléculas se vuelve fundamental para encontrar materiales con un buen desempeño en esta área.

Se han realizado estudios de las trazas de conductividad de la familia de los alcanos a través de mediciones de ruptura de juntura mecánicamente controlada (MCBJ) utilizados para analizar posibles aplicaciones según sus características. Debido a la naturaleza del experimento es muy difícil obtener observaciones precisas, existe una aleatoriedad inherente a cada realización, por lo que en virtud de obtener información veraz sobre características de las moléculas es necesario aplicar métodos de clasificación no supervisada que permitan extraer información esencial sobre las moléculas y sortear la estocaticidad.

En este trabajo se buscará, a través de la implementación de diferentes métodos de clasificación no supervisada, estudiar las diferentes clases correspondientes a los canales de transporte en una porfirina.

Para ti que vas a leer mi trabajo,
espero lo disfrutes.

Agradecimientos

No puedo no comenzar agradeciendo a mi familia, quienes me dieron la oportunidad de estudiar y llegar a este punto de la carrera dándome todas las facilidades para lograrlo. Muchas gracias.

Mi trayectoria universitaria fue intervenida por distintas personas quienes hicieron este proceso más ameno, divertido y soportable. Después de cada prueba o clase sin los K de la E esperando para sacarte una sonrisa nada hubiese sido igual, quizás hasta imposible. Les resumo en apoyo y apañe infinito. Pero quiero hacer una mención especial a quien estuvo ahí desde el día 1 que pisé Beauchef, mi amiga, Isi, gracias por todo el cariño, el apañe, los carretes, las risas, las penas y las estupideces, eres mi mejor recuerdo de la universidad y espero que sigamos haciendo más recuerdos estando fuera.

A mi pareja, que sin él este último año, aún estando en pandemia, supo como distraerme del estrés y ver el lado bueno de las cosas.

Por último, y nunca menos importante, quien me dio la oportunidad de realizar este trabajo, ser auxiliar, aprender más allá de las salas de clases, quien me dio apoyo y aliento haciéndome saber que siempre se puede, la mejor profesora que pude encontrarme en una clase de zumba: Diana. Gracias por todo y por la confianza.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Objetivos

Este trabajo tiene como objetivo general diseñar e implementar herramientas de clasificación no supervisada en MATLAB para trazas de conductancia de Bulky P3 para realizar comparaciones con estudios anteriores de moléculas de P3 y analizar sus resultados.

Objetivos específicos

- Estudio de la literatura actual para obtener resultados preliminares.
- Diseñar y evaluar un método de clasificación no supervisada no utilizado anteriormente en trazas de conductancia.
- Presentar observaciones y conclusiones al respecto.

1.2. Motivación

El uso de moléculas aisladas como componente electrónico ha tomado relevancia en los últimos años. Pero su uso depende fuertemente del conocimiento que se tiene sobre ellas y sus características de conductancia.

Debido a que trabajar con moléculas aisladas implica mediciones en escala nanométrica es complicado obtener un modelo certero del comportamiento de estas. Es por esto que, a través de la aplicación de clasificación no supervisada, se busca encontrar un mejor acercamiento al actuar real de éstas moléculas.

Anteriormente, en el artículo titulado *Unravelling the conductance path through single-porphyrin junctions*[1], se buscó estudiar la conductancia de distintos compuestos químicos de porfirinas, donde finalmente se encontraron 3 clases características (A, B y C). Estas clases representan los distintos canales por los cuales se condujo a través de la molécula. Ahora bien, surge la idea de poder reducir el número de clases que se obtienen, de manera de obtener un único canal por el cual de conduce.

Una de las moléculas utilizadas en el estudio fue P3 (Ver Figura 1.1), y para poder disminuir el número de clases obtenidas en P3 se diseñó la molécula Bulky P3 1.2, la cuál posee

dos grupos extra de bencenos a sus costados en pro de disminuir las opciones de canales para conducir, lo que debiese impactar directamente en el nº de clases que se obtienen al momento de clasificar los mediciones de conductancia.

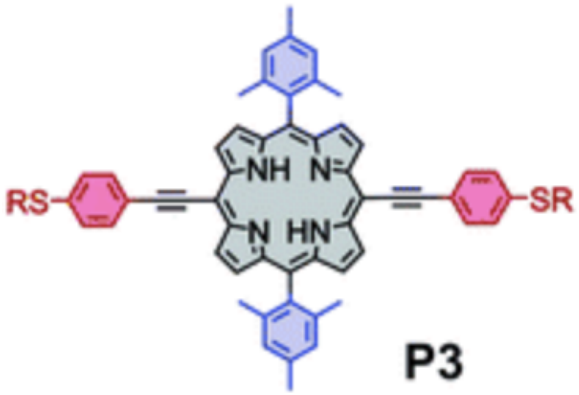


Figura 1.1: P3. Figura extraída de [1]

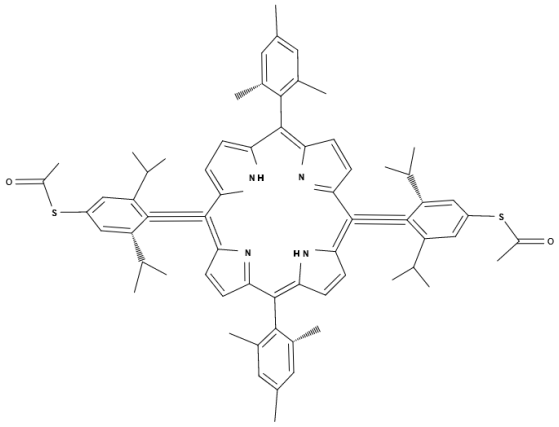


Figura 1.2: Bulky P3

1.3. Estado del Arte

El campo de la electrónica molecular ha crecido increíblemente. Sin embargo, siguen existiendo grandes desafíos al momento de definir aspectos claves de la conductancia de moléculas como la geometría de la unión o su configuración debido a las dificultades que presenta manipular moléculas.

Diversos estudios se han realizado [2][3][4] en el campo de la conductancia de diferentes moléculas, en los cuales se han implementado diferentes herramientas de clasificación [5]-[6] para poder sustraer información relevante de estas. A pesar de que las herramientas existen, aún se encuentran problemas al momento de definir y extraer la información los distintos canales de conducción a través del *clustering*.

Reciente se ha realizado un estudio para definir con qué algoritmos se tienen mejores resultados en la clasificación de datos de trazas de conductancia en moléculas OPE3[7]. Este artículo realiza comparaciones entre distintos métodos para definir el espacio de características en conjunto de distintos métodos de clasificación no supervisada. Como resultado los mejores resultados se dan utilizando un espacio de características reducido a imágenes de 28x28 píxeles en conjunto de UMAP o t-SNE, utilizando los algoritmos GDL (graph degree linkage), GMM (gaussian mixed model) y GAL (graph average linkage). Donde se aprecia que la influencia del espacio característico tiene mayor impacto en los resultados que algoritmo de clasificación en si.

En este trabajo se buscará implementar parte de estas herramientas computacionales en MATLAB para el análisis de datos de trazas de conductancia de moléculas de la familia de las porfirinas, y de esta forma en un futuro poder relacionar su estructura molecular con sus propiedades eléctricas.

Por otro lado, como se mencionó anteriormente, un estudio similar se realizó en una molécula P3 donde se obtuvieron tres clases distintivas (Ver Figura 1.3): clase A, clase B y clase C. Con este trabajo se busca estudiar si es posible encontrar las mismas clases dentro de la nueva molécula diseñada (Bulky P3), si es posible disminuir en número de clases o si se encuentran nuevas clases.

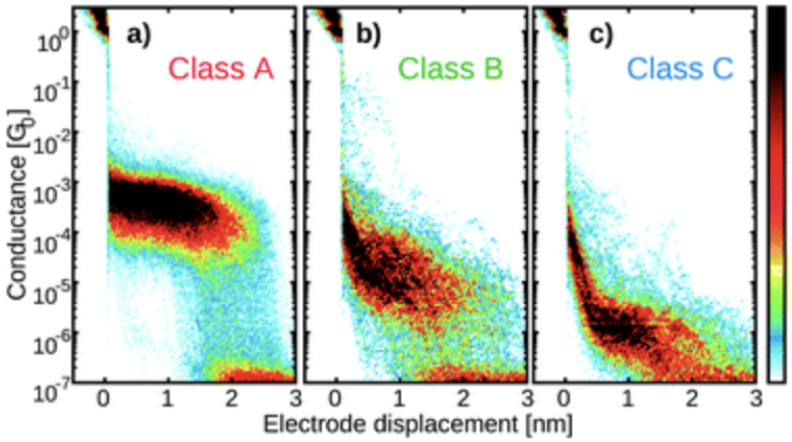


Figura 1.3: Clases A, B y C obtenidas de P3. Figura extraída de [1]

Capítulo 2

Marco Teórico

2.1. Orbitales Moleculares [8]

Si se modelan los electrones dentro de una molécula como un gas de Fermi ideal, en una molécula aislada, los electrones se sitúan en orbitales al rededor del núcleo. Los orbitales moleculares tienen niveles energía discreta desde menor a mayor energía. Los orbitales más importantes para el transporte de carga son los de mayor y los de menor energía, ya que se encuentran cercanos a la energía de Fermi.

En el caso de tener una molécula aislada entre dos piezas de metal con el mismo potencial químico es necesario aplicar un voltaje en sus extremos para que pueda haber traspaso de electrones. El transporte ocurre en el nivel más cercano a la energía de Fermi, y dependiendo de la posición de los orbitales esta ocurrirá en el orbital con nivel más bajo o más alto de energía.

Cuando el metal entra en contacto con la molécula este puede afectar en la tasa de transferencia de electrones, esta interacción se llama acoplamiento electrónico (Γ). Si este acoplamiento es débil los orbitales moleculares pueden no verse afectados o, si el electrón puede ingresar al orbital molecular este se mantiene dentro de la molécula un tiempo suficiente como para perderse y no traspasar a través de la molécula. En cambio, si el acoplamiento es lo suficientemente fuerte se produce un ensanchamiento de los niveles e hibridación de los orbitales moleculares, y por ende la información es capaz de atravesar la molécula. Finalmente, la molécula puede actuar, o no, como puente para los electrones que viajan entre los electrodos.

2.2. Efecto Túnel

Cuando dos electrodos metálicos se encuentran lo suficientemente cerca, dejando un pequeño espacio entre ellos de distancia d, este actúa como una gran barrera de potencial. Según la perspectiva clásica, si un electrón viaja con una energía menor que la barrera de potencial este no podría atravesarla. Sin embargo, desde la perspectiva cuántica existe la probabilidad de que esto ocurra.

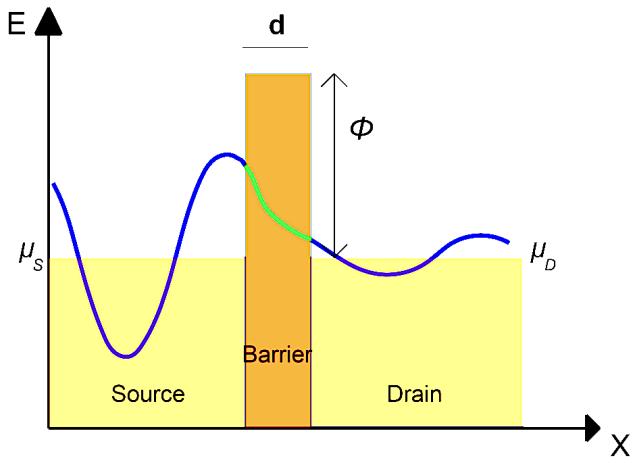


Figura 2.1: Esquemático de efecto túnel. Se utiliza una barrera simplificada rectangular.Las funciones de trabajo de los lados izquierdo y derecho son las mismas, es decir, $\varphi_L = \varphi_R = \varphi$. Fig. extraída de [9]

En la Figura 2.1 la barrera de potencial se modela de forma rectangular con una altura φ y ancho d. La función de onda que atraviesa la barrera (linea azul derecha) tiene baja probabilidad, pero su energía es igual a la energía del electrón entrante (linea azul derecha).El potencial químico de la fuente y el drenaje puede expresarse como:

$$U(x) = \begin{cases} \phi & 0 < x < d \\ o & x < 0 \vee x > d \end{cases} \tag{2.1}$$

La función de onda ψ está dada por la ecuación de Schrödinger:

$$E\psi(x) = \left(-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2} + U(x)) \right) \psi(x). \tag{2.2}$$

La solución a la ecuación 2.2 está dada por:

$$\psi(x) = \begin{cases} A_1 e^{ikx} + A_2 e^{-ikx} & x < 0 \\ B_1 e^{k'x} + B_2 e^{-k'x} & 0 < x < d \\ C_1 e^{ikx} & x > d \end{cases} \tag{2.3}$$

Donde A_1,A_2,B_1,B_2 y C_1 son coeficientes que se determinan a partir de las condiciones de borde. Los vectores de onda están dados por:

$$k = \sqrt{\frac{2m_e E}{\hbar^2}},\ k' = \sqrt{\frac{2m_e(\phi - E)}{\hbar^2}} \tag{2.4}$$

La solución que corresponde a cuando el electrón traspasa la barrera es en el caso de $x > d$. Si se imponen las condiciones de continuidad *psi* y $d\psi/dx$, la probabilidad de que esto ocurra se puede expresar como:

$$T(E) = \left(\frac{A_1}{C_1}\right)^2 \propto exp\left(\frac{-2d}{\hbar}\sqrt{2m_e(\phi - E)}\right) \tag{2.5}$$

El modelo de Simmons[10] presenta una fórmula que detalla la corriente de túnel a través de una barrera arbitraria a un voltaje. Esta se obtiene en conjunto con la ecuación de corriente que fluye a través de la unión cuántica ($I = N/\tau$)[11]:

$$I = \frac{eA}{2\pi h d^2}\left(\phi e^{(\frac{-2d}{\hbar}\sqrt{2m_e\phi})} - (\phi + eV)e^{\frac{-2d}{\hbar}\sqrt{2m_e(\phi+eV)}}\right) \tag{2.6}$$

Donde A es el área del electrodo y V el voltaje aplicado. La Fig. 2.2 muestra la curva característica IV utilizando la el modelo Simmons para distintos factores de asimetría.

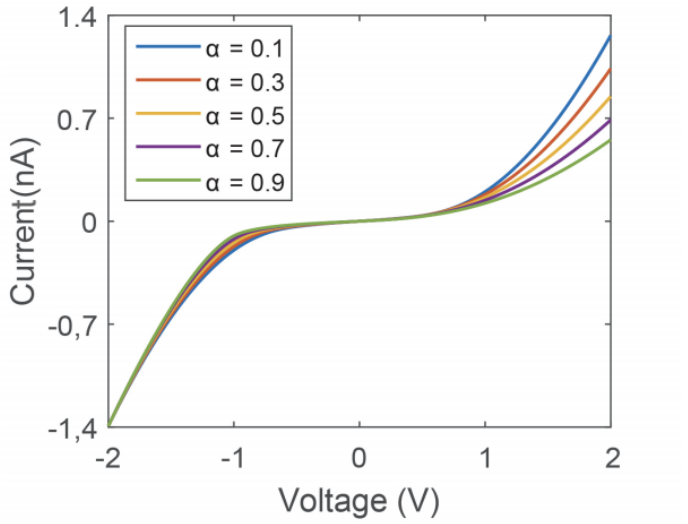


Figura 2.2: Corriente (I) a través una unión de túnel simulada con el modelo de Simmons, programada en MATLAB. Fig. extraída de [9].

2.3. MCBJ

Durante las últimas décadas se han desarrollados distintos métodos para poder estudiar conductancia en moléculas aisladas (STM, ATM, EJB)[12][13], uno de los más populares es la técnica de MCBJ (por sus siglas en ingles *Mechanically Controllable Break Junction*).

Para llevar a cabo el experimento se sitúa la muestra en un mecanismo de flexión de 3 puntos de apoyo sobre un sustrato elástico, como se muestra en la Figura 2.3; y esta se puede deformar de dos maneras: realizando un empuje vertical hacia arriba con la barra o moviendo la muestra hacia abajo.

El empuje vertical es realizado por un piezo eléctrico que empuja el sustrato hasta romper el contacto con los alambres de oro. El movimiento vertical de la barra de empuje se convierte en movimiento lateral con un factor de atenuación que está determinado por la geometría del chip. Basándose en la configuración dispuesta en la Figura 2.4, el factor de atenuación a se estima mediante $a = \Delta l/\Delta z$, donde Δz es el desplazamiento de la barra, Δl es el cambio en la distancia de la abertura.

El segundo método de realiza controlando las abrazaderas, que están conectadas por un tornillo diferencial a un eje de accionamiento, controlado por un servomotor colocado en la parte superior de la configuración. El servomotor es más lento que el piezo eléctrico.

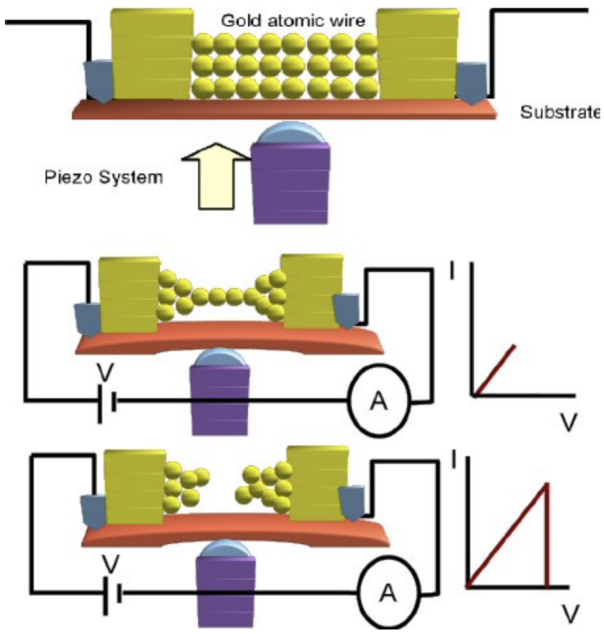


Figura 2.3: Esquema sistemáticos de MCBJ. Figura extraída de [14]

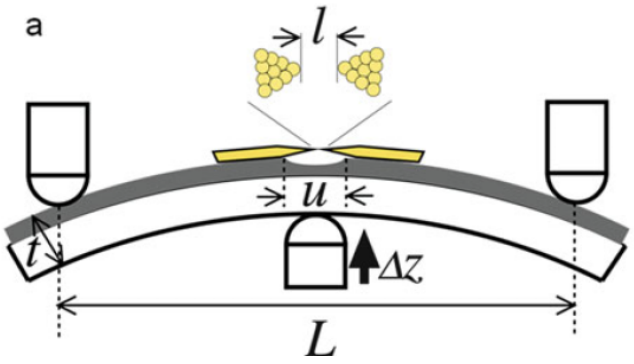


Figura 2.4: Figura extraída de texto [15]

La ventaja de la técnica MCBJ es la alta resolución y la estabilidad de los resultados. Y tiene por objetivo estudiar la ruptura inducida por la corriente de cadenas atómicas de longitudes determinadas.

2.4. Clasificación no supervisada

En esta sección se profundizará en los métodos de clasificación no supervisada a aplicar en los set de datos. La clasificación no supervisa implica que no existe un conocimiento previo de la estructura de los futuras clases.

2.4.1. UMAP

Antes de comenzar la clasificación es necesario realizar un procesamiento de datos creando un espacio n-dimensional el cual represente las características del set datos y reducirlo para ser procesado de manera eficiente y obtener una buena visualización de estos. Para este propósito se utilizará UMAP.

UMAP (*Uniform Manifold Approximation and Projection*)[16] es una nueva técnica de aprendizaje múltiple para la reducción de dimensiones. UMAP se construye a partir de un marco teórico basado en la geometría Riemanniana y la topología algebraica. Además, no tiene restricciones computacionales sobre la incorporación de la dimensión, lo que la hace viable como una técnica de reducción de dimensiones de propósito general para el aprendizaje automático.

En UMAP:

- Existe una variedad en la que los datos se distribuirían uniformemente.
- La variedad subyacente de interés está conectada localmente.
- Preservar la estructura topológica de esta variedad es el objetivo principal.

La primera fase del algoritmo consiste en construir una representación topológica difusa. En la segunda fase hay que optimizar la representación de baja dimensión para tener una representación topológica difusa lo más cercana posible a la medida.

Para la construcción de la representación topológica inicial se puede llevar acabo utilizando el algoritmo de *Nearest-Neighbor-Descent algorithm of Dong et al*[]. Luego para la fase de optimización se utiliza el descenso de gradiente estocástico con una función objetivo diferenciable. Finalmente el algoritmo procede aplicando iterativamente fuerzas atractivas y repulsivas en cada borde o vértice.

2.4.2. Algoritmo K-means

El algoritmo K-means es uno de los métodos más antiguos y utilizados para clasificación. Tiene bajo costo computacional y es escalable.

El algoritmo K-means que intenta encontrar K número de clases o *clusters* que no se sobrepongan ente si. Las clases se representan por sus centroides (típicamente el promedio de puntos dentro de la clase). Primero, se define un número de clases y de manera aleatoria se le asigna un centroide a cada una de ellas. Luego, cada dato es asignado al centroide más cercano formando una clase (*E-Step*) [17] . El siguiente paso es re-evaluar el valor de centroide de cada clase (*M-Step*). Con los nuevos centroides existirán datos que se encuentren más cercanos a otro centroide, por ende se reasignaran a esa clase. Estos pasos repiten hasta que no sea necesario reasignar datos de clase.

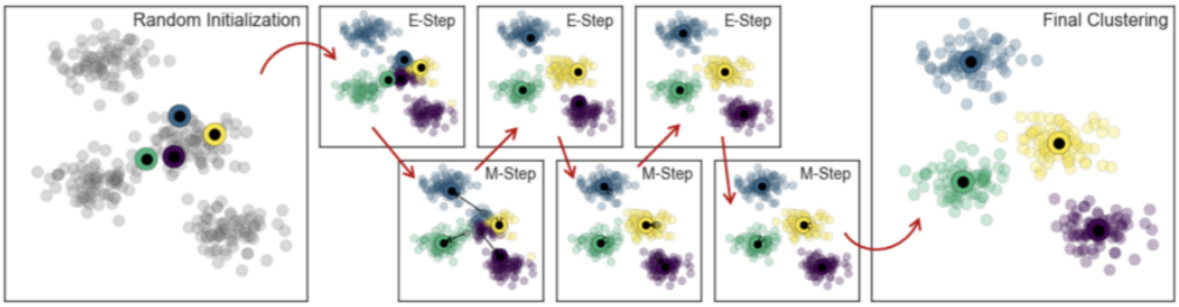


Figura 2.5: Paso a paso de algoritmo K-means (de izquierda a derecha). Para la descripción de cada paso ver texto orginal. Fig. extraída de [18]

Es importante hacer notar que el algoritmo K-Means puede variar su resultado dependiendo de los valores iniciales de los centroides. En pro de obtener un resultado óptimo el algoritmo debe correrse varias veces con distintos parámetros iniciales.

2.4.3. Algoritmo HDBSCAN

HDBSCAN es un algoritmo jerárquico basando en densidad [19]. En HDBSCAN se estima la densidad del espacio en el cual se trabajará utilizando *core distance*, donde a mayor sea este parámetro menor es la densidad de la región. De esta forma se pueden distinguir regiones de alta y baja densidad. En la Figura 2.6 se puede apreciar cómo se vería la función de densidad de un conjunto de datos.

Luego, para seleccionar los *clusters* es necesario definir el umbral jerárquico en el cual se obtienen *clusters* que sean significativos. Para cumplir este propósito HDBSCAN construye un árbol jerárquico para identificar los *peaks* de densidad que se pueden unir y definir si debiesen formar un gran *cluster* o permanecer separados. Para tomar esta decisión se escoge según la estabilidad del *cluster*.

Además, es importante notar que HDBSCAN no asigna necesariamente todos los puntos a un *cluster*, ya que a medida que se mueve en la jerarquía de los datos considera puntos como exceso de masa.

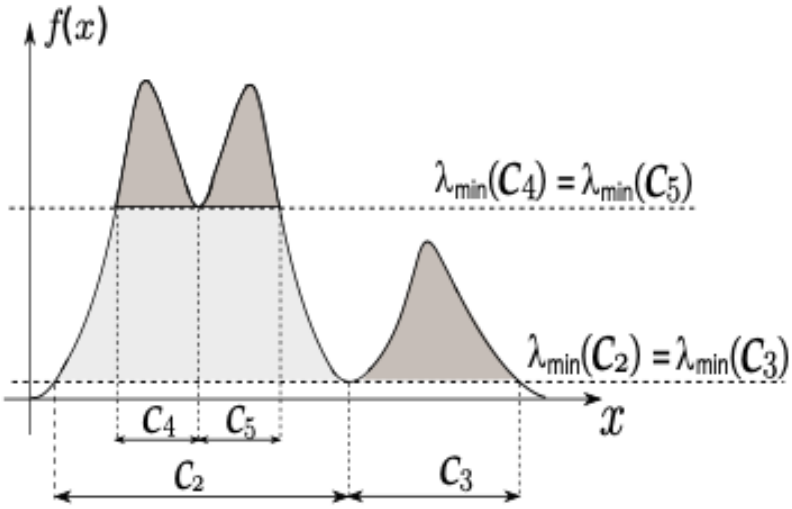


Figura 2.6: Ilustración de función de densidad, *clusters* y exceso de masa. Figura extraída de [19]

Capítulo 3

Metodología

En este capítulo se describirá como se efectuó el proceso de elección de los métodos de clasificación no supervisada para los set de datos y como se llevarán a cabo. Los datos a clasificar fueron recolectados por el equipo de Delf y corresponden a una molécula de la familia de las porfirinas que de ahora en adelante denominaremos como Bulky P3.

Para comenzar, se realizó un estudio de la literatura donde el foco principal haya sido clasificación no supervisada en trazas de conductancia. Tras este proceso se decidió realizar una construcción del espacio de características inspirado en el método MNIST [20], que consiste en en convertir las mediciones en imágenes de 28x28 píxeles. Sin embargo, para evitar la maldición de la dimensión [21], la cual trae problema a la mayoría de los algoritmos de clasificación debido al alto número de dimensiones en los set de datos, se utilizará la técnica de UMAP (*Uniform mainfold approximation and projection*) [16] con medición de distancia por coseno para reducirlo a 3 dimensiones.

La decisión te utilizar MNIST + UMAP para la construcción del espacio de características se hizo a partir de los resultados de su rendimiento presentados en el artículo *Universal approach for unsupervised classification of univariate data* [7], donde presenta un mejor índice de Folwkes-Mallows sobre otros métodos estudiados, en conjunto con el método de clasificación *Graph Average Linkage* (GAL). Además, con el fin de realizar una comparación con el algoritmo ya utilizado en la tesis *Identifying Conductance Pathways in single porphyrin molecules* [8], se utilizará K-means para estudiar su rendimiento y ver posibles mejoras al utilizar UMAP.

Una vez definido el espacio de características se procederá a implementar dos distintos algoritmos de clasificación: K-Means[17] y HDBSCAN [22].

3.1. Creación del espacio de características

Basado en [23]. Es importante que el $feature\ space$, o espacio de características, contenga la información relevante de la forma de cada traza de conductancia. La creación del $feature\ space$ aplicado es parcialmente inspirado por la base da datos MNIST de dígitos escritos a mano, donde el tamaño de la imagen de los dígitos es reducido a imágenes de 28x28 píxeles [24]. En este caso, cada traza de conductancia se convierte en un histograma 2D con

una resolución de MxN bins (Fig. 3.1), donde M y N se definen por el usuario. Cada bin representa una dimensión dentro del espacio de características, y su valor es la coordenada de la traza en dicha dimensión.

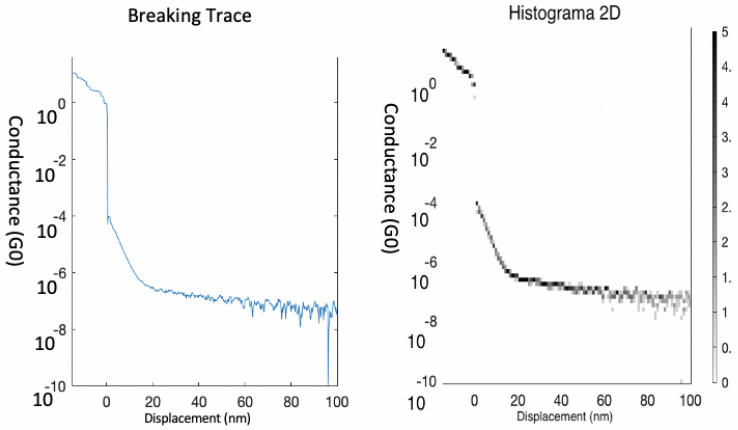


Figura 3.1: Espacio de características. Representación de traza de conductancia como histograma 2D

También es posible representar los datos en forma de histograma 1D con un número L de bins, con el fin de poder extraer mayor información de la unión molecular.

Una vez reducido el tamaño de las imágenes se procederá a aplicar la técnica de UMAP para obtener una mejor visualización del espacio y reducir la dimensión. Los parámetros básicos de UMAP se configuraran de la siguiente manera:

- nº de vecinos = 15.
- distancia mínima = 0.1.
- dimensión = 3.
- métrica = coseno.

3.2. Aplicación algoritmo K-Means

La aplicación del algoritmos K-Means se hará en base al la tesis [8]. El procedimiento a seguir es el siguiente:

Lo primero es separar las trazas que no contienen información de las trazas que tienen un comportamiento de *plateau*, esto se atribuye al efecto túnel. Para lograr esto:

1. Se aplicará algoritmo K-Means con 100 inicializaciones distintas con un número reducido de clases (2-3).
2. Si el resultado no es satisfactorio, el algoritmo se corre nuevamente aumentando el número de clases hasta que se pueda separar claramente las trazas con información relevante de las que no.

3. Una vez identificada la clase con trazas sin información, se aplicará nuevamente el algoritmo K-Means con el fin de asegurar que no se esta pasando por alto algún comportamiento molecular relevante.

4. Finalmente, se presentan las clases resultantes. Si dos clases presentan un patrón similar en los histogramas 1D se pueden fusionar en una misma clase.

Hay que tener el cuenta que el algoritmo K-Means tiene limitaciones[25]. Esto se debe ha que se asume lo siguiente:

1. Las variables tienen la misma varianza

2. La distribución de cada variable es esférica

3. Cada clase tiene aproximadamente el mismo número de trazas.

Si estas suposiciones no se cumplen el algoritmo falla en la clasificación.

Dado que el algoritmo K-Means y UMAP ha sido utilizado anteriormente para trabajos similares, se utilizará la aplicación desarrollada en MATLAB por M. Perrin [26].

3.3. Aplicación algoritmo HDBSCAN

Para la aplicación de HDBSCAN se trabajó principalmente sobre el algoritmo de J. Sorokin[22] siguiendo el siguiente procedimiento:

1. Se aplicará el algoritmo configurando los parámetros de la siguiente manera:

   - min pts: 300
   - outlier tresh: 0.85
   - min cluster size: 500

2. Se evaluaran visualmente los *cluster* obtenidos. Si no es posible separar la clase que contiene trazas con efecto túnel de otras se volverá a aplicar el algoritmo, modificando "min ptsz "min cluster size"disminuyéndolos en un valor de 50 hasta obtener un número de *cluster* entre 3 - 7 donde se puedan distinguir la clase en cuestión.

3. Una vez conforme con los *clusters* obtenidos se presentarán los histogramas 2D de cada *cluster*.

3.4. Revisión

Una vez realizada la clasificación sobre los diferentes set de datos con los distintos algoritmos, se evaluará la posibilidad de realizar diferentes procedimientos para poder obtener clases de interés en caso de ser necesario. Esta decisión se hará en base a una inspección visual de las clases que se obtengan tras la clasificación.

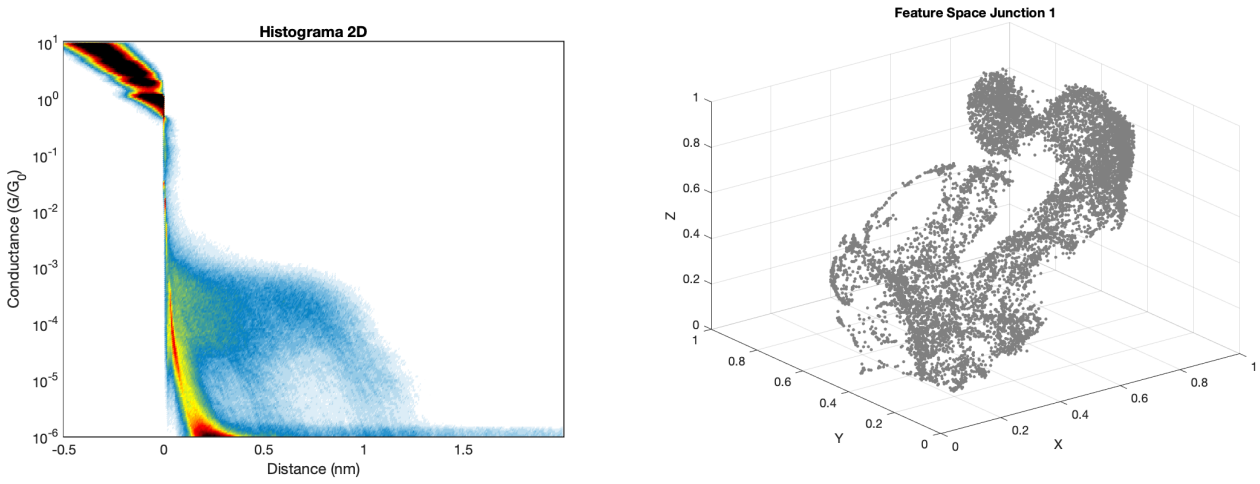
Capítulo 4

Resultados

En esta sección se presentarán los resultados obtenidos a partir de la implementación de los algoritmos mencionados anteriormente en dos set de datos referentes a la misma molécula. Los cuales, a partir de ahora, se denominarán juntura 1 y juntura 2.

4.1. Juntura 1

La juntura 1 consta de 10000 muestras. El histograma 2D de los datos perteneciente a la juntura 1 es el siguiente:

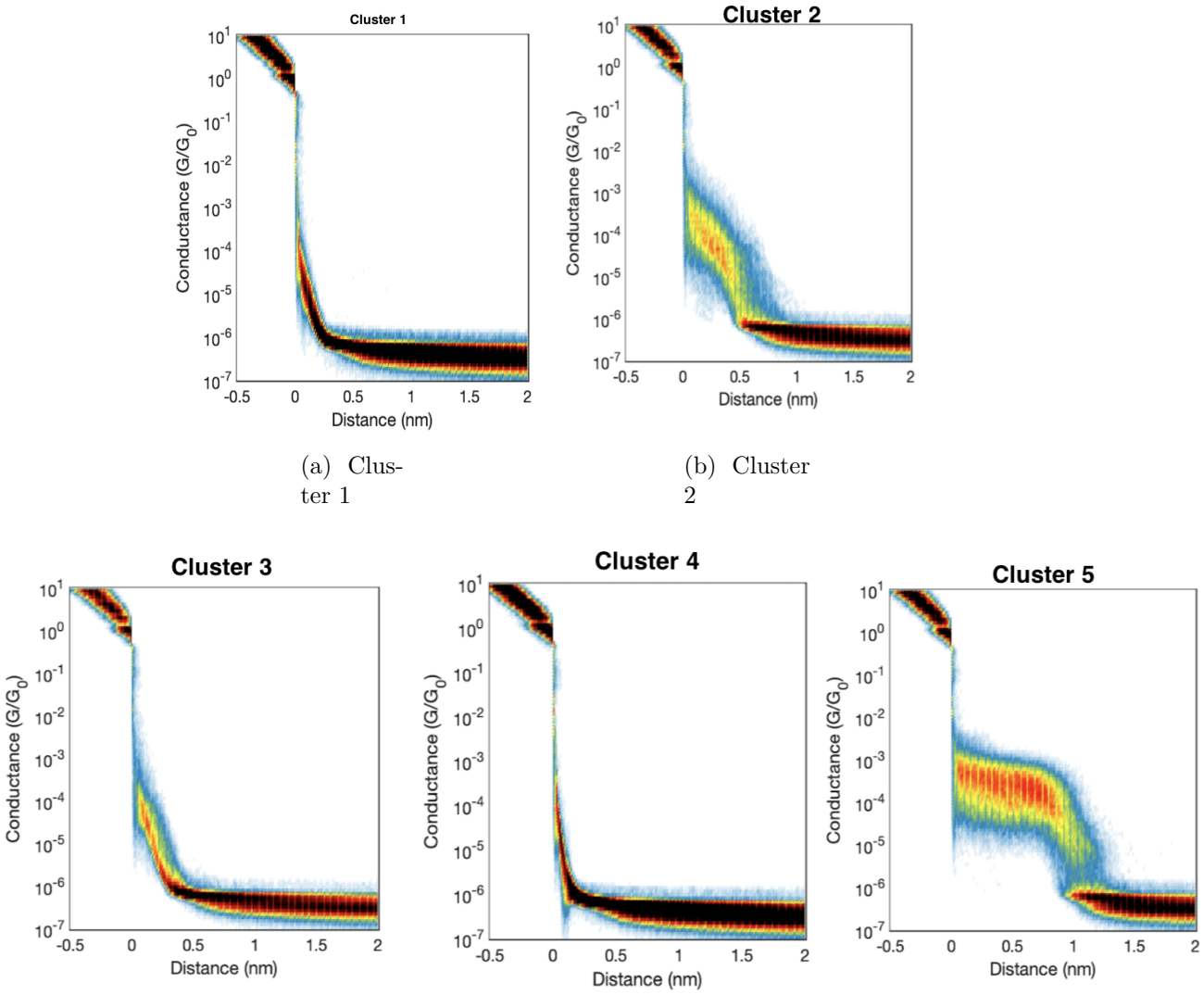


(a) Representación UMAP

(b) Histograma 2D

Figura 4.1: Histograma 2D y representación en espacio 3D de datos juntura 1

Los resultados obtenidos a partir de K-means son los siguientes:



(a) Cluster 1

(b) Cluster 2

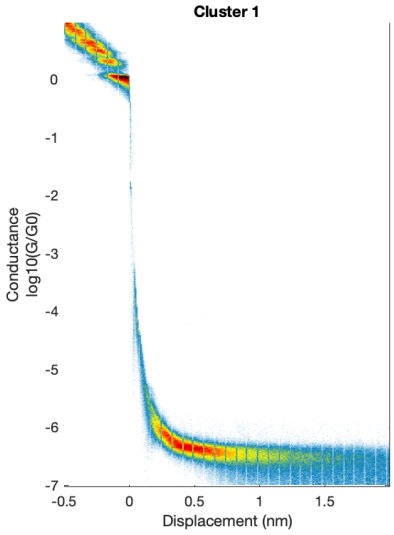
(c) Cluster 3

(d) Cluster 4

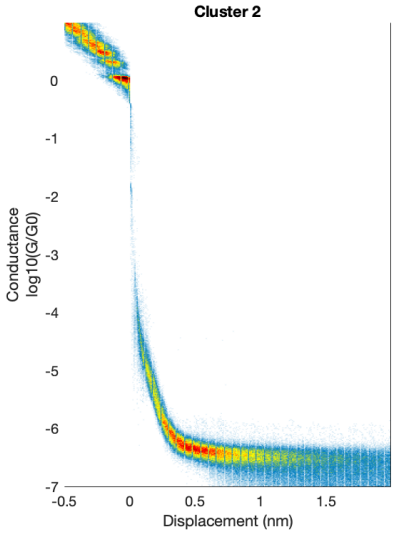
(e) Cluster 5

Figura 4.2: Clústers obtenidos a través de K-means para la juntura 1

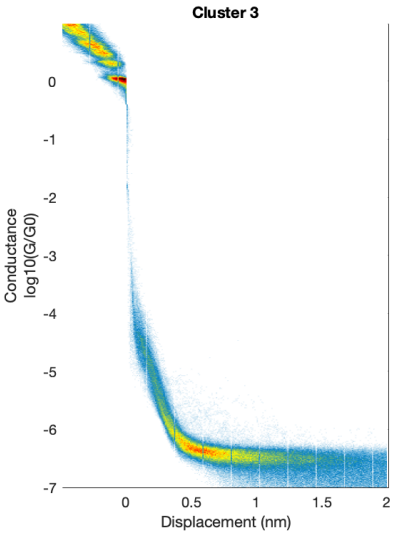
Los resultados obtenidos a partir de HDBSCAN son los siguientes:



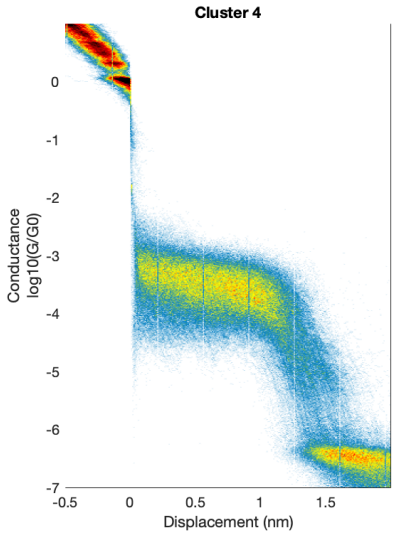
(a) Cluster 1



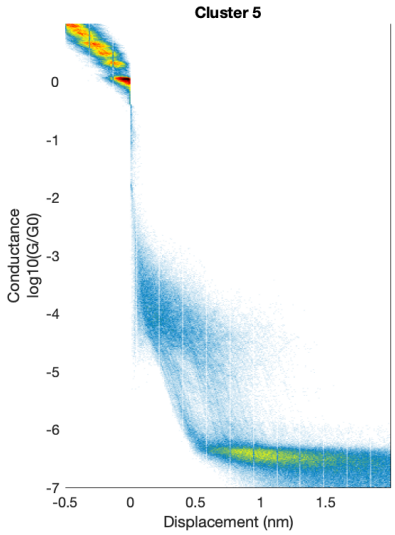
(b) Cluster 2



(c) Cluster 3



(d) Cluster 4



(e) Cluster 5

Figura 4.3: Clústers obtenidos a través de HDBSCAN para la juntura 1

4.2. Juntura 2

La juntura 2 consta de 10000 muestras. El histograma 2D de los datos perteneciente a la juntura 2 es el siguiente:

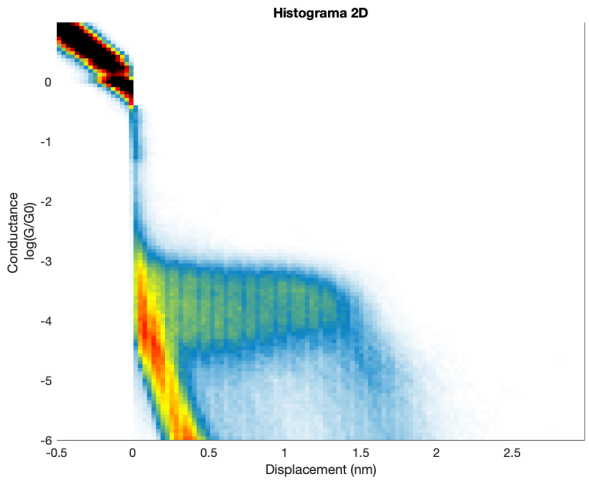
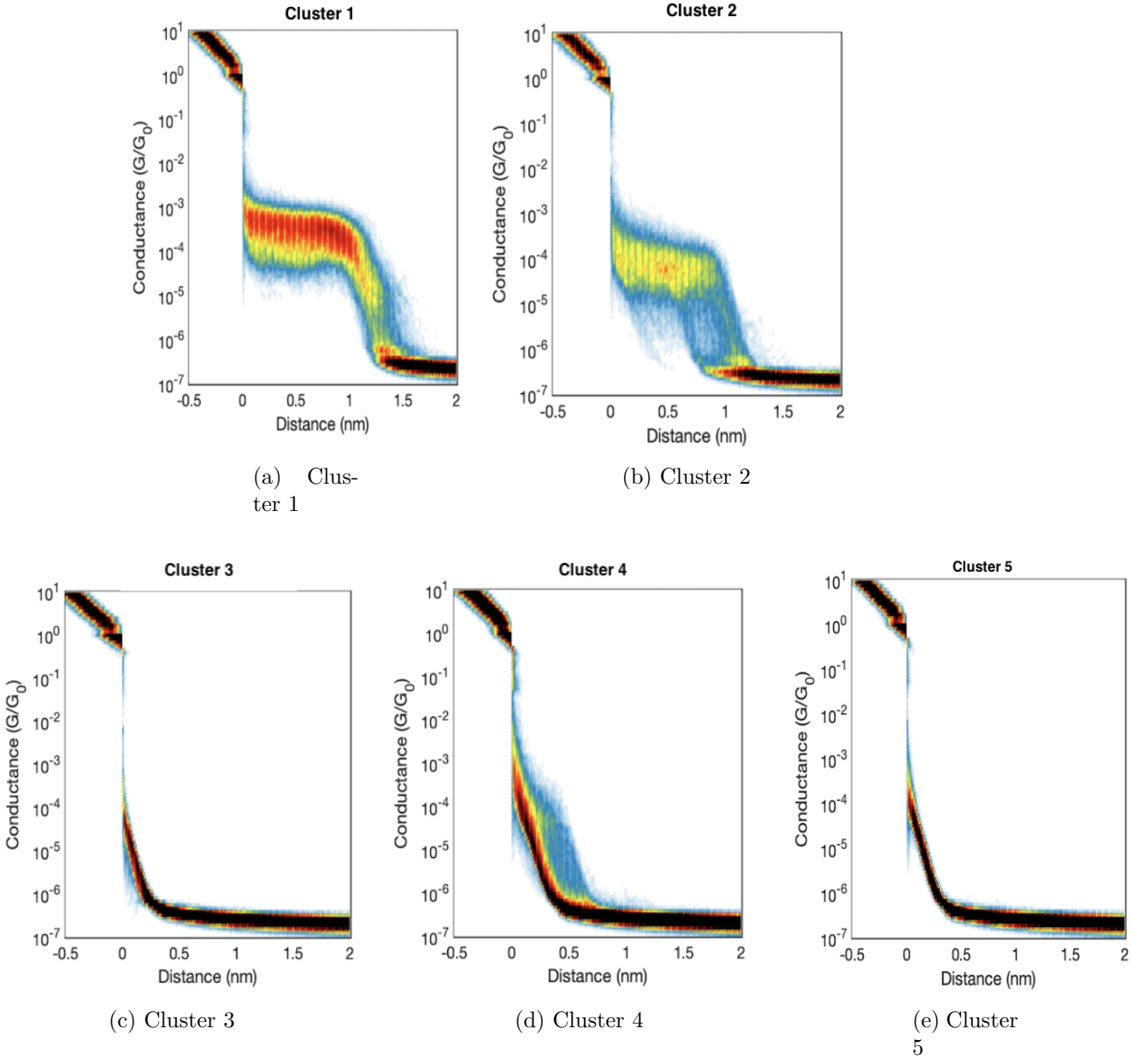


Figura 4.4: Histograma 2D Juntura 2

Y su representación en el espacio 3D:

Figura 4.5: UMAP Juntura 2

Los resultados obtenidos a partir de K-means son los siguientes:



(a) Cluster 1

(b) Cluster 2

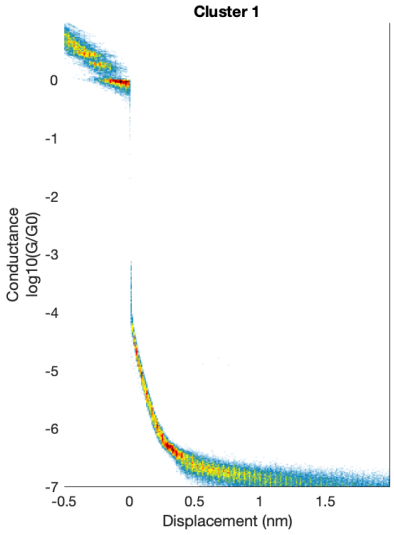
(c) Cluster 3

(d) Cluster 4

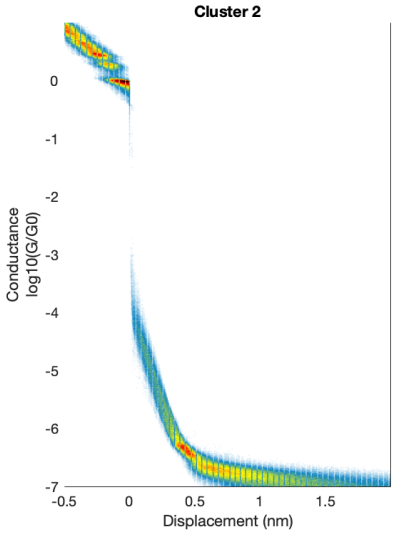
(e) Cluster 5

Figura 4.6: Clusters obtenidos a través de K-Means para la juntura 2

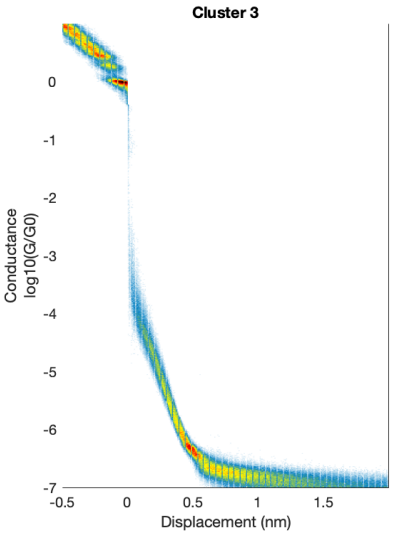
Los resultados obtenidos a partir de HDBSCAN son los siguientes:



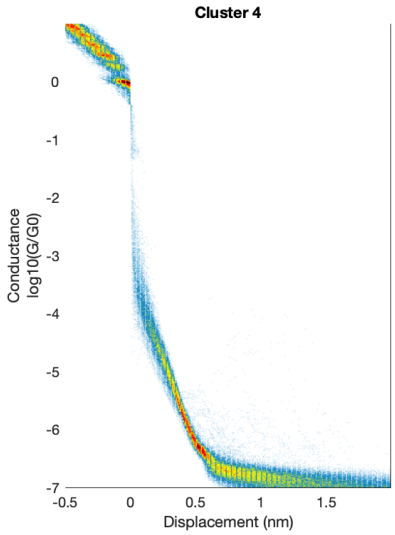
(a) Cluster 1



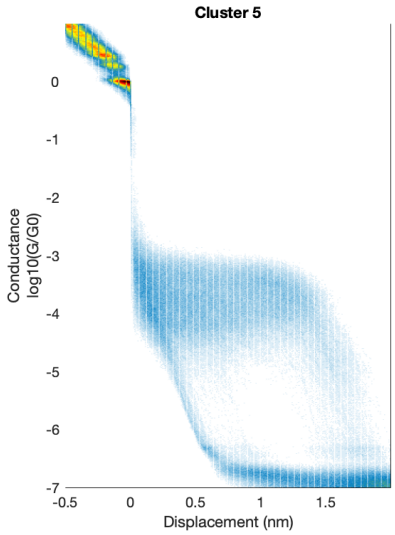
(b) Cluster 2



(c) Cluster 3



(d) Cluster 4



(e) Cluster 5

Figura 4.7: Clusters obtenidos a través de HDBSCAN para la juntura 2

4.3. Bulky P3 + P3

Para esta sección se unificaron los set de datos correspondientes a Bulky P3 y P3. Su histograma 2D se ve de la siguiente manera:

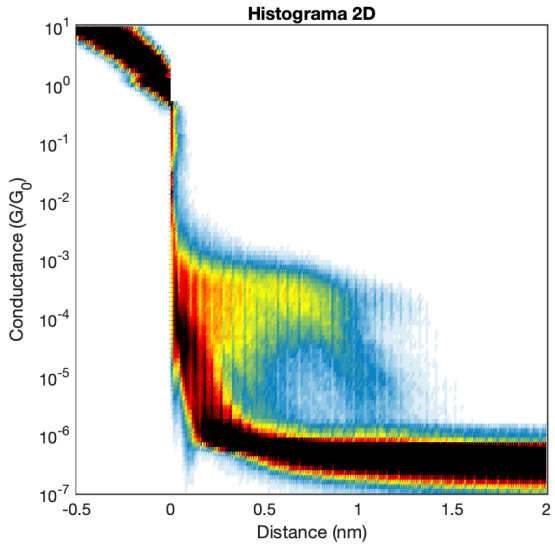


Figura 4.8: Histograma 2D Bulky P3 + P3

Su representación en el espacio 3D es la siguiente:

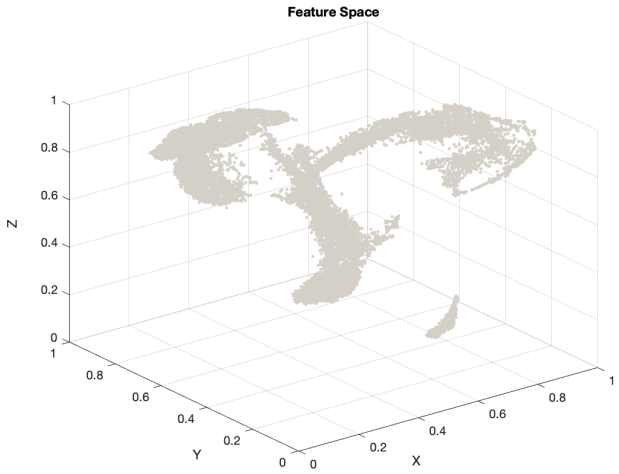
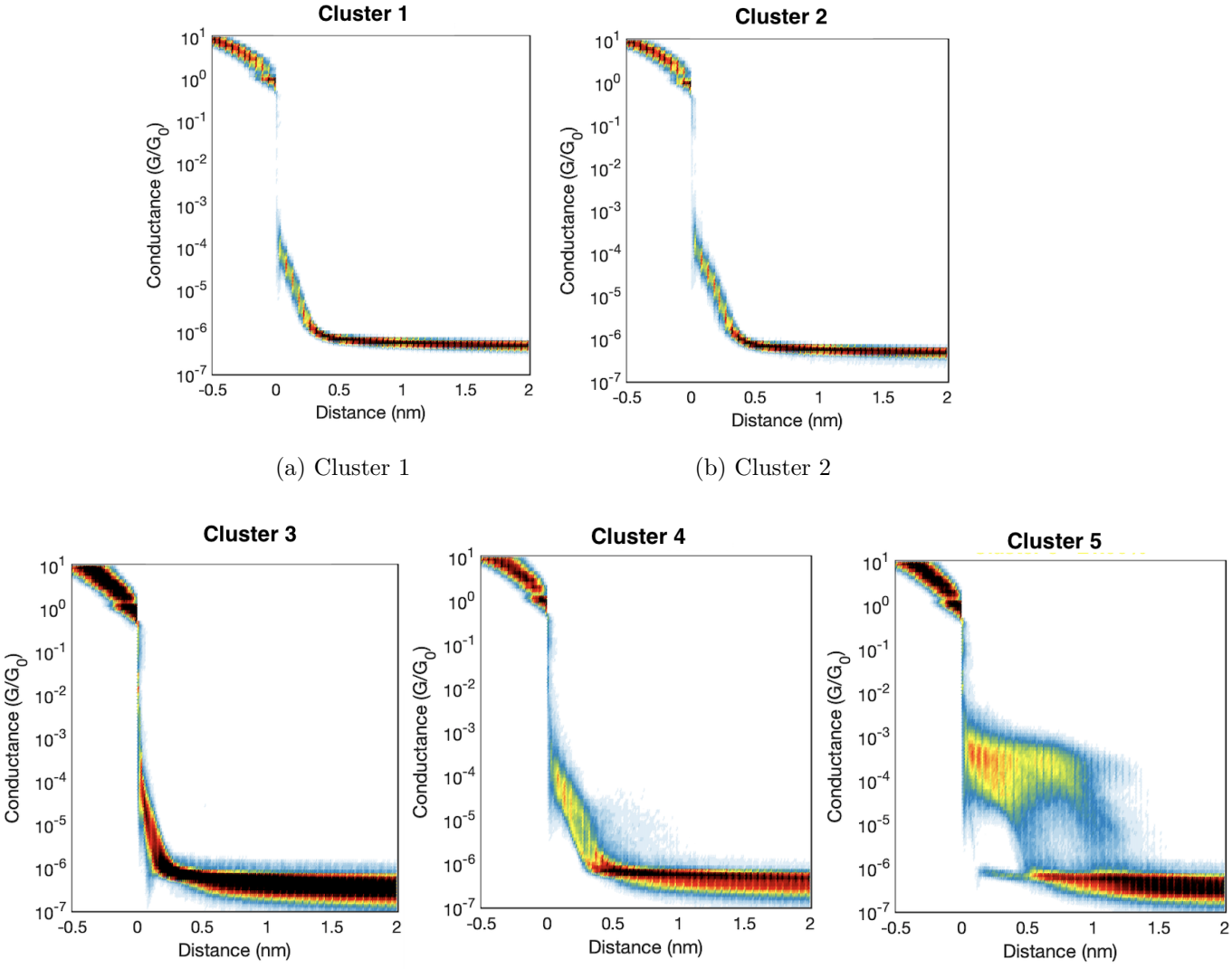


Figura 4.9: Feature Space Bulky P3 + P3

Los resultados obtenidos a partir de K-means son los siguientes:



(a) Cluster 1

(b) Cluster 2

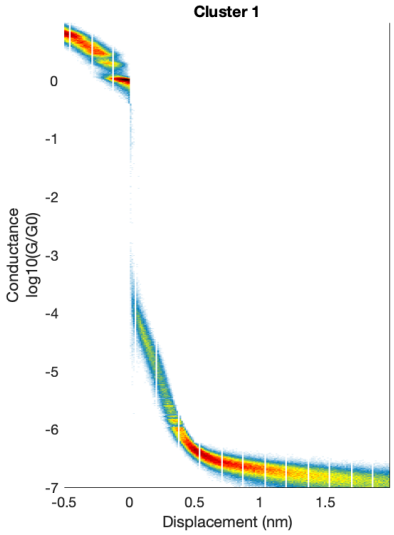
(c) Cluster 3

(d) Cluster 4

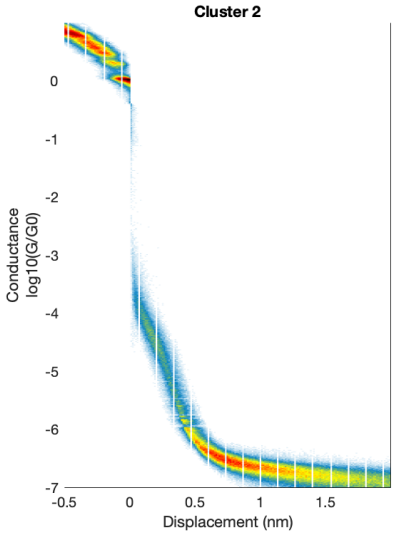
(e) Cluster 5

Figura 4.10: Clusters obtenidos a través de K-Means para datos de Bulky P3 + P3

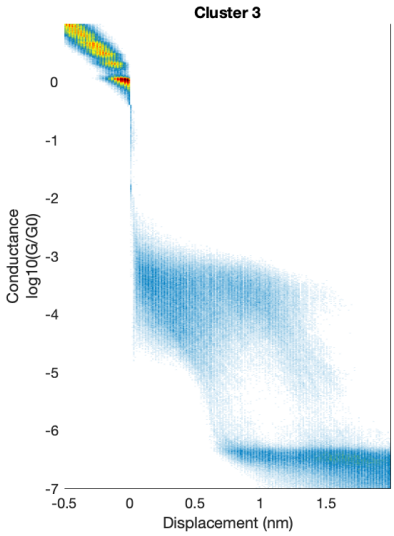
Los resultados obtenidos a partir de HDBSCAN son los siguientes:



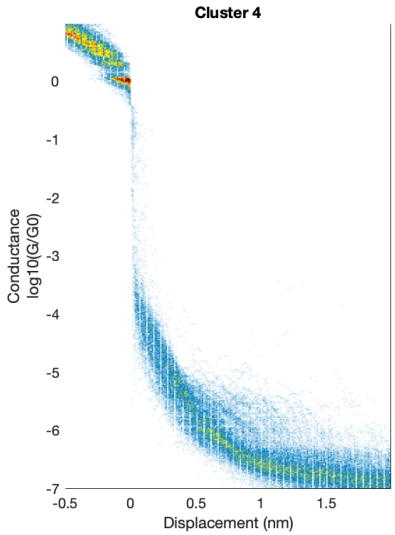
(a) Cluster 1



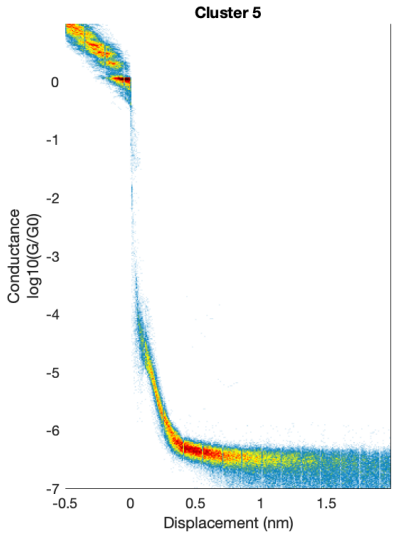
(b) Cluster 2



(c) Cluster 3



(d) Cluster 4



(e) Cluster 5

Figura 4.11: Clusters obtenidos a través de HDBSCAN para datos de bulky P3 + P3

4.4. Resumen de Resultados

Para poder analizar de mejor manera se denominarán las distintas clases significativas con los números I, II y III. Estos resultados abarcarán también información que se encuentra en el Anexo A.

- Clase I: La referencia será el clúster 5 de la Figura 4.2. Con $G = 3 * 10^{-4} G0$
- Clase II: La referencia será el clúster 2 de la Figura 4.2. Con $G = 5 * 10^{-5} G0$
- Clase III: La referencia será el clúster 4 de la Figura 4.6. Con $G = 5 * 10^{-6} G0$

Las clases obtenidos por K-Means se pueden resumir en:

Tabla 4.1: Resultados con K-Means.

	Clase I	Clase II	Clase III
Junction 1	18.24 %	16.66 %	13.90 %
Junction 2	30.83 %	-	28.28 %
Bulky P3 + P3	20.90 %	-	27.28 %

Las clases obtenidos por HDBSCAN se pueden resumir en:

Tabla 4.2: Resultados con HDBSCAN.

	Clase I	Clase II	Clase III
Junction 1	11.90 %	12.99 %	14.85 %
Junction 2	29.47 %	-	10.88 %
Bulky P3 + P3	18.94 %	11.18 %	5.41 %

Capítulo 5

Análisis y Discusión de Resultados

En este capítulo de presentará un análisis de los resultados obtenidos, dando referencia al capítulo anterior. Primero se analizarán los resultados a obtenidos al clasificar los datos de la juntura 1, y luego los de la juntura 2; para finalizar realizando un análisis comparativo de ambas y los resultados presentados en los Anexos.

5.1. Juntura 1

Con respecto a la juntura 1, se puede ver en su histograma 2D (Ver Figura 4.1.a) que esta está conformada por trazas con comportamientos distintivos. La mayoría siguiendo el comportamiento dado por el efecto túnel, ya que por el color rojo se puede ver que esa es la tendencia que sigue la mayoría de las trazas dentro del set de datos de la juntura 1. Por otro lado, se puede visualizar su espacio de características (Ver Figura 4.1.b), o *feature space*, que el espacio 3D que representa estos datos está disperso formando una especie de anillo con algunas zonas más densas que otras.

Tras aplicar K-means, se decidió por 5 clústers, ya que se puede visualizar con facilidad tres clústers que contienen información que no es relevante para el estudio de la conductancia de la molécula en cuestión, dado que representan efecto túnel. Estos clústers son el 1 y 4 de la Figura 4.2. Por otro lado, en los clúster 2 y 5 se aprecia un comportamiento distintivo, en donde el clúster 5 pareciera estar compuesto por trazas que tienen un *plateau* más largo y una conductancia de aproximadamente $3 * 10^{-4}$G0. Por su parte el clúster 2 está compuesto por trazas con una meseta más corta y una conductancia que rodea los $5 * 10^{-5}G0$. Por último, el clúster 3 presenta un comportamiento particular el cuál no se logra identificar fácilmente si corresponde a un juntura o a efecto túnel.

Paralelamente, al clasificar la juntura 1 con HDBSCAN también se obtuvieron 5 clústers (Ver Figura 4.3). De estos, tres clústers presentaron una conducta que se define como efecto túnel, los cuales son irrelevantes para el análisis molecular. Estos son los clústers 1, 2 y 3. Asimismo, los clúster 4 y 5 poseen una forma más distintiva, donde el clúster 4 presenta un *plateau* largo de conductancia $3 * 10^{-4}$G0 y el clúster 5 uno más corto y ancho de conductancia de $6 * 10^{-5}G0$.

La aplicación de ambos algoritmos a un mismo *feature space* concluyó en resultados simi-

lares. En ambos se obtuvieron clústers que representan de manera similar el efecto túnel, y clústers distintivos. El clúster 2 de K-means es similar al clúster 5 de HDBSCAN, y lo mismo ocurre con el clúster 5 de K-means y el 4 de HDBSCAN. Sin embargo, en estos clústers distintivos pareciera que, dentro de ellos, se incluyen distintos tipos de trazas que no pudieron ser representadas en un grupo aparte.

5.2. Juntura 2

Con respecto a la juntura 2, se puede ver en su histograma 2D (Figura 4.4) que está compuesta por trazas con comportamientos similares a los de la juntura 1. Se pueden visualizar trazas que representan el efecto túnel y otras con un *plateau*, donde la mayoría pertenece al primer grupo. Sin embargo, y a diferencia de la juntura 1, su *feature space* (Figura 4.5) es diferente. En este caso se pueden diferenciar grupos densos distribuidos al rededor del espacio característico.

Para mantener consistencia y poder realizar una comparativa se decidió optar por 5 clúster para la clasificación a través de K-Means, al igual que en el caso de la juntura 1. Tras la clasificación (Ver Figura 4.6), nuevamente se obtuvieron 2 clústers que contienen trazas en donde ocurrió efecto túnel (clúster 3 y 5) y 2 clúster con trazas con diferentes *plateaus* (clúster 1 y 2), adicionalmente se encontró un clúster, el 3, que presenta un comportamiento que se puede asociar a una juntura. En esta ocasión, el clúster 1 presenta un *plateau* en largo con conductancia de $3 * 10^{-4}G0$. Por otro lado, el clúster 2 presenta un *plateau* más corto al del clúster 1 donde $G = 8 * 10^{-5}G0$. Por último, se tiene el comportamiento particular del clúster 4, el cual posee una meseta corta con su conductancia baja. Además, el clúster 4 pareciera poseer en su interior trazas correspondientes a efecto túnel, lo que no permite un análisis claro.

En la Figura A.1 se pueden visualizar los clústers obtenidos al utilizar HDBSCAN en el set de datos correspondientes a la juntura 2. Donde esta vez el resultado no es similar a lo obtenido con HDBSCAN en la juntura 1, ni lo obtenido con K-means en este mismo set de datos. Se puede visualizar que 4 de los 5 clústers corresponden a efecto túnel y solamente un único clúster, el clúster 5, contiene información de juntura. El clúster 5 esta conformado por trazas que tienen un *plateau* largo y corto, y además con trazas que tienen una caída de efecto túnel.

Opuesto a lo ocurrido con el set de datos anterior, la aplicación de los distintos algoritmos de clasificación a un mismo *feature space* no concluyó en resultados similares. Sino, más bien pareciera ser que el clúster 5 obtenido por HDBSCAN contiene dentro la información del clúster 1 y 2 obtenido por K-Means.

Para poder alcanzar esas clases que se cree están existentes dentro de este clúster, inicialmente se presentaron dos opciones: realizar sub-clustering u obtener un nº mayor de clústers en HDBSCAN de manera que permita visualizar estás clases. Los resultados del sub-clustering del clúster 5 se encuentra en el Anexo [A.1], y en ellos se puede apreciar que fue posible encontrar las clases que representan los mismos clústers presentes en la clasificación con K-means. El clúster 5-1 es similar al clúster 1 de K-means, con una meseta larga y con $G = 3 * 10^{-4}G0$ y el clúster 5-3 es similar al clúster 2 de K-mean, con una meseta más corta con $G = 8 * 10^{-5}G0$.

Luego, se estudió la posibilidad de obtener más clústers con el algoritmo de HDBSCAN, cortando más abajo en el árbol jerárquico. Sin embargo, las pruebas realizadas mostraron un nº excesivos de clústers (>30) para poder representar estás clases minoritarias, donde en su mayoría eran clústers donde predominaba el efecto túnel, por ende se descartó su uso ya que se determinó en que terminaría siendo poco eficiente el estudio y análisis visual de los clúster cuando hay nº grande de ellos.

Dado, que se descartó la posibilidad de obtener un mayor número de clústers, se buscó otra opción, la cuál consistió en modificar la manera en que se representan los datos en el *feature space* realizando un pre-procesamiento antes de aplicar 28x28 + UMAP, de manera de ayudar a UMAP a poder representar la superficie de los datos dando énfasis en el punto donde se produce el *plateau* (aproximadamente en un desplazamiento de 0nm). Para esto se utilizó una proyección de los datos utilizando una función Gaussiana. Los resultados obtenidos a partir de esto se encuentran en el anexo A.2.

En estos resultados se puede ver que el clúster 5 es similar al clúster 5 de la Figura A.1.No hubo una mejora significativa, y por ende la proyección del eje x de las trazas sobre una función Gaussiana no produjo beneficios importantes al momento de aplicar la clasificación por HDBSCAN en el set de datos. Aún cuando el espacio de característica nos muestra grupos más definidos (Figura A.3) a comparación de cuando no se tenía el pre-procesamiento, el algoritmo tuvo las mismas dificultades de poder distinguir la información.

Finalmente, se puede ver que a través de ambos algoritmos es posible llegar al mismo resultados, aún cuando sea necesario realizar pasos extras como lo es el su-clustering.

5.3. Observaciones generales

Si bien ambos set de datos correspondes a mediciones de conductancia de una misma molécula en un mismo escenario, los resultados obtenidos son diferentes y esto confirma la estocacidad que hay detrás de estas mediciones. Esto se refleja en que los *feature space* de ambos set son diferentes entre sí, aún cuando se esperaría que fueren idénticos o similares dado que corresponden a una molécula.

Por un lado, utilizando K-means los resultados de la juntura 1 y juntura 2 fueron similares, y en ambos casos se pudieron diferenciar dos clases con un *plateau* a un diferente nivel de conductancia. Por otro lado, al aplicar HDBSCAN no se obtuvieron los mismos resultados. Existió dificultad por parte del algoritmo para tratar los datos de la juntura 2, obligando a buscar una solución adicional para obtener resultados concluyentes.

Además, la influencia de la manera en que se presenta el *feature space* tiene una repercusión directa en cómo resulta la clasificación, ya que por ejemplo, con HDBSCAN en la juntura 1 se tiene un *feature space* más disperso y en la juntura 2 se tiene zonas densas de puntos, y es en esas zonas densas donde el algoritmo falla al momento de distinguir clústers sin dejar fuera demasiada información. Esto se debe a que es un algoritmo en que la densidad y la distancia del espacio juegan un rol importante.

Todo lo ocurrido con la juntura 1 y juntura 2 es extrapolable al caso de Bulky P3 + P3, donde ambos algoritmos tuvieron desempeños similares. Y en el caso HDBSCAN se utilizó sub-clustering (Ver Anexo B.1) para diferenciar las distintas clases.

De esta manera, se plantea que una forma efectiva de tratar con este tipo de datos sería limpiarlos, para esto se debería eliminar primero todas las trazas que representan efecto túnel y después realizar clasificación a los datos. De esta forma no se obtendrían un exceso de clústers donde hay información que se no quiere rescatar.

5.4. Bulky P3 y P3

Ahondando en la comparativa que refiere a las clases A, B y C mostradas en la sección de estado del arte en datos de moléculas P3, en este trabajo también se pudieron identificar tres clases: I, II y III.

La clase dominante presentada en la molécula P3 es la clase A, la cual esta constituida por un *plateau* de 2.2nm de largo aproximadamente. Esta clase es la misma que se puede encontrar en los set de datos de Bulky P3, que se encuentra presente tanto en la juntura 1 como en la juntura 2, esta está representada como la clase I.

La clase B tiene un *plateau* más corto pero a su vez más ancho, la cuál tiende a parecerse a la denominada clase II, mas no es exactamente la misma. Además, la clase II no se encuentra en ambas junturas, fue posible identificarla solamente en la juntura 1 y en el set que une juntura 1 con juntura 2 de Bulky P3 en conjunto con el set correspondiente a P3. Se cree que realizando un clustering de mayor número de clústers será posible detectar la clase B pero en un porcentaje menor.

Por último, se tiene la clase C, la cual presenta un *plateau* concentrado en 10^{-6}. Esta clase no fue identificable en las junturas de Bulky P3. Pero si dentro del set Bulky P3 + P3. Lo cual indica que ese canal de conductancia solo esta presente en la molécula P3 y no en Bulky P3.

La clase III no fue identificada anteriormente en la molécula P3. Sobre todo en la juntura 2 aparece en un alto porcentaje. Entonces, si bien se elimino la clase C en este nuevo diseño de molécula se encontró un nuevo canal de transporte.

Finalmente, se puede ver que el diseño de la nueva molécula, si bien cumple con el objetivo de eliminar canales, como lo es parcialmente la clase B y C, se presentaron nuevos canales dentro de su nueva configuración.

Capítulo 6

Conclusión

En este capítulo se abordaran las principales conclusiones obtenidas tras este trabajo y se presentarán los diferentes desafíos que existen aún en el área.

6.1. Conclusión

Durante este trabajo se abordaron dos métodos distintos de clasificación no supervisada, K-Means y HDBSCAN, para tratar un mismo set de datos y ver si era posible obtener mejores resultados con uno que con otro, es decir, poder identificar diferentes clases o más clases que fueran de interés para el estudio de la conductancia de la molécula Bulky P3. Para esto se estudiaron ambos algoritmos y se decidió utilizarlos dado sus buenos resultados en diferentes áreas.

Al momento de aplicarlos, se puede ver que, si bien son algoritmos con diferentes características, como por ejemplo K-Means que es un algoritmo que particiona el espacio en grupos de similar tamaño que, al contrario de HDBSCAN que trabaja bajo el principio de densidad, las clases que se obtuvieron a partir de ellos en las distintas junturas derivó en un mismo resultado. Aún cuando haya sido necesario realizar diferentes variaciones para alcanzar este objetivo, como lo fue el pre-procesamiento de datos o el sub-clustering, se logró identificar las mismas clases en cada set de junturas. Lo que lleva a concluir que el algoritmo que se utiliza en esta clase de daros no juega un rol principal en el resultado final de las clases que se obtienen.

No obstante, en ambos algoritmos se obtuvieron más de una clase con el denominado efecto túnel, es decir, clases que no aportan información de interés y que se busca eliminar del estudio de estas moléculas. Por lo que se aprecia que ninguno de estos algoritmos pudo eliminar dicho comportamiento en una sola clase, sino que entregan más de una clase con información que no se necesita. En el caso particular de la juntura 2 clasificada con HDBSCAN 4 de las 5 clases no aportaban información, por esto se plantea la importancia de realizar una limpieza de las trazas de efecto túnel de los datos antes de realizar la clasificación. Esto haría más fácil el proceso de clasificación, ya que de manera inmediata se obtendrían únicamente clases con la información que se busca: los distintos canales en los cuales se conduce a través de las moléculas.

Como se puede ver, el rol que juega el algoritmo dentro de la clasificación no era el más

importante. Sin embargo, se plantea que la manera en que se representan los datos, es decir, el *feature space* es el que puede facilitar o no la clasificación, esto se debe a que es sobre este espacio donde los algoritmos de clasificación hacen su trabajo, y por ende entre mejor sea la representación de los datos a través de este espacio mejor resultados se pueden obtener. Esto se respalda por el artículo mencionado en la sección de estado del arte [7], donde las mejoras en el desempeño de algoritmos se da de manera más clara según la manera en la que se va a representar el *feature space*. Si bien durante este trabajo su utilizó 28x28 + UMAP, la cual calificaba como uno de los mejores algoritmos para esta tarea, se utilizó la reducción del espacio a 3 dimensiones, las cuales pueden no sean suficientes para representar las trazas de conductancia de esta molécula.

Con lo que a la física de este trabajo respecta, la finalidad de la creación de esta molécula Bulky P3 era poder al menos disminuir el número de canales de transporte. Inicialmente se tenían tres canales de transporte en una molécula P3, los cuales quedaron representados por las clases A, B y C. En la molécula Bulky P3 fueron identificadas tres clases: I, II y III. Donde se vió que la clase A y I son las mismas, es decir la modificación de la molécula P3 permitió que este canal de transporte persistiera. Sin embargo, las clases B y C no pudieron ser identificadas dentro de Bulky P3, pero se espera que al menos la clase II sea posible encontrarla realizando un sub-clustering, y por ende podría encontrarse dentro de Bulky P3 en un porcentaje bajo. La clase C no se identificó, por ende se pudo eliminar este canal de transporte dentro de esta nueva molécula, pero una nueva clase fue identificada: la clase III. La cuál debe ser estudiada para saber su procedencia y cómo poder eliminarla en una próxima modificación a P3.

Si bien finalmente no fue posible disminuir el número de canales de transporte, se confirma que con modificaciones en la constitución de la molécula es posible la eliminar alguno de los canales, y por ende es importante seguir estudiando las trazas de conductancia de estas moléculas y sus futuras variaciones para poder alcanzar un único canal que permita tener certeza de comportamiento eléctrico de estas.

Finalmente, se el objetivo general de este trabajo se cumple. Fue posible diseñar e implementar utilizando MATLAB dos métodos distintos de clasificación no supervisada para trazas de conductancia de una molécula Bukly P3 en pro de analizar sus canales de transporte, realizando una comparativa del desempeño de los algoritmos y del resultado físico detrás de este.

6.2. Trabajo futuro

A futuro se espera que el estudio de trazas de conductancia se siga expandiendo dado que pueden ser de gran utilidad. No obstante, como se mencionó en un principio siguen existiendo desafíos al momento de entender la unión molecular y su geometría, por ende se espera que con este trabajo encaminen diferentes maneras de estudiar estos datos.

Por un lado, cuando se menciona la limpieza de las trazas para poder eliminar el indeseado efecto túnel de estas, se ha planteado *Machine Learning* como una opción que se encuentra en estudio, de manera de que a través de redes neuronales se puedan distinguir rápidamente trazas de interés de trazas con efecto túnel. De esta manera, se obtendría un set de datos

donde las clases resultantes tras la clasificación sean todas, o al menos en su gran mayoría de interés evitando así la revisión de múltiples clases que no brindan información o la necesidad de realizar sub-clustering para identificar clases más pequeñas, ya que se reduciría el tamaño del set de datos significativamente.

Por otro lado, también es importante destacar que UMAP se ha utilizado generalmente para la reducción de espacio a tres dimensiones, y queda propuesto su estudio e impacto su utilización en más dimensiones, donde es posible que ayude a una mejor representación de esta clase de datos.

También quedó en evidencia que la molécula en estudio no fue suficiente para el objetivo que se tenía en mente de poder disminuir el número de canales de transporte. Por lo que deja en claro que es necesario seguir realizando variaciones y modificaciones en el diseño de estas moléculas para poder llegar a un resultado satisfactorio en cuanto la conductancia que estas presenta.

Bibliografía

[1] M. E. Abbassi, P. Zwick, A. Rates, D. Stefani, A. Prescimone, M. Mayor, H. S. J. van der Zant, and D. Dulić, "Unravelling the conductance path through single-porphyrin junctions," *Chemical Science*, vol. 10, no. 36, pp. 8299–8305, 2019.

[2] D. H. Yoon, S. B. Lee, K.-H. Yoo, J. Kim, J. K. Lim, N. Aratani, A. Tsuda, A. Osuka, and D. Kim, "Electrical conduction through linear porphyrin arrays," *Journal of the American Chemical Society*, vol. 125, no. 36, pp. 11062–11064, 2003.

[3] V. Kaliginedi, P. Moreno-García, H. Valkenier, W. Hong, V. M. García-Suárez, P. Buiter, J. L. H. Otten, J. C. Hummelen, C. J. Lambert, and T. Wandlowski, "Correlations between molecular structure and single-junction conductance: A case study with oligo(phenylene-ethynylene)-type wires," *Journal of the American Chemical Society*, vol. 134, no. 11, pp. 5262–5275, 2012.

[4] G. Sedghi, V. M. García-Suárez, L. J. Esdaile, H. L. Anderson, C. J. Lambert, S. Martín, D. Bethell, S. J. Higgins, M. Elliott, N. Bennett, J. E. Macdonald, and R. J. Nichols, "Long-range electron tunnelling in oligo-porphyrin molecular wires," *Nature Nanotechnology*, vol. 6, pp. 517–523, July 2011.

[5] M. Lemmer, M. S. Inkpen, K. Kornysheva, N. J. Long, and T. Albrecht, "Unsupervised vector-based classification of single-molecule charge transport data," *Nature Communications*, vol. 7, Oct. 2016.

[6] J. M. Hamill, X. T. Zhao, G. Mészáros, M. R. Bryce, and M. Arenz, "Fast data sorting with modified principal component analysis to distinguish unique single molecular break junction trajectories," *Phys. Rev. Lett.*, vol. 120, p. 016601, Jan 2018.

[7] M. Abbassi, J. Overbeck, O. Braun, M. Calame, H. S. J. Zant, and M. Perrin, "Universal approach for unsupervised classification of univariate data," *arXiv: Mesoscale and Nanoscale Physics*, 2020.

[8] A. Rates, "Identifying conductance pathways in single porphyrin molecules," Master's thesis, Universidad de Chile, Santiago, Chile, 2019.

[9] J. Muñoz, "Electrical characterization of protein networks and inorganic nanoparticles," Master's thesis, Universidad de Chile, Santiago, Chile, 2018.

[10] J. G. Simmons, "Generalized formula for the electric tunnel effect between similar electrodes separated by a thin insulating film," *Journal of Applied Physics*, vol. 34, pp. 1793–1803, June 1963.

[11] M. Baldo, "6.701 introduction to nanoelectronics," in *Introduction to nanoelectronics*, 2010. MIT OpenCourseWare.

[12] G. Binnig, H. Rohrer, C. Gerber, and E. Weibel, "Surface studies by scanning tunneling microscopy," *Phys. Rev. Lett.*, vol. 49, pp. 57–61, Jul 1982.

[13] G. Leatherman, E. N. Durantini, D. Gust, T. A. Moore, A. L. Moore, S. Stone, Z. Zhou, P. Rez, Y. Z. Liu, and S. M. Lindsay, "Carotene as a molecular wire: conducting atomic force microscopy," *The Journal of Physical Chemistry B*, vol. 103, pp. 4006–4010, May 1999.

[14] C. Sabater, C. Untiedt, and J. M. van Ruitenbeek, "Evidence for non-conservative current-induced forces in the breaking of au and pt atomic chains," vol. 6, pp. 2338–2344, Dec. 2015.

[15] K. M., *Single-Molecule Electronics [electronic resource]: An Introduction to Synthesis, Measurement and Theory*. Singapour: Springer, 2016.

[16] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, p. 861, Sept. 2018.

[17] J. Wu, *Advances in K-means Clustering*. Springer Berlin Heidelberg, 2012.

[18] J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media Inc., 2016.

[19] D. M. Ricardo J.G.B. Campello and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining* (T. S. Hawley and R. G. Hawley, eds.), pp. 170–162, New York, NY: Springer, 2013.

[20] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[21] R. E. Bellman, *Dynamic Programming*. USA: Dover Publications, Inc., 2003.

[22] J. Sorokin, "Jordan/hdbscan." https://github.com/Jorsorokin/HDBSCAN, 2021.

[23] D. Cabosart, M. E. Abbassi, D. Stefani, R. Frisenda, M. Calame, H. S. J. van der Zant, and M. L. Perrin, "A reference-free clustering method for the analysis of molecular break-junction measurements," *Applied Physics Letters*, vol. 114, p. 143102, Apr. 2019.

[24] E. Kussul and T. Baidyk, "Improved method of handwritten digit recognition tested on MNIST database," *Image and Vision Computing*, vol. 22, pp. 971–981, Oct. 2004.

[25] D. Robinson, "K-means clustering is not a free lunch," Jan 2015.

[26] M. E. Abbassi, J. Overbeck, O. Braun, M. Calame, H. S. J. van der Zant, and M. L. Perrin, "Benchmark and application of unsupervised classification approaches for univariate data," *Communications Physics*, vol. 4, Mar. 2021.

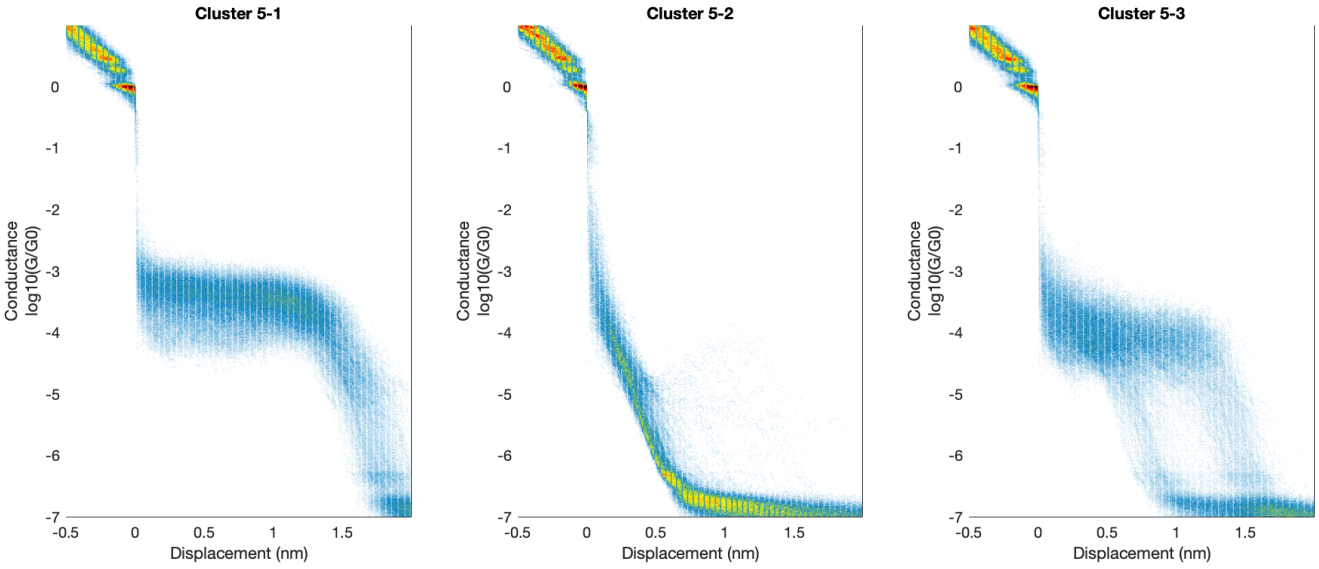
Anexo A

Juntura 2

En este capítulo se muestran los distintos resultados obtenidos con las variaciones aplicadas.

A.1. Sub-Clustering

A continuación se presentan los resultados del sub-clustering aplicado al clúster 5 de HDBSCAN de la juntura 2.



(a) Cluster 5-1

(b) Cluster 5-2

(c) Cluster 5-3

Figura A.1: Sub-clústers obtenidos a través de HDBSCAN para la clúster 5

A.2. Clasificación utilizando otro *feature space*

Se utilizó un pre-procesamiento de los datos, donde a cada traza conformada por su eje x (distancia/desplazamiento) y eje y (conductancia) se ponderó de la manera $x = x * f(x)$ donde $f(x)$ corresponde a la función gaussiana.

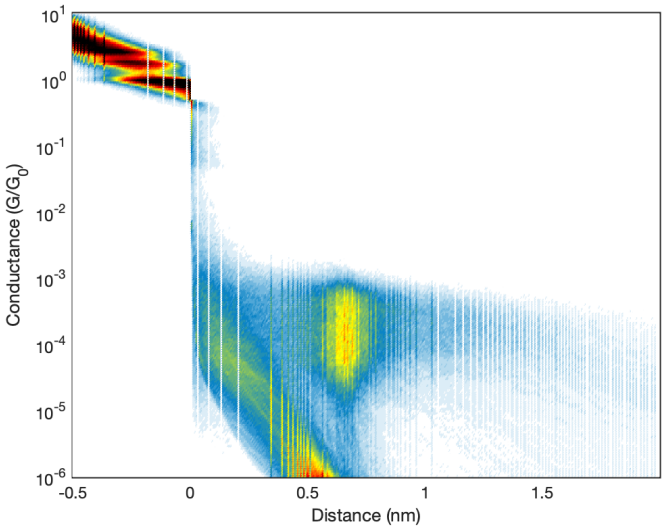


Figura A.2: Histograma 2D de datos pre-procesados

El *feature space* ya clasificado con HDBSCAN, con esta variación es el siguiente:

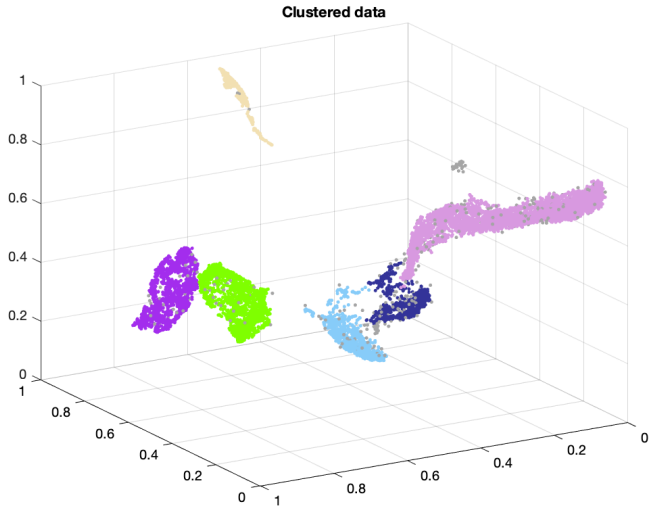
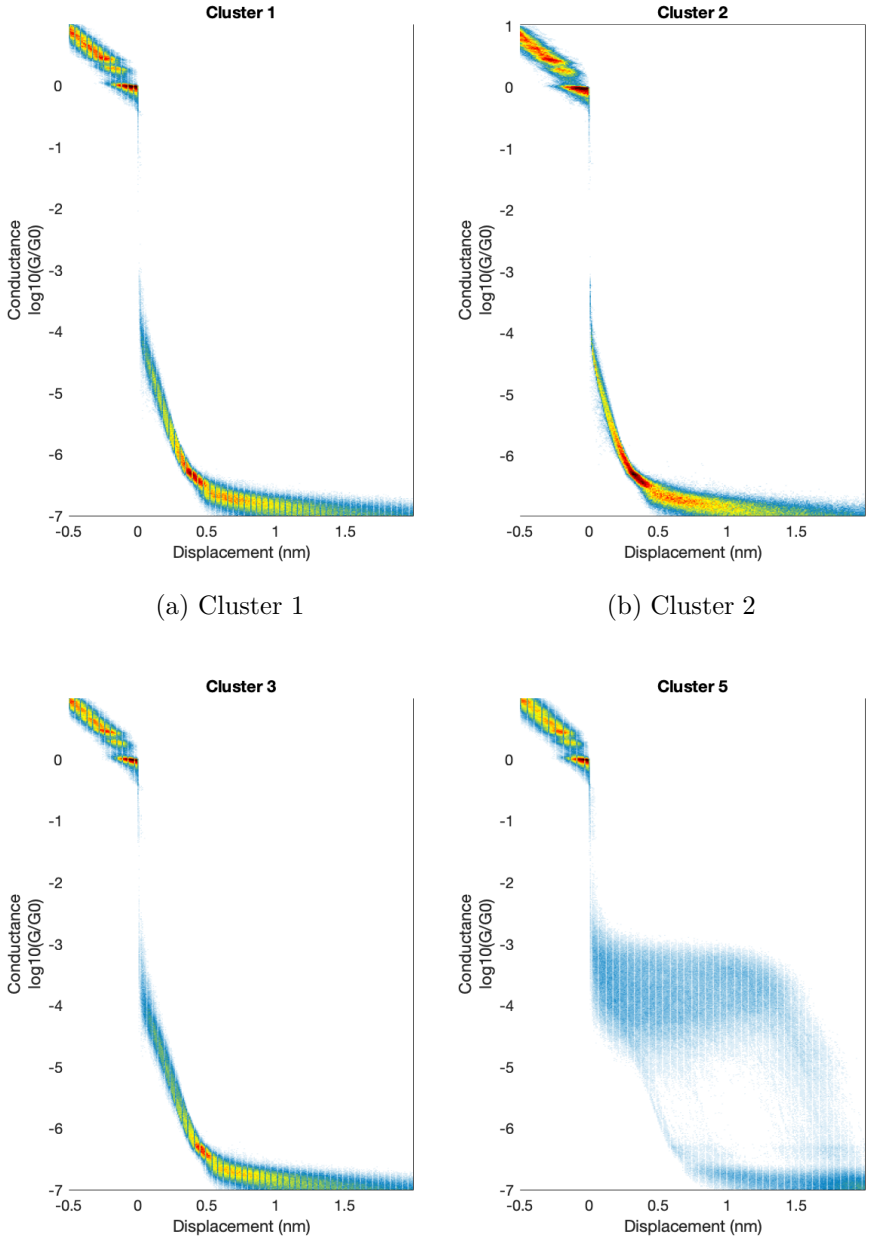


Figura A.3: Espacio característico con clasificación por HDBSCAN.

Los clúster obtenidos son los siguientes:



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

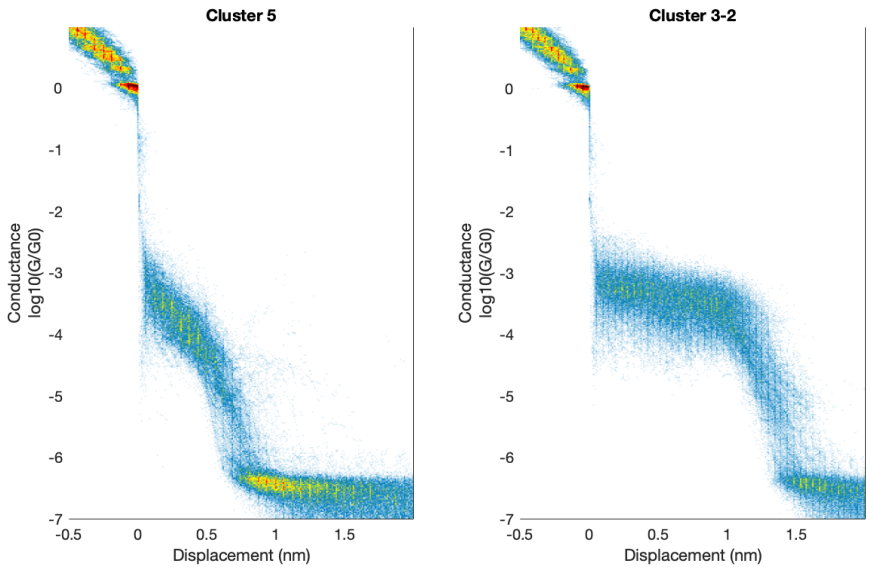
(d) Cluster 5

Figura A.4: Clústers obtenidos a través de HDBSCAN para variación con ponderación Gausiana.

Anexo B

Bulky P3 + P3

B.1. Sub-Clustering



(a) Cluster 3-1

(b) Cluster 3-2

Figura B.1: Clases de interés al realizar sub-clustering en clúster 5 de fig. 4.11