Universidad Autónoma Metropolitana
Unidad Iztapalapa

División de ciencias básicas e ingeniería
Posgrado en ciencias (Química)

"Cálculos de energías de reorganización de compuestos orgánicos nitrogenados utilizando diversas técnicas de Machine Learning"

T E S I S

Para obtener el grado de
Maestro en Ciencias (Química)

P R E S E N T A :
Yaffet Zambrano González
Matricula: 2223803434
Correo: yaffet4556@gmail.com

Director de tesis:
Dr. Humberto Laguna Galindo

Jurado:
Presidente: Dr. Marcelo Enrique Galván Espinosa
Secretario: Dr. Ángel Alejandro García Chung
Vocal: Dra. Adriana Pérez González
Vocal: Dra. Martha Magdalena Flores Leonar

Iztapalapa, Ciudad de México, a 6 de Junio del 2025

Resumen

Este estudio se motiva por la búsqueda de especies químicas candidatas para las baterías de flujo redox acuosas, en particular el interés radica en determinar si la especie de estudio es reversible. Ya que esta propiedad influye en los procesos de carga y descarga de la batería. Este trabajo se divide en dos enfoques; 1) Ajuste: Cálculo de energías de reorganización, que en el ámbito del aprendizaje automático o ML (*Machine Learning* por sus siglas en inglés). 2) Clasificación, en el que se busca determinar si una molécula orgánica es electroquímicamente reversible o no. En el contexto del ML supervisado, los modelos se entrenan con una base de datos, donde las variables son descriptores quimioinformáticos (obtenidos con la biblioteca RDKit de `Python`), mientras que la variable objetivo es, en el primer caso, la energía de reorganización calculada con métodos de estructura electrónica al nivel de teoría B3LYP/6-311+G(d,p)/SMD. Para el primer conjunto de datos se propone una motodología para reducir la dimensionalidad. En el caso 2, la variable objetivo es una propuesta de escala de reversibilidad. Se exploran tres modelos a) Regresión lineal multivariable (RLM con Sklearn), b) Ensambles de árboles de decisión (con XGBoost) y c) Redes neuronales (con Keras), para el problema de ajuste. Los entrenamientos de los modelos se hacen para las familias: Bipiridinas , Bencidinas y Metil Viológeno , y un conjunto que contiene a todas las familias. Mientras que para el problema de clasificación se entrena con el conjunto de todas las familias con el modelo ensambles de árboles de decisión.

Agradecimientos

Le doy, pricipalmente, las gracias a mis padres, Heber y Carina por todo su apoyo, este logro no hubiese sido posible.

De igual manera quiero agradecerle a mi asesor Humberto por su guía en el mundo de la ciencia y por siempre tratarme más como un colega que como estudiante.

También le agradezco al SECIHTI por el otorgamiento de beca 839782 así como al Laboratorio de Supercómputo de UNAM por el tiempo de cálculo otorgado.

Índice

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

Este proyecto de investigación tiene como motivación el estudio de las baterías de flujo redox (RFB, por sus siglas en inglés) [1, 2, 3], las cuales buscan ser una alternativa como dispositivos de almacenamiento de energía que puedan estar conectados o no a la red eléctrica. Estos dispositivos representan una solución prometedora para el abastecimiento de energía eléctrica debido a la creciente demanda y el aumento de las fuentes de energías intermitentes.

Este sistema de almacenamiento electroquímico ofrece diversas ventajas, como lo es su flexibilidad de movilidad, escalabilidad, su funcionamiento a temperatura ambiente y largos ciclos de carga/descarga. En las RFB, las especies activas más comunes están basadas en vanadio pero existen propuestas de compuestos orgánicos en solución acuosa (AO-RFB, por sus siglas en inglés) [4], [5] que pueden tener otras ventajas, como su fácil producción y la manipulación de sus propiedades mediante modificaciones en su estructura química.

Una parte de la investigación se ha centrado en la descripción de propiedades fisicoquímicas que indican si es viable la aplicación de una molécula orgánica como especie activa en una batería de flujo, como por ejemplo, sus propiedades ácido-base y óxido-reducción. Desde esta perspectiva se han estudiado varias familias de compuestos orgánicos nitrogenados.

Uno de los siguientes pasos de estudio que se explora en este trabajo y que complementa al análisis mencionado, es el estudio de la reversibilidad electroquímica. Esta propiedad experimental está relacionada con la tasa de transferencia de carga. Una tasa de transferencia alta indica que el intercambio de carga ocurre de manera efectiva, es decir, con frecuencia sin impedimentos significativos cuando la especie redox se acerca al electrodo. Este proceso se conoce como reversible. Además, un proceso reversible implica estabilidad química, es decir, la especie pueda atravesar por los procesos óxido-reducción sin descomponerse o sin reaccionar con otras especies presentes en el medio, en otras palabras, que las especies redox sean estables. Por otro lado, un proceso irreversible indica que la velocidad de transferencia de carga es lenta y se necesita una diferencia de potencial mayor para estimular la tranferencia.

En este trabajo se aborda el concepto de reversibilidad electroquímica, con un enfoque termodinámico de transferencia de carga. Esta propiedad se propone describir a partir de las energías de reorganización, propuestas por Rudolf A. Marcus, que están relacionadas con la cinética de la

transferencia de un electrón. Esta propiedad tiene un impacto en el rendimiento de las baterías porque se refleja en la vida útil, es decir, el los ciclos de carga y descarga de la batería.

La Teoría de Marcus, propuesta en el año de 1956, explica la tasa de transferencia electrónica, es decir la cinética de una reacción redox, en función de energías termodinámicas; energía libre de reacción (ΔE) y energía de reorganización (λ). La segunda se define como la energía (λ_{Red}) necesaria para llevar los núcleos de la especie oxidada a las posiciones que tendrán en la especie reducida, es decir que mide los efectos de la transferencia electrónica en la geometría molecular. Para el proceso de oxidación, es la energía (λ_{Ox}) necesaria para llevar los núcleos de la especie reducida a las posiciones que tendrán en la especie oxidada. Estas dos energías de reorganización no son iguales en general. Se plantea, como hipótesis, que la relación que guardan sus magnitudes está relacionada con la reversibilidad electroquímica de la molécula, pues si mover los núcleos en ambas direcciones tiene el mismo costo energético, es probable que la molécula sea reversible.

A partir de cálculos de estructura electrónica (TDF, teoría de los funcionales de la densidad) y en el marco de la Teoría de Marcus se hará una determinación de las energías de reorganización[6]; λ_{Red} , λ_{Ox} y el cociente de ambas energías. Esto para distintas familias de especies químicas nitrogenadas (Fig 1.1) tales como; bipiridina (Bpy), bencidina (Bz) y Metil-Viológeno (MV).

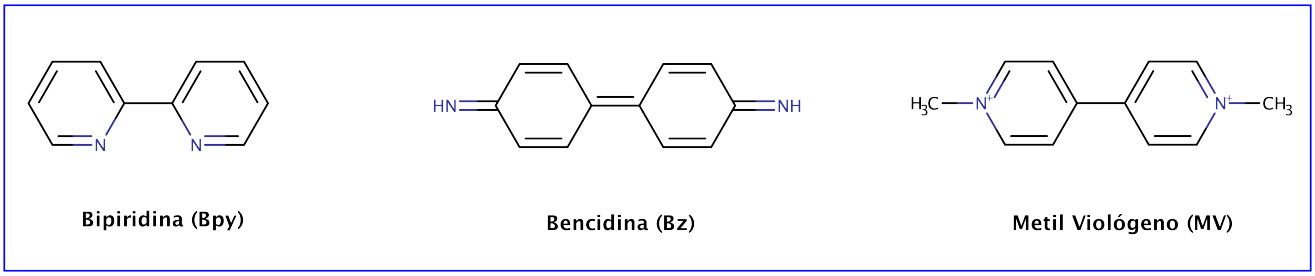


Figura 1.1: Estructuras de las distintas familias que se abordarán en este trabajo.

La predicción de las energías de reorganización y el cociente se abordará en dos vertientes:

- Se utilizarán redes neuronales y regresión lineal (ambas son técnicas de *Machine Learning*) para predecir de manera numérica las dos energías de reorganización antes mencionadas, además se estudiará la pertinencia del cociente como indicador de la reversibilidad.
- Con los ensambles de árboles de decisión se abordará un problema de clasificación que se usará para seleccionar qué especies son reversibles, quasireversibles o irreversibles, dependiendo del valor del cociente.

En ambos casos, los modelos de *Machine Learning* se entrenarán con datos de descriptores quimioinformáticos que proporciona la librería de RDKit en Python y que están relacionados con diversas propiedades y características de las moléculas y que se pueden calcular a partir de los SMILES (que son representaciones de las moléculas que se pueden obtener de manera relativamente sencilla).

La motivación del estudio es doble, por un lado, es muy deseable tener un modelo que prediga si una molécula presenta reversibilidad electroquímica o no, y por otro lado el estudio de las propiedades de las moléculas de interés con métodos de la química cuántica requiere de tiempos de

cómputo largos debido al gran conjunto de moléculas que se estudian. Esto lleva a costos económicos significativos. Por ello usar técnicas de *Machine Learning* tanto para reproducir valores como para clasificar, reduce este tipo de inconvenientes. Además, el estudio de procesos de transferencia de electrones no es exclusivo de las baterías de flujo redox sino que también juega un papel importante en otros campos que pueden ir desde sistemas biológicos hasta industriales, por ejemplo para el proceso de la fotosíntesis.

En los últimos años se han publicado diversos trabajos que utilizan el Aprendizaje Automático (ML, por sus siglas en inglés), una rama de la Inteligencia Artificial (IA). El ML son algoritmos que realizan tareas complejas de forma automática que se aplican a muchas áreas de la actividad humana. El ML ha tenido impacto debido a la gran cantidad de datos que se tiene acceso. De acuerdo al informe[7] del IDC (Internacional Data Corporation), la cantidad de datos a nivel mundial que se estiman para el 2025 alcanzarán los 275 zettabytes (10^{21}bytes). Por otro lado, se ha alcanzado el suficiente poder computacional para producir estos algoritmos.

Las aplicaciones han sido desde la detección y creación de imágenes[8],[9], hasta áreas en la ciencia como la biología molecular con el algoritmo AlphaFold [10] que predice la estructura nativa de una proteína. Por otro lado, en el área de química también han tenido diversas aplicaciones [11] en áreas como: retrosíntesis[12], [13], simulación atómica basada en potenciales ML [14], [15] y catálisis heterogénea [16]. En el contexto de energías de reorganización han habido algunos trabajos; En el trabajo de Omri y Geoffrey [17] muestran que utilizar bosques aleatorios, como modelo, para predecir energías de reorganización para un conjunto de politiofenos, es una opción rápida. Muestran que predecir las energía de reorganización de 31,878 especies es 13 veces más rápida que métodos con métodos convencionales. Mientras que en el trabajo de Cheng y Daniel [18] utilizan redes neuronales de grafos tridimensionales, y muestran un gran beneficio en predicciones rápidas y precisas al introducir información sobre la conformación molecular como característica del grafo y la importancia de la simetría de la arquitectura de un modelo.

Los modelos de ML se pueden clasificar en tres grandes categorías : 1) aprendizaje supervisado que ajusta los datos etiquetados, 2) aprendizaje no supervisado que clasifica los datos no etiquetados y 3) aprendizaje reforzado que utiliza mecanismos de recompensa para guiar el aprendizaje de los datos.

Los datos son tan relevantes que, sin ellos, no existiría el ML, por lo tanto un prerequisito indispensable es tener datos disponibles. La química es única, en el sentido de que tiene la ventaja de que se ha recompilando una gran cantidad de datos a través de los años surgiendo con ello la quimioinformática, incluso antes del surgimiento del ML, desde propiedades macroscópicas, espectros y cálculos que usualmente están basados en teoría de los funcionales de la densidad. Por lo tanto, los datos se pueden obtener de una base de datos o incluso construirla.

Las características, también llamadas representaciones o descriptores, son la información de entrada que procesan los modelos de ML y que generan una salida o respuesta. Tanto los datos como las características son pilares fundamentales del ML. Seleccionar un conjunto adecuado de descriptores puede facilitar una mejor interpretación sobre la influencia de estos en el fenómeno de estudio. Aunque las técnicas de Aprendizaje profundo (Deep Learning, en inglés) pueden extraer caraterísticas importantes por sí mismos, esto suele tener un costo computacional mayor y una pérdida de interpretabilidad.

En química, existen diversas características de entrada, pero la representación de la estructura

molecular suele ser un tema de investigación. Los descriptores se pueden clasificar en tres categorías:

- 1D: Incluyen características generales como la masa molecular, el número de electrones de valencia, entre otros.
- 2D: Se enfocan en los enlaces de las moléculas y se derivan de representar la molécula como un grafo (átomos como nodos y enlaces como aristas). A partir de esta representación, se construyen matrices de adyacencia y otros descriptores relacionados.
- 3D: Representan la disposición espacial de las moléculas. Sin embargo, proporcionar información espacial a un modelo de ML no es una tarea sencilla debido a la invarianza a permutaciones, traslaciones y rotaciones. Para abordar este desafío, se utilizan métodos diseñados para preservar dicha invarianza. Estos métodos suelen basarse en funciones numéricas derivadas de distancias interatómicas y ángulos entre átomos o inspirados en propiedades físicas [19] entre otros enfoques.

Las aplicaciones de estos algoritmos buscan acelerar la investigación y reducir los costos de simulación y experimentación, lo que conduce a una nueva forma de resolver problemas complejos de manera más eficiente.

Capítulo 2

Objetivos

- Calcular las energías de reorganización a partir de cálculos de estructura electrónica con el método de teoría de los funcionales de la densidad para los distintos derivados.
- Diseñar una metodología para encontrar los descriptores que expliquen la mayor variabilidad según el conjunto de datos.
- Explorar tres modelos de *Machine Learning*: 1) Regresión lineal multivariable, 2) Ensambles de árboles de decisión y 3) Redes neuronales para un problema de ajuste.
- Encontrar una metodología que permita predecir de manera confiable las energías de reorganización.
- Construir una escala de reversibilidad electroquímica con un modelo de clasificación.
- Entrenar un modelo de EAD para el problema de clasificación sobre la reversibilidad electroquímica.

Capítulo 3

Marco teórico

3.1. Teoría de Marcus

Este modelo describe la cinética de la transferencia de un electrón entre una especie electrodonadora y una electroaceptora. El proceso se modela con la energía potencial de ambas especies en donde el electrón migra de una superficie a otra. Estas superficies (parábolas) describen la energía en función de un parámetro que es llamado la coordenada de reacción como se muestra en la Fig. 3.1.

La teoría de Marcus[20] se basa en el principio de Franck-Condon y en el efecto túnel. En términos generales, el primero nos dice que la transferencia de electrones es tan veloz que se puede considerar estáticos a los núcleos. Por lo tanto, cuando el electrón pasa de los reactivos a los productos, la disposición espacial de los núcleos se mantiene fija, este proceso sucede cuando se tiene la geometría en el punto q* (Fig. 3.1). En este punto, el orbital más alto ocupado (HOMO) de los reactivos y el orbital más bajo desocupado (LUMO) de los productos se degeneran y a distancias muy pequeñas el efecto túnel es el principal responsable de la transferencia electrónica. La cinética depende de la energía de activación que se define como:

$$E_a = \frac{(\Delta E_r + \lambda)^2}{4\lambda} \tag{3.1}$$

donde ΔE_r es la energía de reacción, y λ es la energía de reorganización que se interpreta como la cantidad energética necesaria para alcanzar las posiciones de los núcleos de la estructura molecular que ha ganado o perdido un electrón sin modificar su geometría, es decir, sin relajarse. La energía de reorganización de la reducción λ_{Red} está relacionada con el proceso de reducción, mientras que λ_{Ox} con el de oxidación. En el contexto de Marcus-Nelsen [6] estos valores se calculan utilizando cuatro puntos de las parábolas, los mínimos y los puntos de la otra parábola cuya coordenada de reacción corresponde con la de los mínimos.

$$\lambda_{Red} = E_{Ox}(g_{Red}) - E_{Ox}(g_{Ox}) \tag{3.2}$$

$$\lambda_{Ox} = E_{Red}(g_{Ox}) - E_{Red}(g_{Red}) \tag{3.3}$$

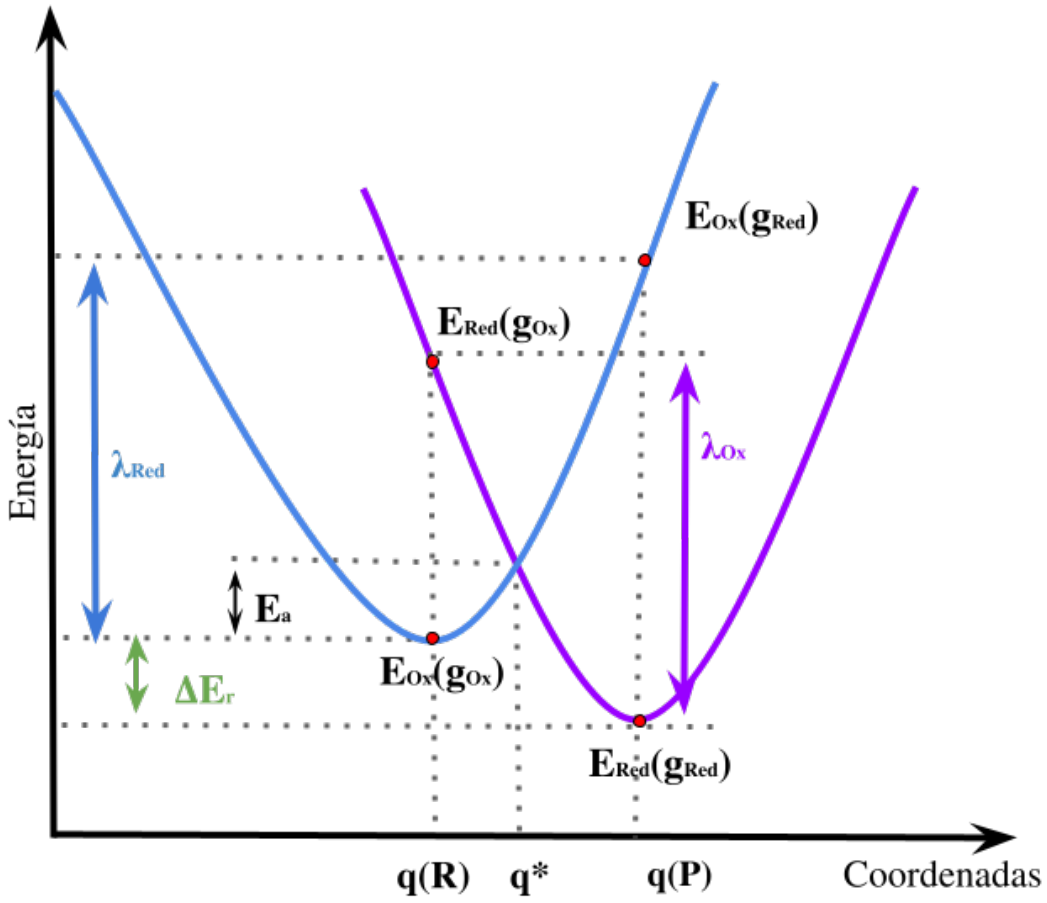


Figura 3.1: Diagrama de superficie de energía potencial de la especie oxidada (Azul) y de la especie reducida (Morado) en una reacción redox monoelectrónica.

donde g es la geometría de la especie correspondiente (oxidada o reducida). Se pueden ver estos cuatro puntos en la Figura 3.1 marcados en rojo.

3.2. Machine Learning

La Inteligencia Artificial (IA) es una disciplina de la informática que busca automatizar procesos, que abarca desde la robótica hasta el procesamiento del lenguaje. En este trabajo, nos enfocaremos en una rama específica de la IA, conocida como aprendizaje de máquina [21, 22] (ML, por sus siglas en inglés). En términos generales, el ML se refiere a algoritmos con la capacidad de aprender conceptos complejos a partir de información más simple. En este contexto "aprendizaje"hace referencia a encontrar una fórmula matemática que aplicada a un conjunto de datos de entrada produce una respuesta o salida deseada.

Una característica clave en el aprendizaje de máquina es la experiencia, ya que determina la

efectividad y calidad del modelo. Se dice que un programa de computadora aprende de la experiencia si, al medir su rendimiento, mejora en la realización de ciertas tareas. Además, el rendimiento de estos algoritmos está vinculado a la representación de los datos; es decir, la calidad de la información con la que se aprende influye en el rendimiento del modelo.

Todos los modelos ML tienen como elementos básicos: 1) Datos de entrada; es un conjunto de información o variables de entrada; 2) Modelo; algoritmo que arroja información; y 3) respuesta del modelo.

Los algoritmos de ML se pueden clasificar en tres categorías globales: 1) Aprendizaje supervisado, 2) Aprendizaje no supervisado y 3) Aprendizaje reforzado. En este trabajo, nos centramos en el primero de estos tipos. El ML supervisado consiste esencialmente en diseñar un algoritmo que usa los datos etiquetados. Existen, en general, dos problemáticas que pueden abordar los modelos de ML supervisado, dependiendo del objetivo que se quiera determinar:

- Clasificación: El algoritmo busca asignar a qué categoría pertenece un vector de entrada. Puede ser expresado mediante una función $f : \Re^n \to \{1, 2, ..., k\}$, donde "n" es la dimensión del vector de entrada y "k" representa el número de categorías. Por ejemplo, cuando la salida se limita a verdadero o falso. También existen variantes, como cuando la salida "y" se presenta como una distribución de probabilidad.
- Regresión o ajuste: Es una técnica que predice valores numéricos reales utilizando un vector como entrada. La función se puede expresar como: $f : \Re^n \to \Re$, donde "n" es la dimensión del vector de entrada. Un ejemplo sería predecir el precio de alquiler de una casa basándose en información como el tamaño, el número de habitaciones, etc.

Existe una gran variedad de modelos o algortimos [23] en ML. Aunque muchos de ellos son aplicables en varios contextos algunos dependen tanto del tipo de problema (clasificación o regresión) como de la naturaleza de la información disponible (aprendizaje supervisado o no supervisado) .

Para modelos de ML supervisado, se utilizan datos etiquedatos para entrenar el modelo. Por ejemplo, una Regresión Lineal es el se considera como el modelo de ML más simple que resuelve esta problemática . Por otro lado, para problemas de clasificación existen algoritmos como el Clasificador Bayesiano Ingenuo, Máquina de Vectores de Soporte. Algunos modelos como Ensambles de Árboles de Decisión, Redes Neuronales, K Vecinos más Cercanos esto, pueden ser aplicados tanto a problemas como clasificación como regresión.

En cambio, para el aprendizaje no supervisado, es decir, cuando no se tiene información sobre las etiquetas de los datos, se suele utilizar modelos como K-Medios, Análisis de Componentes Principales. Muchos modelos son flexibles y se pueden aplicar tanto el aprendizaje supervisado y no supervisado.

3.2.1. Regresión lineal

Es un modelo de ajuste, el cual se puede escribir como una combinación lineal

$$\hat{y} = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + ... + \omega_n x_n \tag{3.4}$$

en donde ω_k son los pesos o coeficientes y que representan el efecto de cada variable independiente en la variable dependiente y. Una manera de medir el error y encontrar el mejor hiperplano que se ajuste a los datos, es decir, encontrar los mejores pesos es por medio de la función de coste (error o pérdida), la más utilizada es el error cuadrático medio (ECM) $C = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

3.2.2. Redes Neuronales

Una red neuronal (RN, o NN por su nombre en inglés, Neural Network) es una técnica del aprendizaje automático que se inspira y simula el funcionamiento de las neuronas del cerebro humano. Las RN están compuestas por nodos o neuronas artificiales y aristas (las conecxiones entre nodos) formando una red (Fig. 3.2). La estructura de una RN se forma por capas, las cuales están compuestas por una o más neuronas. Existen dos tipos de capas: oculta(s) y salida.

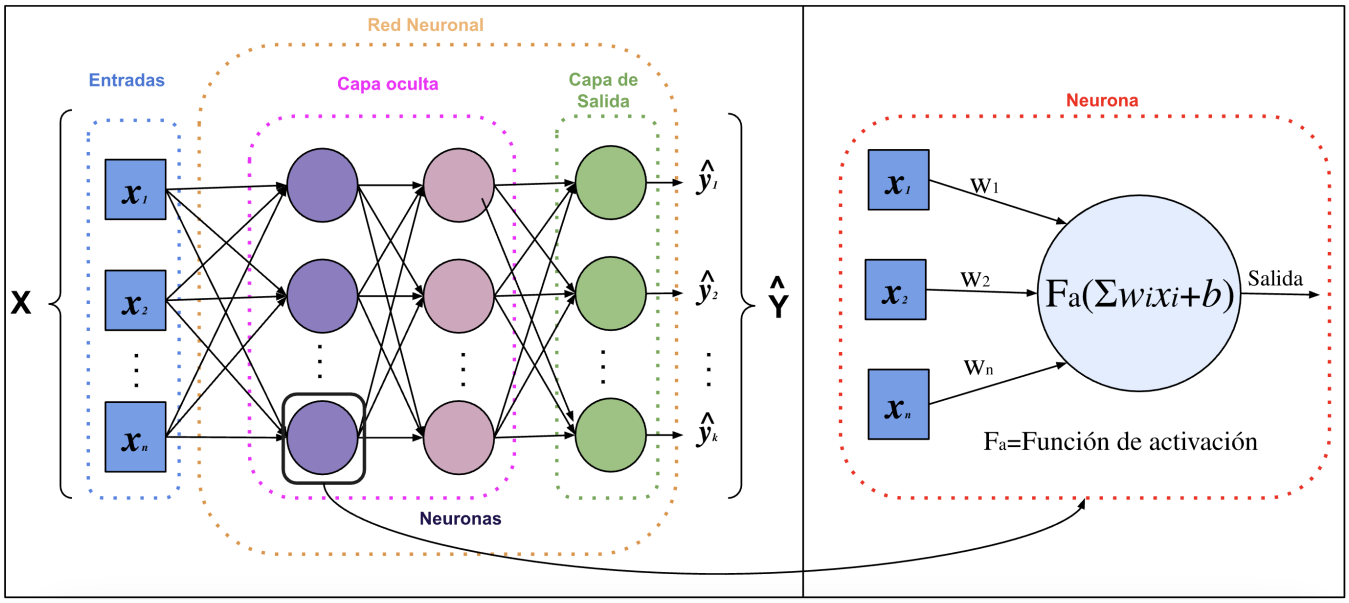


Figura 3.2: Esquema de una red neuronal con sus capas de entrada, ocultas y de salida. Así como la estructura de una neurona; donde a la suma ponderada de todas las entradas se le aplica la función de activación (f_a).

A cada entrada de la neurona se les asignan pesos (ω_i) que indican la importancia de cada entrada, estas entradas pueden ser los descriptores o características, pero también pueden ser las salida de otras neuronas. Durante el entrenamiento de la red, estos pesos se ajustan para mejorar el rendimiento del modelo, minimizando la función de coste (como se presentó en la sección de regresión lineal). La función de activación ($\mathrm{F}_a(z)$) que se encuentra en cada neurona determina el valor de salida en función de la suma ponderada de las entradas y los pesos además del término de sesgo (b).

Todos los modelos de ML supervisado “aprenden"de la misma manera, es decir, se propone una función objetivo o de costo: que a la vez está compuesta por la función de pérdida (L) y la

función de regulación (Ω), por lo tanto el "aprendizaje" en el entrenamiento es minimizarla:

$$\text{Obj}(\theta) = L(y, \hat{y}(\theta)) + \Omega(\theta) \tag{3.5}$$

donde y es la variable objetivo, $\hat{y}$ es la salida del modelo , y θ son los parámetros.

La función de pérdida tiene como propósito evaluar y calcular el error del modelo. Existe una gran variedad de funciones de pérdida, la más popular es el error cuadrático medio:

$$L = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{3.6}$$

dependiendo del problema puede utilizarce una función de costo diferente, como es el caso de la clasificación, se puede utilizar la entropía cruzada.

Por otro lado, la función de regulación controla la complejidad del modelo y tiene como propósito evitar el sobre ajuste. Algunas suelen ser las funciones de regulación llamadas L1 o L2:

$$\text{L1} : \Omega(\theta) = \lambda\sum_{j=1}^{d}|\theta_j|, \tag{3.7}$$

$$\text{L2} : \Omega(\theta) = \lambda\sum_{j=1}^{d}\theta_j^2 \tag{3.8}$$

donde λ es el coeficiente de regulación, θ son los parámetros que se optimizan y d el número de parámetros del modelo.

Existen distintos métodos para minimizar la función de pérdida en problemas de aprendizaje supervisado. Uno de los algoritmos más utilizados es el descenso del gradiente, que busca determinar el conjunto de parámetros (θ) que minimicen la función objetivo. El vector gradiente se define como:

$$\nabla_\theta L = [\frac{\partial L}{\partial \theta_1}, \frac{\partial L}{\partial \theta_2}, \ldots, \frac{\partial L}{\partial \theta_d}] \tag{3.9}$$

donde d es el número total de parámetros. Este vector indica la dirección de máximo crecimiento de la función de pérdida, por lo que desplazarse en el sentido opuesto conduce a un conjunto de valores mínimos locales. Los parámetros se actualizan de acuerdo a la tasa de aprendizaje μ. Se ajusta de tal manera que no sea muy pequeño porque esto tiene un impacto directo en el tiempo para encontrar un mínimo local. Por otro lado, sí es muy grande podría estar oscilando y nunca encontrar un mínimo local.

$$\theta_j = \theta_j - \mu\frac{\partial L}{\theta_j} \tag{3.10}$$

El proceso es iterativo hasta alcanzar un mínimo local o cumplir un criterio de convergencia.

Existen variaciones del método del descenso del gradiente, como el descenso de gradiente estocástico (SGD), Adam, RMSprop entre otros que incluyen distintas maneras de calcular el hiperparámetro μ. Existen otros métodos basados en las segundas derivadas (métodos de Newton). El método de Adam es el más popular por su rápida convergencia.

La manera más popular de calcular dichas derivadas es con el método conocido como propagación hacia atrás o retroprogapación (*backpropagation*, en inglés). El método de retropropagación se resume en determinar los pesos y sesgos de la K-ésima capa, en función de la capa anteriror (K-1).

Existe una gran variedad de redes neuronales con diferentes enfoques. La más sencilla, y que se explicó anteriormente, son las Redes Neuronales Artificiales hacia adelante (RNA), también conocida como Perceptrón Multicapa (PMC).

Las Redes Neuronales Convolucionales (RNC) están basadas en las ARN y forman parte de los métodos de aprendizaje profundo. Este tipo de redes son usadas comúnmente en el procesamiento de imágenes. Las CNN emplean capas convolucionales* que se enfocan en extraer características de pequeñas regiones predefinidas de una imagen. El entrenamiento incluye, además de los parámetros de una ARN los parámetros asociados a las capas convolucionales.

Las Redes Neuronales Recurrentes (RNN, por sus siglas en inglés) son otro tipo de redes neuronales artificiales. Las RNN permiten que las salidas de algunos nodos sean las entradas de los mismos nodos como entradas adicionales. Se suelen utilizar en el procesamiento de lenguaje natural y audio y se han utilizado en la predicción de productos de reacciones orgánicas [24].

Las Redes Neuronales de grafos (GNN, por sus siglas en inglés), es un tipo de aprendizaje profundo, que procesa los datos en una estructura de grafos. En química han tenido éxito en predicción de propiedades de moléculas [25].

Redes de Transformadores (o Tranformers) suele usarse para generar texto, como en el caso de ChatGPT. AlphaFold2 es una aplicación de este tipo de arquitectura.

3.2.3. Funciones de activación

Las funciones de activación se usan para transformar los valores de salida de una neurona, a través de funciones algebraicas. Existen diversas funciones que cumplen con diferentes objetivos en las redes neuronales, cada una con ciertas ventajas, dependiendo del problema. Suelen dividirse en lineales y no lineales.

Lineales:

Identidad: Definida en el intevalo $(-\infty, \infty)$

$$g(x) = x \tag{3.11}$$

Este tipo de funciones genera que la red neuronal se comporte como una sola capa.

No lineales:

- Logística: Definida entre intevalo (0,1)

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.12}$$

- Tangente hiperbólica: Definida en el intervalo (-1,1)

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3.13}$$

*Es un tipo de filtro que extrae diferentes características

- ReLU: Unidad lineal rectificada está definida en el intervalo [0,∞]

$$\text{ReLU}(x) = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{if } x \leq 0. \end{cases} \tag{3.14}$$

- Sigmoide: Definida en el intervalo [0, ∞]

$$\text{SiLU}(x) = \frac{x}{1 + e^{-x}} \tag{3.15}$$

- SoftMax: Toma como entrada un vector $\vec{x}$ de números reales de tamaño K y los normaliza en una distribución de probabilidad de tamaño K. Está definida en el intervalo (0,1)

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e_j^z}, \text{para } i = 1, \dots K \tag{3.16}$$

3.2.4. Ensambles de árboles de decisión

XGBoost (`Extreme Gradient Boosting` [26]) es un algoritmo basado en conjuntos de árboles de decisión (CART, por sus siglas en inglés `Classification and Regression Trees`), que se entrenan de manera secuencial utilizando el método de descenso de gradiente. Los árboles de decisión, o CART, realizan decisiones binarias de forma jerárquica, y están compuestos por "hojas" y "ramas". Las hojas representan un puntaje, mientras que las ramas corresponden a decisiones que llevan a los diferentes puntajes.

La aproximación se puede escribir como:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) \tag{3.17}$$

donde f_k representa un CART y t el número total de árboles.

Si se entiende que es un proceso secuencial, se puede hacer la aproximación a partir del conjunto de árboles anteriores:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3.18}$$

donde $\hat{y}_i^{(t-1)}$ es la aproximación sin el último árbol.

Para entrenar el modelo se usa la función de optimización que consiste en suma de la función de pérdida de entrenamiento y la función de regulación:

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \omega(f_k) \tag{3.19}$$

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_k(x_i)) + \omega(f_t) \tag{3.20}$$

donde l es la función de pérdida, que puede ser el Error cuadrático medio (ECM o MSE *Mean Square Error* por sus siglas en inglés) o pérdida logística n es el número de observaciones y w es el término de regulación que evita el sobreajuste y que mide la complejidad del árbol.

Debido a que hay funciones de pérdida que no suelen tener una forma tan simple como MSE se escribe la función con una aproximación de Taylor:

$$obj^{(t)} = \sum_{i=1}^{n}[l(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i) + \frac{1}{2}h_i * f_t^2(x_i)] + \omega(f_t) \tag{3.21}$$

donde g_i y h_i se definen como la primer y segunda derivada de la función de pérdida:

$$g_i = \frac{\partial}{\partial \hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{3.22}$$

$$h_i = \frac{\partial^2}{\partial^2 \hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{3.23}$$

Por lo tanto, la función de pérdida que optimiza un árbol se escribe de manera general:

$$obj^{(t)} = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + w(f_t) \tag{3.24}$$

si se define un árbol como:

$$f_t(x) = w_{q(x)}, w \in R^T, q : R^d, \{1, 2, ..., T\} \tag{3.25}$$

donde w es un vector de puntaje de cada hoja y T es el número total de hojas en el árbol. $q(x)$ es una función que asigna el puntaje a cada hoja.

Se define a la función de regulación o complejidad como:

$$\omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3.26}$$

Por lo tanto, la función objetivo se puede escribir de la siguiente manera:

$$obj^{(t)} \approx \sum_{i=1}^{n}[g_i w_{1(x_i)} + \frac{1}{2}h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3.27}$$

$$obj^{(t)} = \sum_{j=1}^{T}[G_j w_j + \frac{1}{2}(H_j + \lambda)w_j^2] + \gamma T \tag{3.28}$$

donde $G_j = \sum_{i \in I_j} g_i$ y $H_j = \sum_{i \in I_j} h_i$ y I_j es el conjunto de índices de las observaciones que están en la j-ésima hoja.

Al minimizar la función objetivo con respecto al vector de puntaje de la j-ésima hoja, es decir:

$$\frac{d}{dw_j} obj^{(t)} = 0 \tag{3.29}$$

se encuentra que el conjunto de puntajes mínimo tiene la forma:

$$w_j^{min} = -\frac{G_j}{H_j + \lambda} \tag{3.30}$$

y por lo tanto la función objetivo con el error mínimo tiene la forma:

$$obj^{min} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{3.31}$$

3.3. Análisis de componentes principales

El método de Análisis de Componentes Principales (ACP) en una herramienta exploratoria [27, 28, 29] que se utiliza para reducir la dimensionalidad del conjunto de variables independientes que se utilizan para describir una variable objetivo en un conjunto de datos (dataset), a la vez que incrementa la interpretabilidad y minimiza la pérdida de información, en otras palabras, conserva la "variabilidad". Este problema de conservar la variabilidad de los datos se traduce en encontrar unas nuevas variables, componentes principales (CP), que son funciones lineales de las variables originales. Encontrar estas variables nuevas se reduce a resolver un problema de valores propios.

Un dataset se conforma de n-individuos, cada uno con observaciones de p-variables numéricas $(V_1, V_2, V_3, ..., V_p)$ lo que es equivalente a tener una matriz $\mathbf{X}$ de n×p. PCA busca la combinación lineal de las columnas (V) con la máxima varianza. Este método permite condensar la información que comparten distintas variables.

La componente principal se escribe como

$$\sum_{j=1}^{p} a_j \mathbf{x}_j = \mathbf{Xa} \tag{3.32}$$

donde a es un vector de constantes $a_1, a_2, ..., a_p$ y se conocen como loadings o pesos y estos indican la importancia de cada variable en cada componente. La varianza para cualquier combinación lineal se define como:

$$\mathrm{Var}(\mathbf{Xa}) = \mathbf{a}^T\mathbf{Sa} \tag{3.33}$$

donde $\mathbf{S}$ es la matriz de covarianza asociada al dataset. Obtener la combinación lineal o la CP con la máxima varianza es a través de construir un vector $\mathbf{a}$ que maximice la forma cuadrática de $\mathbf{a}^T\mathbf{Sa}$. Una restricción adicional a esto es que el vector debe estar normalizado y con esto se construyen los multiplicadores de Lagrange

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{a}}\left[\mathbf{a}^T\mathbf{Sa} - \lambda(\mathbf{a}^T\mathbf{a} - 1)\right] = 0 \tag{3.34}$$

$$2\mathbf{Sa} - 2\lambda\mathbf{a} = 0$$

$$\mathbf{Sa} = \lambda\mathbf{a} \tag{3.35}$$

donde λ es el multiplicador de Lagrange. Además que λ corresponde al eigenvalor de la matriz de covarianza $\mathbf{S}$. De tal manera la Ec.3.35 es la ecuación de valores propios que hay que resolver.

La varianza explicada determina la variabilidad que aporta cada CP. Por lo tanto, la varianza acumulada explica la variabiliad dada un conjunto de CP.

3.4. Coeficiente de adecuación KMO

La prueba de adecuación Kaiser-Meyer-Olvin (KMO por sus siglas), es una medida estadística para evaluar si un conjunto de datos son adecuados para hacer un análisis factorial. Henry F. Kaiser propuso en 1970 [30] la medida de adecuación de muestreo MSA (por sus siglas en inglés) y para el año 1974 [31, 32], Meyer e Ingram Olkin la modificaron.

La prueba normalizada se define como:

$$\text{KMO} = \frac{\sum_{j \neq k} r_{j,k}^2}{\sum_{j \neq k} r_{j,k}^2 + \sum_{j \neq k} p_{j,k}^2} \tag{3.36}$$

donde $Q = SR^{-1}S$, y $S = (\text{diag}R^{-1})^{-1/2}$. Por lo tanto el término $\sum \sum_{k>j} r_{j,k}^2$ es la suma de los cuadrados de los elementos superiores fuera de la diagonal de la matriz de correlación R. El término$\sum \sum_{k>j} p_{j,k}^2$ es la suma de los cuadrados de los elementos superiores fuera de la diagonal de la matriz Q. Los elementos diagonales de la matriz diagonal S son las raíces cuadradas de los recíprocos de los elementos diagonales de R^{-1} y Q es R^{-1} con cada fila y cada columna multiplicada por el elemento diagonal correspondiente de S.

La evaluación subjetiva que propone Kaiser sobre la prueba se presenta en la Tabla 3.1. El límite de lo aceptable es en el caso de que la prueba KMO=0.5 . En general, entre más cercano esté a la unidad, se considera que una base de datos ideal para el análisis de factores.

Intervalo	Evaluación
$[0.90, 1]$	maravilloso
$[0.80, 0.90)$	meritorio
$[0.70, 0.80)$	medio
$[0.60, 0.70)$	mediocre
KMO$\geq$ 0.50	miserable
KMO$<$ 0.50	inaceptable

Tabla 3.1: Índices de de la prueba KMO.

3.5. Medidas de correlación

3.5.1. Coeficiente de correlación Pearson

El coeficiente de correlación de Pearson [33, 34], nombrada así por Karl Pearson, mide la correlación lineal entre dos variables y está definida en términos de la covarianza $\mathrm{Cov}(X, Y)$ y la desviación estándar $\sigma(X)$, tal que:

$$\mathrm{P}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma X \sigma Y} \tag{3.37}$$

Es una medida normalizada y simétrica, por lo que el rango de valores están definidos entre -1 y 1. Si el valor está más cercano a la unidad se encuentran sobre una línea recta, por el contrario si es cercana a cero impica que no hay dependencia lineal entre las variables. El signo de la correlación está asociado a la pendiente. Valores negativos son pendientes negativas, mientras que valores positivos son pendientes positivas

3.5.2. Coeficiente de correlación de Kendall

El coeficiente de correlación de Kendall (τ) mide la asociación ordinaria entre dos variables, nombrada así por Mauricio Kendall[35, 36] quien la desarrolla en 1938. El coeficiente se define en términos del número de pares concordantes y discordantes. Es decir, dado un conjunto de observaciones de dos variables aleatorias donde se tiene cualquier par de observaciones (x_i, x_j) y (y_i, y_j) con $j > i$, se dice que son concordantes si se cumple que $x_i > x_j$ y $y_i > y_j$ o si $x_i < x_j$ y $y_i < y_j$ de lo contrario se dice que son discordantes. Por lo tanto τ se define como:

$$\tau = \frac{(\text{n° de pares concordantes}) - (\text{n° de pares discordantes})}{\text{n° de pares totales}} \tag{3.38}$$

El coeficiente varía entre -1 y 1, debido a que el denominador es el número de pares totales, es decir si todos son discordante el valor es negativo (-1), el caso contrario, todos los pares son concordantes entonces el valor que toma es 1. En el caso que el número de pares discordantes y concordantes es el mismo, τ es cero lo que se interpreta como que no hay correlación.

3.5.3. Coeficiente de correlación de Spearman

El coeficiente de correlación de Spearman (ρ) mide la relación entre dos variables que puede ser descrita por una función monótona[†], sean lineales o no. Nombrada así por Charles Spearman y se define tal que:

$$\rho = \frac{\mathrm{Cov}[R[X], R[Y]]}{\sigma(R[X])\sigma(R[Y])} \tag{3.39}$$

donde $R[X]$ y $R[Y]$ son las variables jerarquizadas.

Está definida en un intervalo [-1, 1]. Este coeficiente indica la dirección de asociación, es decir, si la variable X incrementa cuando la variable Y también, ρ es positivo, en el caso contrario ρ es negativos. Para el caso en que el coeficiente es cero, no hay una respuesta de la variable Y cuando incrementa X.

[†]Una función se dice monótona si su derivada es constante en todo el intervalo.

3.5.4. Coeficiente de distancia de correlación

La distancia de correlación [37] (DC, por sus siglas en inglés) es propuesta por Gábor J. Székely en el 2005. Es una medida de la dependencia entre dos variables, es decir, mide la intensidad de la relación ya sea lineal o no, a diferencia del coeficiente de correlación de Pearson, que solo mide linealidad. La DC se define como:

$$\mathrm{dCor}(X,Y) = \frac{\mathrm{dCov}(X,Y)}{\sqrt{\mathrm{dVar}(X)\mathrm{dVar}(Y)}} \tag{3.40}$$

donde:

- $\mathrm{dCov}(X,Y)$ es la covarianza de la distancia.
- $\mathrm{dVar}(X)$ es la varianza de la distancia.

La covarianza y varianza de la distancia se definen como:

$$\mathrm{dCov}^2(X,Y) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{j,k}B_{j,k}, \tag{3.41}$$

$$\mathrm{dVar}^2(X,Y) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{j,k}^2 \tag{3.42}$$

$$\tag{3.43}$$

donde n es el número total de muestras y A y B son las matrices de distancia doblemente centradas, definidas como:

$$A_{j,k} = a_{j,k} - \bar{a}_j - \bar{a}_{\cdot k} + \bar{a}_{\cdot\cdot}, \tag{3.44}$$

$$B_{j,k} = b_{j,k} - \bar{b}_j - \bar{b}_{\cdot k} + \bar{a}_{\cdot\cdot} \tag{3.45}$$

donde $\bar{a}_{j\cdot}$ es la media de la j-ésima fila, $\bar{a}_{k}$. es la media de la k-ésima columna y $\bar{a}_{\cdot\cdot}$ es la gran media, y $a_{j,k}$ y $b_{j,k}$ son las matrices de distancia que se definen como la norma vectorial entre las muestras:

$$a_{j,k} := ||X_j - X_k||, \tag{3.46}$$

$$b_{j,k} := ||Y_j - Y_k|| \tag{3.47}$$

La DC está definida en el intervalo [0, 1]. Si la correlación es igual a 0, implica que las variables son independientes. Por otro lado, una correlación igual a 1 implica que hay una relación perfecta entre las variables, que puede ser lineal o no lineal. En general, si el coeficiente es más cercano a 1 hay una relación más fuerte entre las variables.

En resumen, las correlaciones presentadas miden diferentes tipos de relación entre dos distribuciones de datos. Mientras que el CCP detecta linealidad, DC detecta no linealidad. Por otro lado los coeficientes de correlación CCK CCS detectan relaciones monótonas. Resalta que los coeficientes más cercanos a la unidad, independiente de la formulación, detecta una correlación más alta.

Capítulo 4

Metodología

Como primer paso se tomaron algunos compuestos orgánicos nitrogenados que se conocen que tienen propiedades oxido-reducción, tal como: Bipiridina (Bpy), Metil Viológeno (MV) y Bencidina (Bz) en la Fig. 4.1, se presentan los esquemas moleculares representativos de cada especie.

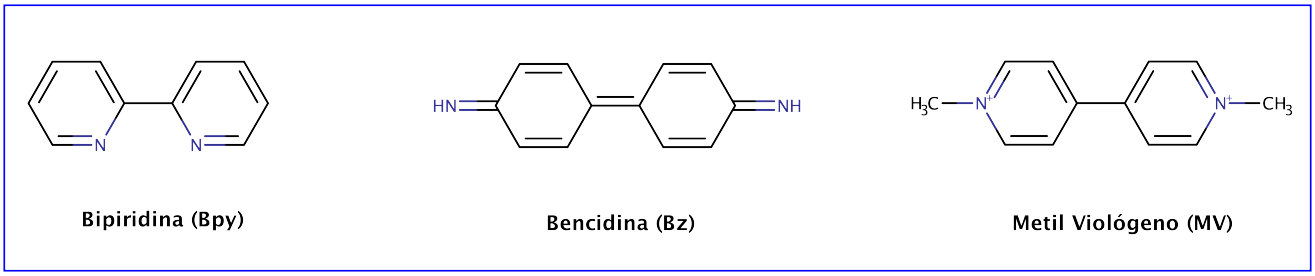
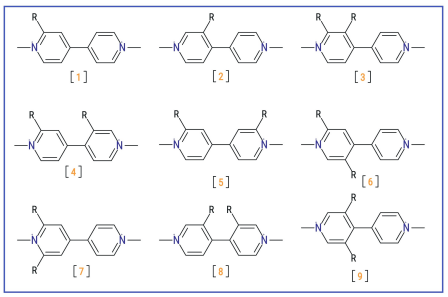


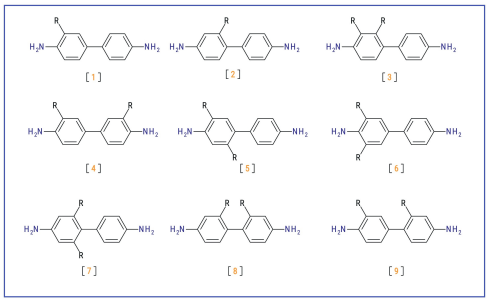
Figura 4.1: Estructuras moleculares de las especies orgánicas nitrogenadas seleccionadas como estructuras base.

A partir de estas familias se obtuvieron diversos derivados con una y dos sustituciones del mismo grupo funcional: para la Bpy, se obtiene un total de 20 derivados (Figura 4.2c); para la Bz, 12 derivavados (Figura 4.2b) ; y para el MV, 9 derivados (Figura 4.2a).

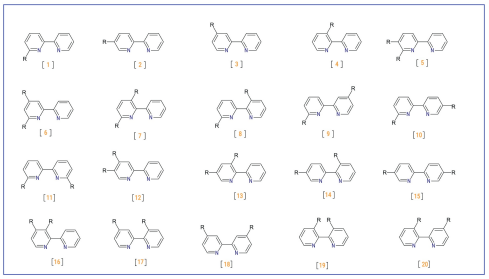
Se utilizan 48 distintos grupos funcionales (Fig. 4.3) sin propiedades ácido-base. Para cada derivado se plantea un proceso de transferencia de dos electrones, y apartir de la Teoría de Marcus se estudian los casos monoelectrónicos, es decir, una reacción óxido-reducción (Redox: $\mathrm{Ox} + \mathrm{e}^- \longrightarrow$ Red) monoelectrónica (intercambio de un electrón), de tal manera que se obtienen dos especies para cada derivado, una que es la especie oxidada (Ox) y la otra como especie reducida (Red). Para la Bpy se plantean 6 reacciones (Figura 4.5) monoelectrónicas, para el MV, 3 reacciones (Figura 4.4).



(a) MV



(b) Bz



(c) Bpy

Figura 4.2: Conjuntos de derivados para las tres familias.

Para el caso de la familia de la Bz, se estimaron algunos pK_as, de acuerdo a la escuación 4.1

$$pK_a = \frac{G(HA) - G(A^-) - G(H^+)}{RT\ln(10)} \tag{4.1}$$

donde G, es la energía libre de Gibbs asociada a las especies: HA, A^- y H^+. R es la constante universal de los gases y T es la temperatura (en este trabajo es T=298.15 K)

La energía libre de Gibbs asociada al protón se obtiene de un ajuste con dos pK_as experimentales [38] que da un resultado de $G(H^+) = -267.8276$Kcal/mol. Una vez obtenidos los valores de pK_a teóricos de todos los procesos ácido-base se seleccionan aquellas especies que están dentro (o muy cercanos) a la ventana del agua, es decir, entre 0 y 14. Esto da un resultado de 4 reacciones redox monoelectrónicas (Figura 4.6) .

Para cada derivado se construye un SMILES (cadena de caracteres que representa la estructura bidimensional de una molécula), considerando que para cada reacción Redox se construyen dos SMILES: uno para la especie Red y otro para la especie Ox. A partir del SMILES se calculan 208 descriptores moleculares o quimioinformáticos (valor numérico de alguna propiedad química) que provee la librería RDKit de Python. Con estos descriptores se construye la base de datos que servirá como variables de entrada para los modelos de ML. Debido a que se tienen dos SMILES por cada reacción, se construyen dos bases de datos: por un lado, los descriptores que se calculan a partir de los SMILES de las especies oxidadas, que para futuras referencias se abreviará como DatOx; y por otro lado, la base de datos construida con los SMILES de las especies reducidas, que se llamará DatRed.

De manera independiente, para cada derivado se hace una optimización de geometría: primero se realiza un análisis conformacional. Para ello, se construyen k confórmeros con la librería de

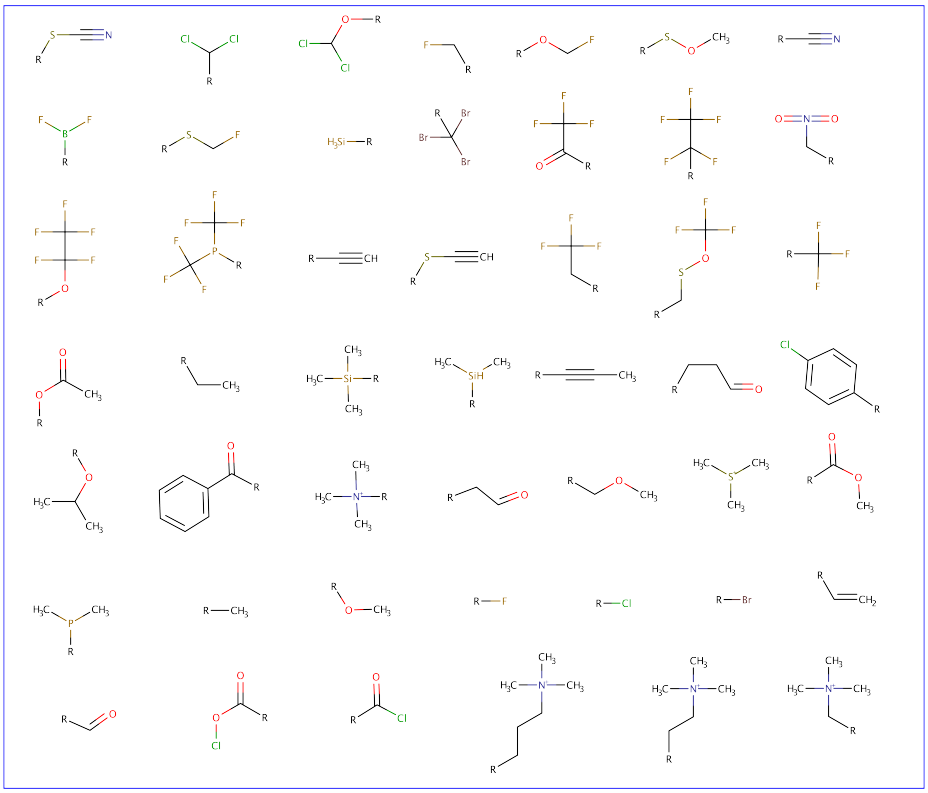


Figura 4.3: Conjunto de 48 distintos grupos funcionales, sin propiedades ácido-base (átomos susceptibles a adherirse un protón H^+).

OpenBabel, se calcula la energía de cada confórmero y se selecciona aquel que tenga la menor energía. Posteriormente, se optimiza la geometría con un cálculo de estructura electrónica (Gaussian 16). Se usa el nivel de teoría B3LYP/6-311+G(d,p)/SMD, utilizando agua como disolvente. Se hace un cálculo termodinámico (aún en Gaussian) para obtener la energía libre de Gibbs. Para confirmar que se obtiene un mínimo de energía local, se hace un análisis de frecuencias. Esto se realiza para cada par de especies de acuerdo a las reacciones redox, es decir, para la especie Ox y Red. Después de obtener las estructuras optimizadas, se procede a hacer cálculos de geometría restringida (Gaussian), es decir, se hace un cálculo únicamente termodinámico, como en el caso anterior.

Una vez obtenidas, para cada reacción redox, las geometrías más estables y sus energías, así como las energías del correspondiente cálculo de geometría restringida, se procede a calcular las energías de reorganización (ER) de acuerdo con el esquema de Marcus-Nelsen. Este conjunto de moléculas y sus ER fungirán como la variable objetivo para los modelos.

El último paso en la metodología consiste en un análisis estadístico, el cual busca identificar las variables de entrada más relevantes. Este análisis se divide en 4 criterios:

- Criterio 0: En esta etapa, se busca tener un sistema de referencia; por ello, se conservan todas las variables de entrada y se entrenan los modelos.
- Criterio 1: Se eliminan aquellas variables con poca variabilidad en el conjunto de las moléculas estudiadas a partir de analisar la varianza de los datos ($\sigma(\mathbf{X})$).

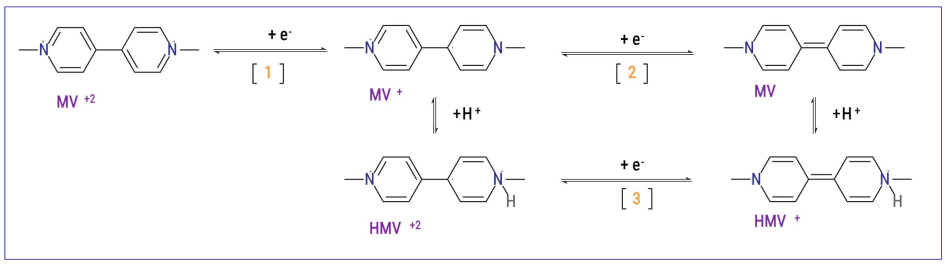


Figura 4.4: Esquema reaccional de los procesos monoelectrónicos para el MV.

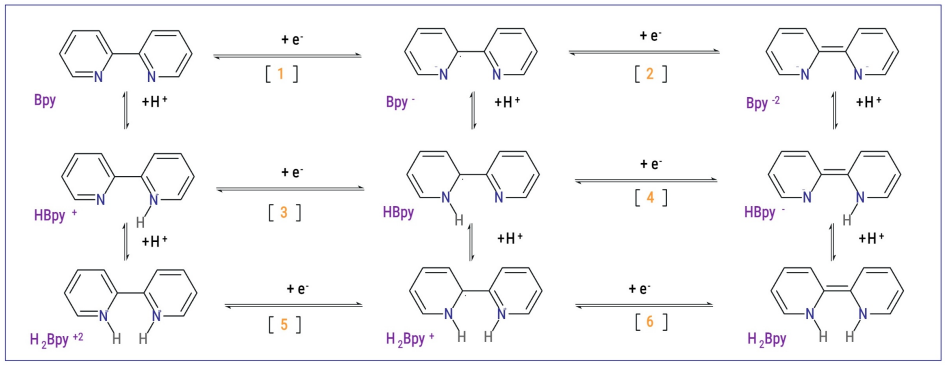


Figura 4.5: Esquema reaccional de los procesos monoelectrónicos para la Bpy.

- Criterio 2: Con el objetivo de no repetir información, se buscan las variables de entrada que sean equivalentes. Para ello, se construye una matriz de correlación Corr($\mathbf{X}, \mathbf{Y}$) (con todas las diferentes formulaciones) entre todas las variables de entrada.
- Criterio 3: Se seleccionan las variables de entrada que tengan mayor correlación con la variable objetivo. Para ello, se genera un vector de correlación utilizando todas las formulaciones propuestas Corr($\lambda, \mathbf{X}$) (Pearson, Kendall, Spearman y Distancia de correlación).
- Criterio 4: Para este último criterio, se usa el método de análisis de componentes principales (PCA), el cual justifica que con el Criterio 3 las variables de entrada tienen importancia estadística y, además, puede ayudar a seleccionar las variables que describan la mayor variabilidad de los datos.

En la Fig. 4.7 se resumen, en un esquema para la familia de las Bipiridinas, los criterios de selección de variables. Donde se expresa el número de variables de conjunto de datos en cada etapa, tanto para el conjunto de las especies oxidadadas como las especies reducidas. En cada estapa se evaluan los modelos de ML para ver el rendimiento del conjunto de datos.

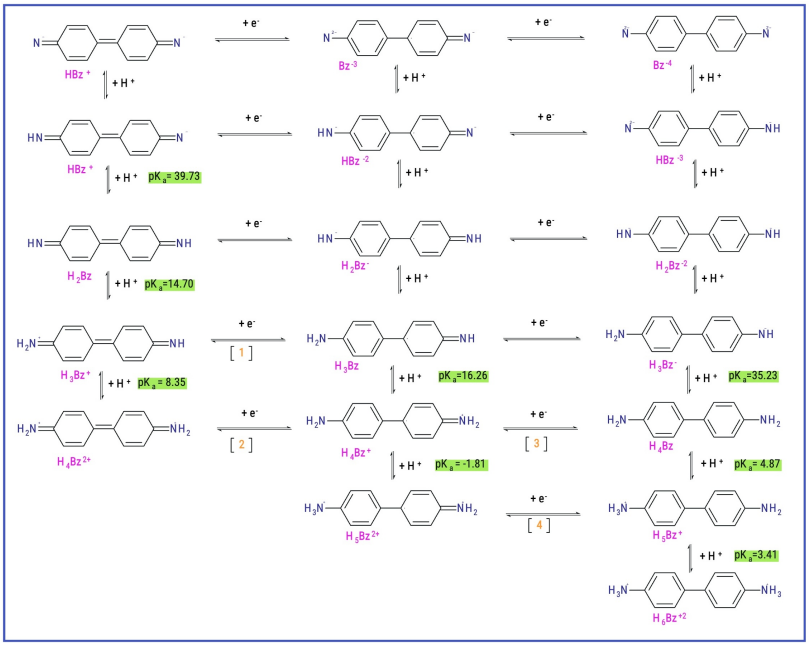


Figura 4.6: Esquema reaccional de los procesos monoelectrónicos para la Bz.

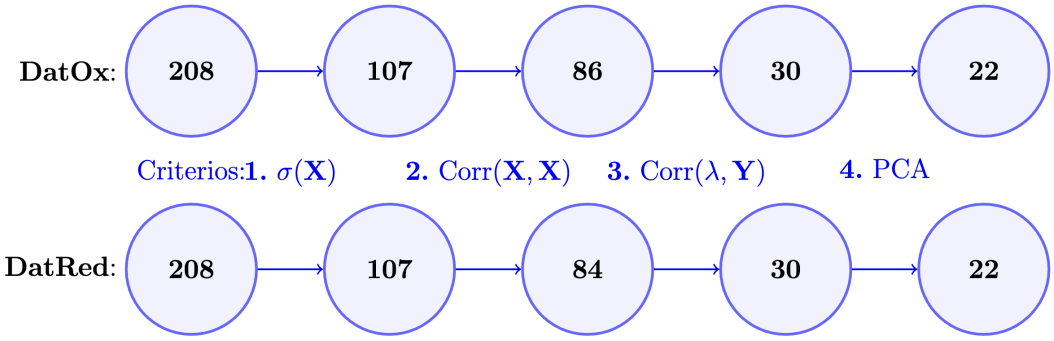


Figura 4.7: Esquema del cambio en el número de variables del conjunto de datos con los diferentes criterios para la familia de las Bipiridinas. DatOx es el conjunto de datos de las especies oxidadas y DatRed el de las reducidas. $\sigma(\mathbf{X})$ es la desviación estándar, Corr$(\mathbf{X},\mathbf{Y})$ es la matrix de correlación del DataSet, Corr$(\lambda,\mathbf{X})$ es el vector de correlación entre el conjunto de datos y la variable objetivo. Para esta figura Corr$(\lambda,\mathbf{x})$ es el CCP.

Capítulo 5

Resultados (Bipiridina)

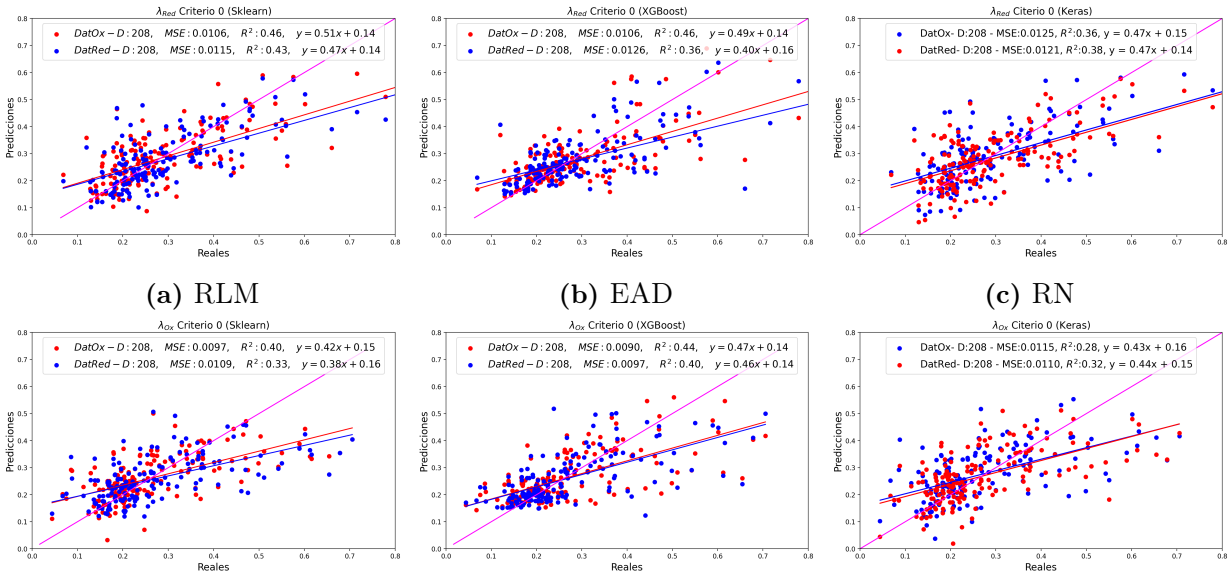
En este capítulo (y todos los siguientes resultados) se presentan los resultados del análisis de los conjuntos de datos y de los modelos de ajuste (regresión lineal multivariable, RLM; ensambles de árboles de decisión, EAD; y redes neuronales, RN) y los resultados de la clasificación (ensambles de árboles de decisión y redes neuronales, RN). La sección de resultados se divide en cuatro capítulos: 1) Bpy, 2) Bz, 3) MV, y 4) todas las familias. A su vez, en cada familia, se abordan los cuatro criterios en los que se describe de manera detallada el tratamiento de los datos o el entrenamiento del modelo, según sea el caso. En todas las familias existen dos bases de datos: DatOx y DatRed. Se presentan los resultados en gráficas de dispersión, para el caso de los modelos de ajuste, en donde la línea rosa (-): es la recta con pendiente 1 y con la ordenada al origen en cero, esta línea sería el caso donde las predicciones son iguales a las variables objetivo. Las otras líneas representan la recta (de ajuste) asociada a cada entrenamiento, con la pendiente (m), la ordenada al origen (b) y el coeficiente de determinación (R^2).

Los datos presentados en las gráficas de dispersión representan al conjunto de prueba (*test* en inglés). En todos los entrenamientos la distribución de los datos es la misma: el 80 % de entrenamiento y el 20 % de prueba. Para la RN se usó una capa oculta con 20 neuronas con la función de activación sigmoide. Para los EAD se proponen 10000 árboles de decisión con una profundidad de 6, pero que al momento de optimizar los árboles, con el método de descenso de gradiente se detendrá si después de 10 árboles (o iteraciones) el parámetro MSE no disminuye. La variable objetivo, es decir, la energía de reorganización (λ_{Red} o λ_{Ox}) están en unidades de eV, por consecuencia las unidades de las predicciones son las mismas.

5.1. Criterio 0: Referencia

En este punto, para establecer un marco de referencia sobre los tres modelos que se exploran para la predicción de energías de reorganización habrá un ajuste, sin ninguna modifcación en las variables de entrada utilizando las 208 (Figura 5.1 tanto para la energía de reroganización λ_{Red} como λ_{Ox}). Se toman en cuenta las 208 variables (obtenidas con la libreria `RDKit` de Python) y se entrenan los modelos con un conjunto de 828 reacciones redox monoelectrónicas.

En los resultados del entrenamiento de los modelos se logra observar la lejanía de los datos de



(a) RLM **(b)** EAD **(c)** RN

(d) RLM **(e)** EAD **(f)** RN

Figura 5.1: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} (primer renglón) y λ_{Ox} (segundo renglón), considerando el Criterio 0. En cada gráfica, se incluyen los coeficientes de determinación (R^2), el error cuadrático medio (MSE) y la ecuación de la línea recta ($y = mx + b$). Donde D representa el número de variables.

prueba con respecto a la línea de referencia (−, con pendiente uno y ordenada al origen 0). Esto muestra que los ajustes no son buenos, en el sentido de que no son reproducibles los datos calculados con mecánica cuántica a partir de los calculados con ML. Por otro lado, la medida de dispersión de los datos (R^2) de la línea de tendencia de cada conjunto de datos, es representativa de los valores de interés, es decir que confirma que el ajuste de los datos calculados a partir de los datos de ML no es bueno. Finalmente los valores del error cuadrático medio (MSE) es otro indicador de la calidad del ajuste.

Los resultados de los parámetros de rendimiento provenientes del conjunto de datos de prueba, tienen como propósito medir el desempeño del modelo de predecir valores a partir de datos que no conoció en el proceso de entrenamiento. Los resultados de estas medidad de rendimiento se muestran en la Tabla 5.1para los tres modelos y para las energías de reorganización λ_{Red} y λ_{Ox}. Los valores resaltados en azul representan el mejor modelo, mientras que los valores en **negritas** señalan las mejores métricas de rendimiento. Es decir, el parámetro MSE marcado en **negritas** destaca como el mejor valor según el renglón correspondiente, al compararlo entre modelos y bases de datos. En cada métrica se indica con una flecha el sentido que debe tomar. Es decir, el objetivo es maximizar a uno tanto R^2 como m , por ello tiene una flecha hacia arriba (↑), mientras que MSE y b se busca minimizar (↓) a cero.

Para λ_{Red} se observa que los mejores rendimientos están entre RLM y EAD, destacando entre estos el primero. Hay que tener en cuenta que el modelo RLM no es tan flexible, es decir, no se pueden modificar los hiperparámetros, simplemente porque no tiene, mientras que los EAD sí lo puede hacer, es decir se puede modificar la tasa de aprendizaje, la profundidad del árbol etc. Por otro lado, para λ_{Ox} el mejor modelo es EAD con el conjunto DatOx.

Energía	Métrica	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
λ_{Red}	MSE(↓)	**0.0106**	0.0115	**0.0106**	0.0126	0.0125	0.0121
	R^2(↑)	**0.46**	0.43	**0.46**	0.36	0.36	0.38
	m(↑)	**0.51**	0.47	**0.49**	0.40	0.47	0.47
	b(↓)	**0.14**	**0.14**	**0.14**	0.16	0.15	**0.14**
λ_{Ox}	MSE(↓)	0.0097	0.0109	**0.0090**	0.0097	0.0115	0.0110
	R^2(↑)	0.40	0.33	**0.44**	0.40	0.28	0.32
	m(↑)	0.42	0.38	**0.47**	0.46	0.43	0.44
	b(↓)	0.15	0.16	**0.14**	**0.14**	0.16	0.15

Tabla 5.1: Métricas de rendimiento de los tres modelos de ML para el criterio 0, con λ_{Red} y λ_{Ox} como variable objetivo. Usando los 208 descriptores quimioinformáticos

A partir de los resultados presentados en las tablas se puede demostrar que la combinación de la variable objetivo, la base de datos y el modelo tienen un impacto en las métricas de validación. Por lo tanto, se puede concluir que para este criterio, es decir, con 208 variables se predice mejor λ_{Ox} con los EAD y la base de datos DatOx mientras que λ_{Red} es mejor una RLM y DatOx.

5.2. Criterio 1: Variabilidad de los datos

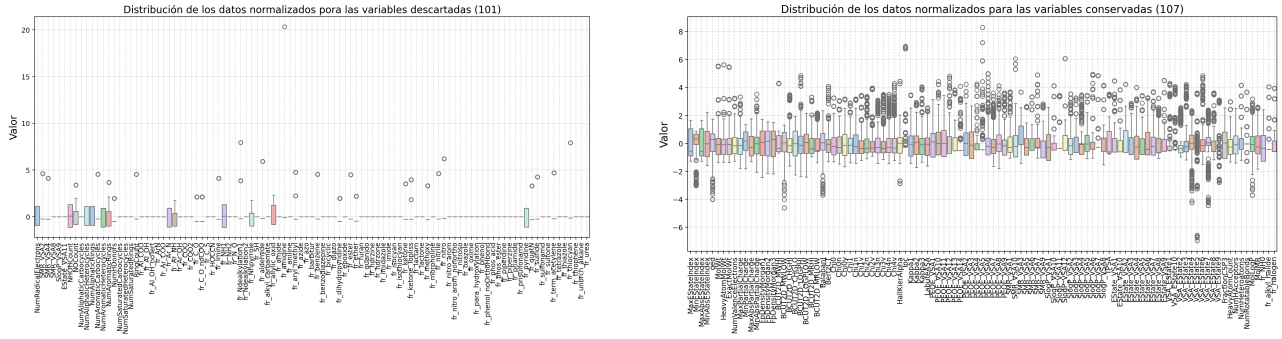
Hay variables que tienen una desviación estándar muy pequeña (o nula), lo que esto puede indicar es que esa variable repite el mismo valor para diferentes moléculas o que difiere muy poco de una a otra especie. Este tipo de variables no aportan información relevante al entrenamiento del modelo porque no permite distinguir una molécula de otra. El criterio es que se descartaron las variables que tienen cuatro o menos distintos valores en todo el conjunto de moléculas.

La Figura 5.2a y 5.3a (DatOx y DatRed respectivamente) muestra la dispersión de los datos (normalizadas: $x = \frac{x-\overline{x}}{\sigma}$) en formato de caja. Hay variables que no tienen variabilidad (no se observa una caja [*] y se descartan. Por otro lado hay algunas variables que muestran cajas, lo que significa que sí tienen dispersión, pero al mirar con detalle el número de valores diferentes que contienen son menos de 4, estas variables típicamente son números enteros y provienen de contar el número de determinados grupos funcionales en una molécula. Para DatOx se encontró que hay 101 que contienen 4 o menos diferentes valores. Mientras que para el conjunto DatRed se descartan 101 variables por lo mismo motivo que en el caso de DatOx. Quedando dos nuevos conjuntos de datos de 107 variables de entrada para cada uno.

Por otro lado, las gráfica de la Figura. 5.2b y 5.3b (DatOx y DatRed respectivamente) presenta la dispersión de las variables que se conservan. Dando un total de 107 variables conservadas para ambos conjuntos de datos. Este tipo de gráficas puede dar más información sobre la composición

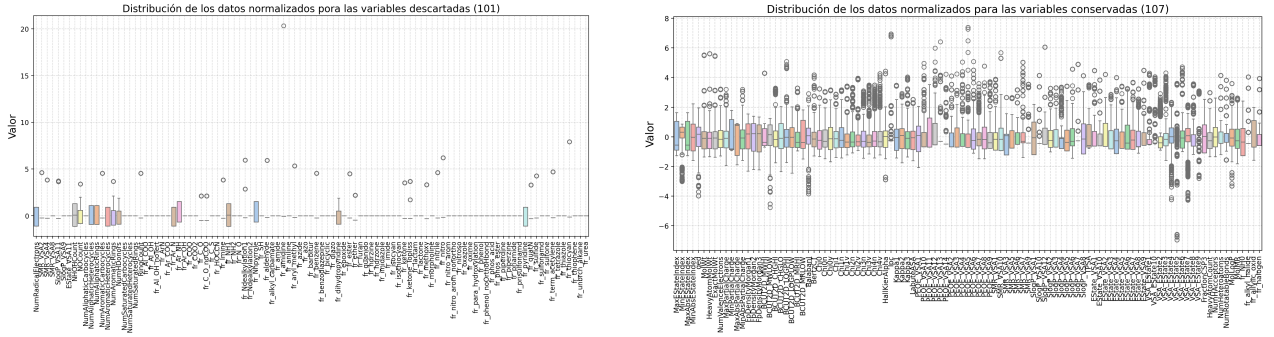
[*]Estas gráficas de cajas o de bigotes muestran la distribución de los datos, las çajasrepresentan los límites del cuartil superior e inferior, mientras que las lineas representa el extremo superior e inferior y los puntos son datos que solo se repiten una vez

del conjunto de datos porque los puntos son valores atípicos (o *outliers* en inglés), es decir, un valor único.



(a) Variables descartadas. **(b)** Variables conservadas.

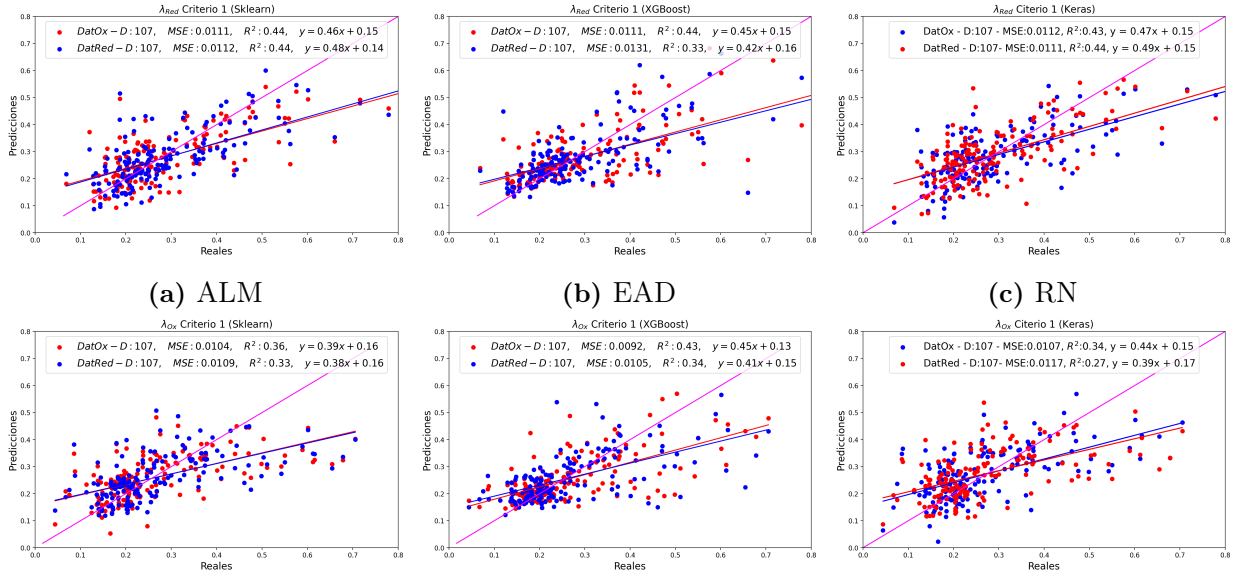
Figura 5.2: Gráficas de cajas para las variables normalizadas de la familia de la Bpy, provenientes del DatOx. En la que describen la varianza de los datos. La Figura (a) contiene a las variables descartadas por su baja variabilidad mientras que la Figura (b) muestra las variables conservadas por su mayor variabilidad.



(a) Variables descartadas. **(b)** Variables conservadas.

Figura 5.3: Gráficas de cajas para las variables normalizadas de la familia de la Bpy, provenientes del DatRed. La Figura (a) contiene a las variables descartadas por su baja variabilidad mientras que la Figura (b) muestra las variables conservadas por su mayor variabilidad.

Descartando dichas variables y entrenando los tres modelos se obtienen los siguientes resultados en la Figura 5.4, para la energía de reorganización λ_{Red} y λ_{Ox} respectivamente. En ambas aproximaciones se logra percibir la dispersión de los datos y que tampoco se logra reproducir los valores que se obtienen con mecánica cuántica. La tendencia de los datos no es la esperada (línea rosa -).



(a) ALM **(b)** EAD **(c)** RN

(d) ALM **(e)** EAD **(f)** RN

Figura 5.4: Gráficas de dispersión para los tres modelos que predicen la energía de reorganización λ_{Red} (primer fila) y λ_{Ox} (segunda fila), considerando el Criterio 1. Se entrenan los modelos con ambos conjuntos de datos.

A partir de los resultados presentados en la Tabla 5.2 se puede observar que la RN tiene mejores parámetros sobre los otros métodos al entrenar el modelo con DatRed para λ_{Red}. El modelo de EAD y DatOx tiene un rendimiento muy similar. Por otro lado, λ_{Ox} parece tener parámetros mejores si se utiliza el modelo EAD con DatOx, este resultado coincide con el criterio 0.

Al comparar los resultados de este criterio con los de referencia tanto para la energía λ_{Red} como λ_{Ox} (Figura 5.5), no se observa una mejora sustancial y sistemática. Por otro lado, la variación es muy pequeña. Esto apoya la idea de que las variables que se descartaron no aportan información relevante a los modelos.

5.2.1. Criterio 2: Análisis de correlación entre variables de entrada

El propósito de esta sección es identificar las variables equivalentes entre sí. El estudio parte del conjunto de variables resultantes del criterio 1, compuesto por 107 variables para DatOx y DatRed. Para este análisis, se emplean tres coeficientes de correlación: CCP, CCK y CCS. Se considera que una correlación mayor 0.90 indica que las variables son equivalentes .

En las Figura 5.6 se presentan las matrices de correlación, usando CCP, donde los renglones corresponden a las variables descartadas y las columnas a las variables conservadas para el

Energía	Métrica	RLM		EAD		RN	
		Ox	Red	Ox	Red	Ox	Red
λ_{Red}	MSE(↓)	**0.0111**	0.0112	**0.0111**	0.0131	0.0112	**0.0111**
	R^2(↑)	**0.44**	**0.44**	**0.44**	0.33	0.43	**0.44**
	m(↑)	0.46	0.48	0.45	0.42	0.47	**0.49**
	b(↓)	0.15	**0.14**	0.15	0.16	0.15	0.15
λ_{Ox}	MSE(↓)	0.0104	0.0104	**0.0092**	0.0105	0.0107	0.0117
	R^2(↑)	0.36	0.33	**0.43**	0.34	0.34	0.27
	m(↑)	0.39	0.38	**0.45**	0.41	0.44	0.39
	b(↓)	0.16	0.16	**0.13**	0.15	0.15	0.17

Tabla 5.2: Métricas de rendimiento de los modelos de ML para el criterio 1, con λ_{Red} y λ_{Ox} como variable objetivo.

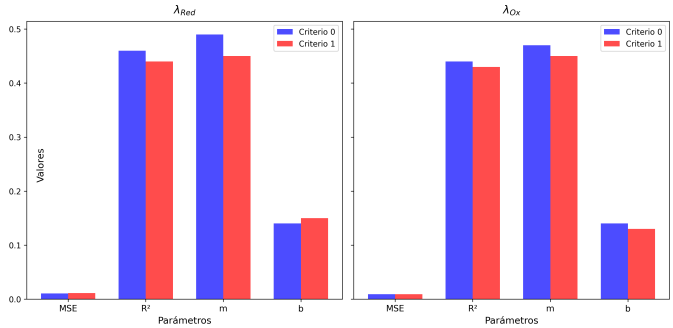
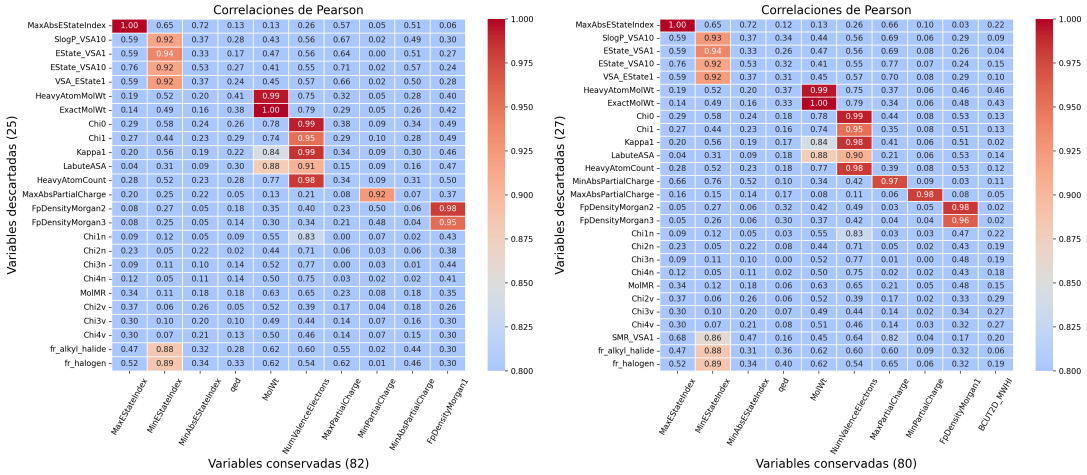


Figura 5.5: Comparación de las métricas de rendimientos entre el criterio 0 y el 1. Las métricas son los valores en azul de las Tablas 5.1 (criterio 0) y 5.2 (criterio 1).

análisis. En particular la Figura 5.6a contiene las correlaciones CCP. Por ejemplo, la variable "`FpDensityMorgan1`" (última columna) tiene una correlación de Corr(X,Y)=0.98, 0.95 con las variables "`FpDensityMorgan2`" y "`FpDensityMorgan3`". Estas tres variables son equivalentes, de acuerdo con el umbral propuesto, por lo que es suficiente con conservar una. Por otro lado, la variable "`qed`" no tiene equivalentes, ya que todas sus correlaciones con el resto de variables son menores a 0.9, por lo que se conserva.

De manera análoga al caso de la matriz de correación del CCP, en la Figura 5.7 se presentan las correlaciones asociadas al CCK y al CCS. El criterio para la deserción de variables es el mismo, es decir, para al valor absoluto de las correlaciones mayores a 0.9 se consideran variables equivalentes. Mientras que para el análisis con CCK se detectaron 10 variables equivalentes pada DatOx y 12 para DatRed, con el CCS hay 23 variables para DatOx y 25 para DatRed. El nuevo conjunto resultante del CCK es de 97 y 95 para DatOx y DatRed respectivamente. Por otro lado, el CCS da un nuevo conjunto de 84 y 82 variables para DatOx y DatRed respectivamente.



(a) DatOx **(b)** DatRed

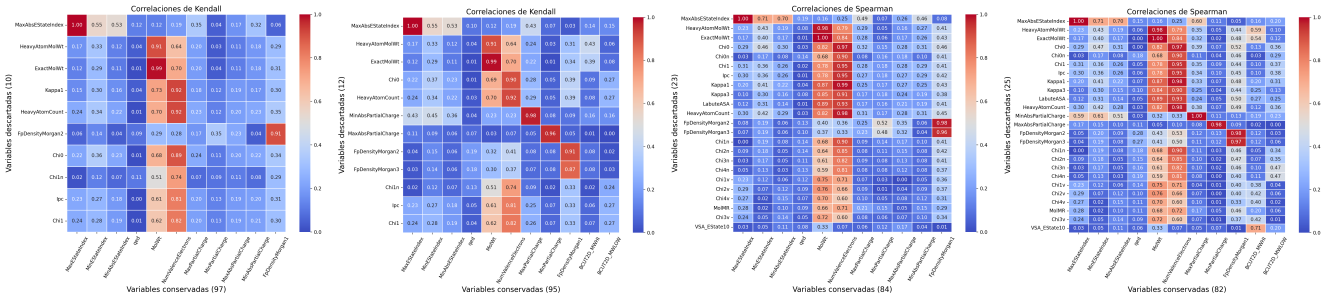
Figura 5.6: Matrices de correlación de Pearson para ambos conjuntos de datos. El código de colores indica que entre mas intenso sea el color rojo hay un correlación alta entre ambas variables. Mientras que la intensidad del color tienda hacia el azul hay una baja correlación.

De acuerdo al análisis de los coeficientes de correlación; Pearson, Kendall y Spearman entre pares de variables, los resultados de los entrenamientos se presentan en la Figura. 5.8 y 5.9 para la energía de reorganización λ_{Red} y λ_{Ox} respectivamente. De manera visual no se aprecian diferencias significativas entre las correlaciones, base de datos o modelos.

La Tabla 5.3 contiene los parámetros de rendimiento de los datos de prueba para los distintos modelos. Los valores en **negritas** indican los mejores parámetros fijando la correlación y variando el modelo y la base de datos, mientras que el mejor modelo es resaltado de color azul. El mejor modelo se identifica por la mayor cantidad de parámetros destacados. Además para cada energía de reorganización los valores subrayados corresponden al mejor modelo entre las diferentes correlaciones. Para el caso del entrenamiento de λ_{Red}, EAD es el mejor modelo para todas las correlaciones, con DatOx como la base de datos que arroja los mejores resultados. Además, los parámetros de EAD para CCP y CCK son muy similares, por lo que se puede seleccionar cualquiera de ellas indistintamente. Para λ_{Ox} se deduce que el mejor modelo es EAD para todas las correlaciones. La base de datos DatOx sigue dando los mejores rendimientos. En este caso, CCK tiene los mejores parámetros de rendimiento.

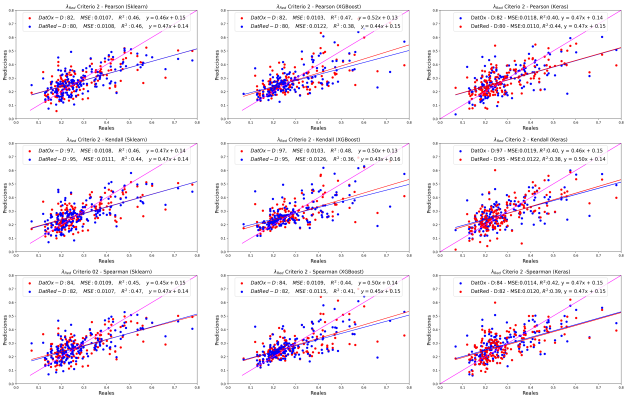
Por lo tanto, con base en la información presentada, se puede concluir que los EAD y la base de datos DatOx ofrecen los mejores resultados para ambas variables objetivo. Al comparar las diferentes correlaciones, no se observan diferencias significativas en los resultados, ya que estos son muy similares: los valores de R^2, m y b varían únicamente en las centésimas, mientras que el MSE muestra diferencias en el orden de las milésimas para ambas energías de reorganización.

Hasta este punto no se ha alcanzado la precisión de los cálculos cuánticos. En la Figura 5.10 se muestran las diferencias entre el criterio 1 y 2. Para λ_{Red} se usa el CCP, mientras que para λ_{Ox} es el CCK ya que son los que mejores rendimientos dieron, ambos con una base DatOx. Por otro lado, aunque la diferencia entre estos coeficientes de correlación no es significativo esto puede interpretarse como que hay variables equivalentes, en el sentido que aportan la misma información.



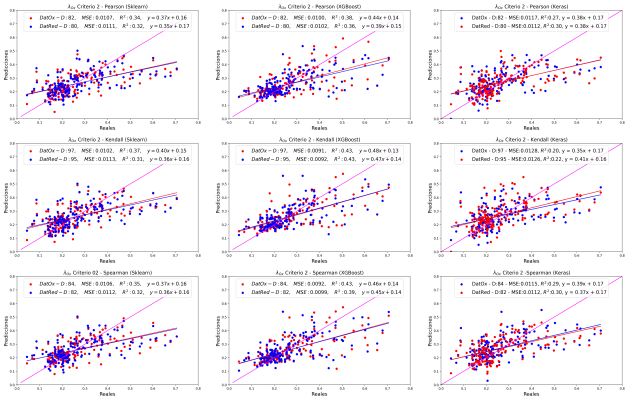
(a) DatOx -Descartadas **(b)** DatOx-Conservadas **(c)** DatRed -Descartadas **(d)** DatRed-Conservadas

Figura 5.7: Matrices de correlación de Kendall en primer renglón y Spearman en el segundo para distintos métodos y base de datos. El código de colores indica que entre mas intenso sea el color rojo hay un correlación alta entre ambas variables. Mientras que la intensidad del color tienda hacia el azul hay una baja correlación



(a) RLM **(b)** EAD **(c)** RN

Figura 5.8: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} considerando el Criterio 2 y las distintas formulaciones de correlación. Pearson (primer renglón), Kendall (segundo renglón) y Spearman (tercer renglón).



(a) RLM **(b)** EAD **(c)** RN

Figura 5.9: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Ox} considerando el Criterio 2 y las distintas formulaciones de correlación. Pearson (primer renglón), Kendall (segundo renglón) y Spearman (tercer renglón).

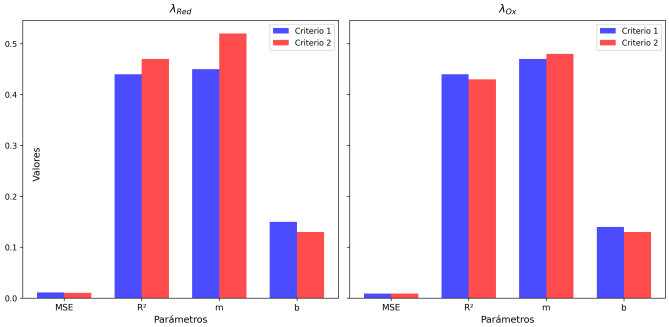


Figura 5.10: Comparación de los parámetros de rendimiento entre el criterio 1 y el criterio 2. Los datos corresponden a los mejores modelos presentados en las Tablas: Criterio 1: 5.2; Criterio 2: 5.3.

Energía	Corr	parámetros	RLM		EAD		RN	
			DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
λ_{Red}	**Pearson**	MSE(↓)	0.0107	0.0108	**0.0103**	0.0122	0.0118	0.0110
		R^2(↑)	0.46	0.46	**0.47**	0.38	0.40	0.44
		m(↑)	0.46	0.47	**0.52**	0.44	0.47	0.47
		b(↓)	0.15	0.14	**0.13**	0.15	0.14	0.15
	Kendall	MSE(↓)	0.0108	0.0111	**0.0103**	0.0126	0.0119	0.0122
		R^2(↑)	0.46	0.44	**0.48**	0.36	0.40	0.38
		m(↑)	0.47	0.47	**0.50**	0.43	0.46	**0.50**
		b(↓)	0.14	0.14	**0.13**	0.16	0.15	0.14
	Spearman	MSE(↓)	**0.0109**	0.0107	**0.0109**	0.0115	0.0114	0.0120
		R^2(↑)	0.45	**0.47**	0.44	0.41	0.42	0.39
		m(↑)	0.45	0.47	**0.50**	0.45	0.47	0.47
		b(↓)	0.15	**0.14**	**0.14**	0.15	0.15	0.15
λ_{Ox}	**Pearson**	MSE(↓)	0.0107	0.0111	**0.0100**	0.0102	0.0117	0.0112
		R^2(↑)	0.34	0.32	**0.38**	0.36	0.27	0.30
		m(↑)	0.37	0.35	**0.44**	0.39	0.38	0.38
		b(↓)	0.16	0.17	**0.14**	0.15	0.17	0.17
	Kendall	MSE(↓)	0.0102	0.0113	**0.0091**	0.0092	0.0128	0.0126
		R^2(↑)	0.37	0.31	**0.43**	**0.43**	0.20	0.22
		m(↑)	0.40	0.36	**0.48**	0.47	0.35	0.41
		b(↓)	0.15	0.16	**0.13**	0.14	0.17	0.16
	Spearman	MSE(↓)	0.0106	0.0112	**0.0092**	0.0099	0.0115	0.0112
		R^2(↑)	0.35	0.32	**0.43**	0.39	0.29	0.30
		m(↑)	0.37	0.36	**0.46**	0.45	0.39	0.37
		b(↓)	0.16	0.16	**0.14**	0.14	0.17	0.17

Tabla 5.3: Parámetros de rendimiento de los modelos de ML para el criterio 2, con λ_{Red} y λ_{Red} como variable objetivo y ambos conjunto de datos así como los tres coeficientes de correlacion(CCP, CCK, CCS).

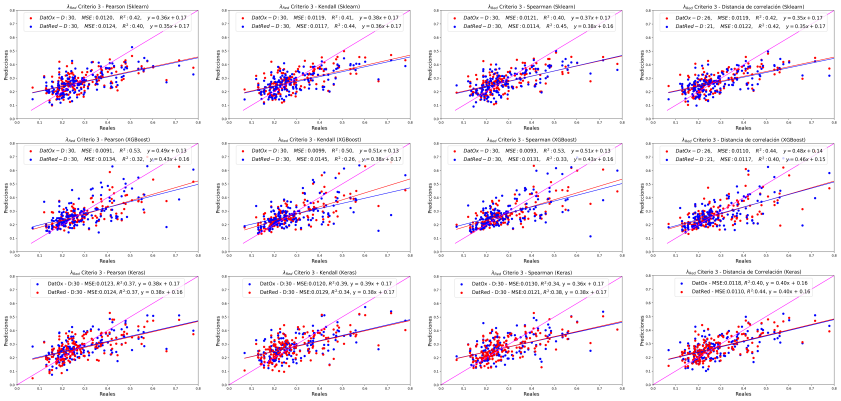
5.3. Criterio 3: Correlación entre variables de entrada vs variable objetivo

En este análisis se trabaja, por primera vez, con la variable objetivo (λ_{Ox} o λ_{Red}). El propósito de esta sección es identificar las variables más representativas dentro del conjunto de datos que conduzcan a un mejor entrenamiento de los modelos. Se emplean las medidas de correlación CCP, CCK y CCS. Se incorpora la medida de correlación no lineal; DC. Esto con el objetivo de no omitir variables que puedan tener un impacto importante en el entrenamiento de los modelos.

Con base en los resultados obtenidos del conjunto de variables obtenidas del criterio 2, se propone seleccionar las 30 variables con la correlación más alta. Para el conjunto de datos se le aplica la misma correlación utilizada en el criterio 2.

Por otro lado, la Distancia de Correlación tomara como partida el conjunto de variables resultantes del criterio 1. Esta medida detecta linealidad y no linealidad. La detección de la relación entre la variable de entrada con la variable objetivo depende de la magnitud de dicha medida, es decir, valores altos corresponden a una dependencia lineal, valores menores a 1 pueden asociarse a la no linealidad y valores cercanos a 0 se interpreta que no hay correlación. Por esta razón se propone que se van a seleccionar las variables que tengan una correlación mayor a 0.1. Por esta razón no se utiliza en el criterio 2 porque no hay manera de asociar un umbral que indique una correlación más fuerte.

El entrenamiento de los tres modelos con las distintas correlaciones se presentan en la Figura 5.11 para λ_{Red}. La Distancia de Correlación detecta que para DatOx hay 27 variables con correlación, mientras que para DatRed detecta 22 variables.

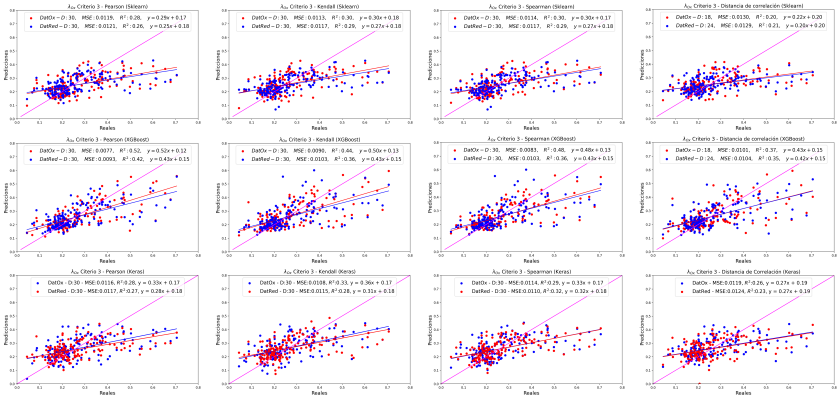


(a) CCP **(b)** Kendall **(c)** Spearman **(d)** DC

Figura 5.11: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} considerando el Criterio 3 y las distintas formulaciones de correlación. Primer renglón contiene al modelo RLM, segundo a EAD y tercer renglón a RN.

Para λ_{Ox} la Figura 5.12 contiene las gráficas de dispersión de los resultados de la validación de los distintos modelos, para todas las correlaciones propuestas en el trabajo. Para la Distancia de

Correlación detecta que si se usa DatOx solo hay 19 variables con correlación no lineal, mientras que para DatRed detecta 25 considerando solo λ_{Ox}.



(a) CCP **(b)** Kendall **(c)** Spearman **(d)** DC

Figura 5.12: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Ox} considerando el Criterio 3 y las distintas formulaciones de correlación. Primer renglón contiene al modelo RLM, segundo a EAD y tercer renglón a RN.

La Tabla 5.4 contiene las métricas de rendimiento de los modelos utilizando todas las medidas de correlación Corr(λ_{Red}, $\mathbf{X}$). Observando los valores se puede detectar que en todas las validaciones el modelo que destaca es el EAD en todos las métricas y en todas las formulaciones de validación. En este criterio los modelos difieren en el orden de décimas para R^2, m y b y MSE en el orden de las diezmilésimas. Es decir, hay una diferencia más notable en este criterio.

Por otro lado, si se comparan las formulaciones, CCP (5.4) contiene la mayoría de los mejores métricas (3 de 4) mientras que las otras correlaciones sólo contienen 2 (hay varios valores que resaltan en azul porque son los mismos en las diferentes correlaciones).

La Tabla 5.5 contiene las métricas de rendimiento de los modelos utilizando todas las medidas de correlación Corr(λ_{Ox}, $\mathbf{X}$). En el caso de λ_{Red} el EAD sigue teniendo el dominio de ser el mejor modelo entre los 3 para cualquier correlación. Ahora si la comparación es entre formulaciones de correlación, destaca CCP porque las 4 métricas (valores en azul) son los mejores, aunque en algunos casos se repiten en la correlación de Spearman. DatOx destaca como la base más util para entrenar los modelos. Si se desease usar DC la mejor base es DatRed y EAD como todos los casos.

En resumen se concluye que el modelo EAD siempre ofrece mejores resultados tanto para λ_{Red} como λ_{Ox} y en particular CCP tiene un mejor rendimiento sobre el resto de coeficientes de correlación. DatOx sigue siendo la base que da mejores resultados en la mayoría de los casos tanto si se decidiese trabajar con alguna correlación o alguna de las dos variables objetivo. Aunque varía muy poco entre bases de datos parece que si se quiere trabajar con λ_{Ox} y DC es mejor usar DatRed.

En la Figura 5.13 se hace una comparación entre el criterio 3 y el pasado. Es notable que hay una mejora en todos las métricas si se aplica el criterio 3.

Corr	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
CCP	MSE	0.0120	0.0124	**0.0091**	0.0134	0.0123	0.0124
	R^2	0.42	0.40	**0.53**	0.32	0.37	0.37
	m	0.36	0.35	**0.49**	0.43	0.38	0.38
	b	0.17	0.17	**0.13**	0.16	0.17	0.16
Kendall	MSE	0.0119	0.0117	**0.0099**	0.0145	0.0120	0.0129
	R^2	0.41	0.44	**0.50**	0.26	0.39	0.34
	m	0.38	0.36	**0.51**	0.38	0.39	0.38
	b	0.17	0.17	**0.13**	0.17	0.17	0.17
Spearman	MSE	0.0121	0.0114	**0.0093**	0.0131	0.0130	0.0121
	R^2	0.40	0.45	**0.53**	0.33	0.34	0.38
	m	0.37	0.38	**0.51**	0.43	0.36	0.38
	b	0.17	0.16	**0.13**	0.16	0.17	0.17
DC	MSE	0.0119	0.0122	**0.0110**	0.0117	0.0118	0.0110
	R^2	0.42	0.42	**0.44**	0.40	0.40	**0.44**
	m	0.35	0.35	**0.48**	0.46	0.40	0.40
	b	0.17	0.17	**0.14**	0.15	0.16	0.16

Tabla 5.4: Resultado de las métricas de rendimiento. Variable objetivo: λ_{Red}, con 30 variables de entrada. En el caso de DC son 27 para DatOx y 22 para DatRed.

5.4. Criterio 4: Análisis de componentes principales

Para este criterio se utilizan los resultados del criterio 3. Para cada conjunto de variables, se calculó el valor de la prueba KMO, cuyos resultados se presentan en la Tabla 5.6. La prueba KMO determina si los datos son o no adecuados para realizar un análisis de componentes principales, mientras más cercano esté a la unidad indica que los datos son adecuados. Los resultados indican que todos los conjuntos son factibles, pero Pearson destaca sobre el resto de maneras de evaluar la correlación. Por otro lado, DC contiene el mejor parámetro de KMO (con un valor de 0.68 con la base DatRed) si se intenta estimar λ_{Ox}, mientras que Pearson y DC tienen la misma magnitud de KMO para predecir λ_{Ox} con la base DatRed.

Energía de reorganización λ_{Red}:

A continuación se detalla el análisis de componentes principales. En el caso del conjunto de datos λ_{Red}-DatRed-Pearson, se aborda el problema de los valores propios. En la Tabla 5.7 se presentan los eigenvalores (μ) y el porcentaje de varianza acumulada ($\%\sum\sigma^2$) en función del número de componentes principales (CP). El orden de los eigenvalores está asociado con la cantidad de variabilidad que cada CP representa, de modo que el eigenvalor más alto corresponde a la mayor variabilidad.

En la Figura 5.14 se muestra un método gráfico conocido como la gráfica de varianza acumulada, que permite identificar visualmente el número mínimo de componentes principales necesarias para explicar al menos el 75 % de la varianza de los datos, manteniento una mínima pérdida de

Corr	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
CCP	MSE	0.0119	0.0121	**0.0077**	0.0093	0.0116	0.0117
	R^2	0.28	0.26	**0.52**	0.42	0.28	0.27
	m	0.29	0.25	**0.52**	**0.43**	0.33	0.28
	b	0.17	0.18	**0.12**	0.15	0.17	0.18
Kendall	MSE	0.0113	0.0117	**0.0090**	0.0103	0.0109	0.0115
	R^2	0.30	0.29	**0.48**	0.36	0.33	0.28
	m	0.30	0.27	0.48	0.43	0.36	0.31
	b	0.18	0.18	**0.13**	0.15	0.17	0.18
Spearman	MSE	0.0114	0.0117	**0.0083**	0.0103	0.0114	0.0110
	R^2	0.30	0.29	**0.48**	0.36	0.29	0.32
	m	0.30	0.27	**0.48**	0.43	0.33	0.32
	b	0.17	0.18	**0.13**	0.15	0.17	0.18
DC	MSE	0.0130	0.0129	**0.0101**	0.0104	0.0119	0.0124
	R^2	0.20	0.21	**0.37**	0.35	0.26	0.23
	m	0.22	0.20	**0.43**	0.42	0.27	0.27
	b	0.20	0.20	**0.15**	0.15	0.19	0.19

Tabla 5.5: Resultado de las métricas de rendimiento. Variable objetivo: λ_{Ox}, con 30 variables de entrada. En el caso de DC son 19 para DatOx y 25 para DatRed.

información. En este caso, se observa que deben ser al menos 7 componentes principales.

Si se seleccionan 7 componentes principales (CP), se alcanza un 77.25 % de la varianza acumulada, lo cual supera el umbral establecido. Estas 7 nuevas variables, que corresponden a las proyecciones de las variables originales en las direcciones definidas por los eigenvectores, podrían emplearse para entrenar diversos modelos de aprendizaje automático. Por otro lado, en este trabajo se opta por analizar los pesos (elementos de los eigenvectores), ya que estos reflejan la importancia de cada variable original en el conjunto de datos, de acuerdo al número de CPs. Este enfoque permite identificar el o los descriptores quimioinformáticos con mayor importancia para el fenómeno que se está estudiando. Por ejemplo, si la variable "`NumRotableBonds`" tuviese el peso más alto (tomando sólo la componente principal 1), ésta sería la variable más representativa de ese CP. Si se toman 7 componentes principales y se suma el peso de la variable ($\mathbf{X}_i$) de cada CP anterior, se obtendrá el peso total de dicha variable en este conjunto de CPs.

En la Tabla 5.8 se presentan las variables ordenadas según su peso al considerar 7 CP. Una propuesta para decidir el número de variables a conservar es retener aquellas que representen el 75 % de la información y descartar el resto. Es decir, valores que están por debajo del valor de 0.8983 (reportado al final de la Tabla 5.8) se descartan, resultando un total de 22 variables.

Energía de reorganización λ_{Ox}:

El método gráfico de la Figura 5.15 muestra que basta con contemplar 8 CPs porque cumplen con la mínina varianza acumulada para el conjunto de variables que provienen del criterio 3 λ_{Ox}-DatOx-Pearson. La Tabla 5.9 contienen en orden los eigenvalores y la varianza acumulada res-

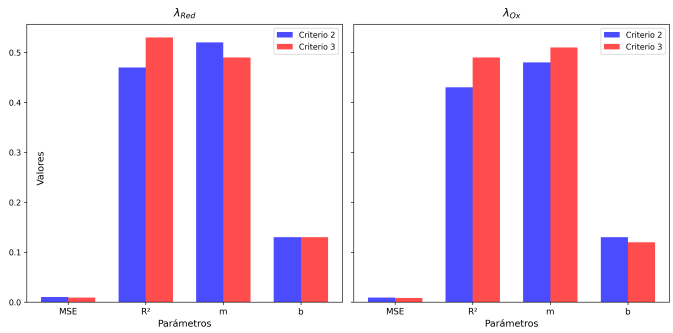


Figura 5.13: Métricas de rendimiento de los mejores modelos del criterio 2 y 3 para ambas energía de reorganización. Los datos corresponden a los mejores modelos presentados en las Tablas: Criterio 2: 5.3; Criterio 3: 5.4 y 5.5.

Energía	Base de Datos	Pearson	Kendall	Spearman	DC
λ_{Ox}	DatRed	0.69	0.65	0.65	0.69
λ_{Red}	DatRed	0.64	0.62	0.65	0.68
λ_{Ox}	DatOx	0.62	0.57	0.58	0.60
λ_{Red}	DatOx	0.65	0.56	0.57	0.64

Tabla 5.6: Valores de la prueba KMO de las 30 variables más correlacionadas con las variables objetivo (λ_{Red} o λ_{Ox}) resultado del Criterio 3.

pectivamente segun la CP para la variable objetivo . Considerar 8 CPs tiene una varianza acumulada de 77.06 %.

CP	μ	% σ^2	% $\sum\sigma^2$	CP	μ	% σ^2	% $\sum\sigma^2$
1	7.843	26.14	26.14	16	0.26	0.85	95.76
2	4.393	14.64	40.79	17	0.22	0.75	96.51
3	3.211	10.70	51.49	18	0.19	0.63	97.14
4	2.43	8.10	59.59	19	0.16	0.53	97.66
5	2.034	6.78	66.37	20	0.146	0.49	98.15
6	1.824	6.08	72.45	21	0.113	0.38	98.53
7	1.438	4.80	77.25	22	0.103	0.35	98.87
8	1.132	3.77	81.02	23	0.098	0.33	99.20
9	0.988	3.29	84.31	24	0.075	0.25	99.45
10	0.784	2.61	86.93	25	0.062	0.21	99.66
11	0.694	2.31	89.24	26	0.043	0.14	99.80
12	0.515	1.72	90.96	27	0.026	0.09	99.89
13	0.477	1.59	92.55	28	0.02	0.07	99.95
14	0.377	1.26	93.80	29	0.011	0.04	99.99
15	0.331	1.10	94.91	30	0.003	0.01	100.00

Tabla 5.7: Eigenvalores (μ) y el porcentaje de varianza (σ^2) para cada componente principal (CP), y el porcentaje de varianza acumulada ($\sum\sigma^2$). Base de datos de las especies reducidas (DatRed).

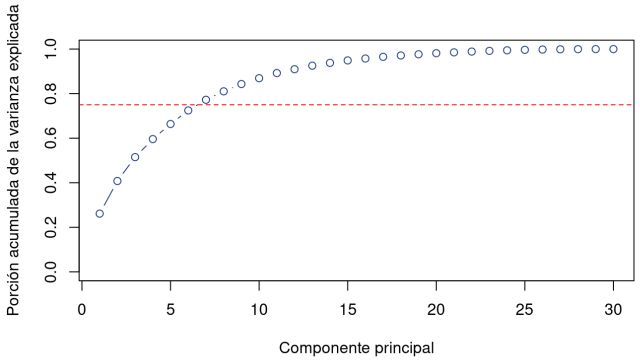


Figura 5.14: Gráfica de varianza acumulada con respecto el número de componentes principales para la base de datos de especies reducidas. La línea punteada roja representa el umbral mínimo aceptable del 75 % (DatRed).

Para elegir el número de variables representativas, se va a utilizar el mismo criterio de los pesos acumulados, en este caso se presentan los resultados en la Tabla 5.10. Dando un nuevo conjunto de 22 variables más representativas de los datos.

Orden	Variable	$\sum\omega$(CP)	Orden	Variable	$\sum\omega(CP)$
1	SlogP_VSA4	1.4963	16	EState_VSA8	0.9928
2	fr_allylic_oxid	1.4659	17	SlogP_VSA1	0.9868
3	VSA_EState6	1.3267	18	PEOE_VSA9	0.9637
4	HallKierAlpha	1.3250	19	SMR_VSA3	0.9636
5	SMR_VSA9	1.2716	20	EState_VSA3	0.9518
6	BCUT2D_CHGHI	1.2299	21	NumRotatableBonds	0.9219
7	PEOE_VSA1	1.1767	**22**	SlogP_VSA8	**0.9192**
8	MolWt	1.1493	23	Chi0n	0.8914
9	BCUT2D_MRLOW	1.1349	24	BCUT2D_CHGLO	0.8895
10	BalabanJ	1.1095	25	SMR_VSA6	0.8801
11	qed	1.0554	26	BCUT2D_LOGPHI	0.8793
12	BCUT2D_LOGPLOW	1.0313	27	SlogP_VSA2	0.7048
13	Chi0v	1.0286	28	FractionCSP3	0.6759
14	NumValenceElectrons	1.0241	29	BCUT2D_MWLOW	0.6659
15	VSA_EState9	0.9972	30	EState_VSA5	0.5269

Descripción estadística de los pesos

Min.	1° Cuarto	Mediana	Media	3° Cuarto	Max.
0.5269	0.8983	0.9950	1.0212	1.1457	1.4963

Tabla 5.8: Orden de prioridad, de acuerdo a sus pesos, de las variables al tomar 7 CP con una varianza acumulada del 77.25 % (DatRed).

Para los distintos conjuntos de datos, tomando en cuenta solo aquellos obtenidos de la correlación de Pearson porque son los que tienen la prueba KMO más alta, se presentan los resultados del anális en la Tabla 5.11. Las variables resultantes proveniente de seleccionar las componentes principales (CP) que explican más del 80 % de la variabilidad de los datos. Las variables resaltadas en color azul indican aquellas que coinciden en ambos análisis. En los cuatro conjuntos de datos coinciden 2 variables (subrayadas). Mientras en los conjuntos que contienen la prueba KMO más alta tanto para λ_{Red} como λ_{Ox} son 11 variables (en azul).

Las gráficas de dispersión correspondientes al conjunto de variables obtenidas de la Tabla 5.11 para los tres modelos se presentan en la Figura 5.16, tanto para la energía de reorganización λ_{Red} como λ_{Ox}.

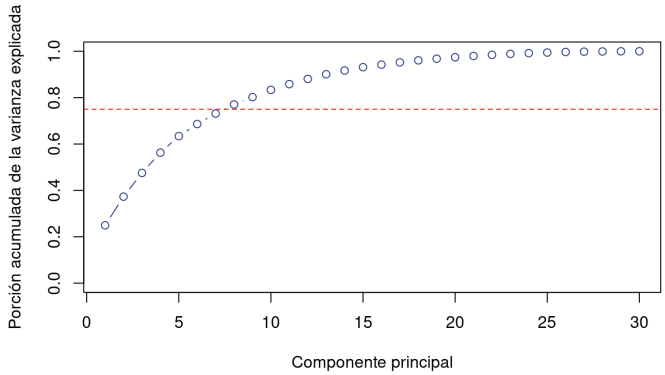


Figura 5.15: Gráfica de varianza acumulada con respecto el número de componentes principales para la base de datos de especies oxidadas. La línea punteada roja representa el umbral mínimo aceptable del 75 %.

CP	μ	% σ^2	% $\sum\sigma^2$	CP	μ	% σ^2	% $\sum\sigma^2$
1	7.481	24.94	24.94	16	0.338	33.80	94.26
2	3.712	12.37	37.31	17	0.288	28.80	95.22
3	3.068	10.23	47.53	18	0.261	26.10	96.09
4	2.623	8.75	56.28	19	0.212	21.20	96.80
5	2.132	7.11	63.39	20	0.189	18.90	97.43
6	1.567	5.22	68.61	21	0.164	16.40	97.97
7	1.364	4.55	73.16	22	0.151	15.10	98.48
8	1.172	3.91	77.06	23	0.119	11.90	98.88
9	0.952	3.17	80.24	24	0.097	9.70	99.20
10	0.932	3.11	83.34	25	0.081	8.10	99.47
11	0.744	2.48	85.82	26	0.071	7.10	99.70
12	0.679	2.27	88.09	27	0.038	3.80	99.83
13	0.598	1.99	90.08	28	0.025	2.50	99.92
14	0.487	1.62	91.71	29	0.02	2.00	99.98
15	0.428	1.43	93.13	30	0.005	0.50	100.00

Tabla 5.9: Eigenvalores (μ) y el porcentaje de varianza (σ^2) para cada componente principal (CP), y el porcentaje de varianza acumulada ($\sum\sigma^2$). Base de datos de las especies oxidadas. Resultado de la Corr(λ_{Ox}, $\mathbf{X}_{\text{DatOx}}$).

En la Tabla 5.12 se resumen los métricas de rendimiento para la energía de reorganización λ_{Red}. Los EAD con la base DatOx son el conjunto que dan los mejores resultados.

Para el caso de la energía de reorganización λ_{Ox} las métricas se muestran en la Tabla 5.13 y los resultados mejores siguen siendo consistentes con lo que se ha observado en estos 4 criterios, es decir, los EAD y la base DatOx conducen al modelo con mejor rendimiento.

En resumen, este criterio reduce la dimensionalidad de la base de datos. La prueba KMO indica que la correlación de Pearson identifica conjuntos de datos más adecuados para PCA. Además, KMO detecta que DatOx es una base datos más adecuada para PCA cuando el conjunto proviene de las variables más correlacionadas con λ_{Red} (de acuerdo al criterio 3), mientras que DatRed es más apropiada cuando el conjunto proviene de las variables más correlacionadas con λ_{Ox}, según el criterio 3. Al entrenar los distintos modelos se encuentra que los EAD destacan sobre el resto. Para λ_{Ox} ambas bases de datos dan resultados similares, pero para λ_{Red} que DatOx es mejor.

Orden	Variable	Peso acumulado	Orden	Variable	Peso acumulado
1	SMR_VSA3	1.5911	16	BCUT2D_LOGPLOW	1.1860
2	BCUT2D_MWLOW	1.5417	17	Chi0v	1.1603
3	MinPartialCharge	1.4925	18	MolWt	1.1455
4	BCUT2D_CHGHI	1.4884	19	VSA_EState9	1.1350
5	PEOE_VSA12	1.4761	20	SlogP_VSA1	1.1267
6	BalabanJ	1.4304	21	BCUT2D_CHGLO	0.9986
7	EState_VSA8	1.4257	**22**	PEOE_VSA11	**0.9905**
8	VSA_EState6	1.4028	23	PEOE_VSA9	0.9413
9	PEOE_VSA1	1.3940	24	NumRotatableBonds	0.9370
10	BCUT2D_LOGPHI	1.3774	25	SMR_VSA6	0.9337
11	HallKierAlpha	1.3659	26	PEOE_VSA8	0.9216
12	SMR_VSA9	1.3441	27	EState_VSA5	0.9166
13	SlogP_VSA8	1.3388	28	Chi0n	0.9166
14	PEOE_VSA13	1.2114	29	SlogP_VSA2	0.8851
15	BCUT2D_MRLOW	1.1884	30	FractionCSP3	0.6983

Descripción estadística de los pesos

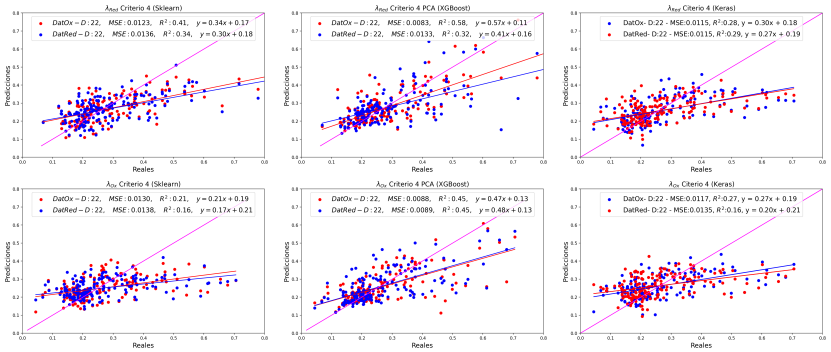
Min.	1° Cuarto	Mediana	Media	3° Cuarto	Max.
0.6983	0.9536	1.1872	1.1987	1.4006	1.5911

Tabla 5.10: Orden de prioridad, de acuerdo a sus pesos, de las variables al tomar 8 CP con una varianza acumulada del 77.06 %. Base de datos de las especies oxidadas.

La Figura 5.17 muestra cómo ha ido evolucionando el rendimiento del mejor modelo de cada criterio, conforme se reduce la dimensionalidad de la base de datos. Este comportamiento es más significativo para λ_{Red}, mientras que para λ_{Ox}, es decir, se observa una mayor mejora sobre las métricas que en λ_{Ox}.

Orden	λ_{Red}-DatOx KMO=0.65 NCP=10 $\%\sum\sigma^2 = 83.34$	λ_{Red}-DatRed KMO=0.64 NCP=10 $\%\sum\sigma^2 = 86.93$	λ_{Ox}-DatRed KMO=0.69 NCP=12 $\%\sum\sigma^2 = 83.40$	λ_{Ox}-DatOx KMO=0.62 NCP=10 $\%\sum\sigma^2 = 81.06$
1	EState_VSA5	EState_VSA5	BCUT2D_MWLOW	EState_VSA4
2	PEOE_VSA8	qed	PEOE_VSA8	EState_VSA5
3	SMR_VSA3	SlogP_VSA4	PEOE_VSA13	SMR_VSA3
4	PEOE_VSA12	SMR_VSA9	SMR_VSA9	TPSA
5	BCUT2D_MWLOW	fr_allylic_oxid	EState_VSA5	SMR_VSA7
6	VSA_EState6	BCUT2D_MRLOW	EState_VSA3	BalabanJ
7	PEOE_VSA13	VSA_EState6	fr_NH0	qed
8	MinPartialCharge	PEOE_VSA1	PEOE_VSA5	NumRotatableBonds
9	BalabanJ	BCUT2D_MWLOW	PEOE_VSA11	BCUT2D_CHGHI
10	PEOE_VSA1	HallKierAlpha	PEOE_VSA3	SlogP_VSA4
11	EState_VSA8	MolWt	BalabanJ	MolLogP
12	BCUT2D_LOGPHI	BalabanJ	qed	MinPartialCharge
13	HallKierAlpha	NumRotatableBonds	PEOE_VSA6	FpDensityMorgan1
14	BCUT2D_CHGHI	VSA_EState9	MolLogP	PEOE_VSA13
15	SMR_VSA9	SMR_VSA3	NumRotatableBonds	PEOE_VSA5
16	VSA_EState9	BCUT2D_CHGHI	PEOE_VSA12	BCUT2D_LOGPLOW
17	BCUT2D_LOGPLOW	EState_VSA3	MinPartialCharge	BCUT2D_CHGLO
18	SlogP_VSA1	BCUT2D_LOGPLOW	NumHAcceptors	PEOE_VSA6
19	PEOE_VSA11	NumValenceElectrons	fr_allylic_oxid	PEOE_VSA3
20	SlogP_VSA8	EState_VSA8	VSA_EState4	PEOE_VSA12
21	BCUT2D_MRLOW	SlogP_VSA1	SlogP_VSA8	EState_VSA3
22	MolWt	BCUT2D_CHGLO	BCUT2D_MRLOW	FractionCSP3

Tabla 5.11: Orden de variables, según su importancia de acuerdo al análisis de PCA .



(a) RLM **(b)** EAD **(c)** RN

Figura 5.16: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} (renglón 1) λ_{Ox} (renglón 2), considerando el Criterio 4.

	Métrica	RLM		EAD		RN	
		Ox	Red	Ox	Red	Ox	Red
λ_{Red}	MSE	0.0123	0.0136	**0.0083**	0.0133	0.0115	0.0115
	R^2	0.41	0.34	**0.58**	0.32	0.28	0.29
	m	0.34	0.30	**0.57**	0.41	0.30	0.27
	b	0.17	0.18	**0.11**	0.16	0.18	0.19

Tabla 5.12: Métricas de rendimiento de los modelos de ML para el criterio 4, con λ_{Red} como variable objetivo

	Métrica	RLM		EAD		RN	
		Ox	Red	Ox	Red	Ox	Red
λ_{Ox}	MSE	0.0130	0.0138	**0.0088**	0.0089	0.0117	0.0135
	R^2	0.21	0.16	**0.45**	0.45	0.27	0.16
	m	0.21	0.17	**0.47**	0.48	0.27	0.20
	b	0.19	0.21	**0.13**	0.13	0.19	0.21

Tabla 5.13: Métricas de rendimiento de los modelos de ML para el criterio 4, con λ_{Ox} como variable objetivo

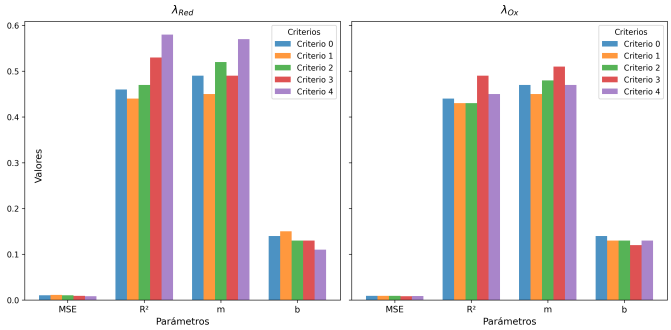


Figura 5.17: Parámetros de validación según el número de criterio y la energía de reorganización. Las métricas destacadas de cada criterio están marcadas en azul y para los criterios que consideran distintas correlaciones están subrayadas. Las tablas correspondientes son: Criterio 0: 5.1; Criterio 1: 5.2; Criterio 2: 5.3; Criterio 3: 5.4 y 5.5; Criterio 4: 5.13 y 5.12.

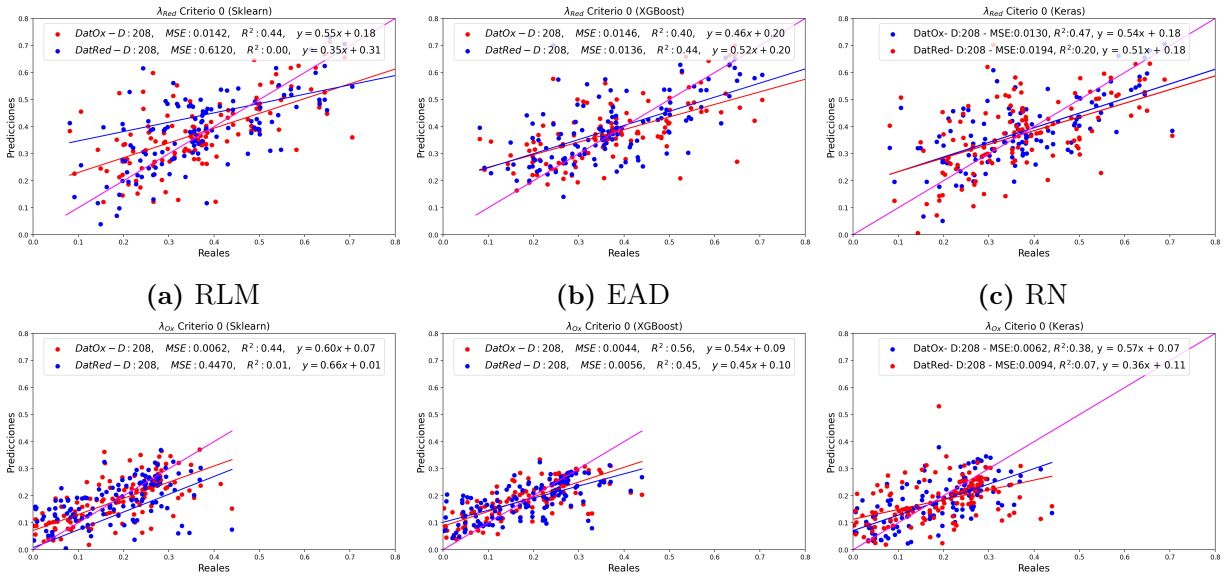
.

Capítulo 6

Resultados (Bencidina)

6.1. Criterio 0: Referencia

Para esta familia se tiene un total de 670 reacciones redox. Los resultados de los modelos de referencia para la familia de las Bencidinas se muestran en la Figura 6.1 para ambas energías de reorganización. El conjunto de valores de λ_{Ox} oscilan entre 0 y 0.5, mientras que λ_{Red} estan entre 0 y 0.8, esto puede interpretarse como que, en el contexto de esta familia, hay un menor costo energético en el sentido de la oxidación, es decir, es más favorable donar un electrón que aceptarlo.



(a) RLM **(b)** EAD **(c)** RN

(d) RLM **(e)** EAD **(f)** RN

Figura 6.1: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} (primer renglón) y λ_{Ox} (segundo renglón), considerando el Criterio 0.

En la Tabla 6.1 se presentan las métricas de rendimiento para las energías de reorganización. Se observa que el modelo de RLM y el conjunto de base de datos DatRed, presenta mucha dispersión

(R^2=0.00) y un error MSE considerablemente alto (0.6120) para predecir λ_{Red}. Además, la pendiente obtenida es de 0.55 y la ordenada al origen es de 0.18. Aunque los valores de m y b caen dentro de la tendencia observada hasta ahora, se le dará mayor peso a las métricas MSE y R^2 porque reflejan la capacidad predictiva del ajuste, que en este caso es muy limitada.

De la Tabla 6.1 se puede observar que el de RN en conjunto con la base de datos DatOx tienen un mayor rendimiento para λ_{Red}. Por otro lado si la intención es predecir λ_{Ox} da un mejor resultado si se trabaja con EAD y DatOx.

Energía	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
λ_{Red}	MSE	0.0142	0.6120	0.0146	0.0136	**0.0130**	0.0194
	R^2	0.44	0.00	0.40	0.44	**0.47**	0.20
	m	**0.55**	0.35	0.46	0.52	0.54	0.51
	b	**0.18**	0.31	0.20	0.20	**0.18**	**0.18**
λ_{Ox}	MSE	0.0062	0.4470	**0.0044**	0.0056	0.0062	0.0094
	R^2	0.44	0.001	**0.56**	0.45	0.38	0.07
	m	0.60	**0.66**	0.54	0.45	0.57	0.36
	b	0.07	**0.01**	0.09	0.10	0.07	0.11

Tabla 6.1: Métricas de rendimiento de los modelos de ML para el criterio 0, con λ_{Red} y λ_{Ox} como variable objetivo.

6.2. Criterio 1: Variabilidad de los datos

A partir de este punto, las gráficas de dispersión se adjuntan en el Anexo **??** para su consulta, pero se presentarán los resultados en las Tablas que se han estado trabajando.

Las Figuras **??**, **??** y **??** y **??** muestran las variables descartadas y conservadas para los conjunto de datos DatRed y DatOx correspondientemente. Para la primer base de datos se eliminan 102 variables dando un nuevo conjunto de 106 variables. Resulta el mismo número de variables si se aplica este criterio al conjunto DatOx.

A partir de estos nuevos conjuntos de datos se entrenan los modelos y en la Figura **??** se muestran los resultados de las pruebas de validación para todos los modelos, tanto para λ_{Red} como λ_{Ox} respectivamente.

En la Tabla 6.2 se resumen las métricas de rendimiento de los distintos modelos y energías de reorganización. Fijando λ_{Red} se puede concluir que el EAD destaca entre los modelos y que al usar DatOx se tienen mejores predicciones. Ahora si se quiere reproducir λ_{Ox} los EAD y combinación con DatOx conducen al mejor modelo. Es decir, en ambas energías es mejor usar DatOx como base de datos y un EAD como modelo.

En la Figura 6.2 muestra el rendimiento de los modelos al estimar la energía de reorganización en función de los criterios 1 y 0. Se observa que el rendimiento no es el esperado.

Energía	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
λ_{Red}	MSE	0.0141	0.0315	**0.0130**	0.0132	0.0150	0.0169
	R^2	0.44	0.22	**0.46**	**0.46**	0.38	0.31
	m	**0.54**	**0.55**	0.50	0.52	0.47	0.50
	b	0.18	**0.17**	0.20	0.19	0.19	0.19
λ_{Ox}	MSE	0.0060	0.0075	**0.0051**	0.0065	0.0056	0.0078
	R^2	0.46	0.33	**0.49**	0.36	0.44	0.22
	m	0.61	0.50	0.53	0.41	**0.59**	0.36
	b	0.07	0.10	0.09	0.11	**0.07**	0.11

Tabla 6.2: Métricas de rendimiento de los modelos de ML para el criterio 0, con λ_{Red} y λ_{Ox} como variable objetivo.

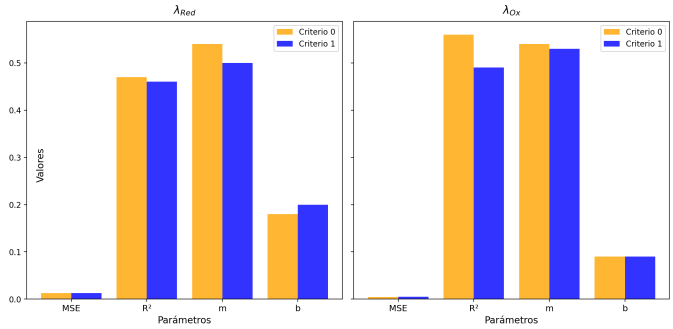


Figura 6.2: Evolución de las métricas de rendimiento en función de los criterios para ambas energía de reorganización. Los datos corresponden a los mejores modelos presentados en las Tablas: Criterio 0: 6.1; Criterio 1: 6.2.

6.3. Criterio 2: Análisis de correlación entre variables de entrada

Se abordan las cuatro formulaciones de correlación ya mencionadas. La Figura **??** muestra la magnitudes de la correlación de todas las correlaciones entre las variables que se van a descartar (renglones) y algunas variables que se conservan (columnas). Para el caso de Pearson y DatOx se descartan 35 variables y se conservan 71 mientras que para la DatRed se descartan 33 y se conservan 73. Con Kendall tanto para DatOx como DatRed se eliminan 10 y se conservan 95, siendo las mismas variables en ambos cojuntos de datos. Spearman también encuentra que en ambos conjuntos de datos hay 27 variables equivalentes y se conservan 79.

La Tabla 6.3 contiene las métricas obtenidas al hacer los ajustes usando el conjunto de prueba de todos los modelos y correlaciones para la variable objetivo λ_{Red}. Al comparar entre modelos y bases de datos (entre columnas) los valores en **negritas** indican las mejores métricas del modelo y base de datos. El mejor modelo resulta ser EAD y la base de datos DatRed. Entre correlaciones Pearson contiene 3 de 4 de las mejores métricas al compararlo con las otras formulaciones de correlación.

Por otro lado, si la variable objetivo es λ_{Ox} la Tabla 6.4 contiene los resultados de validación. Los EAD's siguen teniendo los mejores rendimientos, por otro lado ahora es DatOx quien predice mejor esta energía de reorganización, a la vez que Kendall destaca, sin mucha diferencia, sobre

Corr	Métrica	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0143	0.0183	0.0124	**0.0118**	0.0140	0.0171
	R^2	0.42	0.31	0.49	**0.52**	0.43	0.30
	m	0.46	0.45	0.49	**0.54**	0.49	0.48
	b	0.20	0.22	**0.19**	**0.19**	**0.19**	0.20
Kendall	MSE	0.0134	0.0214	0.0146	**0.0124**	0.0138	0.0171
	R^2	0.46	0.31	0.40	**0.49**	0.43	0.30
	m	0.51	0.54	0.44	**0.54**	0.49	0.48
	b	0.19	0.17	0.22	**0.18**	**0.18**	0.20
Spearman	MSE	0.0143	0.0170	0.0133	**0.0123**	0.0127	0.0176
	R^2	0.42	0.35	0.45	**0.50**	0.48	0.28
	m	0.49	0.48	0.46	**0.52**	0.50	0.46
	b	**0.19**	0.20	0.20	**0.19**	**0.19**	0.21

Tabla 6.3: Métricas de rendimiento de los modelos de ML para el criterio 2, con λ_{Red} como variable objetivo.

Pearson y Spearman.

Corr	Métrica	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0065	0.0082	**0.0050**	0.0064	0.0067	0.0090
	R^2	0.38	0.21	**0.50**	0.37	0.33	0.11
	m	0.46	0.26	**0.54**	0.40	0.53	0.35
	b	0.10	0.13	**0.09**	0.11	**0.09**	0.12
Kendall	MSE	0.0058	0.0077	**0.0048**	0.0064	0.0074	0.0083
	R^2	0.46	0.26	**0.53**	0.37	0.26	0.18
	m	**0.59**	0.34	0.54	0.41	0.49	0.34
	b	**0.07**	0.12	0.08	0.11	0.08	0.12
Spearman	MSE	0.0070	0.0078	**0.0052**	0.0064	0.0066	0.0092
	R^2	0.36	0.23	**0.49**	0.36	0.35	0.09
	m	0.49	0.27	**0.52**	0.41	0.52	0.30
	b	0.10	0.13	**0.09**	0.11	**0.09**	0.13

Tabla 6.4: Métricas de rendimiento de los modelos de ML para el criterio 2, con λ_{Ox} como variable objetivo.

En la Figura 6.3 muestra el rendimiento de los modelos al estimar la energía de reorganización en función del criterio 2 y 1. Se observa una mejora en el rendimiento al aplicar este criterio.

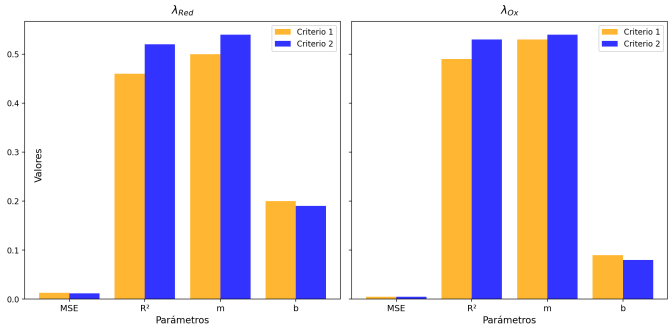


Figura 6.3: Evolución de las métricas de rendimiento en función de los criterios para ambas energía de reorganización. Los datos corresponden a los mejores modelos presentados en las Tablas: Criterio 1: 6.2; Criterio 2: 6.3 y 6.4.

6.4. Criterio 3: Correlación entre variables de entrada vs variable objetivo

Recordando que en este criterio se seleccionan las variables más representativas de acuerdo a la variable objetivo. En este trabajo se exploran cuatro formulaciones. Se seleccionan 30 variables que tengan la correlación más alta con la variable objetivo. Para el caso de la Distancia de Correlación y λ_{Red} indica que solo hay 29 variables con correlación no lineal para la base de datos DatOx mientras que para DatRed son 34 variables, se conserva este número porque éstas son las variables que detecta DC que tienen alguna correlación con la variable objetivo. En este mismo sentido, para λ_{Ox} sólo detecta 18 variables con DatOx y 17 con DatRed.

En la Tabla 6.5 se presentan los resultados de la validación de cada modelo que predice la energía de reorganización λ_{Red}. Si el valor está en **negritas** significa que ese valor es el mejor al compararlo entre modelos. En este caso los EAD tienen las mejores métricas en todas las formulaciones de correlación. Por lo tanto, si ahora se compara entre correlaciones (los valores en azul son las mejores métricas) Spearman contiene las mejores 3 métricas, mientras que Kendall y DC contienen sólo 2 y Pearson ninguno. La base DatOx resulta la mejor entre ambas bases para la mayoría de formulaciones de correlación, en el caso de Pearson la mejor base resulta ser DatRed.

Corr	Métrica	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0209	0.0163	**0.0150**	0.0152	0.0164	0.0173
	R^2	0.17	0.34	**0.38**	**0.38**	0.32	0.29
	m	0.25	0.36	0.42	**0.49**	0.37	0.34
	b	0.29	0.26	0.22	**0.21**	0.24	0.26
Kendall	MSE	0.0181	0.0182	**0.0135**	0.0166	0.0157	0.0187
	R^2	0.27	0.27	**0.45**	0.32	0.36	0.23
	m	0.33	0.31	**0.50**	0.45	0.41	0.37
	b	0.26	0.28	**0.20**	0.22	0.21	0.25
Spearman	MSE	0.0169	0.0178	**0.0128**	0.0165	0.0139	0.0185
	R^2	0.31	0.28	**0.48**	0.32	0.43	0.24
	m	0.35	0.28	**0.49**	0.44	0.44	0.33
	b	0.26	0.28	**0.20**	0.22	0.21	0.27
DC	MSE	0.0175	0.0190	**0.0128**	0.0170	0.0167	0.0192
	R^2	0.29	0.25	**0.47**	0.30	0.31	0.21
	m	0.33	0.30	**0.54**	0.45	0.38	0.33
	b	0.26	0.28	**0.19**	0.22	0.23	0.27

Tabla 6.5: Resultado de métricas de rendimiento. Variable objetivo: λ_{Red}, con 30 variables de entrada. En el caso de DC son 29 variables para DatOx y 34 para DatRed.

En la Tabla 6.6 están los resultados correspondientes a λ_{Ox}. los EAD son los mejores para estas condiciones, es decir que tienen rendimientos por encima del resto de modelos. La DatOx predomina como la base de datos con mejores resultados. La DC contiene las 4 mejores métricas si se compara entre formulaciones si se fija DatOx y EAD como se menciona anteriormente. Pearson y Spearman tiene valores muy parecidos a DC.

En resumen, si se trabaja con DatOx y EAD resulta como la mejor opción, para este caso la mejor correlación resultó ser DC, pero Pearson tiene rendimientos muy similares (si se observa las aproximaciones de λ_{Ox}). En la Figura 6.4 muestra el rendimiento de los modelos al estimar la energía de reorganización en función del criterio 3 y 2. No se observa una mejora sustancial.

Corr	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0066	0.0072	**0.0040**	0.0070	0.0061	0.0086
	R^2	0.35	0.29	**0.61**	0.31	0.40	0.14
	m	0.36	0.27	**0.59**	0.38	0.47	0.28
	b	0.12	0.13	**0.08**	0.11	0.10	0.13
Kendall	MSE	0.0072	0.0073	**0.0050**	0.0068	0.0069	0.0078
	R^2	0.29	0.28	**0.51**	0.32	0.31	0.23
	m	0.32	0.27	**0.52**	0.43	0.39	0.29
	b	0.13	0.13	**0.09**	0.10	0.11	0.13
Spearman	MSE	0.0072	0.0075	**0.0044**	0.0075	0.0057	0.0086
	R^2	0.29	0.27	**0.57**	0.26	0.43	0.15
	m	0.32	0.28	**0.56**	0.35	0.44	0.24
	b	0.13	0.13	**0.08**	0.11	0.10	0.14
DC	MSE	0.0092	0.0069	**0.0051**	0.0076	0.0088	0.0071
	R^2	0.12	0.33	**0.50**	0.24	0.13	0.30
	m	0.18	0.28	**0.52**	0.35	0.22	0.29
	b	0.16	0.13	**0.09**	0.12	0.15	0.13

Tabla 6.6: Resultado de las métricas de rendimiento. Variable objetivo: λ_{Ox}, con 30 variables de entrada. En el caso de DC son 18 variables para DatOx y 17 variables para DatRed.

6.5. Criterio 4: Análisis de componentes principales

Recordando los resultados de la familia de las Bencidinas (Bz), la prueba KMO fue utilizada para seleccionar el conjunto de datos derivados del criterio 3, asociado a alguna medida de correlación. La Tabla 6.7 contiene los resultados de esta prueba, para esta familia. Se puede concluir que la base de datos más adecuada para hacer un análisis de componentes principales es la que proviene del criterio 3 a partir de la Distancia de correlación, es decir, de las combinaciones posibles entre variable objetivo y variables de entrada, tres de ellas DC tiene la prueba más alta entre las cuatro medidas de correlación. En este mismo sentido, Kendall y Spearman tienen el mejor conjunto de datos para predecir λ_{Red} con una base DatRed. Los valores subrayados indican que correlación y que base de datos son la mejor combinación para predecir la energía de reorganización, por ejemplo DC con DatRed es la combinación con la prueba KMO más alta para λ_{Red} , mientras que para λ_{Ox} es DatOx con Kendall o Spearman.

Haciendo el anális de componentes principales para el conjunto de datos con la prueba KMO más alta, se reportan las variables más representativas en la Tabla 6.8 que cumplen al menos el 80 % de la varianza acumulada y conservando el 75 % de variables, descartando las menos representativas. Existen tres variables (subrayados) que coinciden en los tres conjuntos de datos. Mientras que en los conjuntos con una prueba KMO más grande para la energía de reorganización λ_{Red} y λ_{Ox} coinciden 6 variables (en azul).

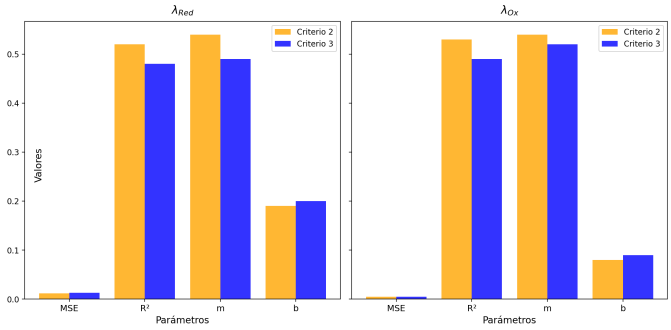


Figura 6.4: Evolución de las métricas de rendimiento en función de los criterios para ambas energía de reorganización. Los datos corresponden a los mejores modelos presentados en las Tablas: Criterio 2: 6.3 y 6.4;; Criterio 3: 6.5 y 6.6.

Energía	Base de Datos	Pearson	Kendall	Spearman	DC
λ_{Ox}	DatRed	0.52	0.62	0.60	0.63
λ_{Red}	DatRed	0.61	0.65	0.60	0.72
λ_{Ox}	DatOx	0.54	0.57	0.54	0.61
λ_{Red}	DatOx	0.59	0.64	0.64	0.63

Tabla 6.7: Valores de la prueba KMO de las 30 variables más correlacionadas con las variables objetivo (λ_{Red} o λ_{Ox}) resultado del Criterio 3.

De acuerdo a los resultados de la Tabla 6.8 se entrenan los distintos modelos y se presentan las métricas de rendimiento en la Tabla 6.9. En color azul, se concluye que el modelo EAD es el mejor para entrenar con una base de datos DatOx y predecir ambas energías de reorganización.

En la Figura 6.5 se muestra la evolución de los métricas de rendimiento de acuerdo a los criterios y la combinación del modelo y base de datos. Para esta familia no se ve un impacto sustancial en la reducción de dimensionalidad, sin embargo la diferencia de magnitudes no es significativa. Por otro lado, la mayoría de los modelos resultantes como mejores se pueden optimizar debido a su flexibilidad.

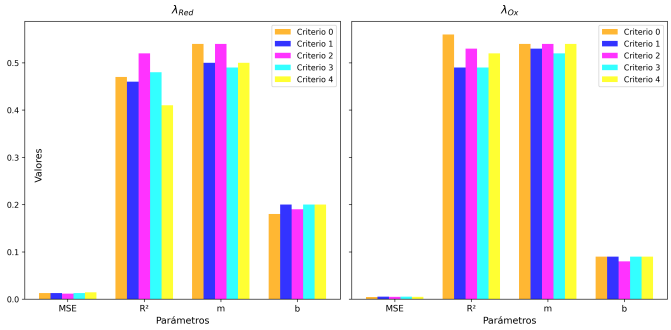


Figura 6.5: Evolución de las métricas de rendimiento en función de los criterios para ambas energía de reorganización. Los datos corresponden a los mejores modelos presentados en las Tablas: Criterio 0: 6.1; Criterio 1: 6.2; Criterio 2: 6.3 y 6.4; Criterio 3: 6.5 y 6.6; Criterio 4:6.9.

Orden	λ_{Red}-DatOx KMO=0.64 NCP=9 $\%\sum\sigma^2 = 81.64$	λ_{Red}-DatRed KMO=0.65 NCP=11 $\%\sum\sigma^2 = 82.48$	λ_{Ox}-DatRed KMO=0.64 NCP=10 $\%\sum\sigma^2 = 82.10$	λ_{Ox}-DatOx KMO=0.57 NCP=10 $\%\sum\sigma^2 = 82.30$
1	qed	PEOE_VSA10	VSA_EState8	PEOE_VSA8
2	SMR_VSA2	SlogP_VSA5	PEOE_VSA13	BCUT2D_LOGPLOW
3	PEOE_VSA10	BalabanJ	PEOE_VSA10	PEOE_VSA10
4	VSA_EState8	EState_VSA9	BCUT2D_LOGPLOW	PEOE_VSA12
5	BCUT2D_MWLOW	NumHAcceptors	Chi1v	SlogP_VSA3
6	SlogP_VSA1	SMR_VSA9	VSA_EState6	EState_VSA8
7	MinPartialCharge	PEOE_VSA11	PEOE_VSA12	SMR_VSA6
8	SMR_VSA5	PEOE_VSA8	PEOE_VSA1	SlogP_VSA6
9	PEOE_VSA11	SMR_VSA7	HallKierAlpha	qed
10	MaxAbsPartialCharge	BCUT2D_LOGPLOW	PEOE_VSA2	VSA_EState6
11	PEOE_VSA2	SlogP_VSA3	PEOE_VSA11	MinPartialCharge
12	VSA_EState6	VSA_EState7	SMR_VSA2	MaxAbsPartialCharge
13	BCUT2D_CHGHI	FpDensityMorgan1	FpDensityMorgan1	SMR_VSA10
14	PEOE_VSA9	VSA_EState2	BCUT2D_MRLOW	VSA_EState2
15	NOCount	SMR_VSA6	EState_VSA4	BalabanJ
16	SlogP_VSA2	SlogP_VSA4	BCUT2D_CHGLO	VSA_EState3
17	TPSA	VSA_EState3	MaxPartialCharge	VSA_EState4
18	Chi1v	BCUT2D_MRLOW	TPSA	EState_VSA9
19	FractionCSP3	SlogP_VSA1	NOCount	BCUT2D_LOGPHI
20	MaxEStateIndex	TPSA	PEOE_VSA9	MolLogP
21	VSA_EState2	EState_VSA8	PEOE_VSA7	TPSA
22	SMR_VSA6	SlogP_VSA8	VSA_EState2	SMR_VSA2

Tabla 6.8: Orden de variables, según su importancia de acuerdo al análisis de PCA.

Energía	Métrica	RLM		EAD		RN	
		Ox	Red	Ox	Red	Ox	Red
λ_{Red}	MSE	0.0202	0.0172	**0.0144**	0.0158	0.0172	0.0167
	R^2	0.19	0.30	**0.41**	0.35	0.29	0.31
	m	0.26	0.30	**0.50**	0.47	0.33	0.37
	b	0.28	0.28	**0.20**	0.21	0.25	0.24
λ_{Ox}	MSE	0.0075	0.0078	**0.0049**	0.0082	0.0069	0.0080
	R^2	0.26	0.23	**0.52**	0.19	0.31	0.21
	m	0.29	0.24	**0.54**	0.30	0.36	0.28
	b	0.13	0.14	**0.09**	0.13	0.12	0.14

Tabla 6.9: Métricas de rendimiento de los modelos de ML para el criterio 4, con λ_{Red} como variable objetivo

Capítulo 7

Resultados (Metil Viológeno)

7.1. Criterio 0: Referencia

Para familia de los Metil Viológenos hay un total de 721 reacciones redox monoelectrónicas. Las gráficas de dispersión de los modelos de referencia para esta familia se muestran en el Anexo ?? Figura ?? para energías de reorganización.

Se observa que los modelos de EAD tienen un mayor rendimiento. En combinación con DatRed predicen mejor λ_{Red}. Por otro lado si ahora se intenta reproducir λ_{Ox}, es mejor usar DatOx pero con un modelo de EAD.

Energía	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
λ_{Red}	MSE	0.0069	0.0077	0.0072	**0.0061**	0.0085	0.0080
	R^2	0.43	0.38	0.38	**0.47**	0.27	0.30
	m	0.47	**0.48**	0.46	**0.48**	0.44	**0.48**
	b	0.16	0.16	0.17	**0.17**	0.16	0.16
λ_{Ox}	MSE	0.0109	0.0117	**0.0091**	0.0092	0.0122	0.0124
	R^2	0.29	0.26	**0.037**	**0.37**	0.23	0.14
	m	0.38	0.39	**0.41**	**0.41**	0.36	0.34
	b	0.19	0.19	0.19	**0.18**	0.20	0.21

Tabla 7.1: Métricas de rendimiento de los modelos de ML para el criterio 0, con λ_{Red} y λ_{Ox} como variable objetivo.

7.2. Criterio 1: Variabilidad de los datos

Las Figuras ??, ?? y ?? y ?? muestran las variables descartadas y conservadas para el conjunto de datos DatOx y DatRed correspondientemente. Para DatOx se eliminan 100 variables dando un

nuevo conjunto de 108 variables. Por otro lado para el conjunto DatRed se descartan 101 variables y se forma un nuevo conjunto con 107 variables.

A partir de estos nuevos conjuntos de datos se entrenan los modelos y en la Figura **??** se muestran los resultados usando los conjuntos de prueba de los modelos, tanto para λ_{Red} como λ_{Ox} respectivamente.

Energía	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
λ_{Red}	MSE	0.0066	0.0070	0.0075	**0.0062**	0.0086	0.0074
	R^2	**0.46**	0.43	0.35	**0.46**	0.25	0.35
	m	0.46	0.46	0.44	**0.48**	0.39	0.46
	b	**0.16**	0.17	0.18	0.17	0.19	0.17
λ_{Ox}	MSE	0.0108	0.0114	0.0095	**0.0090**	0.00104	0.0125
	R^2	0.29	0.27	0.34	**0.38**	0.28	0.14
	m	0.37	0.38	0.39	**0.41**	0.39	0.32
	b	0.19	0.19	0.19	**0.18**	0.19	0.22

Tabla 7.2: Métricas de rendimiento de los modelos de ML para el criterio 0, con λ_{Red} y λ_{Ox} como variable objetivo.

En la Figura 7.1 muestra el rendimiento de los modelos al estimar la energía de reorganización en función del criterio 1 y 0. Se observa que no varía sustancialmente.

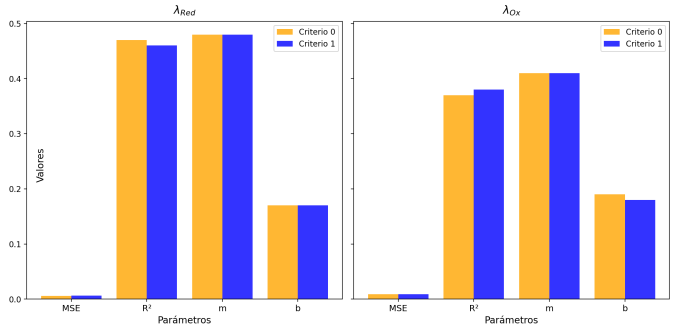


Figura 7.1: Comparación de las métricas de rendimiento entre el criterio 0 y el 1.Las métricas son los valores en azul de las Tablas 7.1 (criterio 0) y 7.2 (criterio 1).

7.3. Criterio 2: Análisis de correlación entre variables de entrada

Se abordan las cuatro formulaciones de correlación ya mencionadas. La Figura **??** muestra las magnitudes de la correlación de todas las formulaciones. Del criterio 1 el conjunto de datos se redujo a 108 y 107 variables para DatOx y DatRed respectivamente.

Al aplicar este criterio en el caso de Pearson y DatOx se redujeron de 108 variables a 81 descartando 27 variables. Para DatRed, con la misma correlación, se conservan 78 y se descartan 29. Con Kendall, tanto para DatOx como DatRed se eliminan 10 y se conservan 98 y 97 variables respectivamente, siendo las mismas variables eliminadas en ambos cojuntos. Spearman también encuentra que en ambos conjuntos las variables eliminadas son las mismas, hay 27 variables equivalentes y se conservan 82 y 81 variables para DatOx y DatRed respectivamente.

La Tabla 7.3 contiene las métricas de todos los modelos y correlaciones para la variable objetivo λ_{Red}. El mejor modelo resulta ser EAD y la base de datos DatRed. Destacando la correlación de Spearman sobre el resto.

Corr	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0069	0.0067	0.0072	**0.0056**	0.0077	0.0068
	R^2	0.43	0.44	0.38	**0.52**	0.33	0.41
	m	0.41	0.42	0.46	**0.51**	0.41	0.46
	b	0.19	0.19	0.18	**0.16**	0.18	0.17
Kendall	MSE	0.0068	0.0069	0.0074	**0.0055**	0.0072	0.0075
	R^2	0.44	0.43	0.35	**0.52**	0.37	0.35
	m	0.43	0.45	0.44	**0.50**	0.42	0.42
	b	0.17	0.17	0.18	**0.16**	0.18	0.18
Spearman	MSE	0.0068	0.0064	0.0062	**0.0050**	0.0075	0.0073
	R^2	0.44	0.47	0.46	**0.56**	0.35	0.36
	m	0.43	0.45	0.50	**0.53**	0.42	0.42
	b	0.17	0.17	0.16	**0.15**	0.18	0.18

Tabla 7.3: Métricas de rendimiento de los modelos de ML para el criterio 2, con λ_{Red} como variable objetivo.

Por otro lado, si la variable objetivo es λ_{Ox} la Tabla 7.4 contiene los resultados de validación. Los EAD siguen teniendo los mejores rendimientos, por otro lado ahora es DatOx quien predice mejor esta energía de reorganización, a la vez que Pearson destaca. En resumen este criterio indica que es mejor predecir λ_{Red} con DatOx y la correlación de Spearman, mientras que λ_{Ox} con DatRed con Pearson.

En la Figura 7.2 muestra el rendimiento de los modelos al estimar la energía de reorganización en función de los criterios 2 y 1. Se observa una mejora significativa en el rendimiento al aplicar este criterio.

Corr	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0116	0.0114	**0.0090**	0.0096	0.0121	0.0119
	R^2	0.25	0.26	**0.38**	0.34	0.16	0.18
	m	0.33	0.36	**0.42**	0.41	0.36	0.33
	b	0.21	0.20	**0.18**	0.18	0.20	0.21
Kendall	MSE	0.0112	0.0116	**0.0093**	0.0106	0.0104	0.0109
	R^2	0.27	0.26	**0.36**	0.27	0.28	0.24
	m	0.34	0.36	**0.39**	0.36	0.37	0.35
	b	0.20	0.20	**0.19**	0.20	0.20	0.20
Spearman	MSE	0.0112	0.0118	**0.0092**	0.0097	0.0114	0.0118
	R^2	0.26	0.24	**0.36**	0.33	0.21	0.18
	m	0.34	0.34	**0.42**	0.39	0.35	0.33
	b	0.21	0.21	**0.18**	**0.18**	0.20	0.21

Tabla 7.4: Métricas de rendimiento de los modelos de ML para el criterio 2, con λ_{Ox} como variable objetivo.

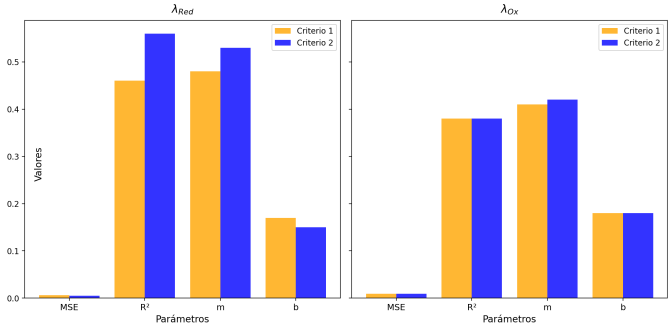


Figura 7.2: Comparación de las métricas de rendimiento entre el criterio 0 y el 1.Las métricas son los valores en azul de las Tablas 7.2 (criterio 1), 7.4 y 7.3 (criterio 2).

7.4. Criterio 3: Correlación entre variables de entrada vs variable objetivo

Se seleccionan las 30 variables más correlacionadas con la variable objetivo λ_{Red} y λ_{Ox} para ambas bases de datos a partir de los resultados del criterio 2. Para el caso de la Distancia de correlación se parte del resultado del criterio 1, con λ_{Red} indica que hay 61 variables con correlación no lineal para la base de datos DatOx y DatRed. En este mismo sentido, para λ_{Ox} se detectan 63 variables con DatOx y 62 con DatRed.

Corr	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0072	0.0075	0.0060	**0.0059**	0.0081	0.0079
	R^2	0.41	0.37	**0.48**	**0.48**	0.30	0.32
	m	0.36	0.33	**0.50**	0.49	0.32	0.32
	b	0.20	0.22	**0.16**	**0.16**	0.21	0.22
Kendall	MSE	0.0077	0.0073	0.0073	**0.0058**	0.0081	0.0075
	R^2	0.36	0.39	0.36	**0.50**	0.30	0.34
	m	0.35	0.35	0.42	**0.51**	0.31	0.34
	b	0.21	0.21	0.19	0.16	0.22	0.21
Spearman	MSE	0.0069	0.0077	**0.0060**	0.0066	0.0078	0.0077
	R^2	0.43	0.36	**0.48**	0.43	0.32	0.33
	m	0.37	0.30	**0.49**	0.47	0.34	0.23
	b	0.20	0.22	**0.16**	0.17	0.21	0.21
DC	MSE	0.0068	0.0069	**0.0059**	0.0061	0.0083	0.0082
	R^2	0.43	0.43	**0.49**	0.47	0.28	0.29
	m	0.40	0.39	**0.51**	0.48	0.37	0.36
	b	0.19	0.19	**0.16**	0.17	0.20	0.20

Tabla 7.5: Resultado de las métricas de rendimiento. Variable objetivo: λ_{Red}, con 30 variables de entrada. En el caso de DC son 58 (DatOx) y 57 (DatRed).

En la Tabla 7.6 están los resultados correspondientes a λ_{Ox}. Los EAD son los mejores. Se observa que DatOx predomina como la base de datos con mejores resultados. Por otro lado, Spearman tiene las mejores métricas.

Corr	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0106	0.0105	**0.0086**	0.0104	0.0120	0.0101
	R^2	0.27	0.29	**0.40**	0.28	0.17	0.30
	m	0.30	0.34	**0.44**	0.36	0.28	0.36
	b	0.23	0.21	**0.18**	0.20	0.23	0.20
Kendall	MSE	0.0118	0.0118	**0.0094**	0.0099	0.0119	0.0113
	R^2	0.21	0.22	**0.35**	0.32	0.18	0.22
	m	0.26	0.29	**0.42**	0.32	0.28	0.29
	b	0.24	0.22	**0.18**	0.20	0.23	0.23
Spearman	MSE	0.0105	0.0104	**0.0086**	0.0095	0.0102	0.0105
	R^2	0.28	0.30	**0.41**	0.34	0.30	0.27
	m	0.29	0.34	**0.43**	0.38	0.33	0.32
	b	0.23	0.21	**0.18**	0.19	0.21	0.21
DC	MSE	0.0116	0.0111	**0.0094**	0.0106	0.100	0.0116
	R^2	0.24	0.27	**0.35**	0.27	0.31	0.20
	m	0.32	0.35	**0.41**	0.36	0.39	0.31
	b	0.21	0.21	**0.18**	0.20	0.19	0.22

Tabla 7.6: Resultado de las métricas de rendimiento. Variable objetivo: λ_{Ox}, con 30 variables de entrada. En el caso de DC son 61 (DatOx) y 51 (DatRed).

En resumen, si se trabaja con DatRed con EAD y Kendall resulta como la mejor opción, para estimar de λ_{Red}. Para λ_{Ox} es mejor usar la correlación de Spearman.

En la Figura 7.3 muestra el rendimiento de los modelos al estimar la energía de reorganización en función del criterio 3 y 2. No se observa una mejora sustancial, en algunos casos las métricas son cercanas.

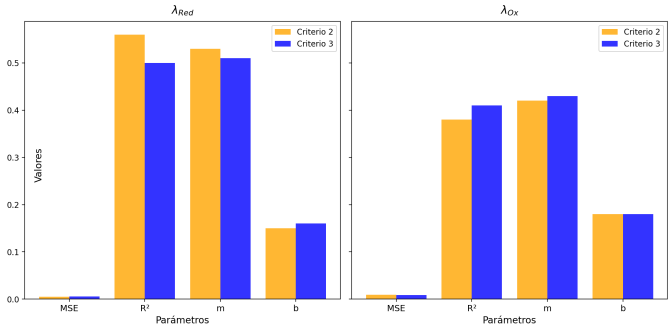


Figura 7.3: Comparación de las métricas de rendimiento entre el criterio 0 y el 1.Las métricas son los valores en azul de las Tablas 7.4 y 7.3 (criterio 2) y 7.5 y 7.6 (criterio 3).

7.5. Criterio 4: Análisis de componentes principales

EN la Tabla 7.7 se presentas los valores de la prueba KMO para los distintos conjuntos de datos. Esto sugiere que, para predecir λ_{Ox}, el conjunto de datos más adecuado es la base de datos DatRed, utilizando como variables aquellas obtenidas con la correlación DC(λ_{Ox}, DatRed). Por otro lado, para la energía λ_{Red}, el conjunto de datos más adecuado es la base de datos DatOx y utilizando las variables obtenidas con la correlación DC(λ_{Red},DatOx).

Energía	Base de Datos	Pearson	Kendall	Spearman	DC
λ_{Ox}	DatRed	0.76	0.73	0.77	0.80
λ_{Red}	DatRed	0.64	0.75	0.59	0.80
λ_{Ox}	DatOx	0.75	0.70	0.67	0.71
λ_{Red}	DatOx	0.74	0.80	0.66	0.81

Tabla 7.7: Valores de la prueba KMO de las 30 variables más correlacionadas con las variables objetivo (λ_{Red} o λ_{Ox}) resultado del Criterio 3. Para el caso de DC, se consideran las variables cuya correlación supera una correlación de 0.2.

En la Tabla 7.9 se presentan los resultados de aplicar PCA a los resultados obtenidos de la Tabla 7.7 tanto para el mejor conjunto de datos para predecir λ_{Red} como λ_{Ox}. Las variables en azul representan que coincide un total de 27 en ambos conjuntos de datos.

En la Tabla 7.8 se presentan las métricas de rendimiento del conjunto de datos de prueba, de acuerdo a PCA para ambas energías de reorganización. El modelo EAD predomina sobre el resto de modelos. Con respecto a λ_{Red} ambas bases de datos tiene tanto pruebas KMO y número de variables similares (con DatRed KMO=0.80 y 42 variables, mientras que para DatOx: KMO=0.81 y 43 variables) y los resultados de las métricas son muy similares. Mientras que λ_{Ox} tiene mejores resultados DatOx, para este conjunto de datos tiene una prueba KMO=0.75 con 22 variables, mientras que con la base de datos DatRed tiene una prueba KMO=0.80 con 45 variables.

En la Figura 7.4 se muestra la evolución de las métricas de acuerdo a los criterios y la combinación del modelo y base de datos. Para esta familia se ve un impacto sobre los rendimientos al reducir la dimensionalidad para la energía de reorganización λ_{Red}. Para la energía λ_{Ox} no tiene el mismo impacto, sin embargo los rendimientos son similares. La mayoría de los modelos resultantes como mejores son EAD y con ello se pueden optimizar.

Energía	Métrica	RLM		EAD		RN	
		Ox	Red	Ox	Red	Ox	Red
λ_{Red}	MSE	0.0075	0.0077	0.0056	**0.0055**	0.0083	0.0082
	R^2	0.37	0.31	0.51	**0.52**	0.28	0.29
	m	0.37	0.31	**0.53**	0.52	0.33	0.32
	b	0.20	0.21	**0.15**	0.16	0.21	0.21
λ_{Ox}	MSE	0.0117	0.0111	**0.0093**	0.0103	0.0109	0.0115
	R^2	0.21	0.27	**0.36**	0.29	0.25	0.20
	m	0.25	0.34	**0.40**	0.35	0.30	0.31
	b	0.24	0.21	**0.18**	0.20	0.23	0.22

Tabla 7.8: Métricas de rendimiento de los modelos de ML para el criterio 4, con λ_{Red} como variable objetivo

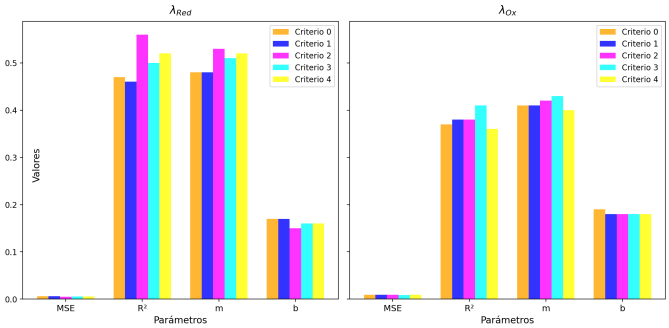


Figura 7.4: Evolución de las métricas de rendimiento en función de los criterios para ambas energía de reorganización. Los datos corresponden a los mejores modelos presentados en las Tablas: : 7.1; Criterio 1: 7.2; Criterio 2: 7.3 y 7.4 ; Criterio 3: 7.5 y 7.6; Criterio 4:7.8.

Orden	λ_{Ox}-DatRed KMO=0.80 DC NCP=7 $\%\sum\sigma^2$ =81.73	λ_{Red}-DatOx KMO=0.81 DC NCP=6 $\%\sum\sigma^2$ =82.27
1	SlogP_VSA7	HallKierAlpha
2	Chi4v	PEOE_VSA1
3	SlogP_VSA12	BCUT2D_MWLOW
4	PEOE_VSA9	MinPartialCharge
5	PEOE_VSA1	VSA_EState6
6	Chi3v	Chi2v
7	Chi1v	Chi4v
8	VSA_EState6	SlogP_VSA1
9	Chi2v	EState_VSA5
10	SlogP_VSA6	EState_VSA7
11	HallKierAlpha	Chi1v
12	qed	FractionCSP3
13	NumRotatableBonds	EState_VSA8
14	SlogP_VSA3	Chi3v
15	SMR_VSA10	MaxAbsPartialCharge
16	BCUT2D_MWLOW	NumRotatableBonds
17	BCUT2D_MRLOW	fr_allylic_oxid
18	BertzCT	SlogP_VSA10
19	Chi4n	SMR_VSA6
20	BCUT2D_MWHI	SlogP_VSA2
21	Kappa2	BalabanJ
22	EState_VSA5	SMR_VSA3
23	EState_VSA7	Chi3n
24	BCUT2D_LOGPLOW	MolMR
25	BalabanJ	VSA_EState1
26	fr_halogen	BCUT2D_MRLOW
27	Kappa3	PEOE_VSA9
28	EState_VSA1	SMR_VSA9
29	VSA_EState4	VSA_EState9
30	SlogP_VSA4	Chi4n
31	NumHeteroatoms	EState_VSA1
32	MinEStateIndex	BCUT2D_LOGPHI
33	BCUT2D_MRHI	BCUT2D_CHGLO
34	SlogP_VSA1	BCUT2D_CHGHI
35	EState_VSA10	Chi2n
36	PEOE_VSA7	Chi1n
37	LabuteASA	NumHeteroatoms
38	HeavyAtomMolWt	MolLogP
39	PEOE_VSA10	Kappa2
40	FpDensityMorgan1	SlogP_VSA8
41	MolWt	HeavyAtomMolWt
42	ExactMolWt	Chi0v
43	PEOE_VSA14	
44	PEOE_VSA11	
45	BCUT2D_CHGHI	

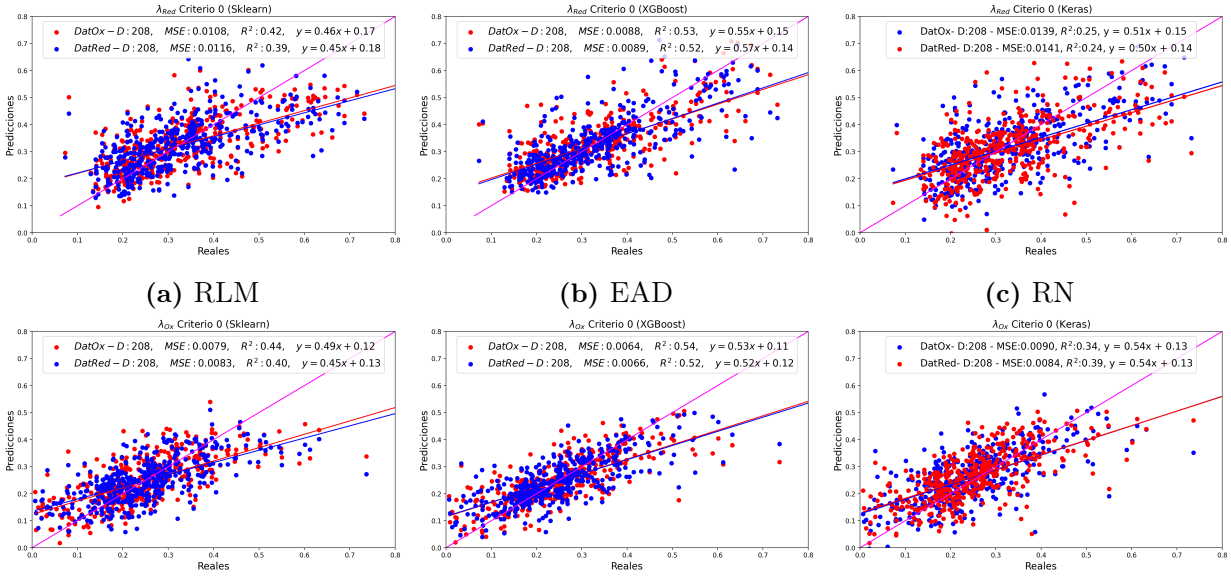
Tabla 7.9: Orden de variables, según su importancia de acuerdo al análisis de PCA.

Capítulo 8

Resultados (Las tres familias)

8.1. Criterio 0: Referencia

Para esta sección se prensentan los resultados del conjunto de las tres familias (Bpy, Bz y MV), en donde hay un total de 2219 reacciones redox monoelectrónicas. Las gráficas de dispersión de los modelos de referencia para esta familia se muestran en la Figura 8.1 para ambas energías de reorganización.



(a) RLM **(b)** EAD **(c)** RN

(d) RLM **(e)** EAD **(f)** RN

Figura 8.1: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} (primer renglón) y λ_{Ox} (segundo renglón), considerando el Criterio 0. En cada gráfica, se incluyen los coeficientes de determinación (R^2), el error cuadrático medio (MSE) y la ecuación de la línea recta ($y = mx + b$). Donde D representa el número de variables.

En la Tabla 8.1 se presentan los resultados de las métricas del conjunto de datos de prueba.

Se observa que los modelos de EAD tienen un mayor rendimiento en conjunto con la base de datos DatOx.

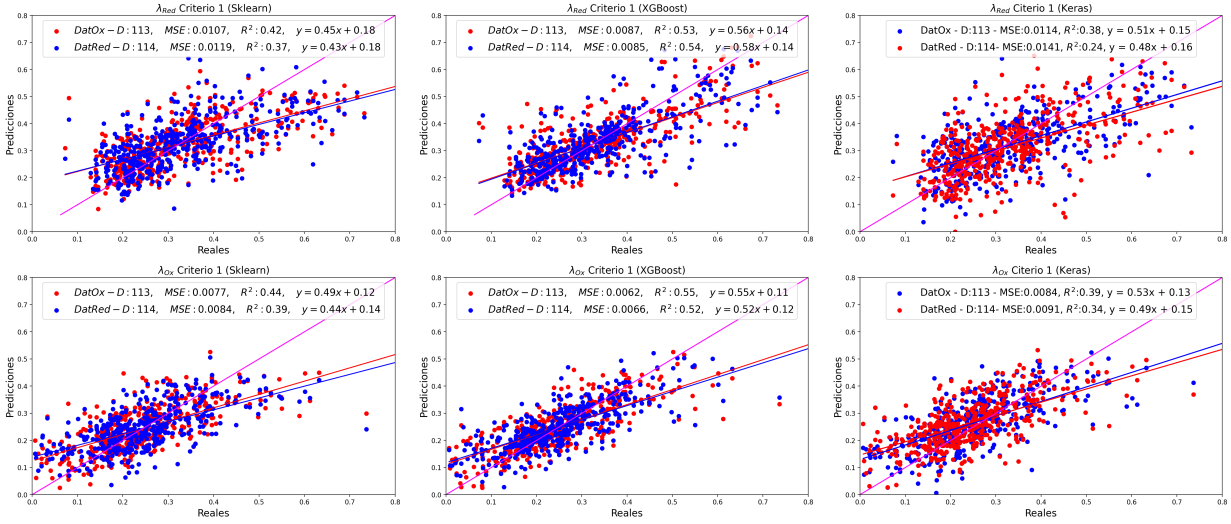
Energía	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
λ_{Red}	MSE	0.0108	0.0116	**0.0088**	0.0089	0.0139	0.0141
	R^2	0.42	0.39	**0.53**	0.52	0.25	0.24
	m	0.46	0.45	0.55	**0.57**	0.51	0.50
	b	0.17	0.18	0.15	**0.14**	0.15	**0.14**
λ_{Ox}	MSE	0.0079	0.0083	**0.0064**	0.0066	0.0090	0.0084
	R^2	0.44	0.40	**0.54**	0.52	0.34	0.39
	m	0.49	0.45	0.53	0.52	**0.54**	**0.54**
	b	0.12	0.13	**0.11**	0.12	0.13	0.13

Tabla 8.1: Métricas de rendimiento de los modelos de ML para el criterio 0, con λ_{Red} y λ_{Ox} como variable objetivo.

8.2. Criterio 1: Variabilidad de los datos

Las Figuras **??**, **??** y **??** y **??** se muestran las variables descartadas y conservadas para el conjunto de datos DatOx y DatRed correspondientemente.Para DatOx se eliminan 95 variables dando un nuevo conjunto de 113 variables. Por otro lado para el conjunto DatRed se descartan 94 variables y se forma un nuevo conjunto con 114 variables.

A partir de estos nuevos conjuntos de datos se entrenan los modelos y en la Figura 8.2 se muestran los resultados de las pruebas de rendimiento de los modelos, tanto para λ_{Red} como λ_{Ox} respectivamente.



(a) RLM **(b)** EAD **(c)** RN

Figura 8.2: Gráficas de dispersión (para el conjuto de todas las familias) para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} y λ_{Ox} repectivamente considerando el Criterio 1.

La Tabla 8.2 resume los resultados con las métricas de validación de los distintos modelos y energías de reorganización. Para predecir ambas energía de reorganización los EAD dan mejores rendimientos, aunque los rendimientos son muy cercanos, para predecir λ_{Red} es mejor utilizar DatRed, mientras que para λ_{Ox} es DatOx.

Para este conjunto, en el criterio 4 se presentarán las mejores métricas de rendimiento de acuerdo a cada criterio.

8.3. Criterio 2: Análisis de correlación entre variables de entrada

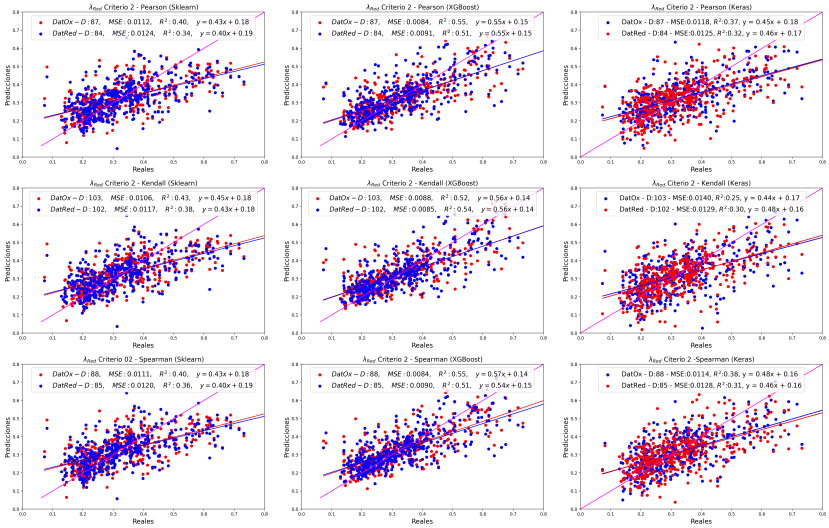
La Figura **??** muestra las correlaciones de todas las formulaciones entre las variables descartadas (renglones) y algunas variables que se conservan (columnas) tanto para la base de datos DatOx como DatRed. Para el caso de Pearson y DatOx se descartan 26 variables y se conservan 87 mientras

Energía	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
λ_{Red}	MSE	0.0107	0.0119	0.0087	**0.0085**	0.0114	0.0141
	R^2	0.42	0.37	0.53	**0.54**	0.38	0.24
	m	0.45	0.43	0.56	**0.58**	0.51	0.48
	b	0.18	0.18	**0.14**	**0.14**	0.15	0.16
λ_{Ox}	MSE	0.0077	0.0084	**0.0062**	0.0066	0.0084	0.0091
	R^2	0.44	0.39	**0.55**	0.52	0.39	0.34
	m	0.49	0.44	**0.55**	0.52	0.53	0.49
	b	0.12	0.14	**0.11**	0.12	0.13	0.15

Tabla 8.2: Métricas de rendimiento de los modelos de ML para el criterio 1, con λ_{Red} y λ_{Ox} como variable objetivo.

que para la DatRed se descartan 30 y se conservan 84. Con Kendall con DatOx se eliminan 10 y se conservan 103, para DatRed se descartan 12 y se conservan 102 variables. Spearman descarta 25 y conserva 88 variables para DatOx, para DatRed descarta 29 y conserva 85.

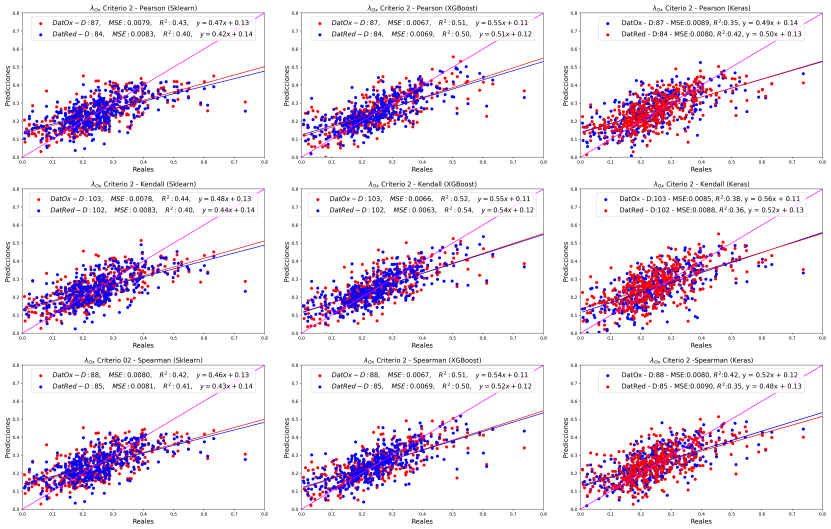
Los resultados, de esta selección de variables, con las métricas de rendimiento para las energías de reorganización 8.3



(a) RLM **(b)** EAD **(c)** RN

Figura 8.3: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} considerando el Criterio 2 y las distintas formulaciones de correlación.

La Tabla 8.3 contiene las métricas del conjunto de prueba de todos los modelos y correlaciones para la variable objetivo λ_{Red}. El mejor modelo resulta ser EAD y la base de datos depende de la correlación. Pearson y Spearman tienen los mismos rendimiento en MSE y R^2, pero la segunda correlación mejora en la pendiente y la ordenada al origen, usando una base de datos DatOx.



(a) RLM **(b)** EAD **(c)** RN

Figura 8.4: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Ox} considerando el Criterio 2 y las distintas formulaciones de correlación.

Por otro lado, si la variable objetivo es λ_{Ox} la Tabla 8.4 contiene los resultados de validación. Los EAD siguen teniendo los mejores rendimientos, como en el caso anterior, la base depende de la correlación. Los rendimientos son muy similares, Kendall tiene un error menor sobre el resto, esto en combinación con la base de datos DatRed.

En resumen para λ_{Red} el mejor modelo resulta de usar Spearman con una base Datox. Mientras que para λ_{Ox} es mejor usar la correlacióm de Kendall con la base DatRed. Ambos con el modelo EAD.

Corr	Métrica	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0112	0.0124	**0.0084**	0.0091	0.0118	0.0125
	R^2	0.40	0.34	**0.55**	0.51	0.37	0.32
	m	0.43	0.40	**0.55**	**0.55**	0.45	0.46
	b	0.18	0.19	**0.15**	**0.15**	0.18	0.17
Kendall	MSE	0.0106	0.0117	0.0088	**0.0085**	0.0140	0.0129
	R^2	0.43	0.38	0.52	**0.54**	0.25	0.30
	m	0.45	0.43	**0.56**	**0.56**	0.44	0.48
	b	0.18	0.18	**0.14**	**0.14**	0.17	0.16
Spearman	MSE	0.0111	0.0120	**0.0084**	0.0090	0.0114	0.0128
	R^2	0.40	0.36	**0.55**	0.51	0.38	0.31
	m	0.43	0.40	**0.57**	0.54	0.48	0.46
	b	0.18	0.19	**0.14**	0.15	0.16	0.16

Tabla 8.3: Métricas de rendimiento de los modelos de ML para el criterio 2, con λ_{Red} como variable objetivo.

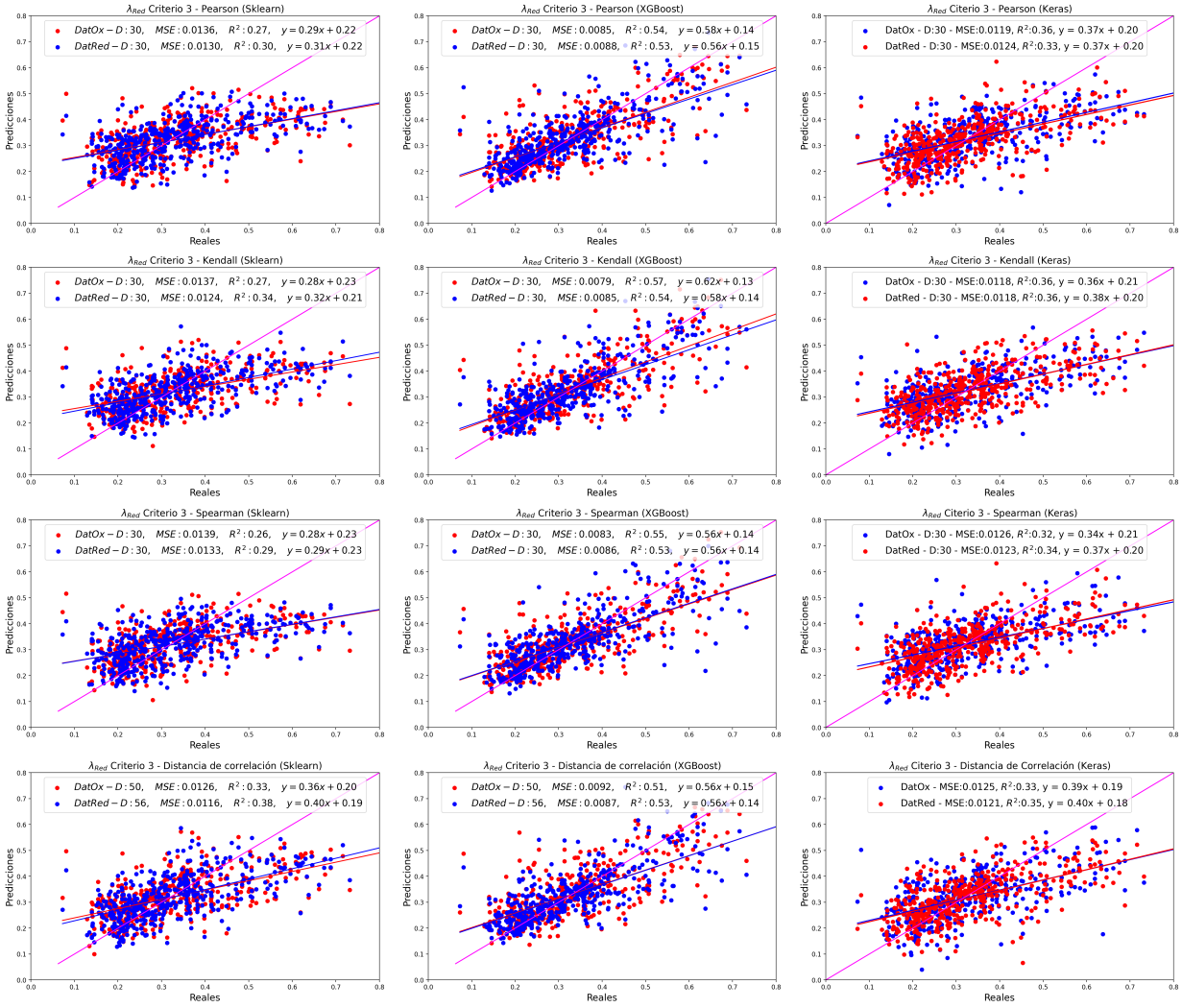
Corr	Métrica	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0079	0.0083	**0.0067**	0.0069	0.0089	0.0080
	R^2	0.43	0.40	**0.51**	0.50	0.35	0.42
	m	0.47	0.42	**0.55**	0.51	0.49	0.50
	b	0.13	0.14	**0.11**	0.12	0.14	0.13
Kendall	MSE	0.0078	0.0083	0.0066	**0.0063**	0.0085	0.0088
	R^2	0.44	0.40	0.52	**0.54**	0.38	0.36
	m	0.48	0.44	0.55	0.54	**0.56**	0.52
	b	0.13	0.14	**0.11**	0.12	**0.11**	0.13
Spearman	MSE	0.0080	0.0081	**0.0067**	0.0069	0.0080	0.0090
	R^2	0.42	0.41	**0.51**	0.50	0.42	0.35
	m	0.46	0.43	**0.54**	0.52	0.52	0.48
	b	0.13	0.14	**0.11**	0.12	0.12	0.13

Tabla 8.4: Métricas de rendimiento de los modelos de ML para el criterio 2, con λ_{Ox} como variable objetivo.

8.4. Criterio 3: Correlación entre variables de entrada vs variable objetivo

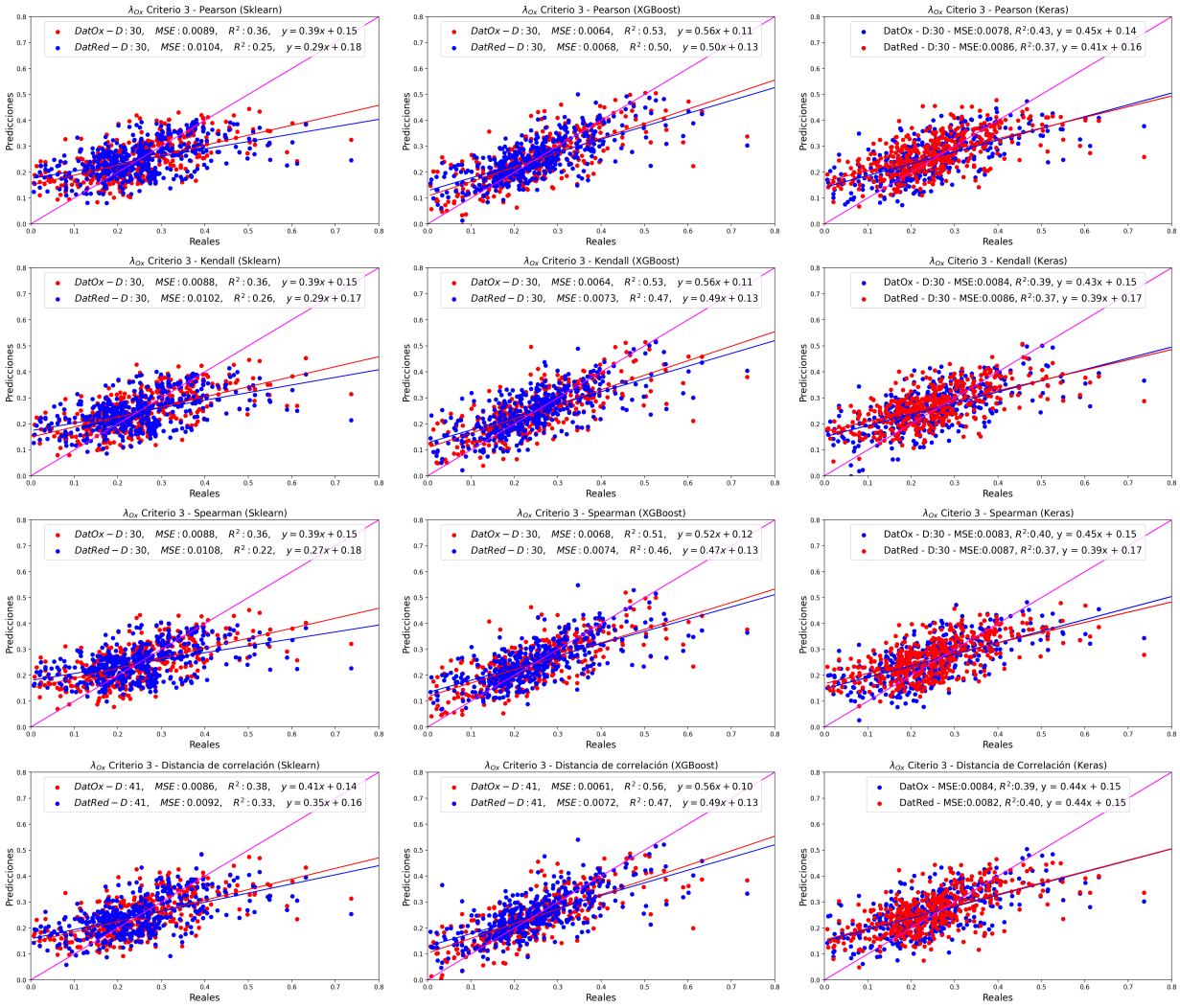
Se seleccionan 30 variables que tengan la correlación más alta con la variable objetivo. Para el caso de la Distancia de Correlación y λ_{Red} y DatOx indica que hay 50 variables con correlación no lineal, mientras que para DatRed son 56 variables. En este mismo sentido, DC detecta que con λ_{Ox} y DatOx o DatRed hay 41 variables, aunque no exactamente las mismas variables.

Las gráficas de dispersión, resultado del critero 3 se muestran en las Figuras 8.5 y 8.6 para la energía de reorganización λ_{Red} y λ_{OX} respectivamente.



(a) RLM **(b)** EAD **(c)** RN

Figura 8.5: Gráficas de dispersión para el conjunto de todas la familias con los tres modelos de regresión que predicen la energía de reorganización λ_{Red} considerando el Criterio 3 y las distintas formulaciones de correlación.



(a) RLM **(b)** EAD **(c)** RN

Figura 8.6: Gráficas de dispersión para el conjunto de todas la familias con los tres modelos de regresión que predicen la energía de reorganización λ_{Ox} considerando el Criterio 3 y las distintas formulaciones de correlación.

En la Tabla 8.5 se presentan los resultados de la validación de cada modelo que predice la energía de reorganización λ_{Red}. El modelo EAD tienen las mejores métricas en todas las formulaciones de correlación. Kendall en combinación con la base de datos DatOx da los mejores resultados, aunque el resto de las maneras de calcular la correlación, tienen rendimientos muy parecidos. La base DatOx resulta como la mejor entre ambas para la mayoría de formulaciones de correlación, excepto para DC.

Corr	Métrica	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0136	0.0130	**0.0085**	0.0088	0.0119	0.0124
	R^2	0.27	0.30	**0.54**	0.53	0.36	0.33
	m	0.29	0.31	**0.58**	0.56	0.37	0.37
	b	0.22	0.22	**0.14**	0.15	0.20	0.20
Kendall	MSE	0.0137	0.0124	**0.0079**	0.0085	0.0118	0.0118
	R^2	0.27	0.34	**0.57**	0.54	0.36	0.36
	m	0.28	0.32	**0.62**	0.58	0.36	0.38
	b	0.23	0.21	**0.13**	0.14	0.21	0.20
Spearman	MSE	0.0139	0.0133	**0.0083**	0.0086	0.0126	0.0123
	R^2	0.26	0.29	**0.55**	0.53	0.32	0.34
	m	0.28	0.29	**0.56**	**0.56**	0.34	0.37
	b	0.23	0.23	**0.14**	**0.14**	0.21	0.20
DC	MSE	0.0126	0.0116	0.0092	**0.0087**	0.0125	0.0121
	R^2	0.33	0.38	0.51	**0.53**	0.33	0.35
	m	0.36	0.40	**0.56**	**0.56**	0.39	0.40
	b	0.20	0.19	0.15	**0.14**	0.19	0.18

Tabla 8.5: Resultado de las métricas de rendimiento. Variable objetivo: λ_{Red}, con 30 variables de entrada. En el caso de DC son 50 variables para DatOx y 56 para DatRed.

En la Tabla 8.6 están los resultados correspondientes a λ_{Ox}. Los EAD son los mejores para estas condiciones, es decir que tienen rendimientos por encima del resto de modelos. La base de datos DatOx predomina como la base de datos con mejores resultados. La DC contiene los mejores rendimientos. El resto de formulaciones de correlación tienen rendimientos similares.

Corr	Métricas	RLM		EAD		RN	
		DatOx	DatRed	DatOx	DatRed	DatOx	DatRed
Pearson	MSE	0.0089	0.0104	**0.0064**	0.0068	0.0078	0.0086
	R^2	0.36	0.25	**0.53**	0.50	0.43	0.37
	m	0.39	0.29	**0.56**	0.50	0.45	0.41
	b	0.15	0.18	**0.11**	0.13	0.14	0.16
Kendall	MSE	0.0088	0.0102	**0.0064**	0.0073	0.0084	0.0086
	R^2	0.36	0.26	**0.53**	0.47	0.39	0.37
	m	0.29	0.29	**0.56**	0.49	0.43	0.39
	b	0.15	0.17	**0.11**	0.13	0.15	0.17
Spearman	MSE	0.0088	0.0108	**0.0068**	0.0074	0.0083	0.0087
	R^2	0.36	0.22	**0.51**	0.46	0.40	0.37
	m	0.39	0.27	**0.52**	0.47	0.45	0.39
	b	0.15	0.18	**0.12**	0.13	0.15	0.17
DC	MSE	0.0086	0.0092	**0.0061**	0.0072	0.0084	0.0082
	R^2	0.38	0.33	**0.56**	0.47	0.39	0.40
	m	0.41	0.35	**0.56**	0.49	0.44	0.44
	b	0.14	0.16	**0.10**	0.13	0.15	0.15

Tabla 8.6: Resultado de las métricas de rendimiento. Variable objetivo: λ_{Ox}, con 30 variables de entrada. En el caso de DC son 41 variables para DatOx y DatRed.

En resumen, si se trabaja con DatOx y EAD resulta como la mejor opción, para λ_{Red} es mejor usar las variables que vengan de la correlación de Kendall, mientras que para λ_{Ox} da mejores resultados usar la correlación DC. Por otro lado, se muestra que 30 variables puedan dan un mejor rendimiento que un conjunto más grande. Por ejemplo, en DC, para la base DatOx se usan 50, mientras que para DatRed son 56, y resulta mejor usar 50 para predecir λ_{Red}, este análisis se puede hacer con el resto de correlación y suelen dar resultados mejores o muy cercanos, pero con un total de 30 variables.

8.5. Criterio 4: Análisis de componentes principales

En la Tabla 8.8 se presentan la variables en el orden de relevancia de acuerdo a PCA para cada energía de reorganización, según la prueba KMO más alta. Para ambos conjuntos de variables hay 11 descriptores (se muestran en azul) que son relevantes para predecir ambas energía de reorganización. En el Apendice A 12 se describen los descriptores, resultado del criterio 4 para el conjunto de datos de todas las familias.

Energía	Base de Datos	Pearson	Kendall	Spearman	DC
λ_{Ox}	DatRed	**0.72**	0.68	0.66	0.70
λ_{Red}	DatRed	0.64	0.74	0.59	**0.79**
λ_{Ox}	DatOx	**0.70**	0.63	0.62	0.64
λ_{Red}	DatOx	0.65	0.72	0.64	**0.81**

Tabla 8.7: Valores de la prueba KMO de las 30 variables más correlacionadas con las variables objetivo (λ_{Red} o λ_{Ox}) resultado del Criterio 3.

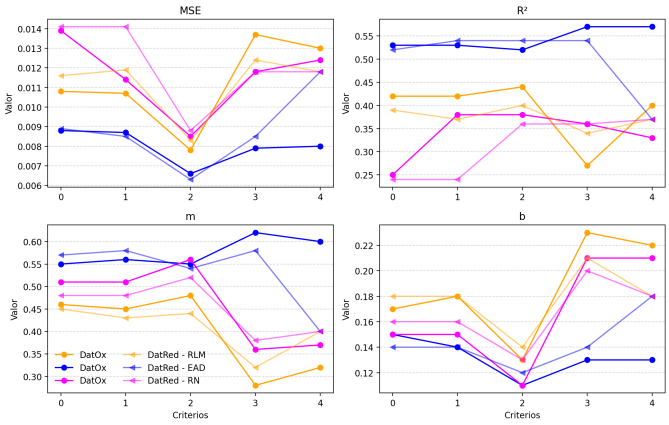


Figura 8.7: Métricas de los modelos, para predecir λ_{Red}, en función del criterio.

En la Tabla 8.9 se presentan las métricas del conjunto de datos de prueba. El modelo EAD es el mejor en ambas energía de reorganización y ambas bases de datos. Los valores en azul corresponden a la combinación modelo-base de datos y resulta ser mejor usar DatOx en ambos cojuntos, aunque los rendimientos son similares en ambas bases de datos. Para el caso de λ_{Red} y la base de datos DatOx tiene 37 variables y una prueba KMO=0.81, mientras que con DatRed tiene 41 variables y una prueba KMO=0.79. La primer combinación con menos variables y una mayor KMO condujo a un mejor rendimiento. Ahora, si el objetivo es λ_{Ox} ambas bases de datos 22 variables, pero mientras DatOx tiene una prueba KMO=0.70 y DatRed una prueba KMO=0.72, tiene ligeramente mejores la base de datos asociada a la prueba KMO más alta.

En la Figura 8.7 se muestran las cuatro métricas de cada modelo en función del criterio. El modelo EAD se destaca en todas las métricas. Por ejemplo, MSE tiene el error más bajo que el resto de modelos. De manera general, aplicar los criterios, para el conjunto de todas las familias de moléculas, tiene un impacto positivo si se utiliza una base de datos DatOx para predecir λ_{Red}.

En la Figura 8.8 se muestran las métricas de los modelos asociados a λ_{Ox}. Los EAD siguen teniendo mejores rendimientos. Para esta energía es mejor usar DatOx, porque en el último criterio DatRed da más error teniendo resultados similares a RLM.

Orden	λ_{Ox}-DatRed KMO=0.72 Pearson NCP=9 $\%\sum\sigma^2$ =80.34	λ_{Red}-DatOx KMO=0.81 DC NCP=8 $\%\sum\sigma^2$ =82.26
1	SlogP_VSA3	EState_VSA9
2	PEOE_VSA8	fr_NH0
3	EState_VSA9	qed
4	VSA_EState9	SMR_VSA3
5	BCUT2D_MRLOW	VSA_EState2
6	PEOE_VSA1	PEOE_VSA1
7	FpDensityMorgan1	PEOE_VSA3
8	BCUT2D_CHGHI	BCUT2D_MWLOW
9	PEOE_VSA6	VSA_EState9
10	SMR_VSA10	BalabanJ
11	SlogP_VSA8	FractionCSP3
12	fr_NH0	BCUT2D_CHGHI
13	VSA_EState7	Chi1v
14	EState_VSA3	SMR_VSA9
15	TPSA	NHOHCount
16	FractionCSP3	Chi2v
17	BCUT2D_MWLOW	BCUT2D_CHGLO
18	PEOE_VSA11	VSA_EState3
19	SlogP_VSA1	BCUT2D_LOGPLOW
20	VSA_EState6	VSA_EState6
21	BCUT2D_CHGLO	VSA_EState4
22	SlogP_VSA2	Chi3v
23	-	BCUT2D_MRLOW
24	-	BCUT2D_LOGPHI
25	-	NumAromaticRings
26	-	Chi4v
27	-	HeavyAtomMolWt
28	-	PEOE_VSA9
29	-	EState_VSA5
30	-	Chi3n
31	-	SlogP_VSA8
32	-	MolWt
33	-	Chi4n
34	-	ExactMolWt
35	-	SlogP_VSA2
36	-	Chi1n
37	-	HeavyAtomCount

Tabla 8.8: Orden de variables, según su importancia de acuerdo al análisis de PCA.

8.6. Clasificación

Para esta sección se aborda un problema de clasificación, es decir, que el modelo prediga, a partir de los descriptores quimioinformáticos, si una molécula es reversible electroquímicamente. En otras palabras, que la reacción redox monoelectrónica asociada vaya en ambos sentidos (oxidación y reducción). Dado que el enfoque es determinar si la reacción es viable para su uso en baterías de flujo redox, solo se necesita saber si es o no reversible para considerarla como candidata. Por ello, el problema se reduce a una clasificación binaria.

Para este problema, la variable objetivo debe contener información sobre si la especie química está asociado a un proceso redox reversible, quasireversibles o irreversible. Una manera de saber esta información es buscar en la literatura. Sin embargo, en este trabajo se propone una escala de reversibilidad a partir de las energías de reorganización. Dicha escala consiste en un cociente entre λ_{Red} y λ_{Ox} y a partir de este encontrar los límites en el que los procesos se clasifican. Por ejemplo, se propone que uno de los límites sea 0.3 y que valores menores a este el proceso es irreversible, porque ya sea el proceso de oxidación o reducción el que tiene un costo energético mayor al del sentido opuesto. Por otro lado si el cociente es cercano a la unidad o mayor a 0.8 (umbral propuesto) el costo energético en ambos sentidos es similar. Por último, los cocientes entre los límites 0.3 a 0.8 se pueden clasificar como quasirreversible. Por lo tanto, la intención de este capítulo es usar las energías de reorganización como un indicador de reversibilidad.

Energía	Métrica	RLM		EAD		RN	
		Ox	Red	Ox	Red	Ox	Red
λ_{Red}	MSE	0.0130	0.0128	**0.0080**	0.0083	0.0124	0.0118
	R^2	0.40	0.32	**0.57**	0.55	0.33	0.37
	m	0.32	0.31	**0.60**	0.54	0.37	0.40
	b	0.22	0.22	**0.13**	0.15	0.21	0.18
λ_{Ox}	MSE	0.0091	0.0117	**0.0072**	0.0073	0.0082	0.0100
	R^2	0.34	0.17	**0.48**	0.47	0.41	0.27
	m	0.36	0.22	**0.50**	0.47	0.43	0.33
	b	0.16	0.19	**0.12**	0.14	0.15	0.18

Tabla 8.9: Métricas de rendimiento de los modelos de ML para el criterio 4, con λ_{Red} como variable objetivo

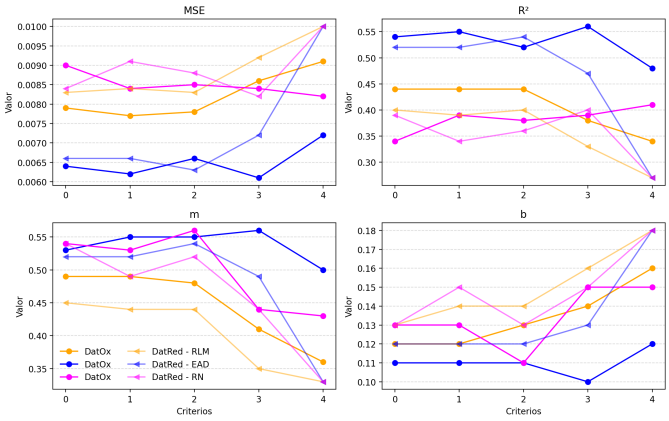


Figura 8.8: Métricas de los modelos, para predecir λ_{Ox}, en función del criterio.

Una vez obtenida la clasificación propuesta, se tienen los conjuntos de datos presentados en la Tabla 8.10. Dado que el enfoque es determinar si la reacción es reversible o no, se combinan las clases quasirreversible e irreversible y se obtiene una clasificación binaria. Otro característica que contribuye a utilizar un modelo binario es que, de acuerdo a los datos, se tienen muy pocas reacciones asociadas a la clase irreversible, lo que puede conducir a sesgos.

Por otro parte, la información de entrada se va a tomar la base de datos DatOx y las 22 variables resultates del criterio 4 al utilizar la correlación de Pearson. En la Figura 8.10 se presentan gráficas de distribución entre variables de entrada, en donde se observa el comportamiento de los descriptores de acuerdo con las dos clases. Con estas gráficas se muestra la complejidad del problema; por ejemplo, no se puede utilizando planos para clasificar.

Como se aprendió en el problema de regresión o ajuste, el mejor modelo de entrenamiento fue un EAD, por este motivo se presentarán resultados con este modelo. En la ecuación 8.1 se presenta la matriz de confusión, la cuál consiste en el número de aciertos para cada clase y se colocan en el elemento $a_{i,i}$ llamados verdaderos positivos o negativos, mientras que los errores asociados a una

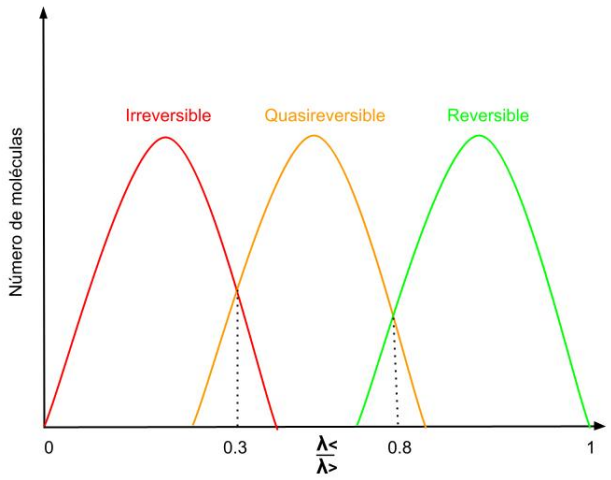


Figura 8.9: Escala de reversibilidad en función de coeficiente de energías de reorganización, donde $\lambda_<$ es la energía mas baja entre λ_{Red} y λ_{Ox}, mientras que $\lambda_>$ es la energía más alta.

Clase	**Tamaño**		
	Total	Entrenamiento	Prueba
Quasirreversible	1074	-	-
Reversible	934	-	-
Irreversible	211	-	-
Quasirreversible + Irreversible (QI)	1285	1028	257
Reversible (R)	934	747	187

Tabla 8.10: Información de la distribución de los datos para una clasificación de tres clases y la de dos.

clase pero que pertenecen a la otra están en los elementos $a_{i,j}$ llamados falsos positivos y negativos. En este ejemplo, las clases son positivos y negativos.

$$\begin{pmatrix} \text{Verdaderos Positivos (VP)} & \text{Falsos Positivos(FP)} \\ \text{Falsos Negativos (FN)} & \text{Verdaders Negativos (VN)} \end{pmatrix} \tag{8.1}$$

Esta matriz se construye a partir de los datos de prueba en donde la suma de las filas representan el total de datos que se están probando. Existen distintas métricas que parten de la matriz de confusión. Por ejemplo, la exactitud, llamada y F1, estas métricas se utilizan para evaludar el modelo. Se definen como:

- Exactitud: Mide la proporción entre las predicciones correctas (verdaderos positivos y verdaderos negativos) y el total de predicciones.

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN}$$

- Precisión: Mide la proporción entre las predicciones verdaderos positivos, y todas las predic-

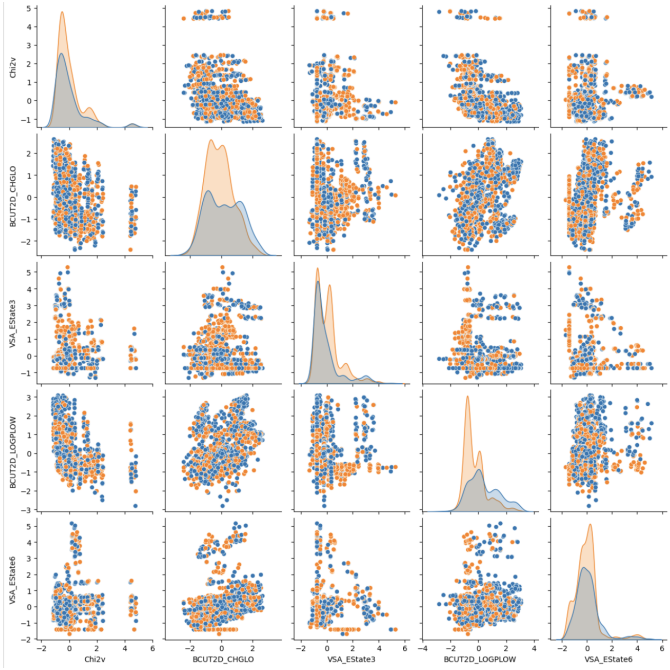


Figura 8.10: Distribución de la base de datos para 5 variables de DatOx. • QI, • R.

ciones predichas como positivas. Para las predicción de los negativos tiene el equivalente.

$$\text{Precisión} = \frac{a_{i,i}}{a_{i,i} + a_{i,j}}$$

- Llamada:Mide la proporción de verdaderos positivos entre todas las predicciones como positivas.

$$\text{Llamada} = \frac{a_{i,i}}{a_{i,i} + a_{j,i}}$$

- F1-Score: Es la media armónica entre la Precisión y la Llamada.

$$F1 = \frac{2\text{Llamada} * \text{Precisión}}{\text{LlamadaPrecisión}}$$

- Métrica promedio:

$$\text{Macro avg} = \frac{1}{N}\sum_{i=1}^{N} \text{Métrica}_i$$

donde N es el número de clases y Métrica_i es la métrica de la clase i.

- Métrica pesada:

$$\text{Weighted avg} = \frac{\sum_{i=1}^{N} \text{Métrica}_i * \text{Soporte}_i}{\sum_{i=1}^{N} \text{Soporte}_i}$$

donde Soporte_i el número de datos para la clase i.

La Figura 8.11 muestra la matriz de confusión correspondiente del conjunto de datos de prueba. El caso ideal, es que el modelo acierte las 257 especies etiquetadas como R, en vez de 90 moléculas. De igual manera, el modelo identifica 127 como QI , en vez de 187 especies.

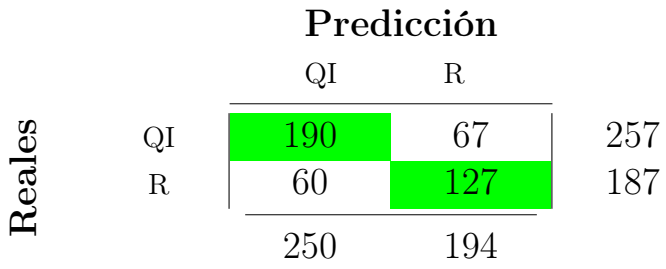


Figura 8.11: Matriz de confusión para la clasificación de reversibilidad electroquímica para el conjunto de todas las familias.

Para el entrenamiento del conjunto de datos obtenido del análisis de componentes principales con la prueba KMO más alta se presentan los resultados en la Tabla 8.11. La precisión global del modelo es del 71.40 %, mientras que para predecir moléculas reversibles es del 74 %.

Clase	Precisión	Recall	F1-Score	Soporte
QI	0.74	0.74	0.75	257
R	0.65	0.68	0.67	187
Macro avg	0.71	0.71	0.71	444
Weighted avg	0.72	0.71	0.71	444
Precisión: 71.40 %				

Tabla 8.11: Métricas de rendimiento para el modelo de XGBoost, con la base de datos DatOx y las variables obtenidas con la correlación DC(λ_{Red}, DatOx)

Este resultado aunque es aceptable, aún se puede y debe optimizar. Hacer este proceso implica hacer una búsqueda de la combinación de hiperparámetros que dé un mejor rendimiento y la vez garantizar que el entrenamiento no esté sobreajustado.

8.7. Optimización

Esta sección tiene como objetivo discutir sobre cómo obtener un mejor modelo, tanto para problemas de regresión como de clasificación. Se discuten tres enfoques para explorar los distintos hiperparámetros que contengan las mejores métricas.

1. Búsqueda exhaustiva (Grid Search): Consiste en contruir un mallado (grid) de combinaciones de hiperparámetros y entrenar el modelo con todas las combinaciones. Se seleccionan los hiperparámetros que tenga las mejores métricas.
2. Búsqueda aleatoria (Random Search): Este método selecciona algunas combinaciones aleatorias del mallado de hiperparámetros, esto reduce el costo computacional que puede costar explorar todas las combinaciones posibles.
3. Búsqueda Bayesiana (Bayesian Optimization): Utiliza un enfoque probabilístico para estimar qué combinaciones de hiperparámetros tienen mayor probabilidad de mejorar el rendimiento del modelo.

4. Exploración sistemática (One-Parameter Tuning): Este enfoque consiste en ajustar un hiperparámetro a la vez, manteniendo constantes los demás. Se entrena el modelo con diferentes valores para ese hiperparámetro y se selecciona el valor que produzca las mejores métricas.

Para tener una mayor confianza de las métricas, se puede entrenar el modelo con el método de validación cruzada, que consiste en dividir el conjunto de entrenamiento en 10 (o menos) subconjuntos. Se Utilizan 9 para entrenar y 1 para validar (con este conjunto de calculan las métricas), repitiendo este proceso 10 veces, cambiando siempre el conjunto de entrenamiento y validación, por último se promedian las métrica de cada ciclo. De esta manera, el rendimiento es más realista sobre la capacidad predictiva del modelo. Por otro lado, es una forma de evitar el sobreajuste.

Capítulo 9

Conclusiones

- Se logró obtener modelos que para predecir las propiedades planteadas, dichos modelos son susceptibles de mejora: el modelo de regresión, que predice la energía de reorganización λ_{Red} y λ_{Ox} y el modelo la clasificación que indica si la molécula es reversible o no. Ambos modelos entrenados con descriptores quimioinformáticos y como información de entrada para predecir ambas respuestas.
- Se logró reducir la dimensionalidad utilizando herramientas estadísticas, identificando las variables más representativas de los conjuntos de datos, tomando en cuenta una serie de criterios. Aunque los rendimiento de modelos son bajos, estos pueden ser optimizados para mejorar su desempeño.
- Del análisis del problema de regresión, se concluye de manera contundente que el modelo basado en XGBoost es el más adecuado y fue extendido al problema de clasificación.
- Los modelos de regresión y clasificación parecen ser complementarios entre sí. Es decir, si se logra justificar experimentalmente el modelo de clasificación, esto indicaría que es posible explorar la reversibilidad electroquímica a partir de la energía de reorganización. Por lo tanto, el modelo de regresión que predice la energía, también podría determinar si la molécula es reversible o no al calcular el cociente.

9.1. Perspectivas

- Realizar las optimizaciones tanto los modelos de regresión como el de clasificación para mejorar su rendimiento y confiabilidad.
- Construir una escala de reversibilidad con información experimental, como los radicales libres, que ampliarían el conjunto de moléculas irreversibles.
- Se explorará el modelo de redes neuronales de grafos.

Capítulo 10

Referencias

[1] Adam Z Weber, Matthew M Mench, Jeremy P Meyers, Philip N Ross, Jeffrey T Gostick, and Qinghua Liu. Redox flow batteries: a review. *Journal of applied electrochemistry*, 41:1137–1164, 2011.

[2] Piergiorgio Alotto, Massimo Guarnieri, and Federico Moro. Redox flow batteries for the storage of renewable energy: A review. *Renewable and Sustainable Energy Reviews*, 29:325–335, 2014.

[3] Jian Luo, Bo Hu, Maowei Hu, Yu Zhao, and T Leo Liu. Status and prospects of organic redox flow batteries toward sustainable energy storage. *ACS Energy Letters*, 4(9):2220–2240, 2019.

[4] Solene Gentil, Danick Reynard, and Hubert H Girault. Aqueous organic and redox-mediated redox flow batteries: a review. *Current Opinion in Electrochemistry*, 21:7–13, 2020.

[5] Puiki Leung, Akeel A Shah, L Sanz, C Flox, JR Morante, Q Xu, MR Mohamed, C Ponce De León, and FC Walsh. Recent developments in organic redox flow batteries: A critical review. *Journal of Power Sources*, 360:243–283, 2017.

[6] Eduardo Martínez-González, Humberto G Laguna, Mariano Sánchez-Castellanos, Sergio S Rozenel, Víctor M Ugalde-Saldivar, and Carlos Amador-Bedolla. Kinetic properties of aqueous organic redox flow battery anolytes using the marcus–hush theory. *ACS Applied Energy Materials*, 3(9):8833–8841, 2020.

[7] David Reinsel-John Gantz-John Rydning, John Reinsel, and John Gantz. The digitization of the world from edge to core. *Framingham: International Data Corporation*, 16:1–28, 2018.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[9] Jordi de la Torre. Modelos generativos basados en mecanismos de difusión, 2023.

[10] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[11] Yun-Fei Shi, Zheng-Xin Yang, Sicong Ma, Pei-Lin Kang, Cheng Shang, P Hu, and Zhi-Pan Liu. Machine learning for chemistry: basics and applications. *Engineering*, 2023.

[12] Marwin HS Segler and Mark P Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal*, 23(25):5966–5971, 2017.

[13] Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science*, 3(10):1103–1113, 2017.

[14] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

[15] Yaolong Zhang, Ce Hu, and Bin Jiang. Embedded atom neural network potentials: Efficient and accurate machine learning with a physically inspired representation. *The journal of physical chemistry letters*, 10(17):4962–4967, 2019.

[16] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B—Condensed Matter and Materials Physics*, 87(18):184115, 2013.

[17] Omri D Abarbanel and Geoffrey R Hutchison. Machine learning to accelerate screening for marcus reorganization energies. *The Journal of Chemical Physics*, 155(5), 2021.

[18] Cheng-Han Li and Daniel P Tabor. Reorganization energy predictions with graph neural networks informed by low-cost conformers. *The Journal of Physical Chemistry A*, 127(15):3484–3489, 2023.

[19] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B—Condensed Matter and Materials Physics*, 87(18):184115, 2013.

[20] P. Atkins, J. de Paula, and J. Keeler. *Atkins Physical Chemistry 11e: Volume 3: Molecular Thermodynamics and Kinetics*. Oxford University Press, 2019.

[21] Pedro Ponce. *Inteligencia artificial: con aplicaciones a la ingeniería*. Alpha Editorial, 2010.

[22] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[23] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386, 2020.

[24] William Bort, Igor I Baskin, Timur Gimadiev, Artem Mukanov, Ramil Nugmanov, Pavel Sidorov, Gilles Marcou, Dragos Horvath, Olga Klimchuk, Timur Madzhidov, et al. Discovery of novel chemical reactions by deep generative recurrent neural network. *Scientific reports*, 11(1):3178, 2021.

[25] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[26] XGBoost Developers. Xgboost documentation: Model introduction, 2024. Accessed: 2024-12-18.

[27] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[28] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[29] Cosma Shalizi. *Advanced data analysis from an elementary point of view*. Citeseer, 2013.

[30] Henry F. Kaiser. A second generation little jiffy. *Psychometrika*, 35(4):401–415, 1970.

[31] Henry F Kaiser and John Rice. Little jiffy, mark iv. *Educational and psychological measurement*, 34(1):111–117, 1974.

[32] Edward E Cureton and Ralph B D'Agostino. *Factor analysis: An applied approach*. Psychology press, 2013.

[33] Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895.

[34] Joseph Lee Rodgers and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[35] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[36] Ron N Forthofer and Robert G Lehnen. Rank correlation methods. In *Public program analysis: a new categorical data approach*, pages 146–163. Springer, 1981.

[37] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769 – 2794, 2007.

[38] John G Graveel, Lee E Sommers, and Darrell W Nelson. Decomposition of benzidine, α-naphthylamine, and p-toluidine in soils. Technical report, Wiley Online Library, 1986.

Capítulo 11

Anexo

11.1. Criterio 1 (Bz)

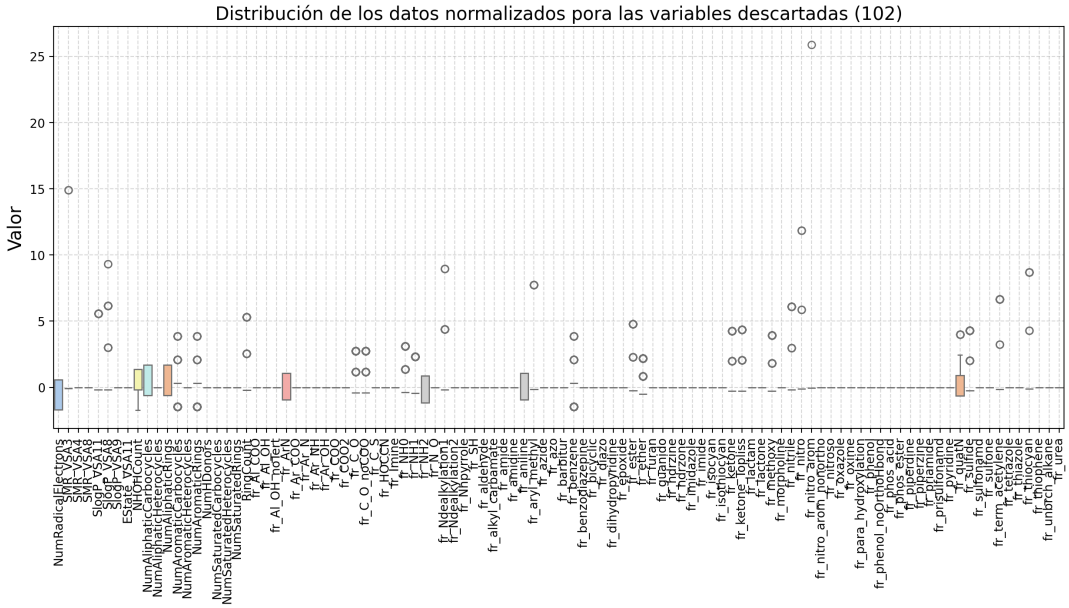


Figura 11.1: Bz-DatOx

11.2. Criterio 2 (Bz)

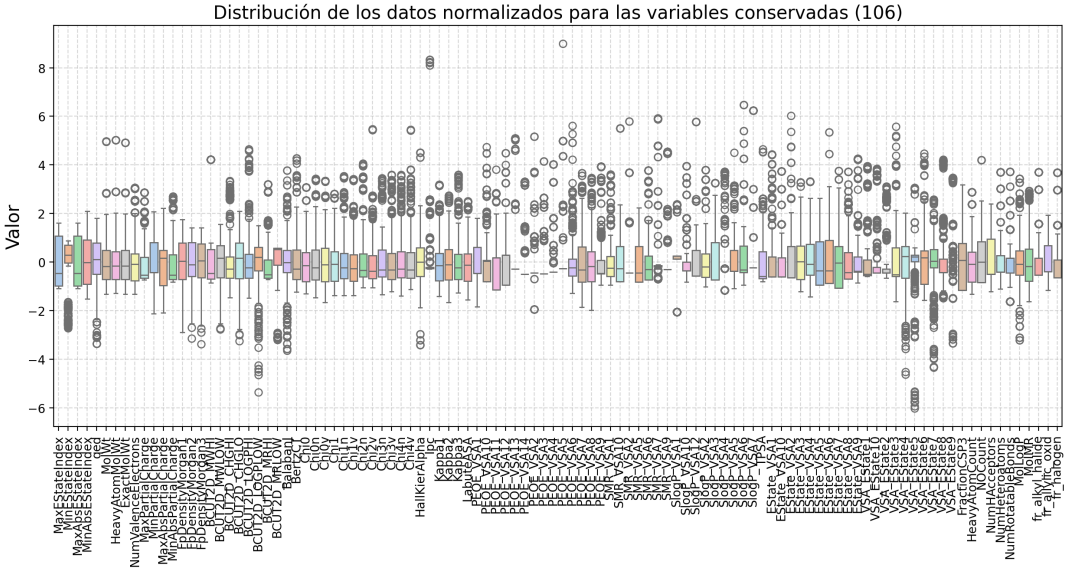


Figura 11.2: Bz-DatOx

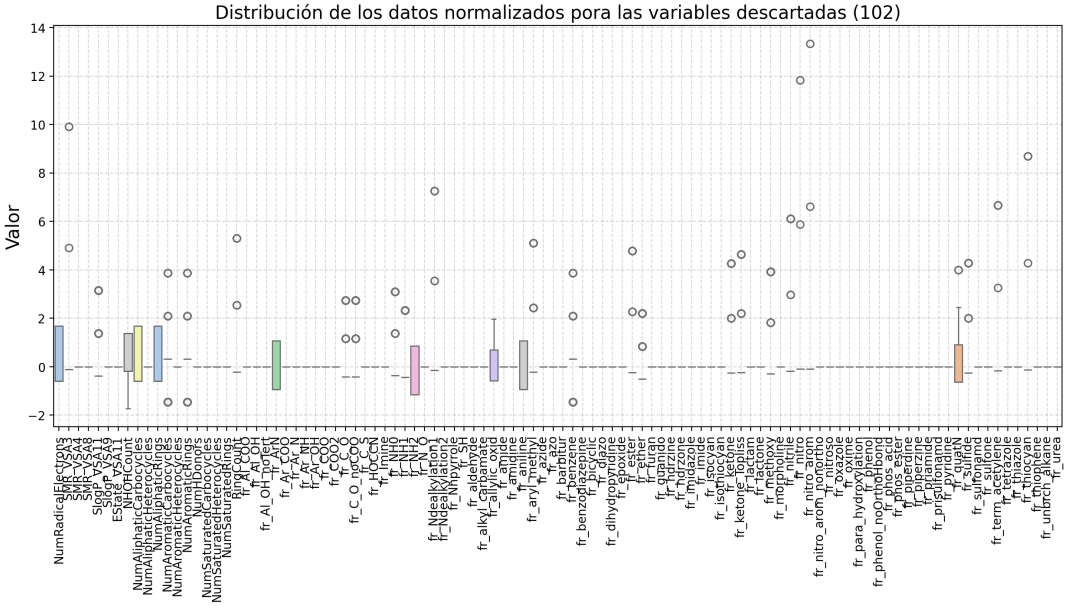


Figura 11.3: Bz-DatRed

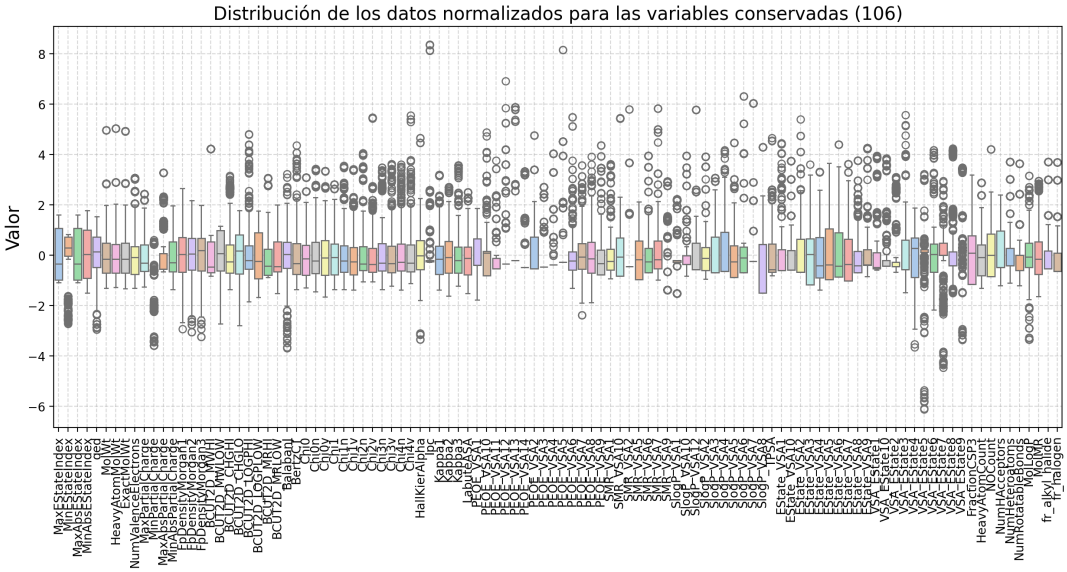
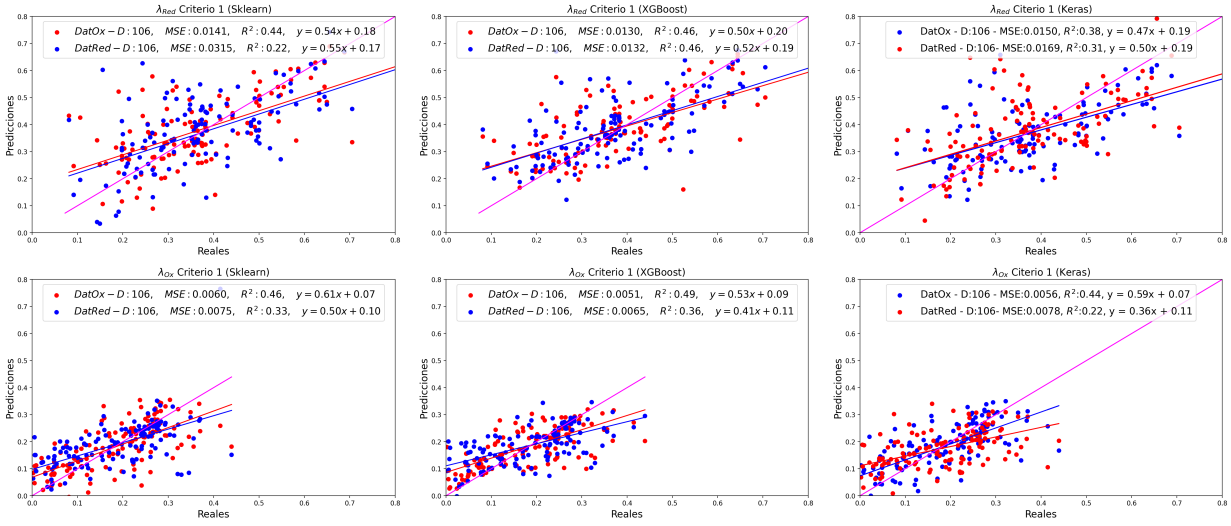
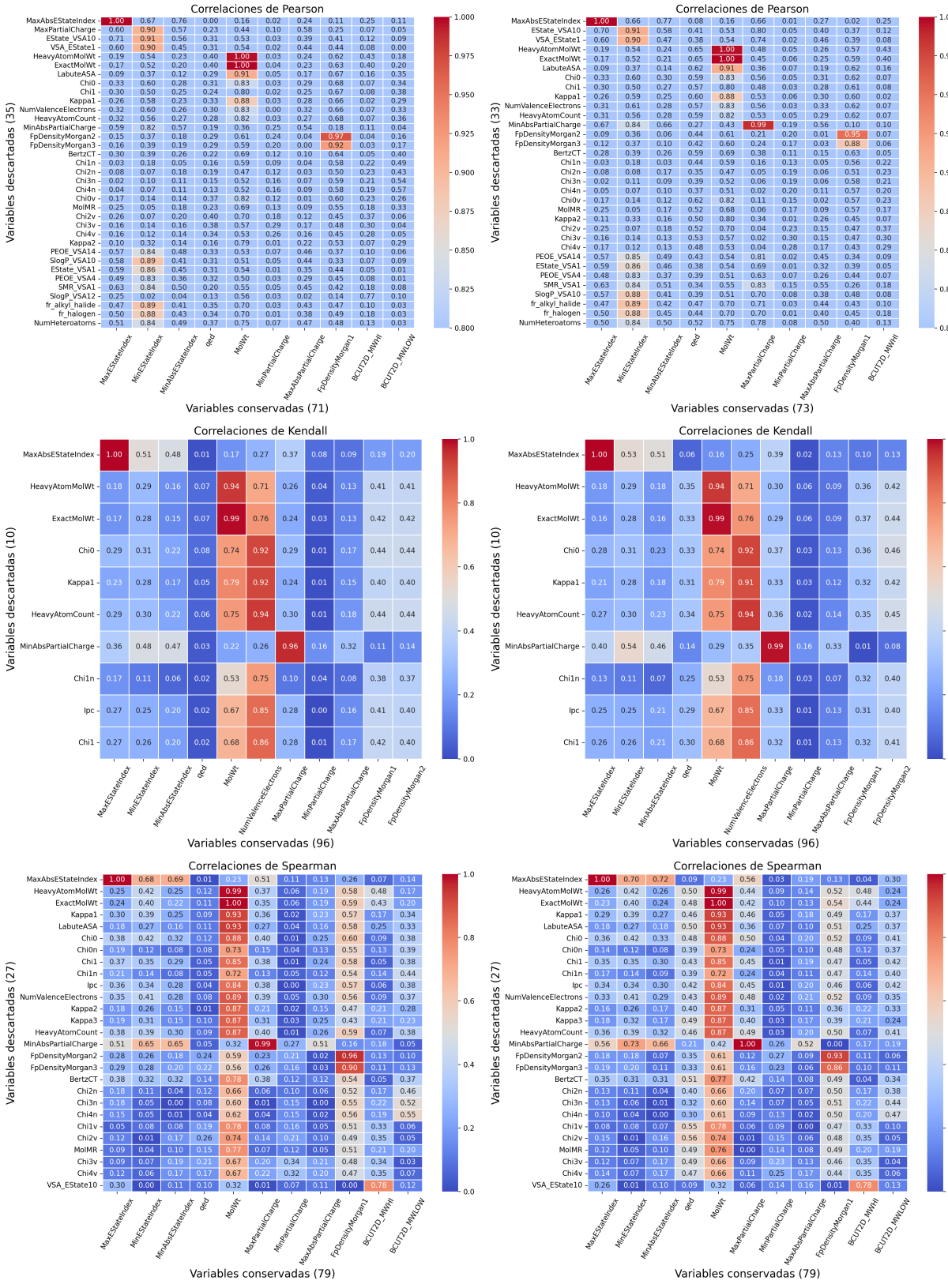


Figura 11.4: Bz-DatRed



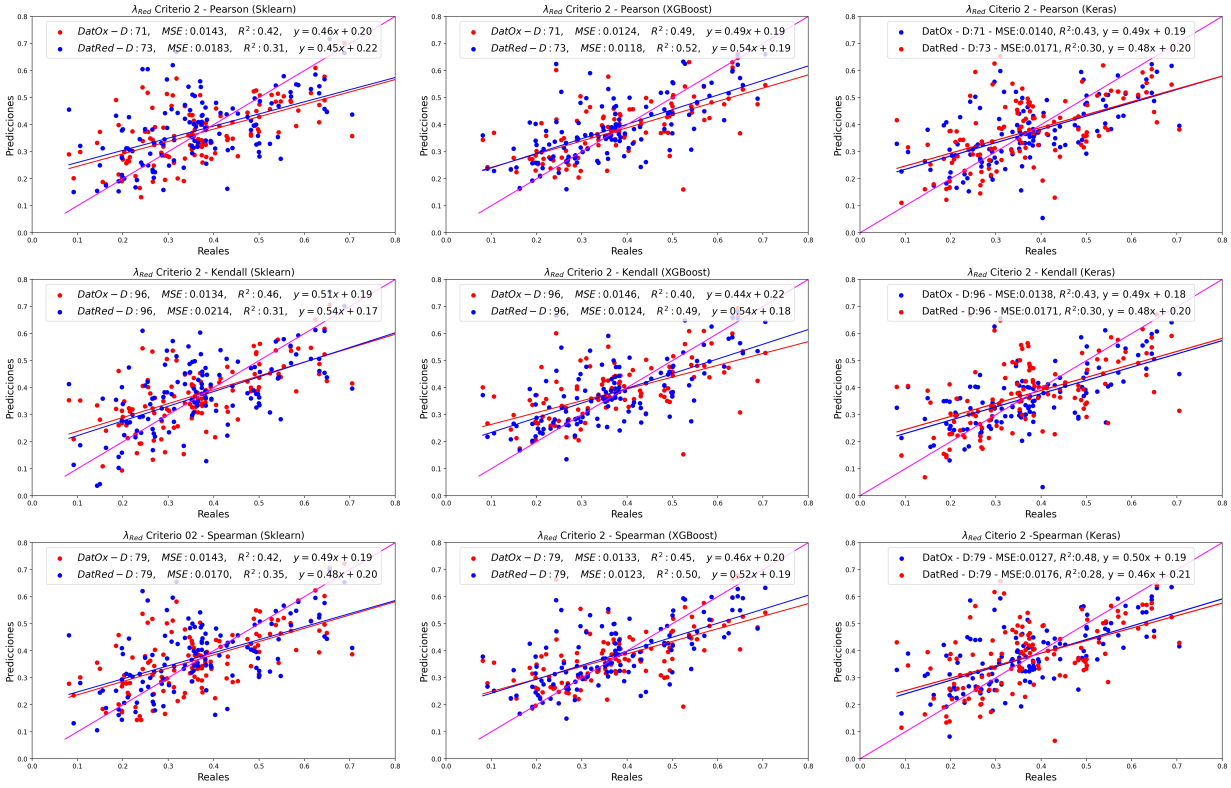
(a) RLM **(b)** EAD **(c)** RN

Figura 11.5: Gráficas de dispersión (Bz) para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} y λ_{Ox} repectivamente considerando el Criterio 1.



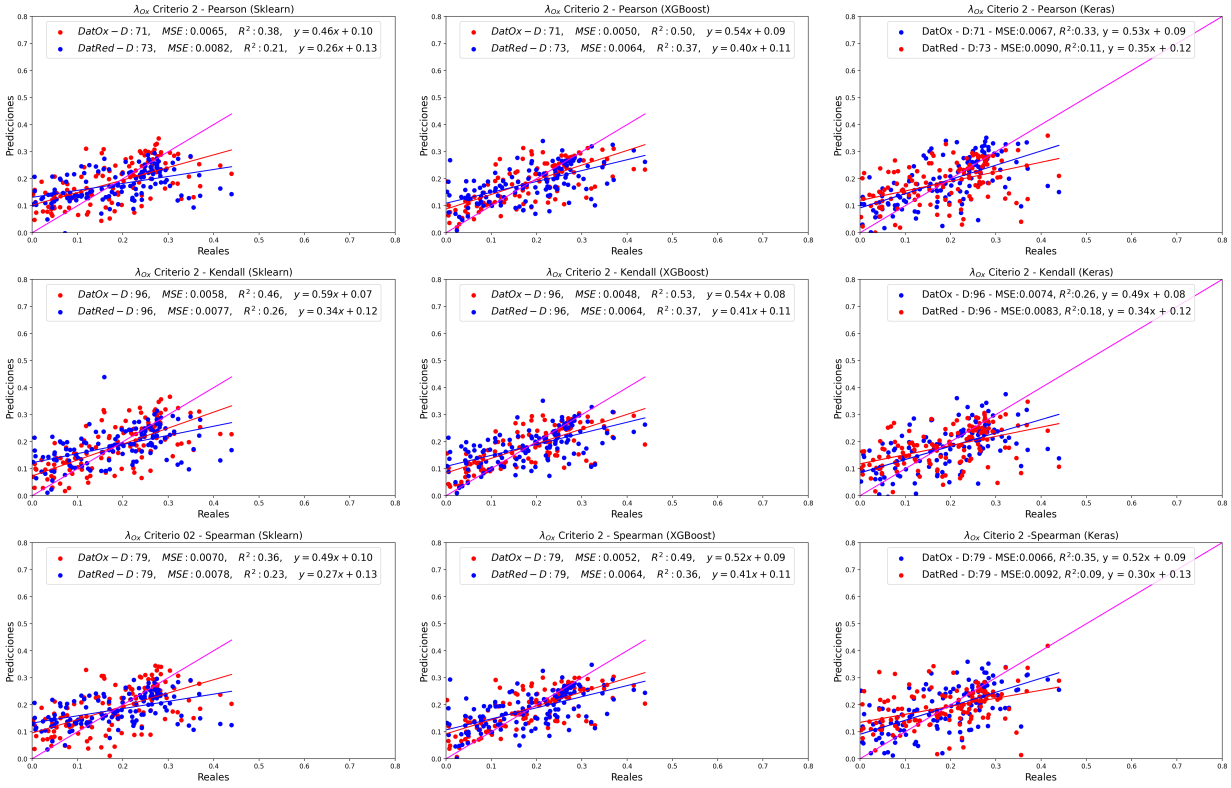
(a) DatOx **(b)** DatRed

Figura 11.6: Matrices de correlación (para la Bz) para distintos métodos y base de datos.



(a) RLM **(b)** EAD **(c)** RN

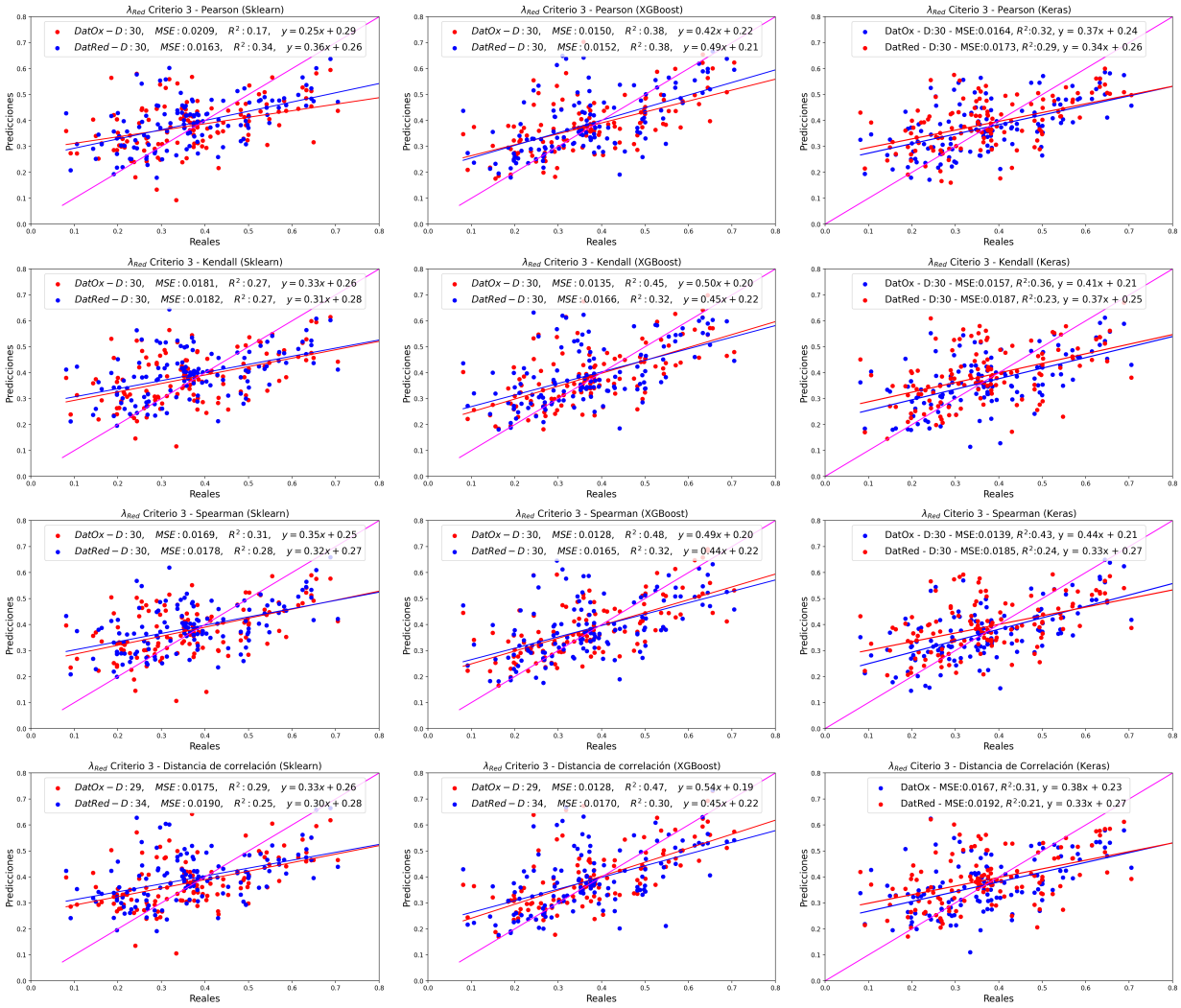
Figura 11.7: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} considerando el Criterio 2 y las distintas formulaciones de correlación.



(a) RLM **(b)** EAD **(c)** RN

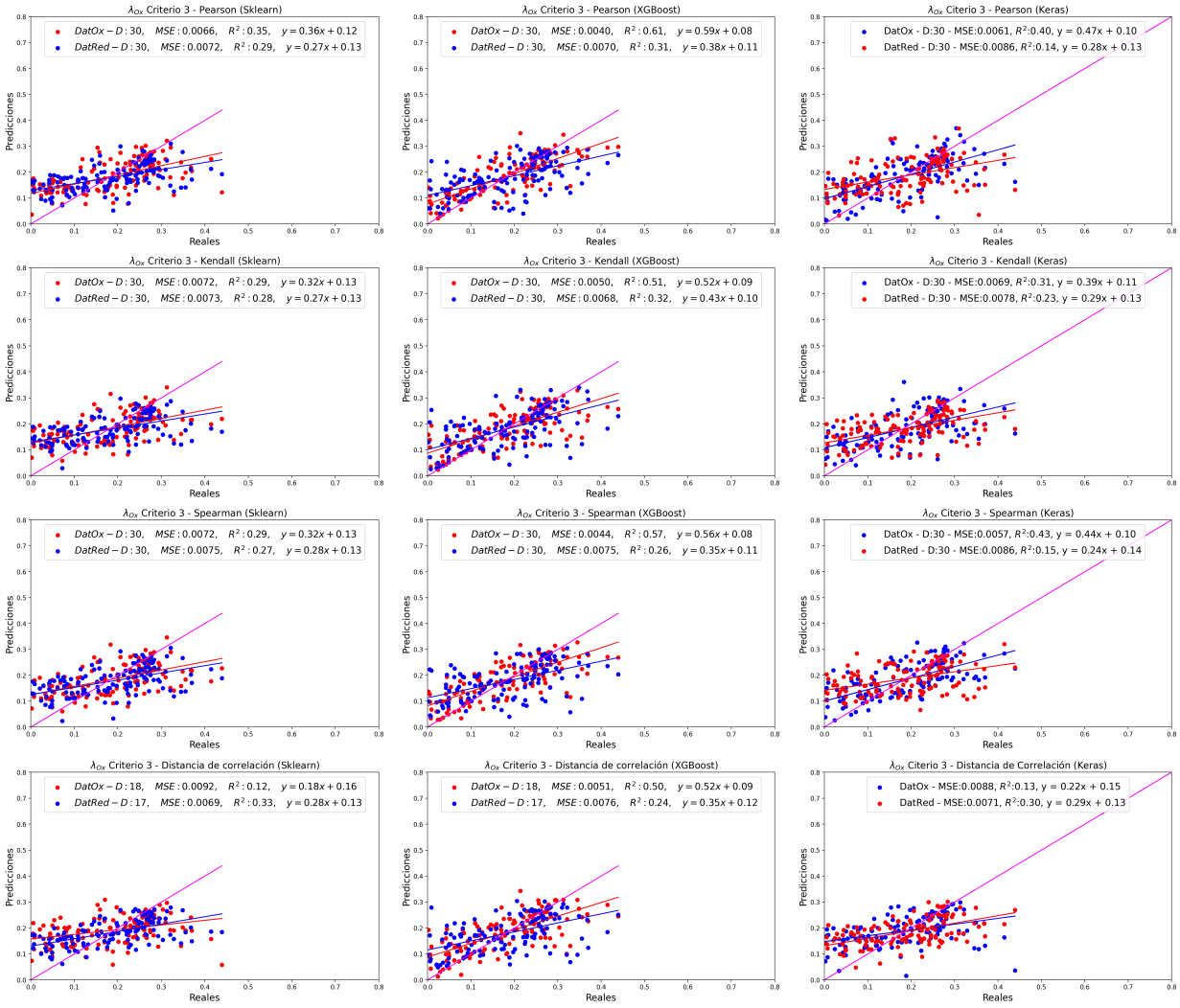
Figura 11.8: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Ox} considerando el Criterio 2 y las distintas formulaciones de correlación.

11.3. Criterio 3 (Bz)



(a) RLM **(b)** EAD **(c)** RN

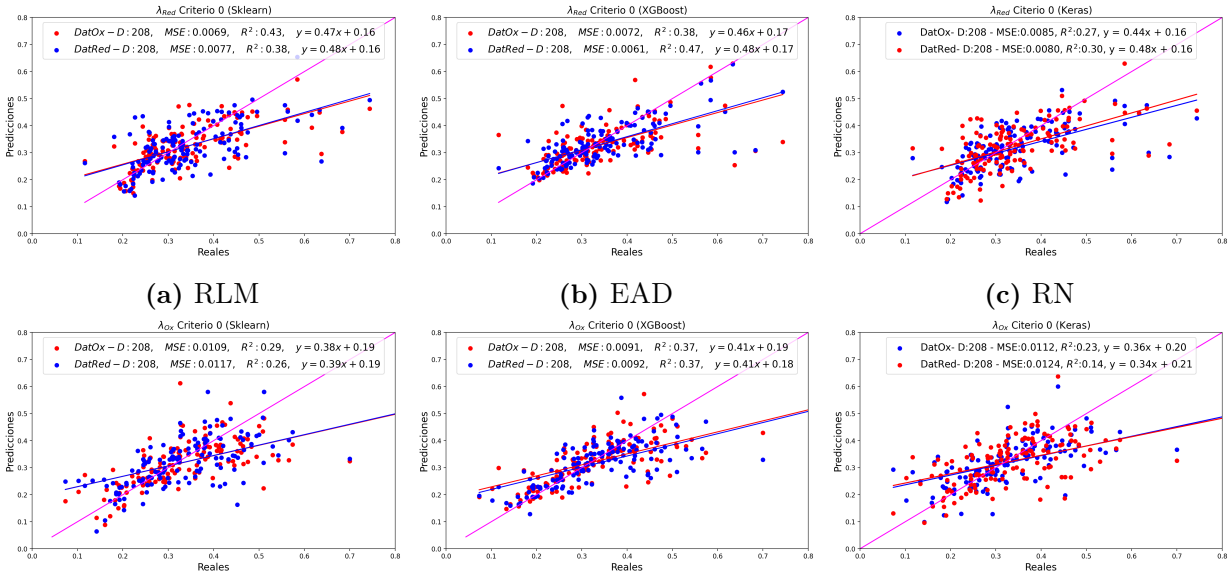
Figura 11.9: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} considerando el Criterio 3 y las distintas formulaciones de correlación.



(a) RLM **(b)** EAD **(c)** RN

Figura 11.10: Gráficas de dispersión (de la familia de las Bz's) para los tres modelos de ajuste que predicen la energía de reorganización λ_{Ox} considerando el Criterio 3 y las distintas formulaciones de correlación.

11.4. Criterio 0 (MV)



(a) RLM **(b)** EAD **(c)** RN

(d) RLM **(e)** EAD **(f)** RN

Figura 11.11: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Ox}, considerando el Criterio 0.

11.5. Criterio 1 (MV)

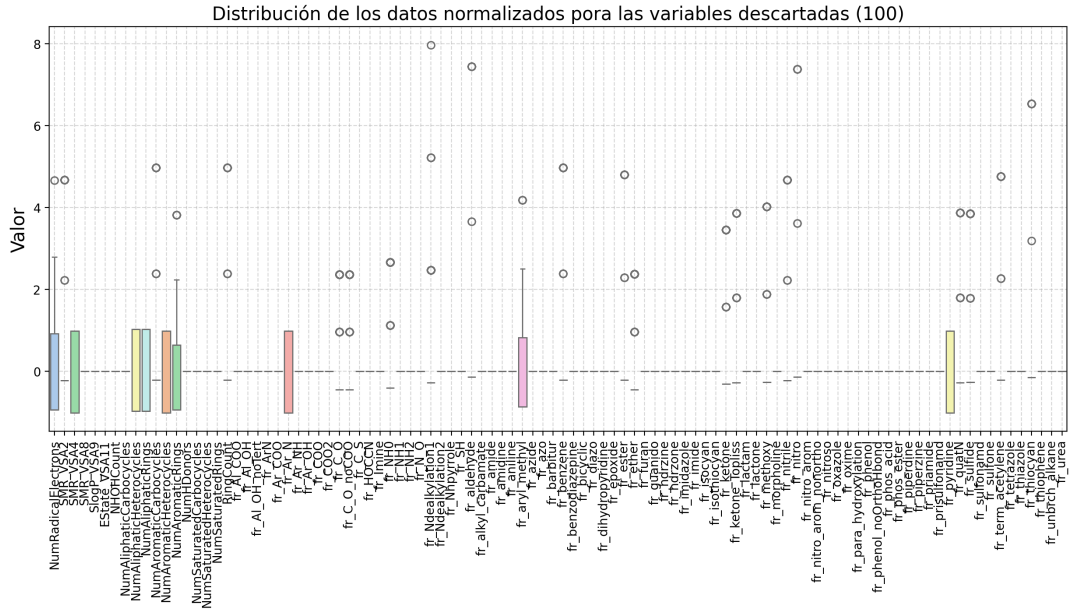


Figura 11.12: MV-DatOx

11.6. Criterio 2 (MV)

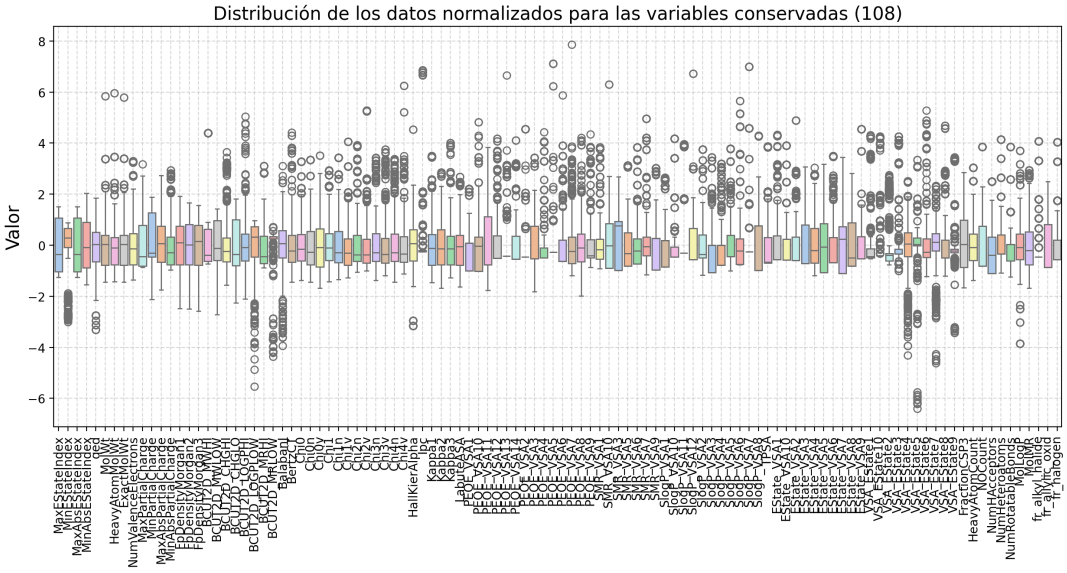


Figura 11.13: MV-DatOx

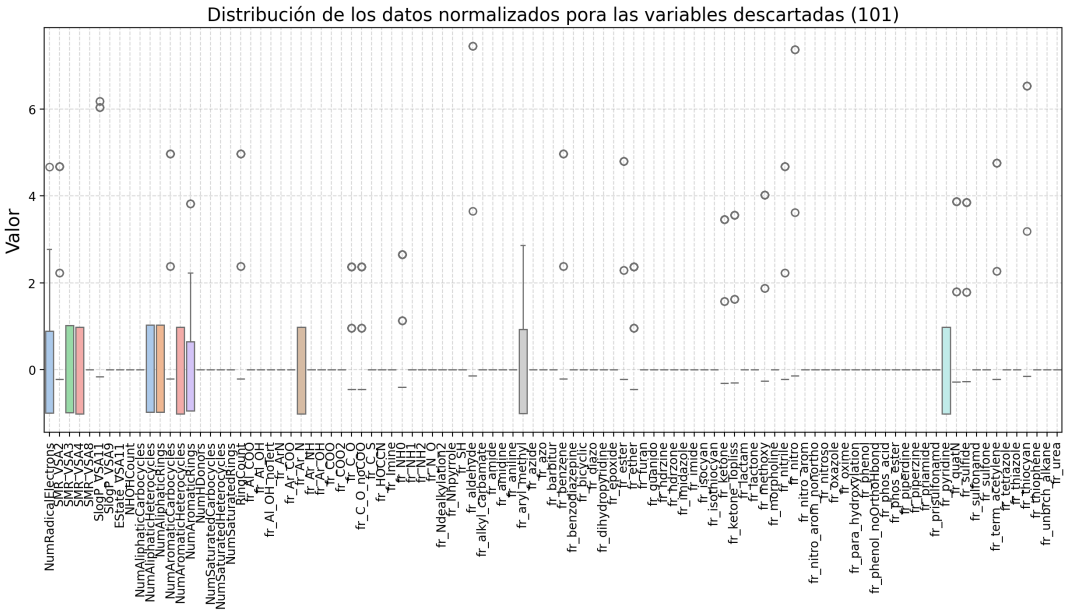


Figura 11.14: MV-DatRed

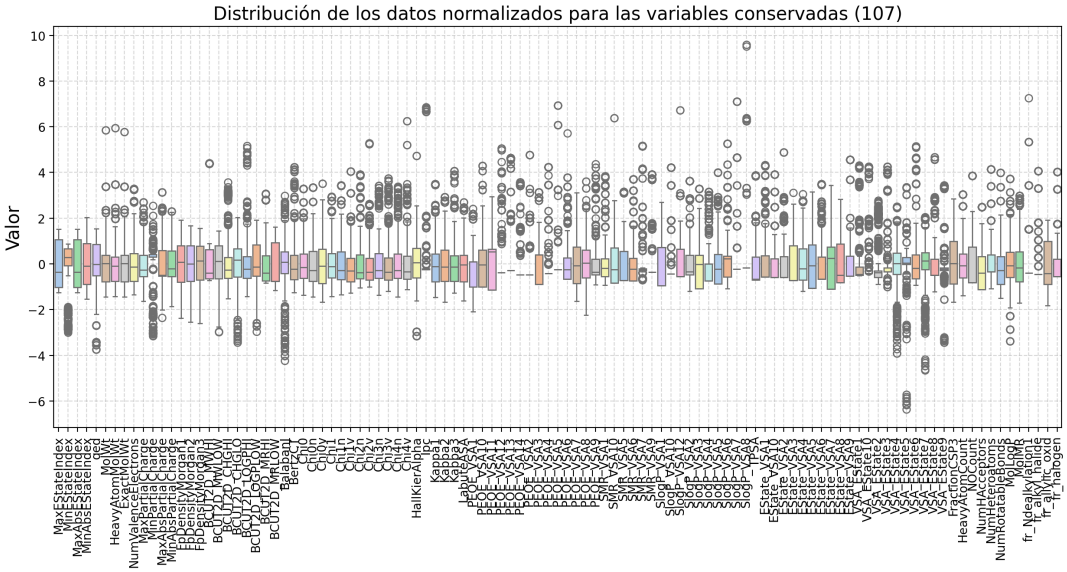
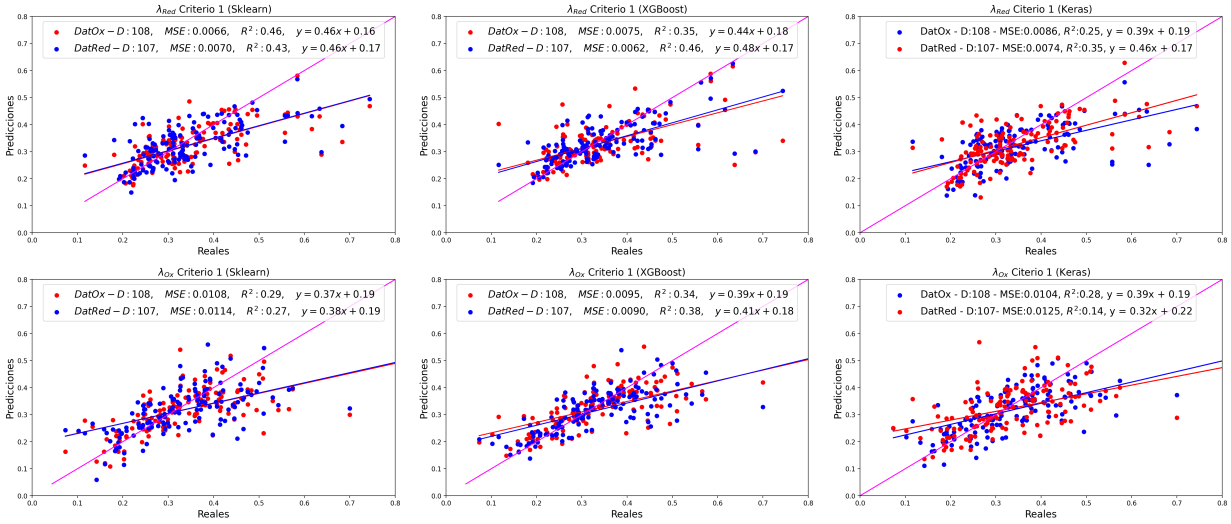
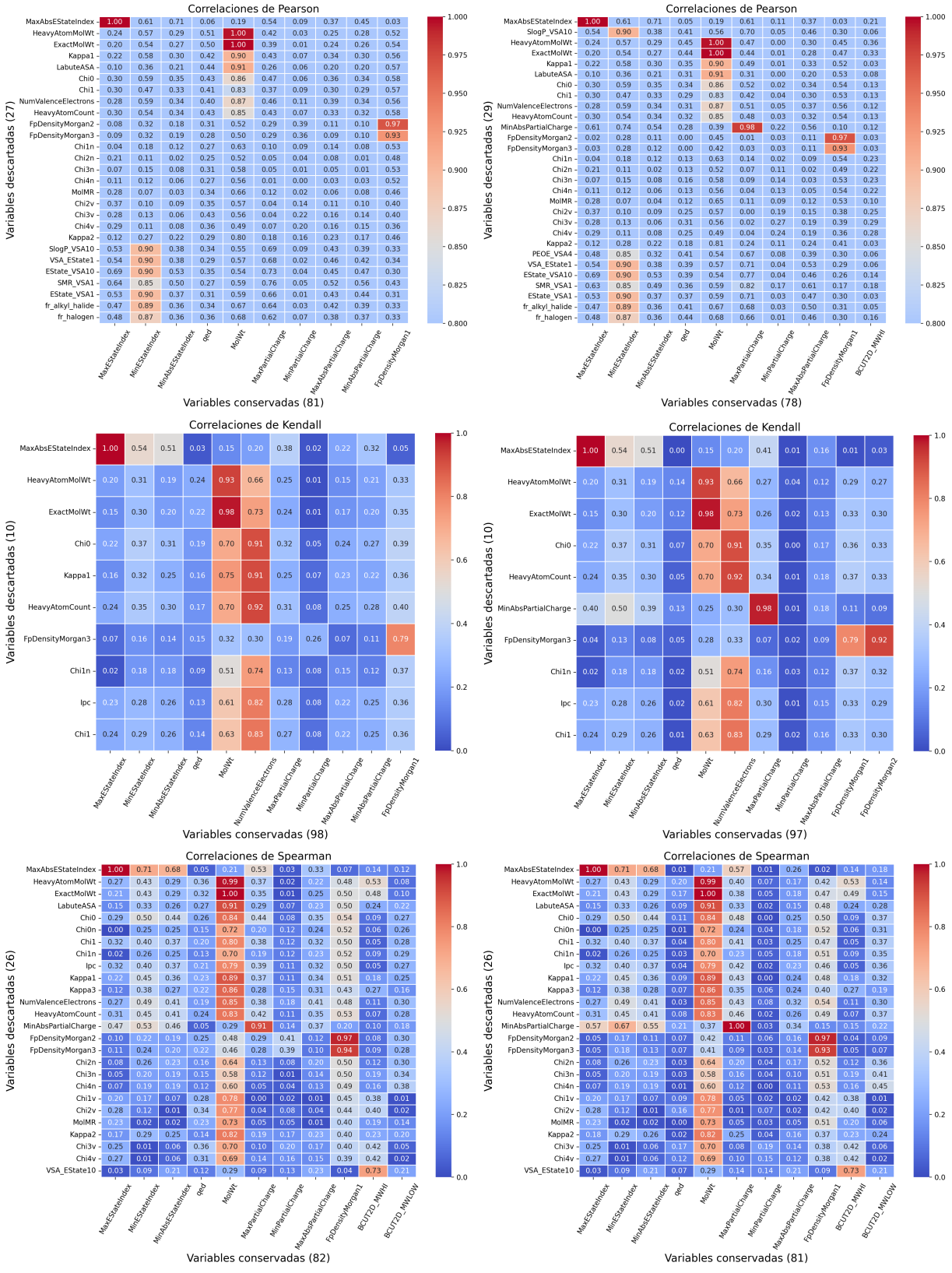


Figura 11.15: MV-DatRed



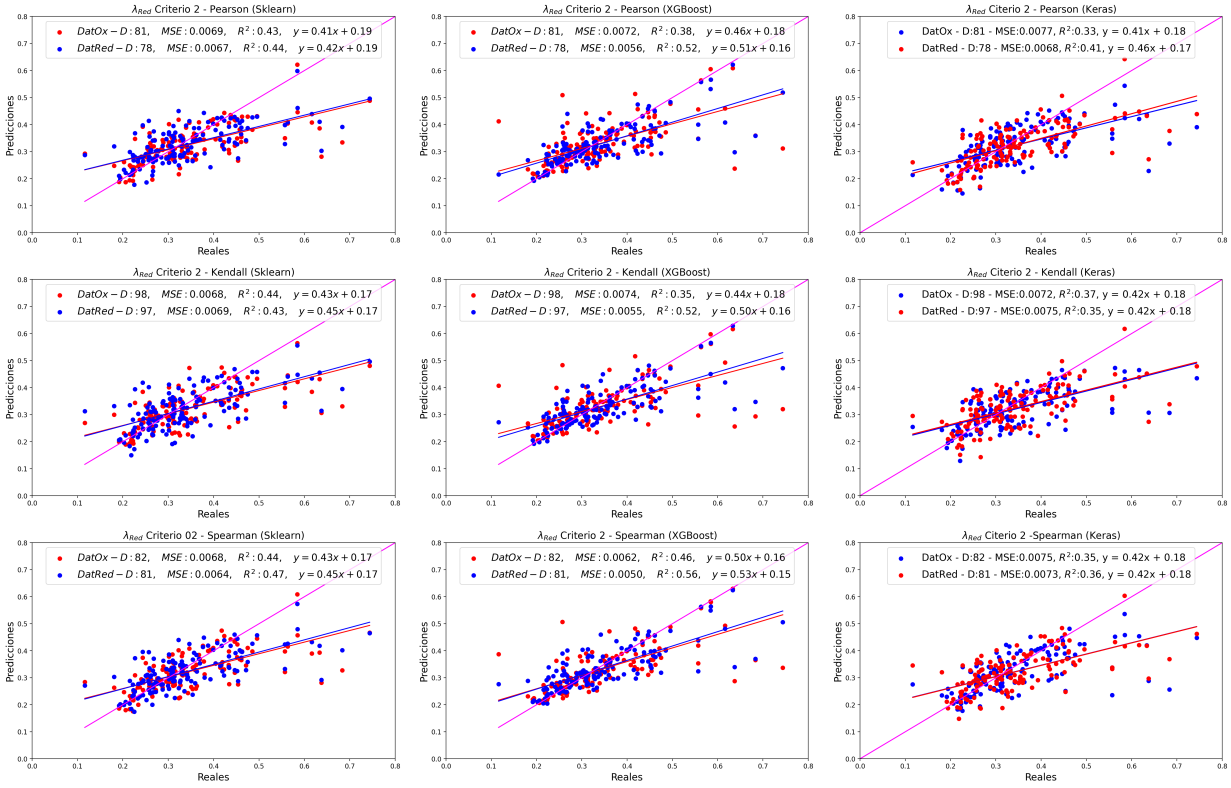
(a) RLM **(b)** EAD **(c)** RN

Figura 11.16: Gráficas de dispersión (MV) para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} y λ_{Ox} repectivamente considerando el Criterio 1.



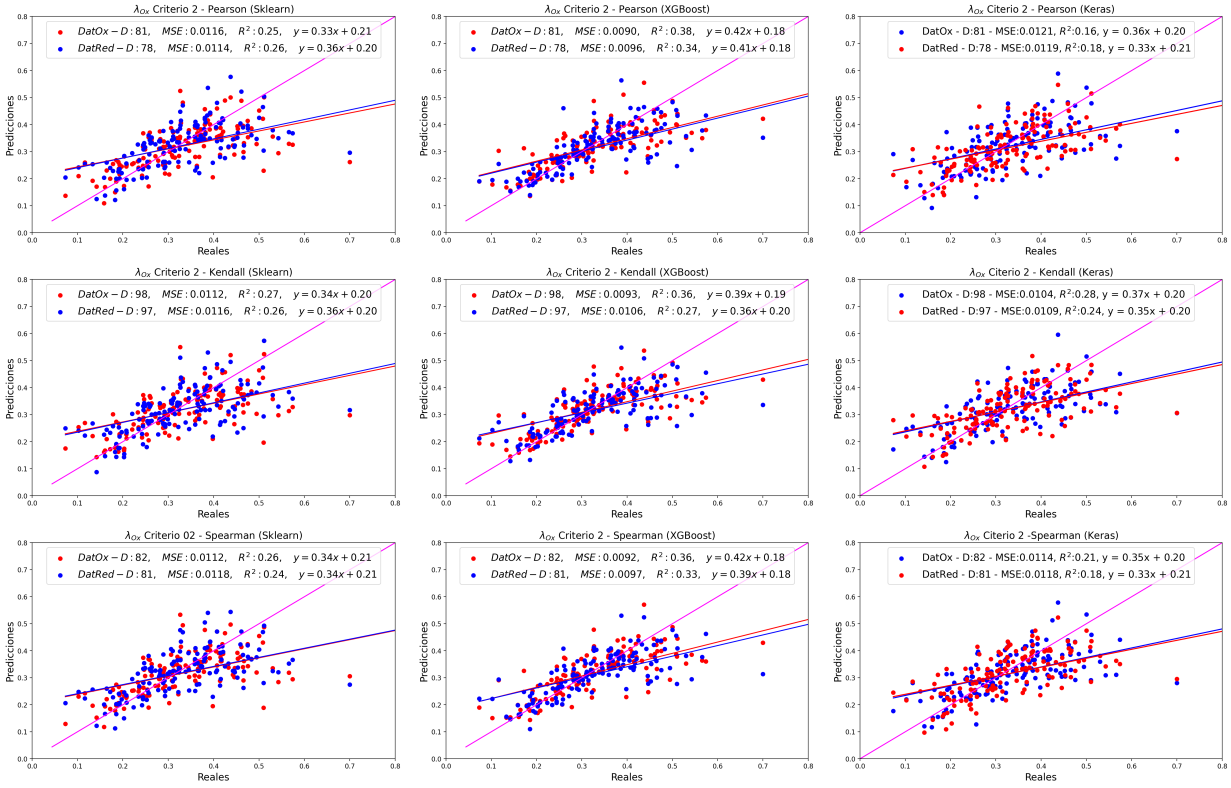
(a) DatOx **(b)** DatRed

Figura 11.17: Matrices de correlación (para la MV) para distintos métodos y base de datos.



(a) RLM **(b)** EAD **(c)** RN

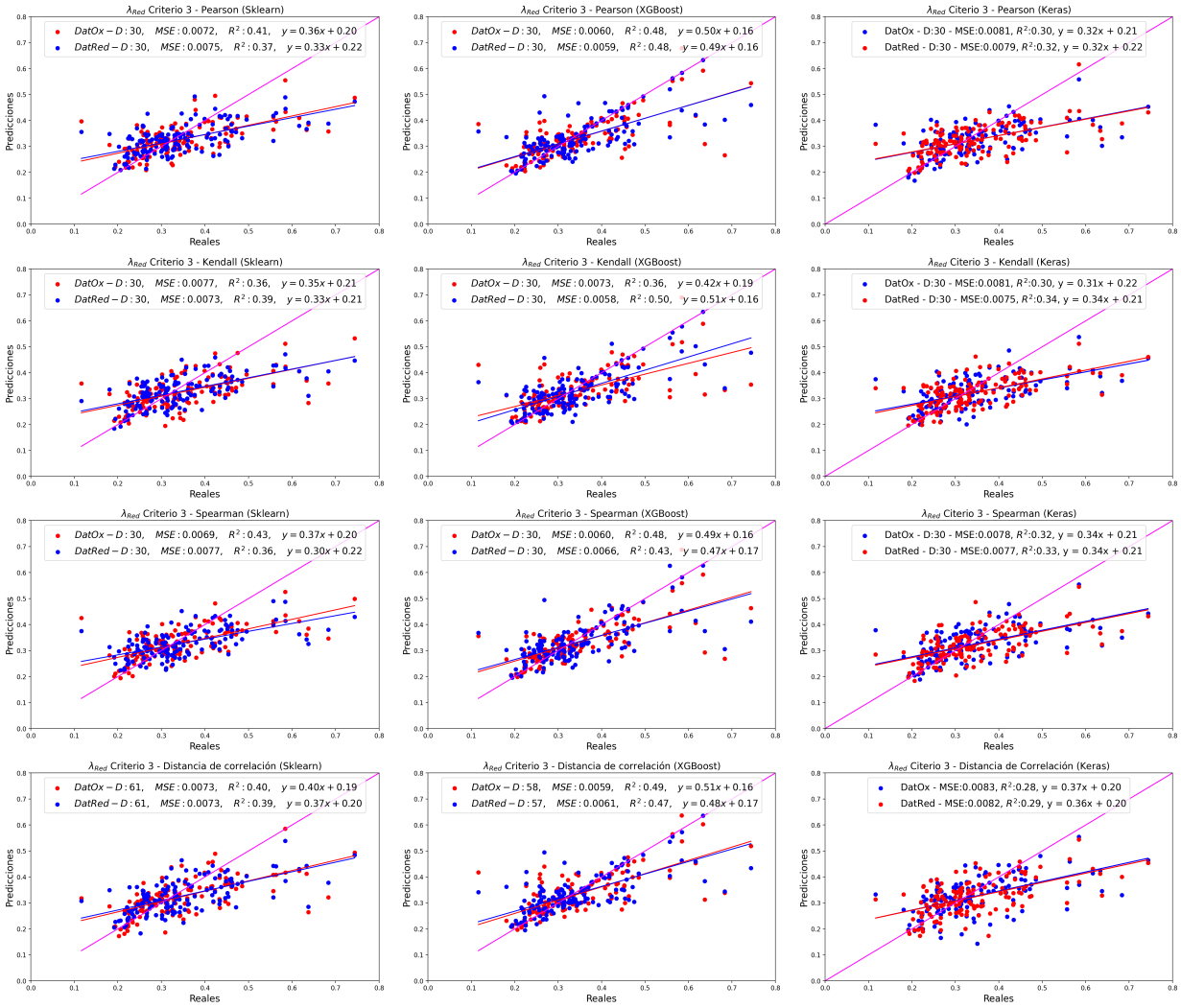
Figura 11.18: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} considerando el Criterio 2 y las distintas formulaciones de correlación.



(a) RLM **(b)** EAD **(c)** RN

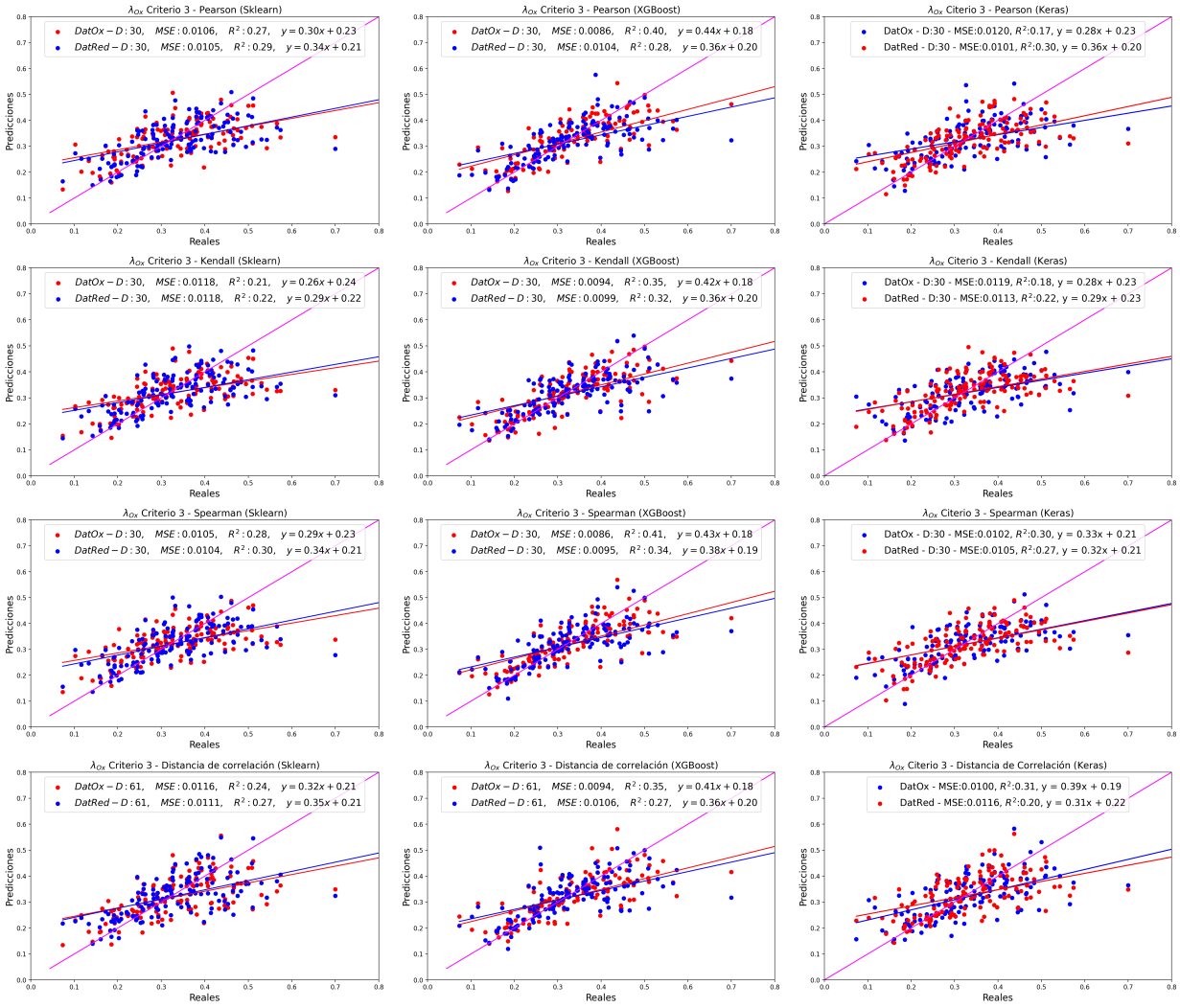
Figura 11.19: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Ox} considerando el Criterio 2 y las distintas formulaciones de correlación.

11.7. Criterio 3 (MV)



(a) RLM **(b)** EAD **(c)** RN

Figura 11.20: Gráficas de dispersión para los tres modelos de ajuste que predicen la energía de reorganización λ_{Red} considerando el Criterio 3 y las distintas formulaciones de correlación.



(a) RLM **(b)** EAD **(c)** RN

Figura 11.21: Gráficas de dispersión (de la familia de las MV's) para los tres modelos de ajuste que predicen la energía de reorganización λ_{Ox} considerando el Criterio 3 y las distintas formulaciones de correlación.

11.8. Todas las familias

11.9. Criterio 1 (Todas)

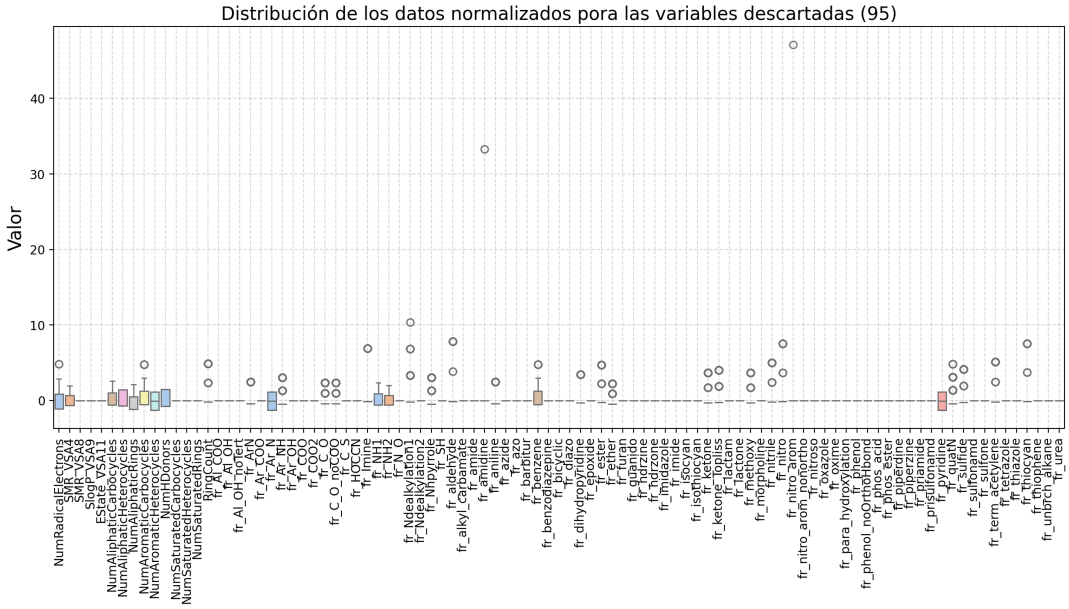


Figura 11.22: ALL-DatOx

11.10. Criterio 2 (Todas)

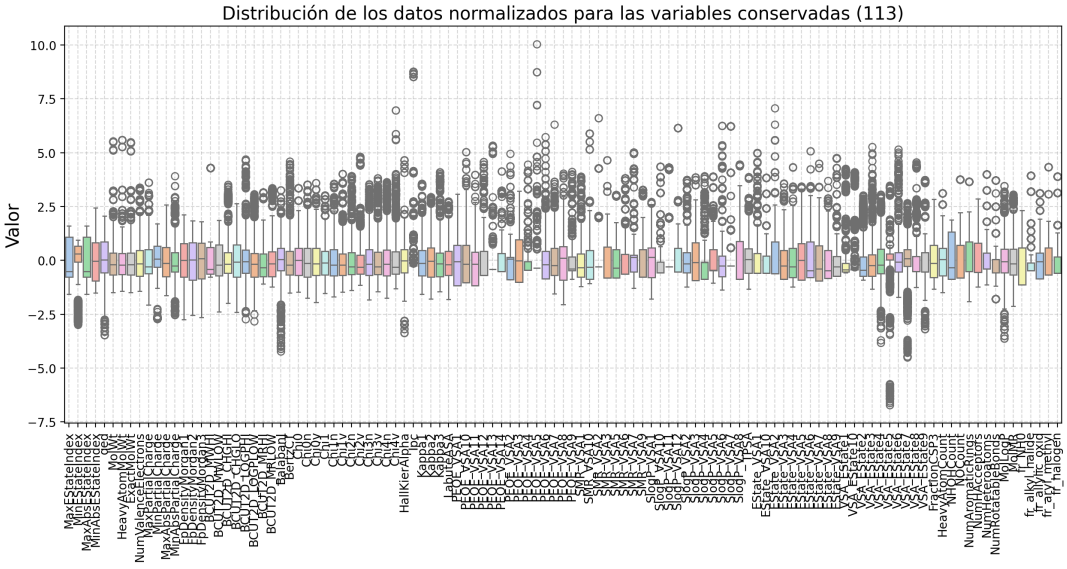


Figura 11.23: ALL-DatOx

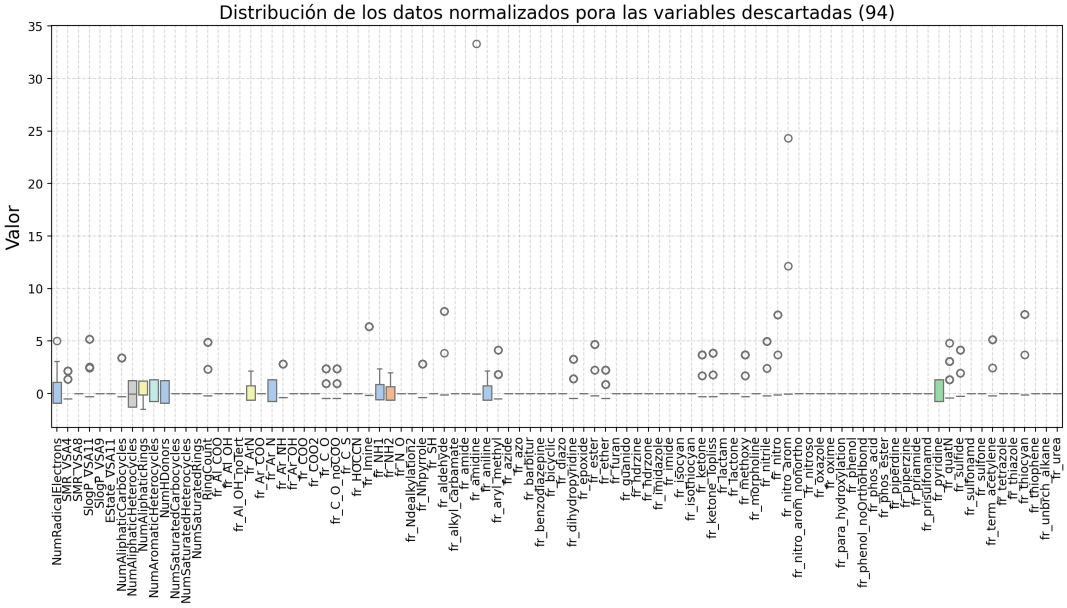


Figura 11.24: ALL-DatRed

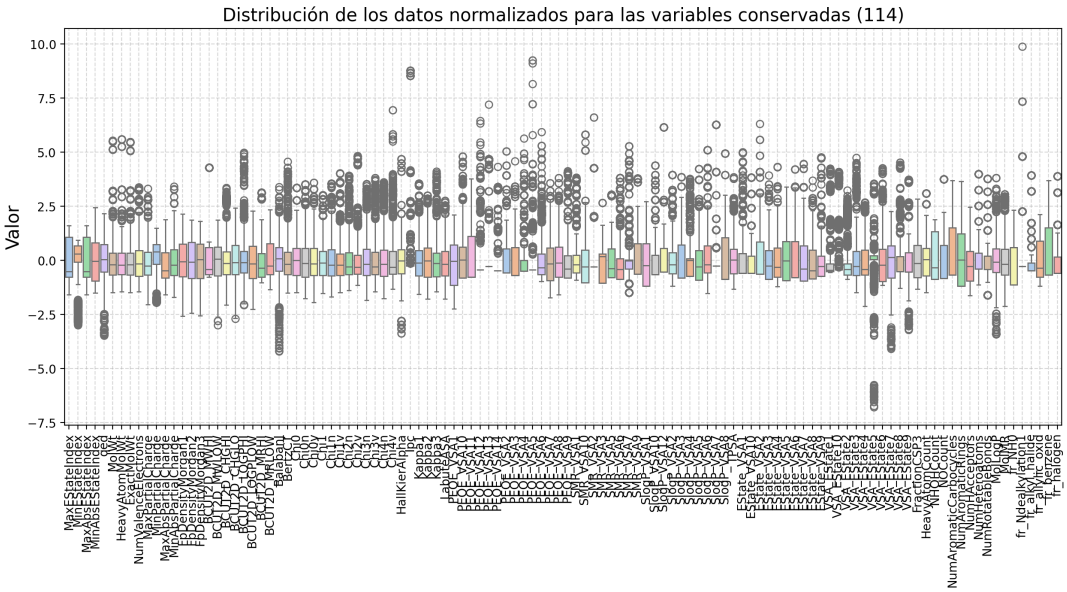
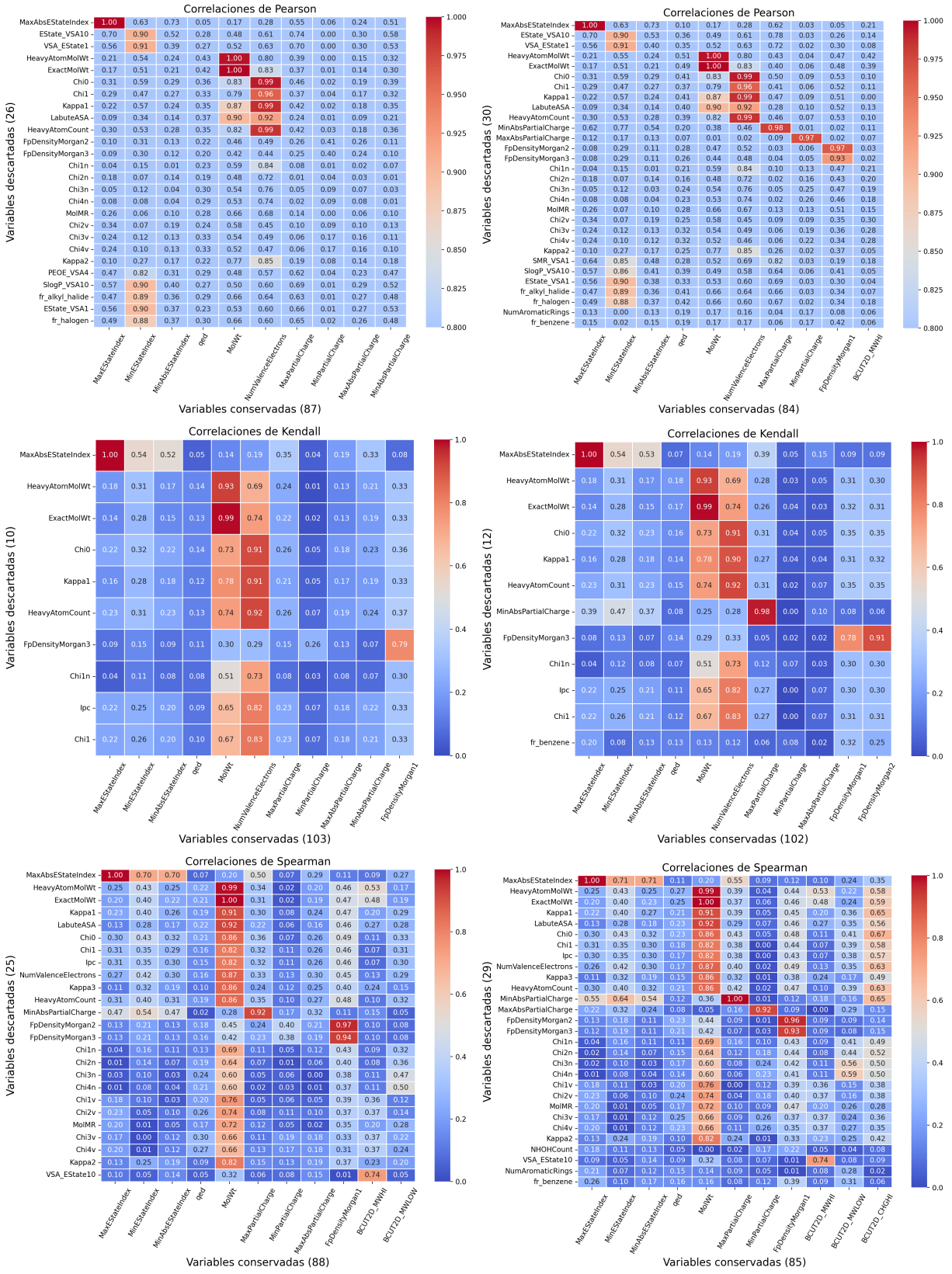


Figura 11.25: ALL-DatRed



(a) DatOx **(b)** DatRed

Figura 11.26: Matrices de correlación para el conjunto de todas las familias para distintos métodos y base de datos.

Capítulo 12

Apéndice A

Descriptores generales

No.	Descriptor	Definición
1	MaxEStateIndex	Representa la máxima electronegatividad entre dos átomos de una molécula.
2	MinEStateIndex	Representa la mínima electronegatividad entre los átomos de la molécula.
3	MaxAbsEStateIndex	Representa la máxima electronegatividad absoluta entre dos átomos de una molécula.
4	MinAbsEStateIndex	Representa la mínima electronegatividad absoluta entre los átomos de la molécula.
5	qed	Indica la extensión de la deslocalización de la energía en la molécula.
6	MolWt	Peso molecular promedio de la molécula.
7	HeavyAtomMolWt	Peso molecular promedio de la molécula ignorando los hidrógenos.
8	ExactMolWt	Peso molecular exacto considerando los isótopos.
9	NumValenceElectrons	Número de electrones de valencia que tiene la molécula.
10	NumRadicalElectrons	Número de electrones radicales que tiene la molécula.
11	MinAbsPartialCharge	Carga parcial absoluta mínima.
12	MaxAbsPartialCharge	Carga parcial absoluta máxima.
13	MinPartialCharge	Carga parcial mínima.
14	MaxPartialCharge	Carga parcial máxima.
15	FpDensityMorgan1	Huella de Morgan, radio 1.
16	FpDensityMorgan2	Huella de Morgan, radio 2.
17	FpDensityMorgan3	Huella de Morgan, radio 3.

Tabla 12.1: Definición de los descriptores de RDKit.

18	FractionCSP3	Fracción de átomos de carbono que están hibridizados como SP^3.
19	HeavyAtomCount	Número de átomos pesados en una molécula.
20	NHOHCount	Número de grupos NH o OH.
21	NOCount	Número de átomos de nitrógeno y oxígeno.
22	NumAliphaticCarbocycles	Número de carbociclos alifáticos (que contienen al menos un enlace no aromático) en una molécula.
23	NumAliphaticHeterocycles	Número de heterociclos alifáticos (que contienen al menos un enlace no aromático) en una molécula.
24	NumAliphaticRings	Número de anillos alifáticos (que contienen al menos un enlace no aromático) en una molécula.
25	NumAromaticCarbocycles	Número de carbociclos aromáticos en una molécula.
26	NumAromaticHeterocycles	Número de heterociclos aromáticos en una molécula.
27	NumAromaticRings	Número de anillos aromáticos en una molécula.
28	NumHAcceptors	Número de aceptores de enlaces de hidrógeno.
29	NumHDonors	Número de donadores de enlaces de hidrógeno.
30	NumHeteroatoms	Número de heteroátomos.
31	NumRotatableBonds	Número de enlaces rotables.
32	NumSaturatedCarbocycles	Número de carbociclos saturados en una molécula.
33	NumSaturatedHeterocycles	Número de heterociclos saturados en una molécula.
34	NumSaturatedRings	Número de anillos saturados en una molécula.
35	RingCount	Número total de anillos en una molécula.
36	MolLogP	Logaritmo del coeficiente de partición octanol-agua de la molécula.
37	MolMR	Masa molar relativa.

Tabla 12.2: Definición de los descriptores de RDKit.

Descriptores topológicos

El grado pesado, d_i, de un átomo es el número de átomos pesados a los que está unido al átomo i.

Para los átomos pesados i se define

$$v_i = \frac{p_i - h_i}{Z_i - p_i - 1}$$

donde p_i es el número de electrones de valencia s y p del átomo i y h_i es el número de hidrógenos a los que está (o debería estar) unido, este incluye todos los átomos de hidrógeno que son necesarios para completar la valencia.

Los índices de conectividad chi de Kier y Hall se calculan a partir del grado del átomo pesado d_i (es decir, el número de vecinos pesados) y v_i.

Los índices de forma molecular kappa de Kier y Hall comparan el grafo molecular con grafos moleculares mínimos y máximos, y pretenden capturar diferentes aspectos de la forma molecular. En esta descripción:

- n denota el número de átomos en el grafo de hidrógeno suprimido.
- m es el número de enlaces en el grafo de hidrógeno suprimido
- a es la suma de $(r_i/r_c - 1)$ donde r_i es el radio covalente del átomo i y r_c es el radio covalente de un átomo de carbono.
- p_2y p_3 representan el número de caminos de longitud 2 y 3, respectivamente.

38	BertzCT	Número de enlaces de resonancia de Bertz. Cuantifica el número de enlaces que pueden participar en una resonancia.
39	Chi0	Suma de $\frac{1}{\sqrt{d_i}}$ sobre todos los átomos pesados.
40	Chi0n	Descriptor con peso en la conectividad de átomos de nitrógeno.
41	Chi0v	Índice de conectividad de valencia atómica (orden 0). Se calcula como la suma de $\frac{1}{\sqrt{v_i}}$ sobre todos los átomos pesados.
42	Chi1	Índice de conectividad atómica (orden 1) Se calcula como la suma de $\frac{1}{\sqrt{d_i d_j}}$sobre todos los enlaces entre los átomos pesados $i - j$
43	Chi1n	Descriptor de primer orden con peso en la convectividad de átomos de nitrógeno.
44	Chi1v	Índice de conectividad de valencia atómica (orden 1) .Se calcula como la suma de $\frac{1}{\sqrt{v_i v_j}}$ sobre todos los enlaces entre los átomos pesados i y j.
45	Chi2n	Índice de segundo orden con peso en la conectividad de átomos de nitrógeno.
46	Chi2v	Índice de segundo orden con peso en la conectividad de átomos de oxigeno y otros átomos pesados.
47	Chi3n	Índice de tercer orden con peso en la conectividad de átomos de nitrógeno.
48	Chi3v	Índice de tercer orden con peso en la conectividad de átomos de oxigeno y otros átomos pesados.
49	Chi4n	Índice de cuarto orden con peso en la conectividad de átomos de nitrógeno.
50	Chi4v	Índice de cuarto orden con peso en la conectividad de átomos de oxigeno y otros átomos pesados.
51	HallKierAlpha	El valor alfa de Hall-Kier para una molécula.
52	Ipc	El contenido de información de los coeficientes del polinomio característico de la matriz de adyacencia de un grafo de molécula suprimido de hidrógeno.
53	Kappa1	Relación entre la forma molecular y la primer dispersión angular. $k_1 = (n-1)^2/m^2$
54	Kappa2	Relación entre la forma molecular y la segunda dispersión angular.
55	Kappa3	Relación entre la forma molecular y la tercer dispersión angular. Para im par:$\kappa = (n-1)(n-3)^2/p_3^2$, para n par $(n-3)(n-2)^2/p_3^2$

Tabla 12.3: Definición de los descriptores de RDKit.

Descriptores de matrices de adyacencia y distancia

A partir de representar una molécula como un grafo donde las nodos son los átomos y los enlaces aristas se contruye la matriz de adyacencia A, donde los elementos $a_{i,j}$ represtán los enlaces 1 o 0 si no existen. La matriz de distancia, D, donde los elementos $d_{i,j}$ es la longitud del camino más corto entre los átomos i y j; vale cero si los átomos no forman parte del mismo componente conectado.

Petitjean define a:

- La **excentricidad** de un vértice como el camino más largo desde ese vértice a cualquier otro vértice del grafo.
- El **radio** del grafo es la excentricidad de vértice más pequeña del grafo.
- El **diámetro** del grafo es la excentricidad de vértice más grande.

Estos valores se calculan a partir de la matriz de distancia y se utilizan para varios descriptores que se describen a continuación.

Los siguientes descriptores se calculan a partir de las matrices de distancia y adyacencia de los átomos pesados:

56	`BalabanJ`	Valor J de Balaban.
57	`BCUT2D_MWHI`	Los descriptores BCUT se calculan a partir de los valores propios de una matriz de adyacencia modificada. Cada elemento $a_{i,j}$ toma el valor $\frac{1}{\sqrt{b_{ij}}}$, donde $b_{i,j}$ es el orden de enlace formal entre los átomos enlazados i y j. La diagonal toma el valor de distintas propiedades (masa molecular, carga, polarizabilidad, etc.). LOW o HI representan el valor más alto y más bajo de los valores propios, respectivamente.
58	`BCUT2D_MWLOW`	
59	`BCUT2D_CHGHI`	
60	`BCUT2D_CHGLO`	
61	`BCUT2D_LOGPHI`	
62	`BCUT2D_LOGPLOW`	
63	`BCUT2D_MRHI`	
64	`BCUT2D_MRLOW`	

Tabla 12.4: Definición de los descriptores de RDKit.

Descripotores de carga parcial

Existen diversas maneras de estimar cargas parciales. PEOE es el método de Igualación Parcial de Electronegatividades Orbitales (PEOE) para calcular cargas parciales atómicas es un método en el que la carga se transfiere entre átomos enlazados hasta el equilibrio. Para garantizar la convergencia, la cantidad de carga transferida en cada iteración se amortigua con un factor de escala exponencialmente decreciente. La cantidad de carga transferida, $dq_{i,j}$, entre los átomos i y j cuando $X_i > X_j$ es:

$$dq_{i,j} = \frac{1}{2^k} \frac{(X_i - X_j)}{X_j^+}$$

donde x_j^+ es electronegatividad del ion positivo del átomo j; X_i es la electronegatividad del átomo i (dependiente cuadráticamente de la carga parcial); y k es el número de iteración del algoritmo. Los valores de electronegatividad se determinan mediante la parametrización. Las cargas PEOE dependen únicamente de la conectividad de las estructuras de entrada: elementos, cargas formales y órdenes de enlace. Los descriptores que utilizan las cargas PEOE llevan el prefijo PEOE.

Las área de superficie subdivididas (VSA por sus siglas inglés) estan basados en un cálculo aproximado de la superficie accesible de Van der Waals (en $Å^2$) para cada átomo, v_i junto con alguna otra propiedad atómica, p_i. Los v_i se calculan utilizando una aproximación de tabla de conexiones. Cada descriptor de una serie se define como la suma de los v_i de todos los átomos i en los que p_i se encuentra en un intervalo especificado (a,b).

L_i indica la contribución a logP(o/w) del átomo i. R_i indica la contribución a la refractividad molar del átomo i. Los rangos se determinaron por subdivisión percentil sobre una amplia colección de compuestos.

Sea q_i la carga parcial del átomo i, tal como se definió anteriormente. Sea v_i el área superficial de van der Waals (Å2) del átomo i (calculada mediante una aproximación de tabla de conexión). Se calculan los siguientes descriptores:

65	LabuteASA	Área de superficie aproximada de Swan (ASA del MOE).
66	PEOE_VSA1	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 1.
67	PEOE_VSA10	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 10.
68	PEOE_VSA11	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 11.
69	PEOE_VSA12	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 12.
70	PEOE_VSA13	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 13.
71	PEOE_VSA14	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 14.
72	PEOE_VSA2	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 2.
73	PEOE_VSA3	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 3.
74	PEOE_VSA4	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 4.
75	PEOE_VSA5	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 5.
76	PEOE_VSA6	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 6.
77	PEOE_VSA7	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 7.
78	PEOE_VSA8	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 8.
79	PEOE_VSA9	Descriptor que cuantifica la superficie accesible aproximada por el método PEOE, enfocándose en la región 9.

Tabla 12.5: Definición de los descriptores de RDKit.

80	SMR_VSA1	Descriptor que cuantifica la superfice móvil rapida accesible por el método SMR, enfocado en la región 1.
81	SMR_VSA10	Descriptor que cuantifica la superfice móvil rapida accesible por el método SMR, enfocado en la región 10.
82	SMR_VSA2	Descriptor que cuantifica la superfice móvil rapida accesible por el método SMR, enfocado en la región 2.
83	SMR_VSA3	Descriptor que cuantifica la superfice móvil rapida accesible por el método SMR, enfocado en la región 3.
84	SMR_VSA4	Descriptor que cuantifica la superfice móvil rapida accesible por el método SMR, enfocado en la región 4.
85	SMR_VSA5	Descriptor que cuantifica la superfice móvil rapida accesible por el método SMR, enfocado en la región 5.
86	SMR_VSA6	Descriptor que cuantifica la superfice móvil rapida accesible por el método SMR, enfocado en la región 6.
87	SMR_VSA7	Descriptor que cuantifica la superfice móvil rapida accesible por el método SMR, enfocado en la región 7.
88	SMR_VSA8	Descriptor que cuantifica la superfice móvil rapida accesible por el método SMR, enfocado en la región 8.
89	SMR_VSA9	Descriptor que cuantifica la superfice móvil rapida accesible por el método SMR, enfocado en la región 9.
90	SLogP_VSA1	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 1.
91	SLogP_VSA10	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 10.
92	SLogP_VSA11	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 11.
93	SLogP_VSA12	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 12.
94	SLogP_VSA2	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 2.
95	SLogP_VSA3	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 3.
96	SLogP_VSA4	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 4.
97	SLogP_VSA5	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 5.
98	SLogP_VSA6	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 6.
99	SLogP_VSA7	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 7.
100	SLogP_VSA8	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 8.
101	SLogP_VSA9	Descriptor que cuantifica la contribución de la lipofilicidad a la superficie accecible, enfocado en la región 9.
102	TPSASMR_VSA1	Área de superficie polar total.

Tabla 12.6: Definición de los descriptores de RDKit.

103	EState_VSA10	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 1.
104	EState_VSA11	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 10.
105	EState_VSA2	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 11.
106	EState_VSA3	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 2.
107	EState_VSA4	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 3.
108	EState_VSA5	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 4.
109	EState_VSA6	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 5.
110	EState_VSA7	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 6.
111	EState_VSA8	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 7.
112	EState_VSA9	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 8.
113	VSA_EState1	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 9.
114	VSA_EState10	Descriptor que cuantifica la contribucióm de la superfice accecible al método EState, enfocado en la regio 1.
115	VSA_EState2	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 10.
116	VSA_EState3	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 2.
117	VSA_EState4	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 3.
118	VSA_EState5	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 4.
119	VSA_EState6	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 5.
120	VSA_EState7	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 6.
121	VSA_EState8	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 7.
122	VSA_EState9	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 8.
123	VSA_EState9	Descriptor que cuantifica la superficio accecible por el método EState, enfocado en la región 9.

Tabla 12.7: Definición de los descriptores de RDKit.

Descriptores de grupos funcionales

Número de grupos funcionales en la molécula.

Ligas de los documentos consultados:

- Documentación RDkit
- DataGrok
- Chemical Computing Group
- Tesis-

124	fr_Al_COO	ácidos carboxílicos alifáticos	167	fr_guanido	Número de grupos guanidina
125	fr_Al_OH	hidroxilo alifáti	168	fr_halogen	Número de halógenos
126	fr_Al_OH_noTert	hidroxilo alifáticos excluyendo terc-OH	169	fr_hdrzine	Número de grupos de hidrazina
127	fr_ArN	Número de N grupos funcionales unidos a los aromáticos	170	fr_hdrzone	Número de grupos hidrazona
128	fr_Ar_COO	Ácido carboxilico aromático	171	fr_imidazole	Número de anillos de imidazol
129	fr_Ar_N	Nitroógenos aromaticos	172	fr_imide	Número de grupos imida
130	fr_Ar_NH	Amina aromática	173	fr_isocyan	Número de isocianatos
131	fr_Ar_OH	hidroxilo aromáticos	174	fr_isothiocyan	Número de isotiocianatos
132	fr_COO	carboxilo	175	fr_ketone	Número de cetonas
133	fr_COO2	oxalato	176	fr_ketone_Topliss	Número de cetonas excluyendo diaryl, a,b-unsat
134	fr_C_O	Número de carbonilo O	177	fr_lactam	Número de betalactámicos
135	fr_C_O_noCOO	Número de carbonilos O, excluido COOH	178	fr_lactone	Número de ésteres cíclicos (lactonas)
136	fr_C_S	Número de tiocarbonilo	179	fr_methoxy	Número de grupos metoxi -OCH3
137	fr_HOCCN	Número de C(OH)CCN-Ctert-alquilo o C(OH)CCNcíclico	180	fr_morpholine	Número de anillos de morfolina
138	fr_Imine	Número de iminas	181	fr_nitrile	Número de nitrilos
139	fr_NH0	Número de aminas terciarias	182	fr_nitro	Número de grupos nitro
140	fr_NH1	Número de aminas secundarias	183	fr_nitro_arom	Número de sustituyentes del anillo de nitrobenceno
141	fr_NH2	Número de aminas primarias	184	fr_nitro_arom_nonortho	Número de sustituyentes del anillo de nitrobenceno
142	fr_N_O	Número de grupos hidroxilamina	185	fr_nitroso	Número de grupos nitrosos, excluyendo NO_2
143	fr_Ndealkylation1	Número de grupos XC-CNR	186	fr_oxazole	Número de anillos de oxazol
144	fr_Ndealkylation2	Número de aminas terciarias-alicíclicas (sin heteroátomos, sin N con puente tipo quinina)	187	fr_oxime	Número de grupos oxima
145	fr_Nhpyrrole	Número de nitrógenos del H-pirrol	188	fr_para_hydroxylation	Número de sitios de para-hidroxilación
146	fr_SH	Número de grupos tiol	189	fr_phenol	Número de fenoles
147	fr_aldehyde	Número de aldehídos	190	fr_phenol_noOrthoHbond	Número de sustituyentes fenólicos OH excluyendo enlaces de hidrógeno intramoleculares orto
148	fr_alkyl_carbamate	Número de carbamatos de alquilo (sujetos a hidrólisis)	191	fr_phos_acid	Número de grupos de ácido fosfórico
149	fr_alkyl_halide	Número de haluros de alquilo	192	fr_phos_ester	Número de grupos éster fosfórico
150	fr_allylic_oxid	Número de sitios de oxidación alílica excluyendo el esteroide dienona	193	fr_piperdine	Número de anillos de piperdina
151	fr_amide	Número de amidas	194	fr_piperzine	Número de anillos de piperzina
152	fr_amidine	Número de grupos amidina	195	fr_priamide	Número de amidas primarias
153	fr_aniline	Número de anilinas	196	fr_prisulfonamd	Número de sulfonamidas primarias
154	fr_aryl_methyl	Número de sitios arilmetilo para la hidroxilación	197	fr_pyridine	Número de anillos de piridina
155	fr_azide	Número de grupos azida	198	fr_quatN	Número de nitrógenos cuaternarios
156	fr_azo	Número de grupos azo	199	fr_sulfide	Número de tioéter
157	fr_barbitur	Número de grupos de barbitúricos	200	fr_sulfonamd	Número de sulfonamidas
158	fr_benzene	Número de anillos de benceno	201	fr_sulfone	Número de grupos sulfona
159	fr_benzodiazepine	Número de benzodiazepinas sin anillos fusionados adicionales	202	fr_term_acetylene	Número de acetilenos terminales
160	fr_bicyclic	Bicíclico	203	fr_tetrazole	Número de anillos de tetrazol
161	fr_diazo	Número de grupos diazo	204	fr_thiazole	Número de anillos de tiazol
162	fr_dihydropyridine	Número de dihidropiridinas	205	fr_thiocyan	Número de tiocianatos
163	fr_epoxide	Número de anillos de epóxido	206	fr_thiophene	Número de anillos de tiofeno
164	fr_ester	Número de ésteres	207	fr_unbrch_alkane	Número de alcanos no ramificados de al menos 4 miembros (excluye los alcanos halogenados)
165	fr_ether	Número de oxígenos del éter (incluido el fenoxi)	208	fr_urea	Número de grupos de urea
166	fr_furan	Número de anillos de furano			

Tabla 12.8: Definición de los descriptores de RDKit.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00128
Matrícula: 2223803434

Cálculos de energías de reorganización de compuestos orgánicos nitrogenados utilizando diversas técnicas de Machine Learning.

YAFFET ZAMBRANO GONZALEZ
ALUMNO

REVISÓ

MTRA. ROSALIA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

En la Ciudad de México, se presentaron a las 11:00 horas del día 6 del mes de junio del año 2025 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DR. MARCELO ENRIQUE GALVAN ESPINOSA
DRA. ADRIANA PEREZ GONZALEZ
DRA. MARTHA MAGDALENA FLORES LEONAR
DR. ANGEL ALEJANDRO GARCIA CHUNG

Bajo la Presidencia del primero y con carácter de Secretario el último, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRO EN CIENCIAS (QUÍMICA)

DE: YAFFET ZAMBRANO GONZALEZ

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

Aprobar

Acto continuo, el presidente del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

DIRECTOR DE LA DIVISIÓN DE CBI

Román Linares Romero
DR. ROMAN LINARES ROMERO

PRESIDENTE

DR. MARCELO ENRIQUE GALVAN ESPINOSA

VOCAL

DRA. ADRIANA PEREZ GONZALEZ

VOCAL

DRA. MARTHA MAGDALENA FLORES LEONAR

SECRETARIO

DR. ANGEL ALEJANDRO GARCIA CHUNG