UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA

FACULTAD DE CIENCIAS QUÍMICAS

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

UNIVERSIDAD AUTÓNOMA DE
CHIHUAHUA

PREDICCIÓN DE PIGMENTOS: UN ENFOQUE DESDE EL APRENDIZAJE PROFUNDO.

POR:

I. Q. LISSETE QUINTANA RICARDO

TESIS PRESENTADA COMO REQUISITO PARA OBTENER EL GRADO DE

MAESTRÍA EN CIENCIAS EN QUÍMICA

CHIHUAHUA, CHIH., MÉXICO **ENERO DE 2022**

UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA

Chihuahua, Chih., a 31 de enero de 2022.
Oficio: 11/CA/SIP/22

Dr. Ildebrando Pérez Reyes
Secretario de Investigación y Posgrado
Facultad de Ciencias Químicas
Universidad Autónoma de Chihuahua
P R E S E N T E

Los integrantes del comité, informamos a Usted que efectuamos la revisión de la tesis intitulada: **"Predicción de pigmentos: un enfoque desde el aprendizaje profundo",** presentada por la **I.Q. Lissete Quintana Ricardo**, alumna del programa de Maestría en Ciencias en Química.

Después de la revisión, indicamos a la tesista las correcciones que eran necesarias efectuar y habiéndolas realizado, manifestamos que la tesis, de la cual adjuntamos un ejemplar, ha cumplido con los objetivos señalados por el Comité de Tesis, por lo que puede ser considerada como adecuada para que se proceda con los trámites para la presentación de su Examen de Grado.

A t e n t a m e n t e
"Por la ciencia para bien del hombre"

Dr. Adrián Hernández Becerril
Asesor de tesis

Dra. María Elena Fuentes Montero
Asesora de tesis

Dra. María de Lourdes Ballinas Casarrubias
Co-Directora de tesis

Dra. Erika Salas Muñoz
Asesora de tesis

Dr. José Manuel Nápoles Duarte
Director de tesis

Dr. Ildebrando Pérez Reyes
Secretario de Investigación y Posgrado

FACULTAD DE CIENCIAS QUÍMICAS
SECRETARIA
INVESTIGACIÓN Y POSGRADO

*El que suscribe certifica que las firmas que aparecen en esta acta, son auténticas, y las mismas que utilizan los C. Profesores mencionados.

FACULTAD DE CIENCIAS QUÍMICAS
Circuito Universitario
Campus Universitario #2 C.P. 31125
Tel. +52 (614) 236 6000
Chihuahua, Chihuahua, México
http://www.fcq.uach.mx

AGRADECIMIENTOS

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT), por el apoyo económico otorgado, a través de la beca para estudios de posgrado en la Universidad Autónoma de Chihuahua; sin la cual no hubiese sido posible la realización de este proyecto.

Agradezco a la Universidad Autónoma de Chihuahua y a la Facultad de Ciencias Químicas por permitirme continuar con mi preparación académica; así como a los maestros que contribuyeron a mi formación durante este período.

A los Doctores María Elena Fuentes Montero, María de Lourdes Ballinas Casarrubias, Erika Salas Muñoz y Adrián Hernández Becerril por el tiempo dedicado a la revisión del documento y las sugerencias para enriquecer el trabajo.

Agradezco al Dr. José Manuel Nápoles Duarte por dirigir este proyecto y brindarme la oportunidad de trabajar en esta área de investigación, contando con su experiencia y conocimientos relacionados con el Aprendizaje Automático.

DEDICATORIA

Este trabajo está dedicado a mi familia, que con su apoyo y ejemplo han sido mi motivación cada día para superarme, perseverar y seguir adelante.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTA DE ABREVIATURAS

HOMO: Orbital molecular de más alta energía ocupado.

LUMO: Orbital molecular de más baja energía desocupado.

UV-Vis: Ultravioleta-visible.

LC-DAD: Cromatografía Líquida de Alta Resolución con Matriz de Detección de Diodos.

UPLC-MS: Cromatografía Líquida de Ultra Presión con Espectrometría de Masas.

FTIR: Espectroscopía de Transformada de Fourier.

DFT: Teoría Funcional de la Densidad.

MAE: Error Absoluto Medio.

CNN: Redes Neuronales Convolucionales.

InChI: Identificador Químico Internacional (IUPAC International Chemical Identifier).

SMILES: Especificación de Introducción Lineal Molecular Simplificada (Simplified Molecular Input Line Entry Specification).

Colab: Google Colaboratory.

ECFP4: Huella Dactilar de Conectividad Extendida con diámetro igual a 4 enlaces.

RESUMEN

El Azul Maya es un pigmento de color intenso, elaborado y utilizado principalmente por culturas mesoamericanas. Este se forma como resultado de la fuerte unión de la molécula colorante índigo con la estructura inorgánica de la zeolita paligorskita, formando un complejo orgánico-inorgánico de elevada estabilidad. Las zeolitas son minerales cristalinos, cuya estructura está formada por unidades tetraédricas conectadas a través de esquinas alternas hasta generar un arreglo tridimensional. Sobre la base del número de tetraedros conectados en estas estructuras, se definen aberturas de poros de dimensiones moleculares, lo cual les proporciona propiedades de selectividad y gran capacidad absorbente. El análisis de estas estructuras y del pigmento Azul Maya, indican que el empleo de zeolitas como matriz base para alojar moléculas colorantes orgánicas, puede ser una vía efectiva para obtener pigmentos estables. El presente trabajo propone una metodología, que permita proponer moléculas candidatas a formar pigmentos híbridos estables; así como predecir propiedades relevantes para el efecto colorante, utilizando algoritmos de aprendizaje automático. Para ello, se forma un conjunto de 1153 compuestos colorantes a partir de diferentes repositorios, teniendo en cuenta la aplicación de criterios para el curado de bases de datos moleculares. Se realiza una selección de 171 compuestos de acuerdo a la similitud con el índigo y dehidroíndigo, basada en el uso de huellas dactilares moleculares. Se realizan determinaciones de Energías de Atomización para moléculas pequeñas, empleando el método de Kernel Ridge Regression y QM7 como conjunto de entrenamiento. Este modelo realiza predicciones de energías con un Error Absoluto Medio de 3.944 kcal/mol para el conjunto de prueba y 16.869 kcal/mol para un conjunto de moléculas de estudio, cuando se emplea la Función de Kernel Laplaciano. Además, se realiza la predicción de la energía de orbitales HOMO y LUMO, así como la brecha energética existente entre estos; usando modelos de Redes Neuronales Convolucionales.

Palabras clave: Pigmentos híbridos, similitud molecular, Aprendizaje Automático.

ABSTRACT

Maya Blue is an intense color pigment, made and used mainly by Mesoamerican cultures. This is formed as a result of the strong union of the indigo dye molecule with the inorganic structure of the palygorskite zeolite, forming an organic-inorganic complex with high stability. Zeolites are crystalline minerals, whose structure is made up of tetrahedral units connected through alternate corners to generate a three-dimensional arrangement. On the basis of the number of connected tetrahedra in these structures, pore openings of molecular dimensions are defined, which provides them with selectivity properties and great absorbent capacity. The analysis of these structures and the Maya Blue pigment indicate that the use of zeolites as a base matrix to host organic dye molecules can be an effective way to obtain stable pigments. The present work proposes a methodology that allows proposing candidate molecules to form stable hybrid pigments; as well as predicting properties relevant to the coloring effect, using Machine Learning algorithms. For this, a set of 1153 coloring compounds is formed from different repositories, taking into account the application of criteria for curing molecular databases. A selection of 171 compounds is made according to the similarity with indigo and dehydroindigo, based on the use of molecular fingerprints. Atomization Energies determinations are made for small molecules, using the Kernel Ridge Regression method and QM7 as a training set. This model makes energy predictions with a Mean Absolute Error of 3,944 kcal/mol for the test set and 16,869 kcal/mol for a set of study molecules, when the Laplacian Kernel Function is used. In addition, the prediction of the energy of the HOMO and LUMO orbitals is carried out, as well as the energy gap between them; using models of Convolutional Neural Networks.

Keywords: Hybrid pigments, molecular similarity, Machine Learning.

I. INTRODUCCIÓN

Los pigmentos son definidos como sustancias que contienen partículas orgánicas o inorgánicas, insolubles en el medio; cuya finalidad es la de aportar color, brillo, opacidad o ciertas propiedades físico-químicas requeridas (Bender 2013).

Las principales aplicaciones de los pigmentos incluyen: la industria cosmética, automotriz, colorantes alimenticios, tintas para impresión, materiales de construcción, decoraciones artísticas, entre otras (Stratmann et al. 2020).

Los colorantes orgánicos comprenden un conjunto de compuestos, naturales o sintéticos, que por lo general ofrecen colores intensos y gran variedad de tonalidades. Sin embargo, su aplicación práctica se encuentra limitada por su rápida degradación ante diversos factores como la radiación UV, agentes químicos y su baja resistencia térmica (Pfaff 2017).

El Azul Maya es un pigmento desarrollado por las antiguas civilizaciones mesoamericanas, el cual se forma por la unión del colorante orgánico índigo con la estructura inorgánica de la zeolita paligorskita; logrando como resultado un pigmento híbrido de color brillante y elevada estabilidad química (Chen et al. 2019).

La existencia de este tipo de complejos orgánico/zeolita, ofrece una amplia gama de posibilidades en cuanto a la exploración de nuevos materiales, que integren sus propiedades, para crear compuestos de interés para la industria de pigmentos y pinturas.

Investigadores como (Dejoie et al. 2010; Giustetto et al. 2011; Kowalak & Zywert 2011; Zhang et al. 2015; Woodtli et al. 2018; Chen et al. 2019; Yamamoto et al. 2020) han desarrollado estudios experimentales, donde combinan diferentes moléculas colorantes y zeolitas, con el fin de crear pigmentos híbridos de elevada estabilidad.

Estos métodos tradicionalmente empleados requieren gran disposición de materiales y reactivos, por lo que su alcance es limitado en cuanto al número de muestras que pueden evaluar.

El Aprendizaje Automático (Machine Learning) constituye un área de la Inteligencia Artificial que se basa en el empleo de herramientas computacionales, para la extracción de características y detección de patrones en sistemas complejos de datos, con el objetivo de aumentar el rendimiento en diferentes tareas (Jiang, Gradus, and Rosellini 2020). Dentro de este grupo, el término Aprendizaje Profundo o Deep Learning se refiere al empleo de varias capas para el procesamiento de información y específicamente a modelos de Redes Neuronales con múltiples capas ocultas, según se describe en la figura 1 (Alsuliman et al. 2020; Sharma et al. 2021)

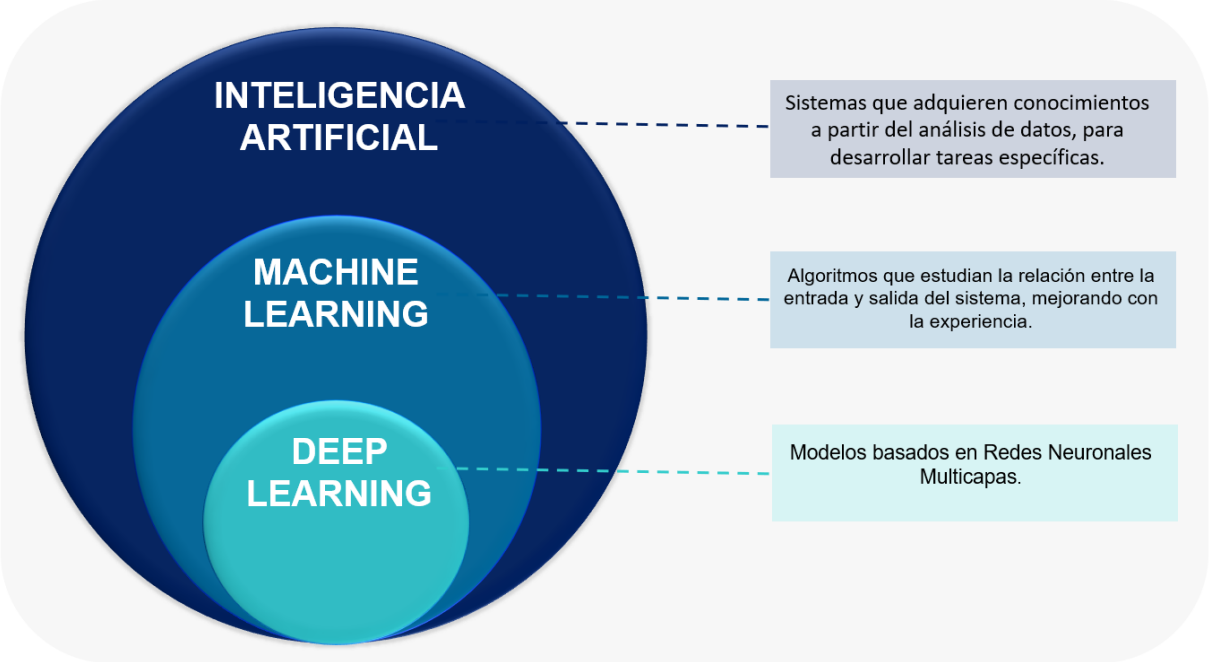


Figura 1. Relación existente entre Inteligencia Artificial, Machine Learning y Deep Learning.

Estos algoritmos forman parte de muchas aplicaciones en el sector comercial y trabajos de investigación; en esferas que incluyen el diseño de materiales, desarrollo de fármacos y el estudio de relaciones de estructura química-propiedades de los compuestos (Müller and Guido 2016).

Mediante la combinación de estas técnicas y herramientas de minería de datos es posible desarrollar una metodología, que basada en la compatibilidad existente entre el índigo y la paligorskita, permita analizar gran cantidad de moléculas presentes en

bases de datos para obtener compuestos candidatos a formar un pigmento híbrido estable; así como predecir propiedades relevantes para el efecto colorante.

II. ANTECEDENTES

II.I. Caracterización de colorantes presentes en el Azul Maya.

En esta sección se realiza una descripción general del colorante índigo, seguidamente se define su estructura, grupo cromóforo y se indican las características geométricas de la molécula. Luego se explica acerca de la transformación que experimenta esta molécula durante el proceso de elaboración del Azul Maya y, por tanto, se procede a describir su forma oxidada (dehidroíndigo); analizando su influencia sobre las características del pigmento.

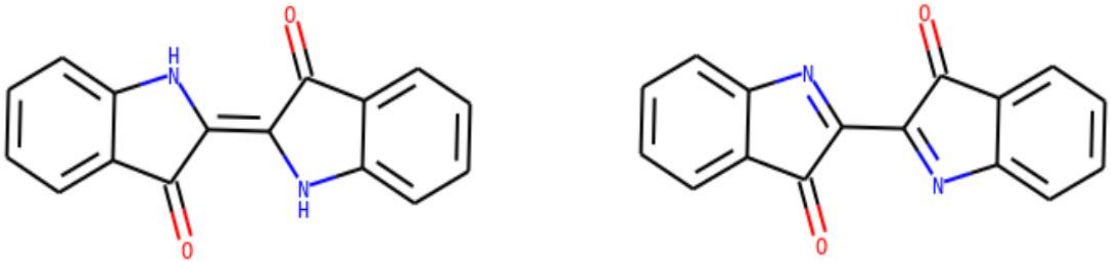
El índigo es un compuesto de fórmula molecular $C_{16}H_{10}N_2O_2$, que tiene forma de polvo de color azul oscuro, siendo uno de los colorantes naturales más usados desde la antigüedad. Este es un compuesto que presenta dimensiones moleculares de 4.83 x 12.3 $Å^2$ (Dejoie et al. 2010), es insoluble en agua y puede ser sintetizado químicamente o extraído de varias fuentes vegetales. Dicha molécula presenta un grupo cromóforo, formado por dos grupos funcionales (N-H y C=O) en direcciones opuestas a ambos lados del enlace central C=C; siendo un compuesto donante-aceptor donde los grupos carbonilo participan como aceptores de electrones y los grupos amino como donadores (Dong and Zhang 2019).

La coloración oscura de este pigmento está dada por una brecha energética de alrededor de 2.5 eV, existente entre el orbital molecular de más baja energía desocupado (LUMO) y el orbital molecular de más alta energía ocupado (HOMO); pudiendo mostrar diferentes colores en el espectro ultravioleta (UV) y visible, dependiendo de cambios químicos y de posición de los grupos laterales de los anillos (Ju et al. 2019; Volkov et al. 2020).

Cuando la molécula de índigo es expuesta a elevadas temperaturas y en presencia de oxígeno, se transforma en un nuevo compuesto de color amarillo llamado dehidroíndigo, que constituye su forma oxidada. Este compuesto se forma como resultado de la pérdida de dos átomos de hidrógeno en la molécula de índigo, con la

consiguiente transformación del doble enlace central carbono-carbono por un enlace simple (He et al. 2010).

En la figura 2 se muestra la estructura de las moléculas colorantes descritas.



Índigo Dehidroíndigo

Figura 2. Representación de las moléculas de índigo y dehidroíndigo.

Por tanto, se afirma que inevitablemente, en el Azul Maya deben estar presentes ambas moléculas y que la temperatura utilizada durante el proceso de fabricación, entre otros factores, determinará el contenido de dehidroíndigo en la muestra, dando como resultado la gran variabilidad de tonalidades en las que se puede encontrar dicho pigmento (Pascual, Carbó, and Carbó 2011).

Doménech en año 2013, realizó un estudio en el que emplea métodos de Cromatografía Líquida de Alta Resolución con Matriz de Detección de Diodos (LC-DAD) y Cromatografía Líquida de Ultra Presión con Espectrometría de Masas (UPLC-MS) acoplada, para separar e identificar componentes de varias muestras de Azul Maya. Para ello reproduce un proceso de fabricación del pigmento, empleando colorante índigo sintético y paligorskita proveniente de la región de Yucatán, en el cual se realiza una molienda en seco de ambos componentes; seguido de una etapa de calentamiento, donde se extraen muestras a varias temperaturas. Estas muestras son analizadas para recopilar datos termoquímicos del pigmento; así como analizar el contenido de índigo y dehidroíndigo. Mediante este estudio no solo se confirma la presencia de dehidroíndigo como componente del pigmento, sino que también se logra establecer una expresión que relaciona la concentración de cada compuesto con la

temperatura de fabricación, mostrando así una dependencia directa de la formación de dehidroíndigo con dicho factor (Doménech-Carbó et al. 2013).

Varios autores han empleado otras moléculas como fuentes de colorantes para preparar pigmentos híbridos a base de zeolitas. Entre estos, Zhou en el año 2014, realizó un estudio en el que sintetiza un pigmento a partir del colorante isatina ($C_8H_5NO_2$) y la zeolita paligorskita. Para ello, realiza una mezcla en seco de ambos componentes, que contiene 10 % del colorante. Luego, se someten a una etapa de calentamiento de 130 ^{0}C, durante 48 horas. A continuación, se realiza un proceso de extracción en un Soxhlet con $CHCl_3$, para eliminar los restos de isatina que no se encuentran unidas a la zeolita. Después de 72 horas de extracción, el 5.9 % del colorante se mantuvo retenido en la estructura de la zeolita. Además, se mantuvo el pigmento a reflujo en una solución acuosa al 30 % de H_2O_2, durante 5 horas, manteniendo su color intenso. Estos resultados indican que las moléculas colorantes penetran en el interior de los canales de la paligorskita, por lo cual obtienen protección ante los agentes oxidantes externos (Zhou et al. 2014).

Zhang en el año 2015, preparó varios pigmentos, mezclando azul de metileno ($C_{16}H_{18}ClN_3S_3H_2O$) con la paligorskita, y otra variante donde crea una capa de SiO_2 sobre la superficie de la zeolita, mediante la condensación de tetrahidroxisilano, para intensificar el proceso de adsorción con el colorante. Para la preparación de los pigmentos, se ponen en contacto ambos componentes en solución acuosa, a temperatura ambiente, y para el caso de la segunda variante, se le añade tetrahidroxisilano y ácido acético. A continuación, la mezcla se mantiene en agitación magnética durante 4 y 1 horas respectivamente; luego, el producto es secado y sometido a un proceso de calentamiento hasta 240 ^{0}C. Una vez obtenidos los pigmentos, se analiza su estabilidad, exponiéndolos a varios agentes químicos como ácido clorhídrico (1 M), hidróxido de sodio (1 M) y etanol anhidro, durante 72 horas a temperatura ambiente; así como a radiación UV a 60 ^{0}C, para simular el proceso de envejecimiento. En el caso del pigmento formado por el colorante y la paligorskita, se

observó una severa decoloración ante la exposición ante los tres reactivos. Para el pigmento que fue preparado añadiendo la capa de SiO_2, al analizar el sobrenadante a 24, 48 y 72 horas; se observó, aunque en menor medida, efectos de decoloración ante todos los agentes químicos. En cuanto al análisis de estabilidad ante el proceso de envejecimiento acelerado, para el caso del compuesto azul de metileno-paligorskita se observó una pérdida del color a los 24 días de exposición a la radiación, mientras que el compuesto azul de metileno-SiO_2-paligorskita no mostró disminución del color en ese tiempo; lo cual indica que la modificación de la superficie de la zeolita puede aumentar considerablemente la fortaleza de las interacciones y por tanto, la estabilidad del pigmento (Zhang et al. 2015).

Chen en el año 2019, preparó un pigmento rojo mezclando moléculas del colorante XL-GRL con la zeolita sepiolita. Este pigmento constituye un colorante sintético catiónico, de fórmula molecular $C_{18}H_{21}BrN_6$ y dimensiones de 5.243 x 6.570 x 16.464 Å. Para este estudio empleó tres métodos de elaboración: aplicando presión de vacío para impulsar el colorante dentro de los poros de la sepiolita, método de mezcla y molienda de los componentes en seco, así como el método de mezclar los compuestos a través de una solución acuosa; seguidos los tres procesos, de una etapa de calentamiento. Al analizar los pigmentos preparados, mediante estudios de Difracción de Rayos X, los patrones indicaron la presencia de moléculas XL-GRL bloqueando los túneles de la zeolita, ubicadas dentro de los canales y adsorbidas en la superficie; destacando que mediante el empleo del método de vacío se logra obtener un mayor número de moléculas en el interior de los canales. Además, se realiza un estudio de estos empleando Espectroscopía de Transformada de Fourier (FTIR) para analizar las interacciones existentes, lo cual arrojó que entre las moléculas colorantes y la zeolita se producen fundamentalmente, atracciones electrostáticas y que no existe formación de enlaces de hidrógeno, ni enlaces coordinados. También se evaluó la estabilidad de los pigmentos, con lo cual se confirmó que los pigmentos obtenidos mediante la

aplicación de vacío, muestran mayor resistencia térmica y química que los obtenidos empleando otros métodos de fabricación (Chen et al. 2019).

Posteriormente, Yamamoto realizó un estudio donde obtiene pigmentos híbridos a partir de la unión de las zeolitas paligorkita y sepiolita, con los colorantes Rojo de Metilo ($C_{15}H_{15}N_3O_2$) y Alizarina ($C_{14}H_8O_4$). Para la preparación de los pigmentos, emplea el método de mezcla y molienda de los componentes en seco, seguido de un proceso de calentamiento de 368 K por 24 horas y 413 K por 24 horas más. Luego se evalúa la estabilidad térmica de los colorantes y de los compuestos híbridos obtenidos, mediante análisis termogravimétricos y térmico diferencial. Para ambos colorantes orgánicos, la descomposición térmica comienza a 480 K, obteniéndose una destrucción total a 600 K. Para el caso de los pigmentos, estos no muestran decoloración en este rango de temperaturas, indicando una resistencia térmica más elevada que las moléculas colorantes. Estos resultados evidencian como el empleo de zeolitas como base para crear pigmentos híbridos, puede mejorar considerablemente las propiedades de los colorantes orgánicos (Yamamoto et al. 2020).

II.II. Caracterización de la zeolita Paligorskita.

En la literatura, se definen las zeolitas como compuestos cristalinos, cuya estructura está formada por la unión de varios marcos de tetraedros (TO_4, siendo T un átomo de silicio, aluminio o fósforo, entre otros), que se repiten periódicamente generando así estructuras microporosas. Dependiendo del modo en que se conectan estos marcos de tetraedros, se identifica una gran variedad de estructuras de zeolitas con sistemas de canales bidimensionales y tridimensionales (Li and Yu 2014).

Estas constituyen materiales que presentan gran capacidad de adsorción, baja densidad, elevada estabilidad química, así como resistencia mecánica y térmica. Por lo que han sido ampliamente utilizadas en áreas que van desde el desarrollo de suplementos alimenticios, tratamiento de aguas residuales, procesos de catálisis, separación de mezclas, entre otras (Li et al. 2020).

La paligorskita es una zeolita fibrosa de composición ideal $Si_8(Mg_2Al_2)O_{20}(OH)_2(OH_2)_4 4H_2O$, con una estructura caracterizada por la combinación de dos capas tetraédricas (SiO_4) y unidas por una octaédrica (MO_6, siendo M iones de aluminio y magnesio), formando así canales con dimensiones de 6.4 x 3.7 Å. En dicha estructura están presentes además, cationes intercambiables y moléculas de agua unidas a los átomos T (agua estructural) o rellenando los canales de la zeolita (agua zeolítica) (Chen et al. 2019).

La gran capacidad de adsorción de esta zeolita, capacidad de intercambio catiónico y elevada área superficial; así como las interacciones de esta con la molécula de índigo, dan lugar a la extrema estabilidad del Azul Maya; por lo que ha sido de interés para varios investigadores, determinar otros estructuras de zeolitas que muestren resultados semejantes (Zhang et al. 2015).

En este sentido, Dejoie en el año 2010, realizó un estudio donde evaluó la estabilidad de pigmentos sintetizados a partir del colorante índigo y varias matrices zeolíticas: zeolita LTA, mordenita y zeolita MFI. Las zeolitas LTA y MFI son materiales sintéticos comerciales. La zeolita LTA posee en su estructura un sistema de poros tridimensionales que contienen cajas esféricas (diámetro ~ 11,4 Å), a las cuales se accede por canales de 4,1 Å de diámetro. La estructura de la zeolita MFI contiene canales de 10 unidades de tetraedros, de dimensiones de aproximadamente 5 Å. De las zeolitas evaluadas, solo la que forma el compuesto índigo-zeolita MFI no experimenta una destrucción total en condiciones oxidantes. Cuando este compuesto se expone a ácido nítrico concentrado experimenta un ligero cambio de azul a violeta. Para una concentración tri-diluida también se observa decoloración, aunque en menor magnitud y 25 horas de duración son requeridas para destruir totalmente el color del pigmento. Además, se sugiere que la presencia de los átomos de aluminio en el marco de la zeolita no constituye un factor que influya sobre la estabilidad del pigmento obtenido, pues la zeolita LTA como la mordenita son ricos en aluminio y no forman compuestos resistentes a las condiciones de oxidación (Dejoie et al. 2010).

Giustetto en el año 2011, realizó la obtención de pigmentos formados por la molécula de índigo y la zeolita sepiolita, con el objetivo de analizar su estabilidad en comparación con pigmentos índigo-paligorskita. La sepiolita es una zeolita de fórmula ideal $Mg_8Si_{12}O_{30}(OH)_4(OH_2)_4 \bullet nH_2O$; $n \leq 8$, cuya estructura está formada por la combinación de capas ectaédricas y tetraédricas (T:O:T), al igual que en la paligorskita; pero en este caso los canales formados presentan diámetro ligeramente superior (10.6×3.7 Å). Para la preparación del pigmento, se realizó la molienda y mezcla de la sepiolita e índigo en seco. A continuación, la mezcla fue calentada a 190 ^{0}C durante 20 horas; seguido de una etapa de extracción con acetona, para eliminar el exceso de colorante que no se encuentra unido a la zeolita. Luego, se procede a analizar la estabilidad de los pigmentos, para ello son puestos en contacto con ácido nítrico a 65%, donde se observa disminución del color después de algunas horas con una pérdida total después de 4 días; mostrando una menor resistencia que el pigmento índigo-paligorskita, el cual no muestra afectación del color tras este ataque ácido. Luego, se expusieron a hidróxido de sodio al 32%, mostrando ambos pigmentos un comportamiento similar al observado anteriormente, pero en este caso fue más evidente la destrucción de la estructura de sepiolita. A continuación, se realizaron análisis termogravimétricos del pigmento índigo-sepiolita, donde se obtiene evidencia de la presencia de las moléculas colorantes en el interior de los túneles de la zeolita. Por tanto, se indica que la menor estabilidad del pigmento a base de sepiolita, se debe a que en los canales más estrechos de la paligorskita (6.4 x 3.7 Å), los grupos funcionales de la molécula de índigo pueden establecer enlaces de hidrógeno a ambas superficies de los túneles; mientras que en los canales mayores de la sepiolita, la molécula solo puede formar enlaces con un extremo de la superficie de estos, quedando el otro extremo de la molécula sin unirse a la estructura de la zeolita (Giustetto et al. 2011).

Woodtli, en el año 2018, desarrolló un pigmento híbrido empleando la molécula de índigo y la zeolita L, con el objetivo de evaluar su aplicación práctica. La zeolita L

consiste en un material sintético con disposición hexagonal, que presenta canales paralelos de diámetro 0,71 nm y 1,26 nm aproximadamente; con una distancia de centro a centro de los túneles de 1,84 nm. Para la obtención del pigmento, se realiza la mezcla de los componentes, seguido de un proceso de calentamiento al vacío, para asegurar que las moléculas se difundan dentro de los poros de la zeolita. A continuación, se realiza el lavado del pigmento para eliminar el exceso de colorante y posteriormente, se realiza el sellado de los canales, mediante la adición de un aminosilano. Luego, se evaluó la aplicación del colorante sobre tela de algodón, mediante la técnica de revestimiento con rodillo y se realizaron análisis de espectroscopía UV-Vis. Dichos análisis no mostraron diferencias entre los espectros de la impresión y los del pigmento índigo-zeolita L. Por tanto, se espera que este material zeolítico; así como el método de obtención y aplicación del pigmento, pueda ser extendido a otros colorantes (Woodtli et al. 2018).

II.III. Interacciones existentes entre la molécula de índigo y la zeolita paligorskita.

El estudio de las interacciones que dan origen a la unión de la molécula de índigo y la paligorskita, así como los sitios de localización del colorante en la estructura de la zeolita han sido ampliamente debatidas en las últimas décadas. En cuanto a los sitios de ubicación de la molécula de índigo en la matriz zeolítica, estudios iniciales sugerían que la molécula colorante podría adsorberse a la superficie externa de la zeolita; ubicarse en el interior de los canales de la misma o situarse a la entrada de estos.

Estudios posteriores indicaban que la formación del pigmento está determinada por la inclusión de las moléculas de índigo en el interior de los túneles de la paligorskita y la transformación de estas a dehidroíndigo, lo que provoca que dichas moléculas queden atrapadas en el interior de los canales, bloqueando la entrada de estos.

En este sentido Doménech en el año 2013, realizó un estudio donde emplea técnicas de Voltametría de micropartículas, acompañada de espectroscopía infrarroja y visible, para evaluar el proceso de formación de pigmentos sintéticos y tener indicios acerca de los sitios de ubicación del colorante en la matriz inorgánica. La elaboración de los

pigmentos se realiza mediante trituración en seco del colorante índigo y la paligorskita, seguido de una etapa de calentamiento a varias temperaturas en un rango de 120 a 180 ^{0}C, donde se toman muestras del compuesto. Los análisis indican que el proceso de formación del Azul Maya tiene lugar mediante dos pasos que ocurren de modo paralelo. Uno de estos consiste en la pérdida de agua zeolítica durante el proceso de calentamiento, como proceso determinante de la velocidad, con la correspondiente unión del índigo a la estructura de la zeolita, a la vez que se va produciendo la oxidación parcial del colorante. El otro paso consiste en la ubicación de las moléculas de índigo y dehidroíndigo dentro de los túneles de la zeolita mediante procesos difusivos, mostrando que la relación dehidroíndigo/índigo máxima no se encuentra en la superficie de la zeolita sino a cierta distancia dentro de los canales de esta (Doménech et al. 2013).

En cuanto a las interacciones, los estudios indican que se producen enlaces de hidrógeno entre los grupos amino y carbonilo del colorante, con los grupos SO_4 de la paligorskita; así como entre los grupos carbonilos y las moléculas de agua presentes dentro de los túneles de la zeolita (agua estructural). Enlaces directos entre el índigo y los cationes octaédricos (Mg^{2+} y Al^{3+}) de la zeolita; así como enlaces con Al^{3+} que sustituye al silicio en centros tetraédricos. Además, se refiere que el colorante se mantiene atrapado dentro de los canales debido a un impedimento estérico existente entre estos (Zhang et al. 2015).

Sánchez-Ochoa en el año 2017, realizó un análisis de las interacciones de la molécula colorante en el interior de los túneles de la paligorskita deshidratada; teniendo en cuenta los cálculos de la energía de enlace mediante Teoría Funcional de Densidad (DFT). Los resultados muestran que la molécula de índigo se difunde dentro de los túneles de la paligorskita y luego cuando se transforma a dehidroíndigo bloquea dichos túneles, quedando atrapada dentro de la estructura. Además, muestra la presencia de enlaces de hidrógeno de los grupos funcionales de la molécula colorante y el agua estructural de la zeolita (Sánchez-Ochoa, Cocoletzi, and Canto 2017).

En el año 2018, Bernardino realizó un estudio del pigmento Azul Maya mediante el empleo de resonancia Raman, espectroscopía UV-Vis y métodos termogravimétricos, con el objetivo de realizar un análisis de interacciones. Los resultados sugieren la formación de enlaces de hidrógeno entre los grupos C=O y N-H del índigo, con las moléculas de agua coordinadas dentro de los túneles de la paligorskita. También indican la presencia de enlaces de hidrógeno entre las moléculas de índigo y la superficie de la zeolita. Además, se obtiene información que indica la presencia de enlaces directos de dicha molécula (N y O) con los grupos Mg^{2+} y Al^{3+} presentes en el interior de los canales de la zeolita, lo cual coincide con la formación de un compuesto complejo índigo/paligorskita (Bernardino, Constantino, and De Faria 2018).

También, Caliandro en el año 2019, desarrolló una investigación en la que prepara pigmentos híbridos paligorskita-índigo, con el objetivo de analizar el proceso de formación del Azul Maya. La obtención del pigmento se realiza mediante la mezcla de ambos componentes y luego estos son sometidos a un proceso de calentamiento a varias temperaturas, 105, 150 y 200 ^{0}C. Durante este proceso se realizan Análisis Termogravimétricos y de Espectroscopía UV, los cuales indicaron que en un rango de temperaturas de 150 a 200 ^{0}C ocurre la transformación del índigo a dehidroíndigo. Se realiza un seguimiento del proceso de formación del pigmento con los cambios de temperatura, lo cual arroja que la formación transcurre durante la etapa de calentamiento y no manifiesta cambios ante temperaturas constantes. Por tanto, una exposición de los componentes a elevadas temperaturas durante tiempo prolongado no aumentará la estabilidad del pigmento. Además, se sugiere que la formación del Azul Maya está determinada por la ubicación de las moléculas de índigo dentro de los canales de la paligorskita y que este proceso ocurre mediante tres etapas. Un primer paso, a temperaturas de hasta 120 ^{0}C, donde se produce la liberación de moléculas de agua dentro de los canales de la paligorskita, dando lugar a la formación de sitios activos en la estructura. Un segundo paso, a temperaturas superiores a 120 ^{0}C, determinado el aumento de la hidrofobicidad del sistema por la presencia de índigo y

por los cambios en la estructura del colorante (alrededor de 150 ^{0}C). El tercer paso, alrededor de 175 ^{0}C, que está dado por la acumulación de las moléculas de índigo dentro de los poros de la zeolita, con la consecuente saturación de los sitios activos de esta (Caliandro et al. 2019).

II.IV. Modelos de Aprendizaje Automático (Machine Learning)

II.IV.I. Método de Kernel Ridge Regression

Los algoritmos de aprendizaje automático comprenden un conjunto de herramientas que pueden ser empleadas para determinar la relación entre ciertas características de compuestos químicos y algunas propiedades de referencia. De este modo es posible generar soluciones aproximadas a los cálculos mecánico-cuánticos de dichas propiedades (Faber and Anatole von Lilienfeld 2019).

Los métodos de Kernel Ridge Regression, constituyen una clase de modelos que se basan en el empleo del llamado ¨Truco de Kernel¨ para realizar una transformación no lineal en las características de estudio, de modo que las traslade a otro espacio de mayor dimensión, donde sea más sencillo el ajuste de un modelo que describa el comportamiento de los datos (Patle and Chouhan 2013).

A continuación, se presenta una descripción del funcionamiento del modelo, según se muestra en la literatura (Stulp & Sigaud 2015; Witten et al. 2017; Faber & Anatole von Lilienfeld 2019; S. Theodoridis 2020).

Si se tiene la función f(X), que representa el comportamiento de los datos estudiados, según la expresión de la ecuación 1.

$$f(X) = \sum_{i=1}^{n} \alpha_i \phi(x_i) \quad (1)$$

siendo X la matriz de características x_i a analizar, α_i los coeficientes o parámetros de pesos y $\phi(x_i)$ la función no lineal de las características.

Se debe minimizar la función de error para determinar los parámetros que describen la función f(X) que mejor se ajuste a los datos. Para ello se emplea el método de ¨Ridge Regression¨, que se basa en el uso de la función de error cuadrático L_2 añadiendo

regularizadores, que controlan los parámetros de pesos y evitan valores extremos que provoquen sobreajuste del modelo (ecuación 2).

$$L = \sum_{i=1}^{n}[f(x_i) - y_i]^2 + \lambda \sum_{i=1}^{n} {\alpha_i}^2 \quad (2)$$

Siendo $f(x_i)$ el valor predicho por la función ajustada, y_i el valor real y λ el parámetro de regularización. Cuando el modelo sea ajustado a los datos de entrenamiento, este parámetro de regularización se encarga de desviar ligeramente su pendiente, de modo que sea introducido una penalización o error adicional (aumenta el bias, que puede ser interpretado como la suma de cuadrados del modelo ajustado) para que este no se ajuste perfectamente a los datos de entrenamiento. Este ligero error introducido entonces durante el entrenamiento, aumentará la generalización del modelo a nuevos datos y, por tanto, al ser evaluado en el conjunto de prueba debe mostrar un desempeño similar al obtenido durante el entrenamiento; lo cual se denomina varianza del modelo (disminuye la varianza). De este modo se obtiene una compensación entre complejidad del modelo y su generalización.

A continuación, durante el proceso de minimización de esta función se obtiene un término que representa una matriz, cuyos elementos están dados por el producto escalar de cada par de vectores de características (en el denominado espacio de características o espacio de Hilbert, es decir, se refiere a la función del problema evaluada en cada característica $\phi(x_i)$).

$$\phi(x_i)\phi(x_j)^T = \begin{bmatrix} \phi(x_1)\phi(x_1)^T & \cdots & \phi(x_1)\phi(x_n)^T \\ \phi(x_2)\phi(x_1)^T & \cdots & \phi(x_2)\phi(x_n)^T \\ \vdots & & \vdots \\ \phi(x_n)\phi(x_1)^T & \cdots & \phi(x_n)\phi(x_n)^T \end{bmatrix} \quad (3)$$

Luego, se hace uso del ¨Truco de Kernel¨ (ecuación 4), el cual establece que el producto escalar de cada par de vectores de características puede ser escrito como una sola función de dichas características, la denominada Función de Kernel.

$$\langle \phi(x_i); \phi(x_j) \rangle = K(x_i, x_j) \quad (4)$$

Por tanto, se sustituye esta matriz (ecuación 3) por la matriz de Kernel de la expresión siguiente:

$$K(x_i, x_j) = \phi(x_i)\phi(x_j)^T = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & \cdots & k(x_2, x_n) \\ \vdots & & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \quad (5)$$

Esta operación evita tener que realizar combinaciones arbitrarias en el espacio de características. A la vez, el producto de dichas características puede dar lugar a vectores de gran complejidad, pero la Matriz de Kernel tiene la propiedad de que cada término en este nuevo espacio (espacio de características), puede ser sustituido por una función de los elementos a multiplicar en el espacio original, generando el mismo resultado.

Realizando esta operación de sustituir K en la función de error a minimizar se obtiene la expresión del parámetro α_i, al cual se le denomina Coeficiente de Regresión (ecuación 6).

$$\alpha_i = (K + \lambda I)^{-1} y_i \quad (6)$$

donde I representa la matriz identidad correspondiente a K.

Esta propiedad hace que no sea necesario conocer la naturaleza de la función inicial que describe el problema de estudio, ya que hasta para funciones complejas es posible realizar predicciones teniendo solamente las características (ecuación 7).

$$f(X) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j) \quad (7)$$

Las Funciones de Kernel pueden variar, dependiendo de la transformación que realicen en los datos de entrada; lo cual va a determinar el espacio hacia el que trasladen las características de estudio. Algunos ejemplos se muestran en la tabla 1 (Sharafi et al. 2016).

Tabla 1. Funciones de Kernel.

Función de Kernel	Ecuación
Kernel Lineal	$K(x_m, x_n) = \langle x_m, x_n \rangle$
Kernel Polinomial	$K(x_m, x_n) = (x_m{}^T * x_n + c)^d$
Kernel Gaussiano	$K(x_m, x_n) = e^{\left(\frac{\|x_m - x_n\|_2{}^2}{2\sigma^2}\right)}$
Kernel Laplaciano	$K(x_m, x_n) = \left(\frac{\|x_m - x_n\|_1}{\sigma}\right)$

Se conoce que los métodos de Kernel Ridge Regression han sido empleados para la determinación de propiedades moleculares, en conjuntos de datos conocidos, manteniendo elevada precisión en los resultados a una fracción del costo computacional de los métodos tradicionales. En este sentido Rupp, en el año 2012 realizó un estudio donde ajusta un modelo de este tipo para realizar evaluaciones de Energías de Atomización sobre 7165 moléculas orgánicas pequeñas (QM7) procedentes del conjunto de datos GDB-13 (Blum and Reymond 2009). Para ello, emplea una función de Kernel Gaussiano que toma como descriptores moleculares la Matriz de Coulomb, la cual contiene información relacionada con las cargas nucleares de cada átomo y sus coordenadas cartesianas. Además, se emplean como etiquetas para el aprendizaje supervisado, valores de Energías de Atomización determinadas mediante métodos mecánico-cuánticos. En este estudio se determina que, al ajustar el modelo sobre un conjunto de entrenamiento compuesto por 500 moléculas, se obtienen predicciones con un Error Absoluto Medio (MAE) de 17 kcal/mol. Si el número de elementos utilizados en el entrenamiento aumentara a 7000 compuestos, el error del modelo pudiese disminuir hasta un valor ligeramente inferior a 10 kcal/mol. Para evaluar la transferibilidad del modelo, se ajusta el modelo en un conjunto de 1000 compuestos que presentan en su estructura hasta 5 átomos de carbono, nitrógeno, oxígeno y azufre (átomos pesados) y se emplea como conjunto de prueba 6000 moléculas que contienen hasta 7 átomos pesados. Al evaluar el desempeño del modelo, se obtiene un MAE durante la etapa de entrenamiento igual a 15.2 kcal/mol,

siendo similar al valor obtenido para el conjunto de prueba; con lo cual se concluye que si se ajusta un modelo empleando el 15% de los datos, se logra realizar predicciones para el resto de los elementos con un MAE de alrededor de 15 kcal/mol (Rupp et al. 2012).

Este mismo autor, en el año 2015 emplea esta clase de modelos para realizar predicciones de Energías de Atomización molecular, utilizando varias funciones de Kernel. El entrenamiento se realiza sobre el conjunto de datos QM7 con el objetivo de evaluar si este tipo de modelos realiza predicciones comparables a un nivel de Teoría de Densidad Funcional (DFT). Para determinar los hiperparámetros del modelo (λ; σ) se emplea el método de Grid Search con valores en la escala logarítmica en base 2. Utlizando valores óptimos de hiperparámetros se obtiene un modelo con Función de Kernel Gaussiano que predice energías con un MAE de 13.5 kcal/mol para el conjunto de entrenamiento y 12.5 kcal/mol para el conjunto de prueba, formado por 6000 elementos. Además, se ajusta un modelo de Kernel Laplaciano que realiza predicciones con un MAE de 8.8 kcal/mol para ambos conjuntos de datos. Estos modelos aplicados mantienen un Coeficiente de Determinación (R^2) alrededor de 0.99 para todos los casos (Rupp 2015).

En el año 2018, Collins desarrolló un trabajo en el que evaluó el desempeño de modelos de Kernel Ridge Regression para la predicción de propiedades moleculares, cuando se emplean diferentes representaciones estructurales como descriptores. Para la etapa de entrenamiento se emplean varios conjuntos conocidos, entre los que se encuentran QM7 y QM9. Las representaciones evaluadas incluyen identificadores SMILES, que indican la conectividad entre los átomos y otras que incluyen información relacionada con la conformación del compuesto, como la Matriz de Coulomb. Su estudio evidencia que el empleo de representaciones de geometría molecular aumenta el rendimiento de los modelos de aprendizaje automático para describir la relación estructura-propiedades, lo cual se refleja en la disminución de los valores de MAE de las predicciones generadas (Collins et al. 2018).

II.IV.II. Método de Redes Neuronales Convolucionales (CNN)

Las CNN constituyen estructuras formadas por un conjunto de capas convolucionales que toman como entrada una representación de los datos, aprovechando la ubicación espacial de cada elemento. Dichas estructuras están compuestas por unidades llamadas nodos o neuronas, conectados entre sí, donde se aplican funciones matemáticas a los datos de entrada.

En la primera capa de la estructura, cada neurona se encarga de procesar la información de los elementos vecinos, realizando una suma ponderada de sus características, cada una de ellas con un parámetro de peso asignado que indica el grado de influencia que esta tiene sobre el resultado; y seguidamente se aplica una transformación no lineal de los datos. Luego en la segunda capa, se realiza nuevamente este proceso de convolución, pero cada nodo recopila información no solo de los elementos que los rodean (primeros vecinos), sino de los segundos vecinos también. Así sucesivamente, en cada capa convolucional se produce una actualización del estado de los nodos y mediante el reajuste de los parámetros de pesos se logra realizar una detección de patrones de características en los datos; obteniéndose así en la última etapa, una nueva representación con un menor número de parámetros (Gao et al. 2020).

Por lo general, a la salida de la CNN es conectada una Red Neuronal Multicapa, donde se realiza la tarea de predicción de valores o clasificación. Esta estructura está compuesta por un conjunto de nodos, donde cada elemento de una capa se encuentra conectado a todos los de la capa anterior. Esta toma como entrada a los vectores de características generados en la CNN y en cada nodo se realiza una transformación de los vectores según los parámetros de pesos asignados. Para el caso de Aprendizaje Supervisado, a la salida de la última capa se realiza una comparación entre la predicción arrojada por el modelo y los valores reales (cálculo de error). Se implementa un método de optimización, que se encarga de ajustar los parámetros de la red para minimizar esta diferencia y que el modelo realice predicciones acertadas, a este

proceso se le refiere como proceso de aprendizaje. Una vez que dicho modelo esté ajustado, puede ser empleado para realizar nuevas predicciones sobre datos desconocidos (Hamilton 2020).

Los modelos CNN son comúnmente empleados en áreas de reconocimiento de imágenes, visión computacional, estudio de redes sociales y diagnósticos médicos. Más recientemente están siendo utilizados en aplicaciones químicas y biológicas, como el descubrimiento de drogas, determinación de propiedades moleculares y desarrollo de materiales (Russo et al. 2020).

Yuan en el año 2019, realizó un estudio sobre la aplicación de un modelo de CNN para la predicción de 12 propiedades de toxicidad molecular, sobre el conjunto Tox21. Este modelo combina descriptores del tipo atómico, posición de las coordenadas, distancia y otras características moleculares. Luego, sobre esta información tienen lugar las operaciones de convolución, se vectorizan las características extraídas y se alimentan a la red multicapa donde se realizan las predicciones. El desempeño de este modelo es comparado con métodos tradicionales de aprendizaje automático como: Random Forest, Support Vector Machine y varios métodos de aprendizaje profundo; superándolos para 9 de las 12 características analizadas (Yuan et al. 2019).

En este mismo año, Jeon realizó un trabajo donde propone un modelo para la predicción de propiedades moleculares, como solubilidad, toxicidad, entre otras; basado en Redes Neuronales Convolucionales. Este modelo obtiene vectores de atributos a partir del uso de huellas dactilares moleculares. Luego, se crean arreglos matriciales con dichos vectores y se realizan las operaciones de convolución, para extraer características significativas de los datos. A continuación, mediante el empleo de capas densas, se realiza la predicción de la propiedad de estudio o la clasificación, según sea el caso. Al evaluar el modelo propuesto, en diferentes conjuntos de datos (Tox21, HIV, BBBP, SIDER, Malaria, CEP, ESOL, FreeSolv, Lipophilicity), se alcanzaron resultados comparables a modelos Convolucionales de Grafos y de Redes

Neuronales Completamente Conectadas (Fully Connected Neural Network), siendo particularmente efectivo para problemas de clasificación (Jeon and Kim 2019).

Torng, en el año 2019 aplicó modelos CNN a la determinación de interacciones ligando-proteína, empleando un conjunto de moléculas pequeñas etiquetadas como aprobadas o no para la investigación de fármacos. El primer paso consiste en construir un modelo (codificador), al cual se le entregan las estructuras de ambos compuestos y sus sitios de unión, para que obtenga una representación de estos. Luego se implementa un modelo de CNN para cada estructura (ligando y proteína), con el objetivo de extraer características de ambos compuestos, que tengan influencia sobre las etiquetas. A continuación, los vectores de salida de ambos modelos, se alimentan a una misma capa densa donde se realiza la clasificación. Mediante la predicción de las etiquetas se puede determinar si el modo de unión determinado por el modelo es el adecuado, así como comprobar el desempeño de clasificación de la molécula. Este estudio logra plantear un sistema complejo, representándolo por estructuras independientes y permite confirmar que este acoplamiento de modelos puede integrar información de ambos compuestos satisfactoriamente (Torng and Altman 2019).

III. JUSTIFICACIÓN

El Azul Maya es un pigmento ancestral que presenta elevada estabilidad, formado por la unión de la molécula orgánica de índigo y una zeolita incolora denominada paligorskita. El análisis de este tipo de compuestos, indica que el empleo de zeolitas mejora considerablemente la estabilidad de los colorantes orgánicos, hasta crear materiales de elevada resistencia. Sin embargo, se carece de una metodología que, aplicada previa a ensayos experimentales, pueda identificar moléculas capaces de formar esta clase de pigmentos híbridos. La combinación de algoritmos de aprendizaje automático y herramientas de análisis de datos, permite extraer y agrupar información de compuestos químicos, obtener representaciones de estructura molecular en forma numérica y realizar operaciones matemáticas para generalizar a partir de casos conocidos. Por tanto, mediante la utilización de estas técnicas y en base a las interacciones existentes entre la molécula de índigo y la paligorskita, es posible diseñar una metodología que proponga moléculas colorantes candidatas para formar pigmentos estables.

IV. PROBLEMA

Se carece de una metodología apoyada en bases de datos, que permita predecir moléculas capaces de formar un pigmento colorante-zeolita estable.

V. HIPÓTESIS

Utilizando herramientas de minería de datos, modelos de Kernel Ridge Regression y Redes Neuronales Convolucionales, es posible diseñar una metodología que permita determinar moléculas candidatas a formar pigmentos híbridos estables; así como predecir propiedades relevantes para el efecto colorante.

VI. OBJETIVOS

VI.I. Objetivo general

Desarrollar una metodología, que permita proponer moléculas candidatas a formar pigmentos híbridos estables; así como predecir propiedades relevantes para el efecto colorante, utilizando modelos de Kernel Ridge Regression y Redes Neuronales Convolucionales.

VI.II. Objetivos particulares

- Construir un conjunto de compuestos colorantes, teniendo en cuenta la aplicación de criterios para el curado de bases de datos moleculares.
- Seleccionar compuestos de acuerdo a la similitud con el índigo y dehidroíndigo, basada en el uso de huellas dactilares moleculares.
- Determinar Energías de Atomización para moléculas pequeñas, utilizando un modelo de Kernel Ridge Regression y QM7 como conjunto de entrenamiento.
- Realizar la evaluación de energía de orbitales HOMO y LUMO, así como la brecha energética existente entre estos, usando modelos de Redes Neuronales Convolucionales.

VII. MATERIALES Y MÉTODOS

VII.I. MATERIALES

En esta sección se indicarán los materiales y equipos a emplear durante el estudio.

- Bases y conjuntos de datos.

Biblioteca de tintes Max Weaver.

La Biblioteca de tintes Max Weaver (Williams 2017), contiene alrededor de 98000 colorantes textiles, donados a la Universidad del Estado de Carolina del Norte; con el objetivo de que la información sea accesible con fines de investigación. Los compuestos que incluye esta colección abarcan el rango de colores correspondientes a todo el espectro visible. Además, el 38% de los 2700 colorantes digitalizados hasta 2017, cumplen la Regla de Lipinski y el 71% del resto presenta solo una violación; lo cual indica que presentan biodisponibilidad al ser consumidos por vía oral (Kuenemann et al. 2017).

Base de datos ChemSpider.

ChemSpider (Antony John Williams 2007) consiste en una base de datos con acceso gratuito, que contiene información de estructuras químicas, varios identificadores entre los que destacan InChI, SMILES y nombre químico; así como algunas propiedades fisicoquímicas de los compuestos (Cavalla 2015).

Este sitio agrupa información proveniente de alrededor de 400 fuentes, que incluyen otras bases de datos públicas o privadas, revistas científicas y catálogos de proveedores. Además, presenta opciones de búsqueda de moléculas de acuerdo a su nombre, representación estructural o determinada propiedad de interés (Silakari and Singh 2021).

Base de datos PubChem.

La base de datos PubChem (Kim et al. 2016) constituye un repositorio público que contiene más de 25 millones de estructuras químicas de moléculas pequeñas; así como sus propiedades químicas y biológicas (Li et al. 2010).

Entre la información que reporta esta base de datos, se incluyen los códigos identificadores de los compuestos, SMILES, elementos que contiene, estructura química, masa molecular, carga de la molécula, actividad bilógica, entre otras (Cheng et al. 2014).

Asociada a esta, se tiene también una librería, denominada PubChemPy (https://pubchempy.readthedocs.io/en/latest/), la cual permite interactuar con la base de datos mediante el lenguaje de programación de Python. Esta librería incluye comandos mediante los cuales se logra extraer información de compuestos y propiedades, directamente desde el entorno de Colab. Además, contiene funciones integradas con la librería pandas, que proporciona acceso a la información requerida, en formato de fácil manejo y análisis por el usuario (Swain 2014).

Base de datos COSMOS.

La base de datos COSMOS (Yang et al. 2017) se crea como parte de un proyecto de investigación, cuya finalidad consiste en elaborar un repositorio que contenga información relacionada con la toxicidad de ingredientes cosméticos, para potenciar el uso de métodos teóricos que sustituyan los ensayos de toxicidad que se realizan en animales. Esta agrupa más de 40000 estructuras químicas únicas, reportadas en cosméticos según organismos regulatorios en la Unión Europea y Estados Unidos. Además, contiene la identificación de los compuestos de acuerdo al área de aplicación del producto: tinte de cabello, fragancia, agente emulsionante, entre otras (Yang et al. 2017).

Base de datos de sustancias de tintes para cabello (HDSD).

La base de datos HDSD (Williams et al. 2017), contiene información de sustancias empleadas para teñir el cabello, incluyendo colorantes temporales, semipermanentes y precursores de tintes permanentes.

Los colorantes temporales son definidos en esta base de datos como moléculas de elevado peso molecular que se unen a la superficie del cabello mediante enlaces débiles, por lo cuales son removidos con facilidad después de varias etapas de lavado.

Los colorantes semipermanentes se refieren a compuestos de menor peso molecular que logran ubicarse en el interior de las fibras del cabello, donde forman enlaces débiles. Estas sustancias son paulatinamente eliminadas, a causa de la expansión de los poros capilares durante el enjuague del cabello (Williams et al. 2018).

Los precursores de tintes permanentes constituyen moléculas pequeñas, que no presentan color, pero se difunden dentro de las fibras capilares y reaccionan, formando oligómeros colorantes que quedan atrapados en el interior de estas. Además, estos compuestos formados se pueden enlazar covalentemente a sitios específicos en el interior del cabello, lo cual les impide que sean removidos por el lavado (Williams et al. 2018).

Conjunto QM7

El conjunto de datos QM7 contiene información de 7102 moléculas orgánicas estables en forma de coordenadas tridimensionales y datos de sus correspondientes energías de atomización. Las moléculas que incluye dicho conjunto contienen en su estructura hasta 23 átomos, contando con un máximo de 7 átomos pesados (carbono, nitrógeno, oxígeno y azufre). La geometría de las moléculas del conjunto fue generada mediante la librería OpenBabel, versión 2.3.2 (R. Guha et al., 2006) y la determinación de las energías fue realizada mediante cálculos de Teoría de Densidad Funcional (DFT) implementados en Gaussian (M. J. Frisch et al., 2009), version 09; usando el método de Perdew–Burke–Ernzerhof (PBE0) y def2-TZVP como conjunto base (Rupp 2015).

Conjunto QM9

Este conjunto agrupa información de 133885 compuestos en forma de coordenadas cartesianas y 19 etiquetas que incluyen energías de orbitales HOMO, LUMO, y la brecha energética entre estos. Las moléculas de este conjunto incluyen en su estructura hasta 9 átomos de carbono, nitrógeno, oxígeno y flúor. La geometría de los compuestos fue obtenida a partir de los SMILES, usando el programa Corina (Version 3.491 2013). Las energías se determinan usando cálculos de Teoría de Densidad

Funcional (DFT) en Gaussian 09 (M. J. Frisch et al. 2009) y el método B3LYP (Ramakrishnan et al. 2014).

- Cómputo en la nube: Google Colaboratory

Colaboratory (http://colab.research.google.com) consiste en una plataforma gratuita a la que el usuario con una cuenta de Google, puede acceder desde navegadores como Firefox y Chrome. Esta realiza la ejecución de cálculos computacionales en la nube, con acceso a una GPU, de modo que no se requiere la adquisición e instalación de equipamientos por parte del programador. Dicha plataforma permite la implementación de código en el lenguaje de programación de Python, mediante la integración de varias librerías como NumPy, Matplotlib y Scikit-Learn. Además, soporta el procesamiento y visualización de archivos de gran tamaño, al importarlos directamente desde Google Drive o una ubicación local, hacia el entorno de trabajo (Prashanth, Mendu, and Thallapalli 2021).

- Librería RDKit

RDKit (https://www.rdkit.org/) incluye un conjunto de herramientas destinadas al estudio y manipulación de moléculas. Esta permite obtener representaciones estructurales de compuestos a partir de sus identificadores (SMILES), así como generar códigos identificadores correspondientes a una representación molecular determinada. Además, contiene funciones que analizan estructuras químicas, reconociendo el tipo y número de átomos que contiene, enlaces químicos, grupos funcionales, entre otros atributos (Landrum 2019).

- Librería QML

La librería de QML (https://github.com/qmlcode/qml) contiene un conjunto de módulos que permiten la aplicación de modelos de aprendizaje automático, en el lenguaje de programación de Python. Entre estos módulos se encuentran disponibles, objetos para realizar la lectura del contenido de archivos, la codificación de estructuras químicas de modo accesible para el modelo; así como la implementación de diferentes

representaciones de características de los compuestos (AS Christensen, FA Faber, B Huang, LA Bratholm, A Tkatchenko, KR Muller 2017).

- Librería SchNetPack

SchNetPack (https://schnetpack.readthedocs.io/en/stable/) constituye una librería que emplea el lenguaje de Python para implementar modelos de aprendizaje profundo (Deep Learning) y específicamente arquitecturas SchNet, de Redes Neuronales Convolucionales. Esta incluye varios conjuntos de datos de referencia, para su fácil acceso. Además, presenta módulos en forma de capas, los cuales se modifican y pueden ser ordenados para configurar la arquitectura de red definida por el usuario (Schütt et al. 2019).

VII.II. MÉTODOS

Construir un conjunto de compuestos colorantes, teniendo en cuenta la aplicación de criterios para el curado de bases de datos moleculares.

Como primer paso de la metodología, se deben explorar varios repositorios y bases de datos para extraer información de compuestos colorantes. A continuación, se identifican las fuentes y los criterios utilizados para la selección en cada caso.

1. Biblioteca de tintes Max Weaver.

De esta extensa colección se encuentra disponible una muestra de 150 estructuras químicas de colorantes textiles, las cuales son utilizadas para la construcción del conjunto de datos de estudio.

2. Base de datos ChemSpider.

Esta base de datos reporta resultados experimentales correspondientes a técnicas espectroscópicas, los cuales se utilizan para seleccionar moléculas que presenten longitud de onda de absorción máxima ($\lambda_{máx}$) correspondiente al espectro visible.

Además, se muestran los compuestos identificados según su aplicación fundamental, con lo cual se extraen estructuras etiquetadas como colorantes (dye).

3. Base de datos PubChem.

Para la selección de compuestos de esta base de datos, se identifican moléculas que en su descripción incluyan color.

4. Base de datos COSMOS.

Para realizar una selección de compuestos, se procede a identificar estructuras etiquetadas como tintes de cabello (Hair dyeing) y colorantes cosméticos (cosmetic colorant).

5. Base de datos de sustancias de tintes para cabello (HDSD).

En cuanto a la selección de moléculas colorantes, se extraen aquellas etiquetadas como tintes permanentes y semipermanentes, debido a que los denominados precursores no presentan color.

Una vez que se obtiene una selección de colorantes de los diferentes repositorios, se agrupan y se procede a realizar el curado de los datos moleculares (Figura 3); con la finalidad de eliminar del conjunto aquellos compuestos que no sean de interés o puedan conducir a errores en cálculos teóricos.

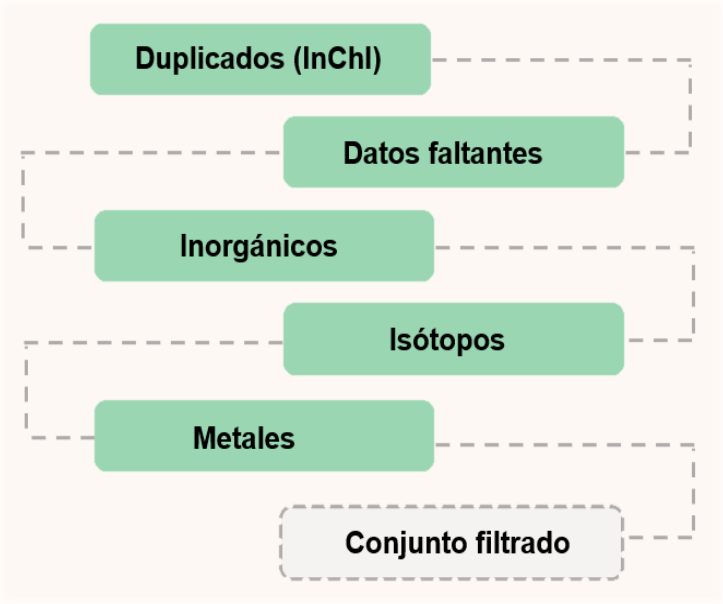


Figura 3. Criterios para el curado del conjunto de datos.

Para ello se desarrolla un proceso que consiste en aplicar varios criterios según se describe a continuación:

I. Eliminar compuestos duplicados, según el código identificador InChI.
II. Remover del conjunto aquellas moléculas a las cuales le falte información.
III. Eliminar sustancias inorgánicas.
IV. Descartar moléculas que presenten isótopos en su estructura.
V. Retirar compuestos metálicos.

Una vez realizado este proceso, se obtiene un conjunto de moléculas colorantes que será utilizado como punto de partida para la selección de compuestos candidatos a formar pigmentos híbridos.

Seleccionar compuestos de acuerdo a la similitud con el índigo y dehidroíndigo, basada en el uso de huellas dactilares moleculares.

La selección de moléculas activas en un área de aplicación, a partir del elevado número de compuestos presentes en bases de datos, se denomina cribado virtual (virtual screening). Un método comúnmente empleado para desarrollar esta tarea consiste en extraer moléculas similares a compuestos de actividades estudiadas, a partir de huellas dactilares (Cereto-Massagué et al. 2015).

Las huellas dactilares moleculares se basan en transformar la estructura de los compuestos a un vector binario de longitud variable. Para el caso de aquellas huellas basadas en subestructuras (Substructure keys-based fingerprints), los caracteres presentes en el vector dependen de la presencia (1) o ausencia (0), en la molécula, de aquellas características o subestructuras predefinidas en un registro. El número de elementos de este listado determina a su vez, la longitud del vector de huella dactilar (Muegge and Mukherjee 2016).

En este estudio, una vez que ha construido un conjunto de compuestos colorantes, se escogen aquellas moléculas que presenten en su estructura grupos característicos de las moléculas de índigo y dehidríndigo: grupos amino, carbonilos y anillos de cinco miembros.

A continuación, dichos compuestos son codificados utilizando la huella dactilar MACCSkeys. En esta clase de huella, las subestructuras se encuentran definidas por SMARTS, que incluyen ciertos grupos, enlaces o elementos descritos en una lista determinada. Este procedimiento transforma a cada molécula en un vector de 166 caracteres binarios (Pattanaik & Coley 2020).

Una vez que se tiene una codificación de la estructura molecular, es posible realizar comparaciones y determinar similitudes entre compuestos. En este caso, se evalúa la similitud de los colorantes del conjunto respecto al índigo y dehidroíndigo, determinando el Índice de Dice según la ecuación 8.

$$\mathrm{Dice} = \frac{2c}{a + b} \tag{8}$$

En esta expresión, el numerador indica el número de elementos comunes en los vectores a comparar (c) y el denominador, el total de elementos de ambos vectores. a y b representan la cantidad de caracteres iguales a 1 para cada vector (A; B). El índice resultante toma valores entre 0 y 1, aumentando el número de elementos comunes entre ambos vectores a medida que el valor de este índice se acerca a la unidad (Cereto-Massagué et al. 2015).

En la figura 4 se muestra el diagrama que resume el procedimiento descrito para la selección de compuestos colorantes.

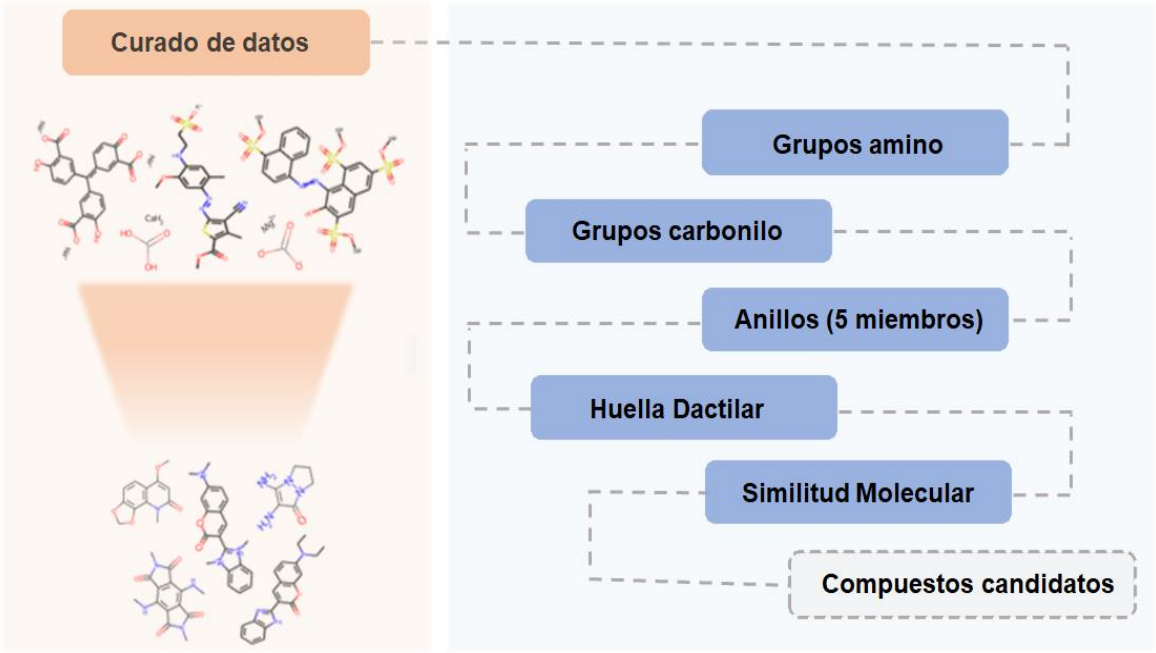


Figura 4. Selección de compuestos de acuerdo a la similitud con el índigo y dehidroíndigo.

Determinar Energías de Atomización para moléculas pequeñas, utilizando un modelo de Kernel Ridge Regression y QM7 como conjunto de entrenamiento.

Se conoce que Modelos de Kernel Ridge Regression pueden ser ajustados con éxito sobre conjuntos de datos conocidos, como son el QM7, para realizar predicciones de Energía de Atomización (Rupp et al. 2012; Rupp 2015). Por tanto, se empleará el QM7 para construir un modelo y se evaluará su capacidad para realizar predicciones en compuestos desconocidos y que no pertenezcan a este conjunto predefinido. Para ello, se siguen los pasos descritos a continuación:

a) Seleccionar el conjunto de datos de estudio.

Las estructuras y propiedades de los compuestos, se obtienen de la base de datos PubChem. Esta base de datos, permite realizar una búsqueda de los compuestos relacionados con una estructura fijada por el usuario, mediante la opción de dibujar y editar su arreglo, para seleccionar los fragmentos de interés. En este caso, se realiza la búsqueda sobre la base de las moléculas de índigo y dehidroíndigo, efectuando transformaciones en sus estructuras hasta obtener varios subconjuntos de datos.

Luego, se procede a realizar el análisis de estos subconjuntos, teniendo en cuenta el número de compuestos que contienen, distribución de peso molecular, homogeneidad, entre otros. Además, se realiza una selección, teniendo en cuenta la presencia de grupos característicos de los colorantes de referencia.

Luego, se determina el porcentaje de similitud de cada compuesto con la molécula de índigo. Para ello, se codifican las estructuras químicas empleando Huellas dactilares de Morgan, equivalentes a las Huellas Dactilares de Conectividad Extendida. Esta clase de huellas topológicas se basan en dividir la estructura del compuesto en varios segmentos, de determinado número de enlaces, siguiendo una trayectoria radial. Cada carácter de la huella representa una de estas secciones, generadas siguiendo diferentes rutas que inician en cada átomo del compuesto. Por tanto, esta clase de huella siempre representa las características de la molécula y la extensión del vector puede ser fijada por el usuario (Capecchi, Probst, and Reymond 2020). En este caso, se utiliza la Huella dactilar de Morgan con radio de 2 enlaces químicos o Huella Dactilar de Conectividad Extendida con diámetro igual a 4 enlaces (ECFP4). Se establece una longitud de 1024 caracteres para cada vector y se construyen empleando la herramienta RDKit.

La similitud molecular se evalúa determinando el Coeficiente de Tanimoto o Jaccard, que constituye la medida de similitud de uso más generalizado. Este parámetro se determina como el cociente del número de caracteres comunes que toman valor 1 para ambas huellas dactilares (c) y la cantidad total de caracteres que ocupan cifras de 1 entre las dos huellas a comparar (ecuación 9). En esta ecuación, a y b representan la cantidad de bits iguales a 1 para cada vector.

$$\text{Tanimoto} = \frac{c}{a+b-c} \tag{9}$$

Este coeficiente tiene valores entre 0 y 1, donde un índice de 0 indica que ambos compuestos no tienen elementos en común y en cambio, un coeficiente igual a la unidad se obtiene para dos compuestos idénticos (Cereto-Massagué et al. 2015).

A continuación, se debe identificar las moléculas que cumplan con los parámetros establecidos para el conjunto QM7. Para ello se determina el número total de átomos de cada molécula, así como el contenido de átomos de carbono, nitrógeno, oxígeno y azufre. Extrayendo solo las muestras que no excedan de 23 átomos y 7 de los átomos especificados.

Se determina la representación de los compuestos, en forma de coordenadas cartesianas tridimensionales, empleando la herramienta OpenBabel (R. Guha et al. 2006) y luego son transformados a Matrices de Coulomb mediante la librería QML (AS Christensen et al. 2017).

Luego se deben obtienen los valores de energías de atomización de las moléculas de estudio, empleando métodos mecánico-cuánticos. Para ello, se toman las coordenadas de los compuestos y su estructura es relajada mediante cálculos de DFT en Orca, utilizando el método de Perdew–Burke–Ernzerhof (PBE0) y def2-TZVP como conjunto base. Las energías de atomización se calculan sustrayendo a la energía de cada molécula, la suma de las energías de sus átomos componentes.

b) Construir un modelo de Kernel Ridge Regression para la determinación de Energías de Atomización.

Este modelo toma como entrada la estructura de las moléculas en forma de Matriz de Coulomb, que contiene información de las fuerzas repulsivas presentes entre sus átomos componentes y realiza la predicción de la propiedad (E), según la ecuación 10. La arquitectura del modelo se construye haciendo uso de la librería Scikit-learn. Los coeficientes de regresión están dados por α_i y la matriz de Kernels se representa por K (Rupp et al. 2012).

$$E(M) = \sum_i \alpha_i K(M, M_i) \quad (10)$$

Para realizar el ajuste de dicho modelo sobre el conjunto QM7 se emplea el método de Validación Cruzada K-fold, reservando el 30% de los datos como conjunto de prueba (2130 moléculas) y dividiendo el resto en 10 bloques (10-fold cross-validation). Se implementa una ¨Grid search¨ para evaluar diferentes combinaciones de

hiperparámetros en base logarítmica, con funciones de Kernel Gaussiano (ecuación 11) y Laplaciano (ecuación 12), hasta encontrar la combinación con la que el modelo muestre mejor desempeño.

$$\mathrm{K}(\mathrm{M},\mathrm{M_i}) = \mathrm{e}^{\left(-\frac{\|\mathrm{M}-\mathrm{M_i}\|^2}{2\sigma^2}\right)} \tag{11}$$

$$\mathrm{K}(\mathrm{M},\mathrm{M_i}) = \frac{\|\mathrm{M}-\mathrm{M_i}\|}{\sigma} \tag{12}$$

Siendo, σ la longitud de escala de las funciones de Kernel aplicadas.

En la tabla 2 se muestra el rango de valores a analizar para cada parámetro de estudio.

Tabla 2. Hiperparámetros del modelo.

Hiperparámetros	**Variantes**						
Kernel	Laplaciano			Gaussiano			
Lambda: λ	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	
Sigma: σ	1	10	10^2	10^3	10^4	10^5	10^6
K	10						

Una vez que se conoce que este modelo es capaz de realizar predicciones de energía de atomización sobre el conjunto QM7, se procede a aplicar el modelo ajustado, sobre el conjunto de moléculas de estudio. Para ello se alimentan al modelo las coordenadas cartesianas correspondientes y se obtienen predicciones utilizando la combinación de hiperparámetros y la función de Kernel, resultante del proceso de validación cruzada.

Luego se pretende evaluar la transferibilidad del modelo a compuestos que presenten más de 7 átomos de carbono, nitrógeno, oxígeno y azufre en su estructura. Por tanto, se toman 57 compuestos de un conjunto conocido, denominado QM9, que incluyen moléculas con hasta 9 de los átomos especificados y sus Energías de Atomización (etiquetas). El modelo toma como variables de entrada, la estructura de los

compuestos y evalúa el Error absoluto Medio de las predicciones generadas, en relación a sus etiquetas.

Realizar la evaluación de energía de orbitales HOMO y LUMO, así como la brecha energética existente entre estos.

Para realizar la determinación de energías de orbitales, se emplea un modelo de Redes Neuronales Convolucionales SchNet, entrenado sobre el conjunto QM9; mediante la librería SchNetPack. La arquitectura de dicho modelo se explica en el Apéndice II.

El algoritmo toma como entrada las coordenadas cartesianas de los compuestos y crea una representación de su estructura, construyendo 130 características para describir los entornos atómicos. El modelo emplea 130 filtros convolucionales y 5 bloques de interacción. Además, se utilizan 40 funciones gaussianas y una función de corte coseno con radio 10.

A continuación, se ubica un módulo de red neuronal de dos capas, que toma la representación generada en el primer bloque y realiza el cálculo de la propiedad a predecir; tomando la media y desviación estándar como primera suposición para inicializar el modelo.

Los datos son alimentados al modelo en lotes de 100 elementos, para un total de 50000 compuestos en la etapa de entrenamiento, 1000 para validación y 82885 para la etapa de prueba. El número de elementos a utilizar para el entrenamiento del modelo está determinado por la capacidad de cómputo requerida para realizar los cálculos.

El entrenamiento del modelo se realiza mediante el optimizador Adam, con una tasa de aprendizaje inicial de $1 * 10^{-3}$ y la evaluación del Error Cuadrático Medio en los cálculos. La tasa de aprendizaje se programa para que sea modificada durante el proceso de optimización. El parámetro que indica el número de épocas que deben transcurrir sin cambios en el valor de la función de error, para que se modifique la taza de aprendizaje, se define como patience. Para este caso, la taza de aprendizaje se reduce en un factor de 0.6 con patience de 10, hasta alcanzar un valor mínimo de

$1 * 10^{-6}$. Se definen 100 épocas de entrenamiento y como métrica se calcula el Error Absoluto Medio de las predicciones.

VIII. RESULTADOS Y DISCUSIÓN.

Construir un conjunto de compuestos colorantes, teniendo en cuenta la aplicación de criterios para el curado de bases de datos moleculares.

Se recopilaron 1759 compuestos colorantes procedentes de varios repositorios y bases de datos (Nápoles and Quintana 2021e). Una inspección visual de algunos elementos del conjunto indica que hay presencia de átomos aislados, metales y colorantes inorgánicos; lo cual se corrobora en las distribuciones de Masa Molecular y Número de Átomos de los compuestos (figura 5).

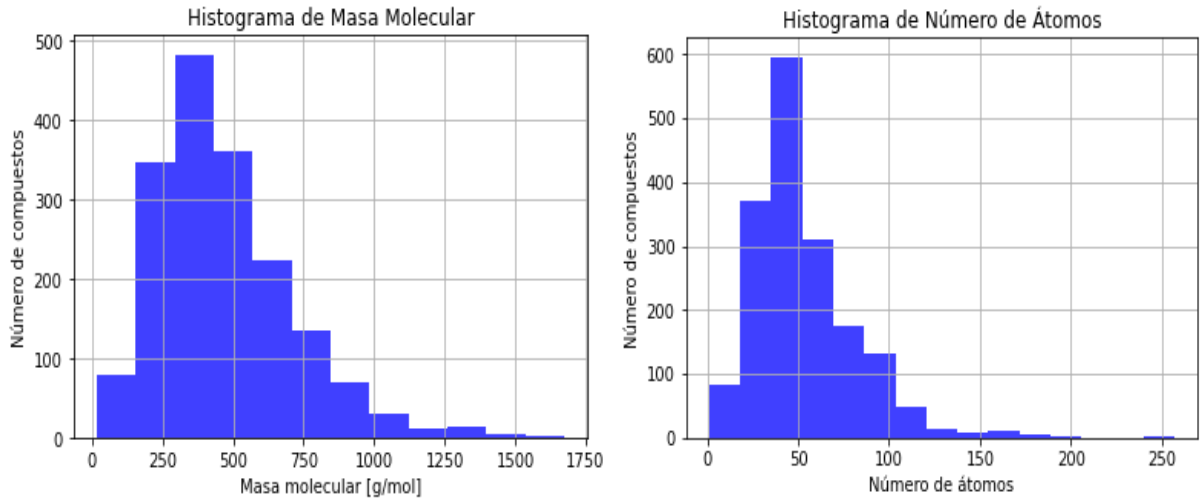


Figura 5. Histogramas de Masa Molecular (izquierda) y Número de Átomos (derecha) para el conjunto inicial.

Por tanto, se deben aplicar criterios para el curado de bases de datos moleculares que permitan eliminar estas sustancias. Una vez aplicados dichos criterios se obtiene un conjunto de 1153 compuestos colorantes (Nápoles and Quintana 2021c).

Seleccionar compuestos de acuerdo a la similitud con el índigo y dehidroíndigo, basada en el uso de huellas dactilares moleculares.

Una vez construido el conjunto de colorantes, se seleccionaron aquellos compuestos que presentan grupos característicos de las moléculas de índigo y dehidroíndigo, generando de este modo 183 moléculas. Luego se realiza una inspección de las

estructuras de dichas moléculas y se retiran aquellos compuestos que presenten irregularidades, quedando un total de 171 moléculas candidatas (Nápoles and Quintana 2021b). En el Apéndice I se muestra el conjunto de moléculas obtenidas y en el Apéndice III, el código correspondiente a la selección de dichos compuestos.

La figura 6 corresponde al histograma que masa molecular del conjunto obtenido, que muestra valores que se distribuyen de 150 a 960 g/mol aproximadamente, con un máximo alrededor de 390 g/mol. Además, se aprecia que el Número de Átomos de este conjunto curado es superior a 16 aproximadamente, con un máximo número de compuestos que contienen alrededor de 50 átomos; lo cual evidencia que han sido eliminados los átomos aislados y metales que se encontraban inicialmente en el conjunto.

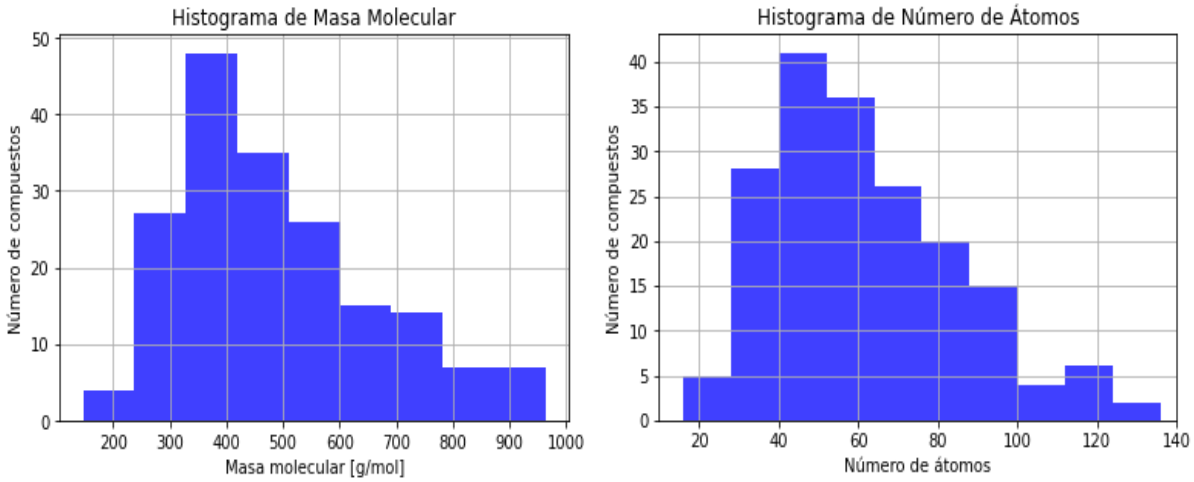


Figura 6. Histogramas de Masa Molecular (izquierda) y Número de Átomos (derecha) para el conjunto de moléculas candidatas.

A continuación, se aprecian los histogramas de similitud de los compuestos, con las moléculas de índigo y dehidroíndigo (figura 7). Estos gráficos evidencian que los elementos del conjunto presentan índices superiores a 0.4 y que la mayoría de los compuestos presentan alrededor del 60% de elementos comunes con los colorantes presentes en el Azul Maya.

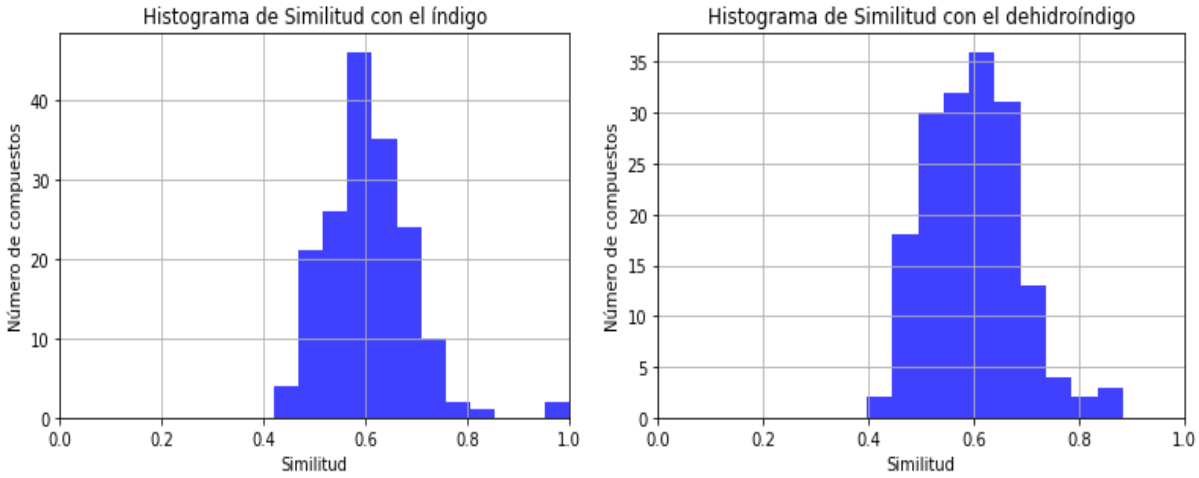


Figura 7. Histogramas de similitud con la molécula de índigo (izquierda) y dehidroíndigo (derecha).

En la figura 8 se muestra la estructura de algunos de los compuestos candidatos obtenidos, que presentan un coeficiente de similitud superior a 0.7 con el índigo y dehidroíndigo. En dicha figura se aprecia que aparecen algunas moléculas de gran tamaño (a, d, g) en comparación con los colorantes de referencia; sin embargo, al mezclarse con la paligorskita, parte de su estructura pudiese ubicarse en el interior de los canales. De este modo pudiesen establecer interacciones con la zeolita, similares a las que se producen el Azul Maya, debido a la presencia en su estructura de grupos amino y carbonilo. Además, se observa la presencia de otras moléculas de menor número de átomos, que pudiesen localizarse en el interior de los canales de la zeolita (b, c, i). Entre estas, para el caso de la molécula de isatina (i), se ha demostrado en trabajos anteriores (Zhou et al. 2014), que al mezclarse con la paligorskita durante un proceso de calentamiento; penetra los canales de la estructura y forma un pigmento resistente a agentes oxidantes.

Se requiere de estudios posteriores, que realicen cálculos de Dinámica Molecular, para analizar el comportamiento de cada caso al interaccionar con la estructura de la paligorskita.

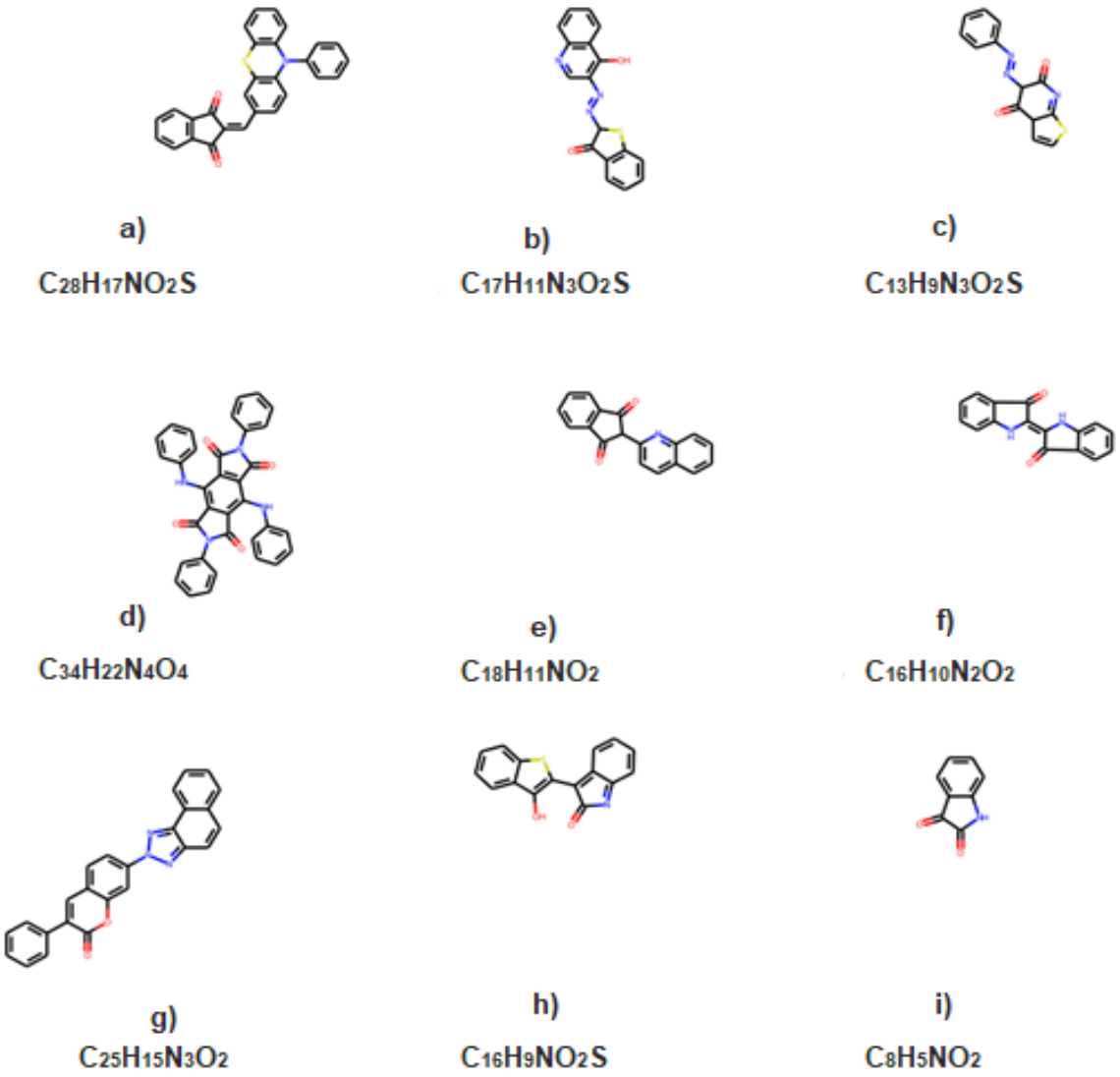


Figura 8. Representación de algunos de los compuestos presentes en el conjunto.

Determinar Energías de Atomización para moléculas pequeñas, utilizando un modelo de Kernel Ridge Regression y QM7 como conjunto de entrenamiento.

a) Seleccionar el conjunto de datos de estudio.

Se obtuvieron 21 subconjuntos de datos de moléculas pequeñas, relacionadas a fragmentos de las moléculas de índigo y dehidroíndigo, de la base de datos PubChem. Para un total de más de 8000 moléculas entre todos los subconjuntos.

Se realizó el análisis, en cuanto a distribución de pesos moleculares y porcentaje de similitud con la molécula de índigo.

Se decide seleccionar dos subconjuntos que contienen compuestos relacionados a fragmentos de la molécula de índigo, que suman un total de 975 moléculas (Nápoles and Quintana 2021d).

El histograma de masa molecular de este conjunto muestra una distribución en el rango de 55 a 229 g/mol, destacando que todas las moléculas de este conjunto presentan menor masa molecular que la molécula de índigo (Figura 9).

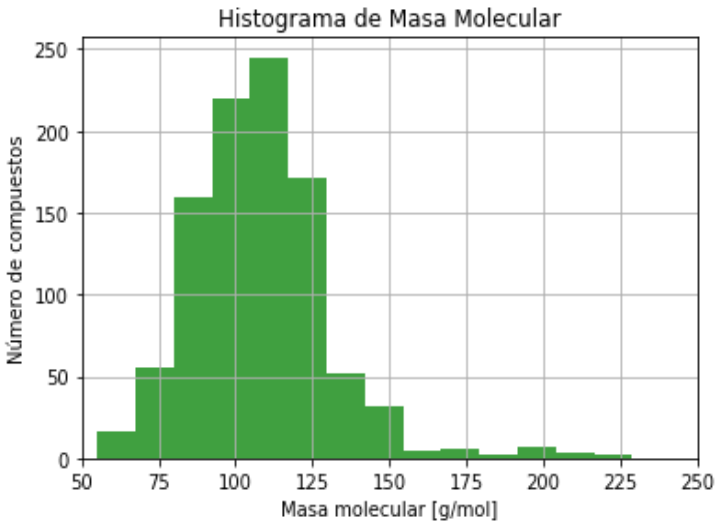


Figura 9. Histograma de masa molecular del conjunto de datos.

Al analizar la distribución del número de átomos se observa que los compuestos presentan entre 7 y 26 átomos, siendo moléculas pequeñas; lo cual indica que este conjunto contiene gran cantidad de muestras que pudiesen ser alimentadas a un modelo entrenado por el conjunto QM7 (Figura 10).

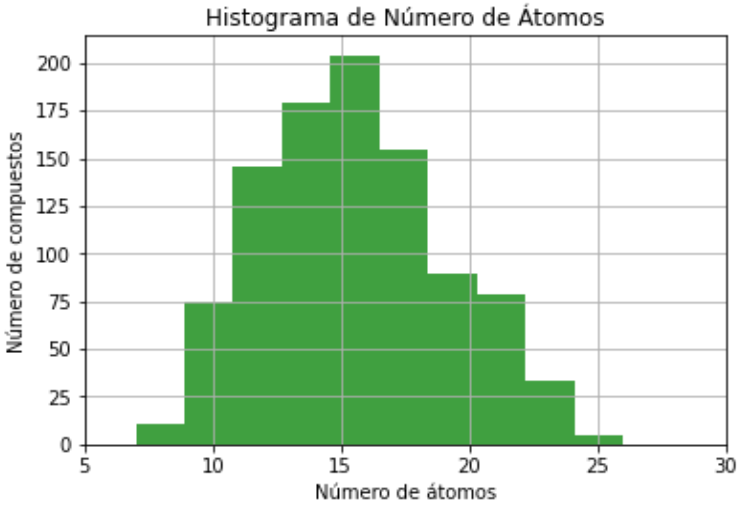


Figura 10. Histograma de número de átomos de las moléculas del conjunto.

Por tanto, se procede a analizar el histograma que representa la distribución de compuestos en cuanto al contenido de átomos de carbono, nitrógeno, oxígeno y azufre. Este gráfico muestra que las moléculas del conjunto contienen entre 4 y 10 átomos de este tipo, con un máximo alrededor de 7; lo cual indica que la mayoría de las muestras cumplen con el segundo parámetro también (Figura 11).

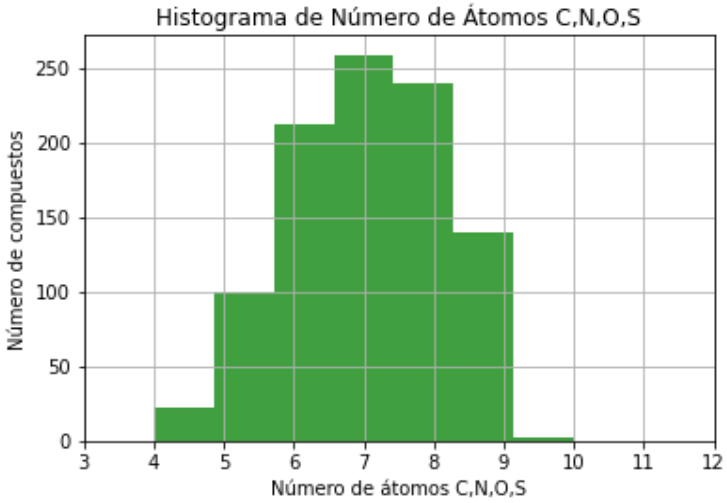


Figura 11. Histograma de número de átomos de C, N, O, S de las moléculas del conjunto.

Seguidamente se decide realizar un filtrado, seleccionando moléculas que presenten los grupos amino, carbonilo y el anillo de 5 miembros característicos de los colorantes de referencia, con lo cual resulta un conjunto de 114 compuestos.

Luego se realiza la distribución del número de átomos que presenta cada molécula (Figura 12), mostrando una distribución de 8 a 21 átomos (todas contienen menos de 23 átomos).

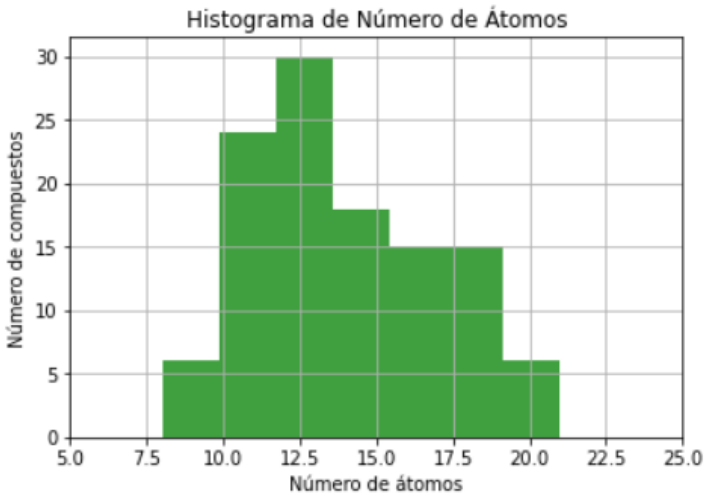


Figura 12. Histograma de número de átomos de las moléculas del conjunto de datos.

Además, se realiza la distribución del número de átomos de carbono, nitrógeno, oxígeno y azufre; manteniéndose en un rango de 6 a 9, por lo que habrá que eliminar del conjunto aquellas moléculas que excedan de 7 de estos átomos (Figura 13).

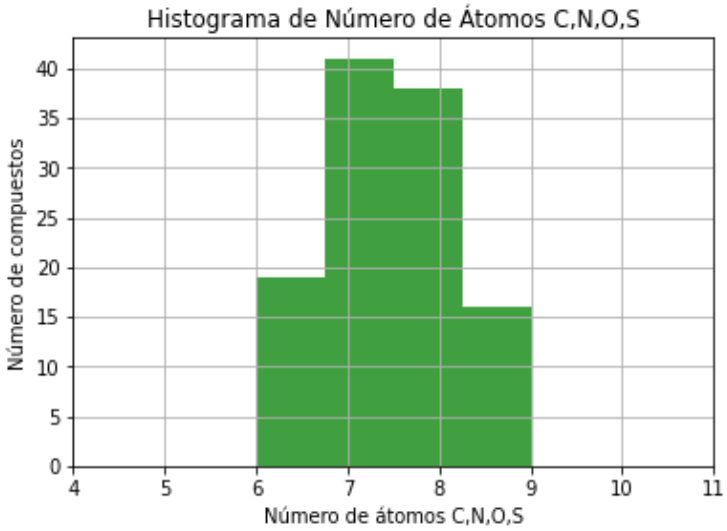


Figura 13. Histograma de número de átomos de C, N, O, S de las moléculas del conjunto.

Luego, se eliminan del conjunto aquellas moléculas que posean átomos de fósforo, flúor, cloro, yodo, bromo y silicio en su estructura, debido a que los compuestos del conjunto QM7 no presentan este tipo de átomos; reduciéndose el conjunto de 114 a 54 moléculas.

Las coordenadas cartesianas de dichos compuestos fueron generadas utilizando OpenBabel (Nápoles and Quintana 2021f). A continuación, se realizó la optimización de las estructuras moleculares a un nivel de Teoría de Densidad Funcional (DFT), utilizando la función Perdew-Burke-Ernzerhof (PBE0) y def2-TZVP como conjunto base. Durante el proceso de relajación de la estructura de las 54 moléculas, el cálculo para 3 de estos compuestos falló. Estos elementos fueron excluidos del conjunto y todos los cálculos posteriores, así como el ajuste del modelo, se realizan sobre un conjunto de 51 compuestos (Nápoles and Quintana 2021a).

Los valores de energías de atomización calculadas, se muestran en la tabla 3 para algunos de los elementos del conjunto.

b) Construir un modelo de Kernel Ridge Regression para la determinación de Energías de Atomización.

El primer paso consiste en el análisis de los datos de entrada. En este caso, los 7102 compuestos que forman el conjunto QM7 presentan valores de energías de atomización que se distribuyen desde -250 hasta -2275 kcal/mol, aproximadamente; donde la mayoría de las moléculas del conjunto presentan energía de alrededor de -1500 kcal/mol, como se muestra en la figura 14. Un modelo ajustado sobre este conjunto de datos debe realizar predicciones confiables en este rango de valores de energías.

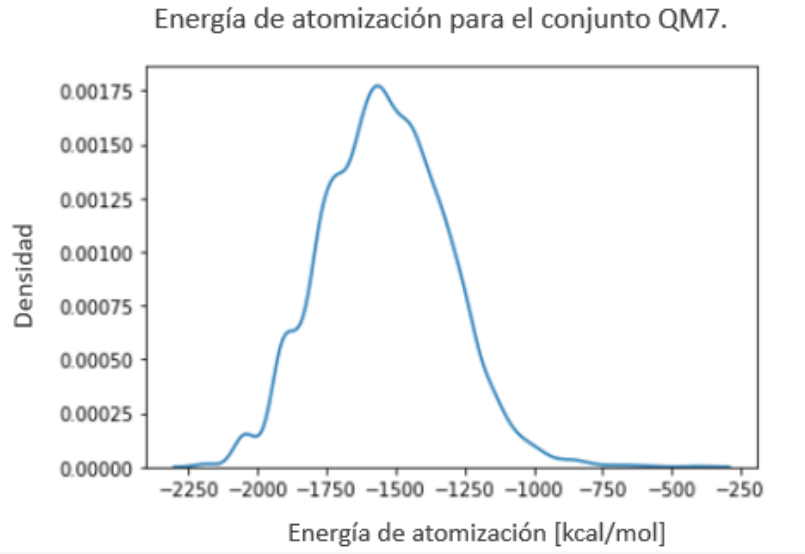


Figura 14. Distribución de Energías de atomización para los compuestos del conjunto QM7.

Una vez que se tiene una idea general del rango de valores por el que están compuestos los datos, se aplica el método de Grid Search para ajustar diferentes modelos de Kernels, utilizando a la vez varias combinaciones de hiperparámetros. Al realizar esta búsqueda, se obtiene mayor desempeño para un modelo de Kernel Laplaciano con valores de α igual a $1 * 10^{-6}$ y σ igual a 100.

Empleando esta óptima combinación de parámetros y la función de Kernel Laplaciano, se realiza nuevamente el ajuste del modelo y se obtienen valores de Error Absoluto

Medio (MAE) de 3.944 kcal/mol con un R^2 de 0.999; para el conjunto de prueba (figura 15). Este modelo mantiene un desempeño superior a otros reportados en la literatura (Rupp, 2015), donde se alcanzan valores de MAE de 8.8 kcal/mol y R^2 igual a 0.998. Dichos resultados indican que el modelo ha realizado un buen ajuste a los datos de entrenamiento y también, ha mostrado buen desempeño sobre datos que no había analizado con anterioridad (conjunto de prueba), por lo que es de esperar que pueda generalizar a nuevos conjuntos de datos.

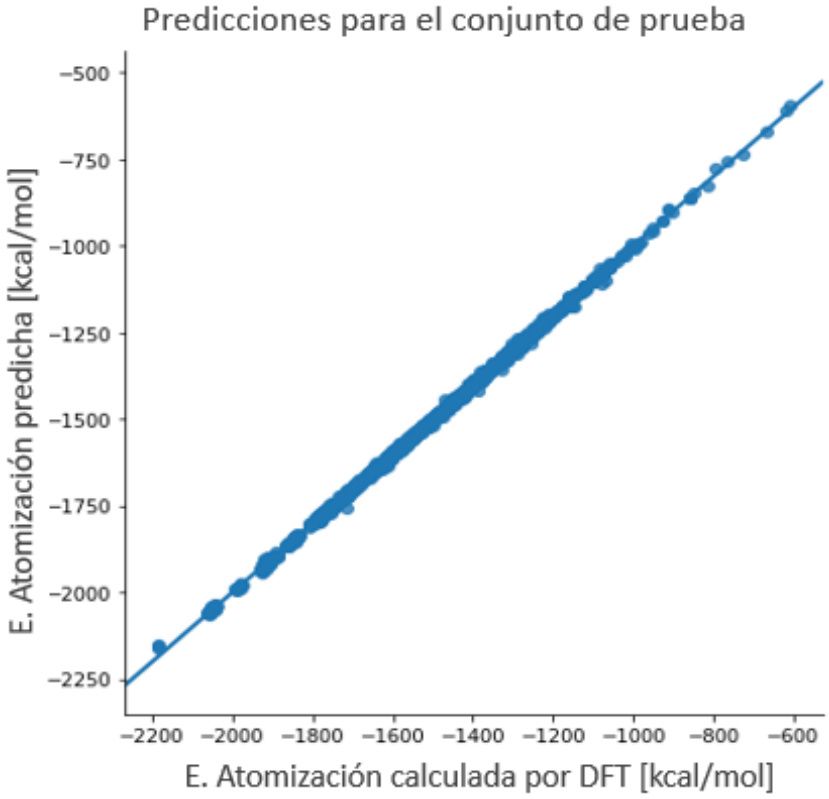


Figura 15. Predicciones del modelo Kernel Ridge Regression con función Laplaciana, para el conjunto de prueba.

Por tanto, se procede a emplear este modelo sobre el conjunto de moléculas de estudio, para realizar predicciones de sus energías de atomización. En la tabla 3 se muestran los resultados de las predicciones para algunos compuestos de este conjunto y los valores de energías de atomización determinados mediante cálculos de DFT.

Tabla 3. Energías de atomización correspondientes al conjunto de moléculas de estudio.

Fórmula general	E. Atom_ DFT [kcal/mol]	E. Atom_ Predicha [kcal/mol]
$C_3H_4N_2O$	-1029.19500	-1023.26553
$C_3H_3NO_2$	-996.957650	-986.232324
C_4H_7NO	-1310.92261	-1301.96491
$C_3H_6N_2O$	-1179.30140	-1166.75310
$C_3H_5NO_2$	-1097.06377	-1095.72358
$C_2H_5N_3O$	-1035.45945	-1045.50502
C_5H_5NO	-1354.04183	-1351.92088
$C_4H_3NO_2$	-1158.09018	-1150.23461
$C_3H_3N_3O$	-1100.40774	-1074.96697
C_5H_5NO	-1341.37385	-1333.14058
$C_3H_2N_2O_2$	-990.178000	-1046.15133
$C_4H_6N_2O$	-1331.67994	-1337.76284
C_3HNO_3	-962.799400	-953.497359
$C_4H_5NO_2$	-1283.43551	-1270.79238
C_5H_9NO	-1603.39589	-1604.45724
$C_3H_4N_2O_2$	-1148.77084	-1151.60244
$C_4H_8N_2O$	-1487.33131	-1471.56325
$C_4H_8N_2O$	-1476.29092	-1470.17564
$C_4H_7NO_2$	-1411.43112	-1408.16409
$C_3H_7N_3O$	-1348.49949	-1358.55946
$C_3H_2N_2OS$	-990.037703	-976.739394
$C_3H_5NO_2S$	-1173.64122	-1169.83149
$C_3H_5NOS_2$	-1141.76135	-1140.80631

En la figura 16 se pueden observar los valores de energías de atomización de todos los compuestos del conjunto, empleando métodos de DFT y el modelo ajustado. Estos resultados indican que las determinaciones por ambos métodos mantienen un comportamiento similar, salvo en algunos casos puntuales donde los valores predichos por el modelo son ligeramente menores.

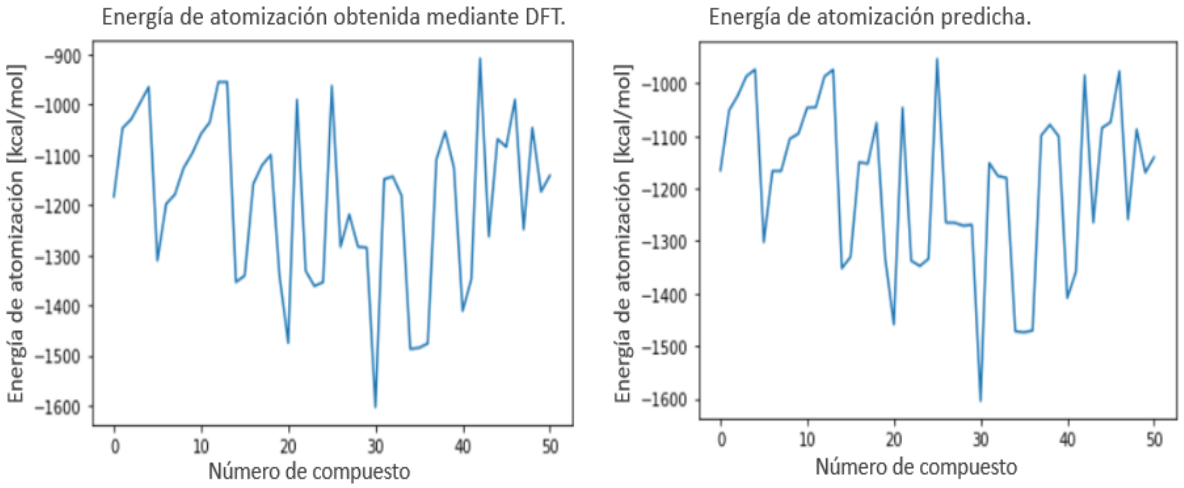


Figura 16. Valores de energías de atomización de las moléculas de estudio. Resultados calculados mediante métodos mecano-cuánticos (izquierda) y valores predichos por el modelo (derecha).

Este comportamiento en los resultados, se puede corroborar con los gráficos de la figura 17, donde se aprecia que, para ambos métodos la distribución de las energías de atomización se encuentra en un rango de -700 y -1900 kcal/mol aproximadamente. Donde la mayoría de los compuestos del conjunto muestran valores alrededor de -1100 kcal/mol, para ambos casos.

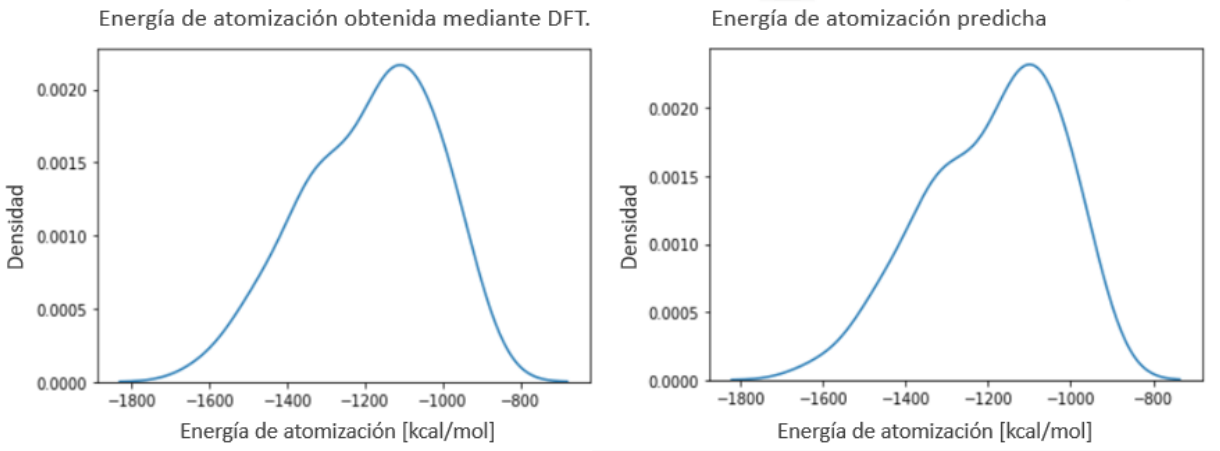


Figura 17. Distribución de Energías de atomización predicha. Resultados calculados mediante DFT (izquierda) y valores predichos por el modelo (derecha).

El correcto desempeño del modelo para este nuevo conjunto de datos, se puede corroborar con el gráfico de la figura 18. En este gráfico se representan los valores de energías de atomización calculados mediante métodos de DFT y los predichos por el modelo. Al determinar las métricas correspondientes a dichos valores, se obtienen estadígrafos de R^2 igual a 0.982 y Error Absoluto Medio igual a 16.869 kcal/mol.

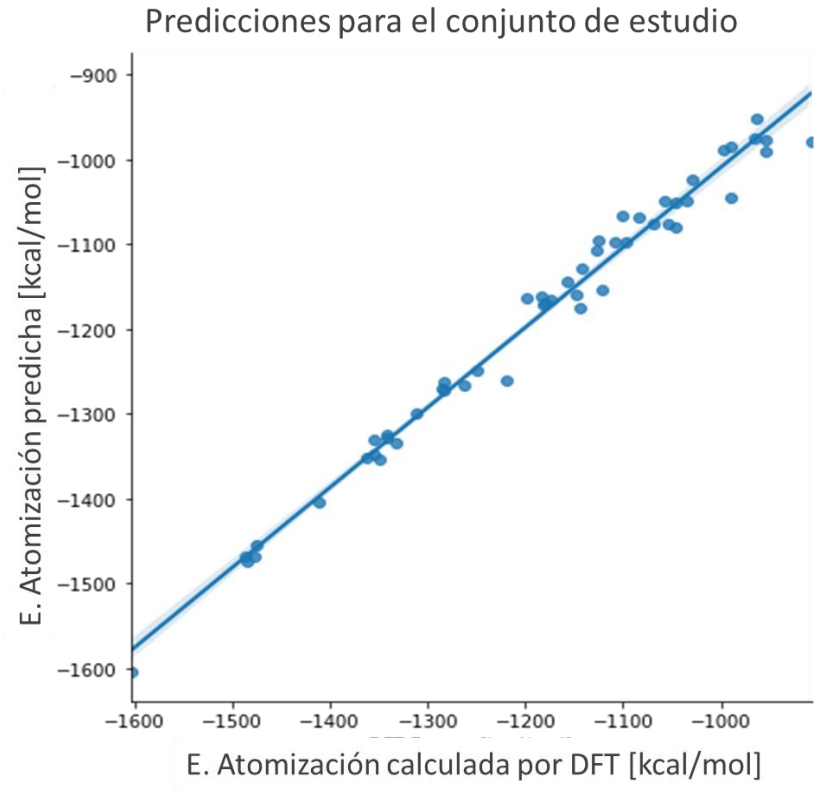


Figura 18. Gráfico de valores de energías calculados y predichos por el modelo.

En cuanto a la evaluación de la transferibilidad del modelo para compuestos que presentan más de 7 átomos de los especificados en su estructura, se muestra en la tabla 4, valores de las energías reales y las predicciones realizadas por el modelo para algunos de los casos de estudio.

Tabla 4. Energías de atomización correspondientes a moléculas de estudio con más de 7 átomos C, N, O, S.

Fórmula general	E. Atom_ DFT [kCal/mol]	E. Atom_ Predicha [kCal/mol]
$C_6H_{13}N_3$	-1562.9699	-1491.5139
$C_8H_{10}O$	-1705.5575	-1693.0242
$C_7H_8O_2$	-1520.9588	-1486.3370
$C_6H_8N_2O$	-1462.3331	-1427.9713
C_9H_{10}	-1718.3430	-1730.0271
$C_7H_{14}N_2$	-1532.5777	-1478.1880
C_8H_9N	-1595.3299	-1616.8245
C_8H_8O	-1532.5495	-1546.5921
C_9H_{12}	-1892.7585	-1872.5570
$C_8H_{10}O$	-1678.0331	-1695.9929
$C_7H_{14}N_2$	-1702.9025	-1637.7309
$C_8H_{11}N$	-1773.8430	-1757.8560
$C_8H_{10}O$	-1704.2767	-1676.0870
$C_7H_8O_2$	-1520.1612	-1492.0849
C_9H_{12}	-1893.2335	-1864.2409
$C_7H_8N_2$	-1667.0454	-1635.6347
C_8H_9N	-1794.6004	-1768.1544
C_7H_7NO	-1603.4718	-1570.6203
$C_8H_{11}N$	-1921.2637	-1908.9834

El gráfico de los valores de energías de atomización calculados mediante métodos mecano-cuánticos y los predichos por el modelo, muestra cierta dispersión de los puntos respecto a la línea; con un estadígrafo R^2 igual a 0.904 y Error Absoluto Medio de 33.013 kcal/mol. Estos resultados evidencian que la transferibilidad del modelo a compuestos de mayor número de átomos es limitada (Figura 19).

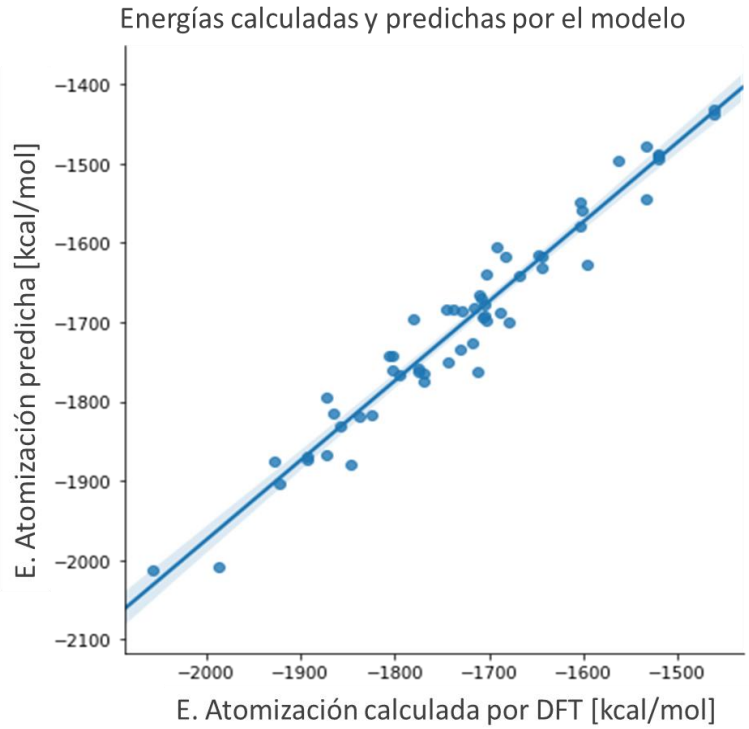


Figura 19. Gráfico de valores de energías calculados y predichos por el modelo para compuestos de más de 7 átomos C, N, O, S.

Realizar la evaluación de energía de orbitales HOMO y LUMO, así como la brecha energética existente entre estos.

Se ajustó un modelo de Redes Neuronales Convolucionales SchNet sobre el conjunto QM9, empleando la librería SchNetPack. Este modelo realiza predicciones de Energías de Atomización, Energía de orbitales HOMO, LUMO y brecha energética; para compuestos que presenten hasta 9 átomos de carbono, nitrógeno, oxígeno y flúor en su estructura.

Como se comentó en la metodología, del conjunto inicial de datos (133885 compuestos), se emplean 50000 elementos para la etapa de entrenamiento, 1000 para la validación y el resto como conjunto de prueba.

La figura 20 muestra la correspondencia entre los valores de Energías de Atomización predichos por el modelo en relación a los datos de referencia, para 1000 de las

moléculas presentes en el conjunto de prueba. Esto se refleja por la coincidencia de los valores graficados con una línea, que indica el correcto ajuste del modelo y su generalización hacia los datos de prueba, con Error Absoluto Medio de 0.03 eV.

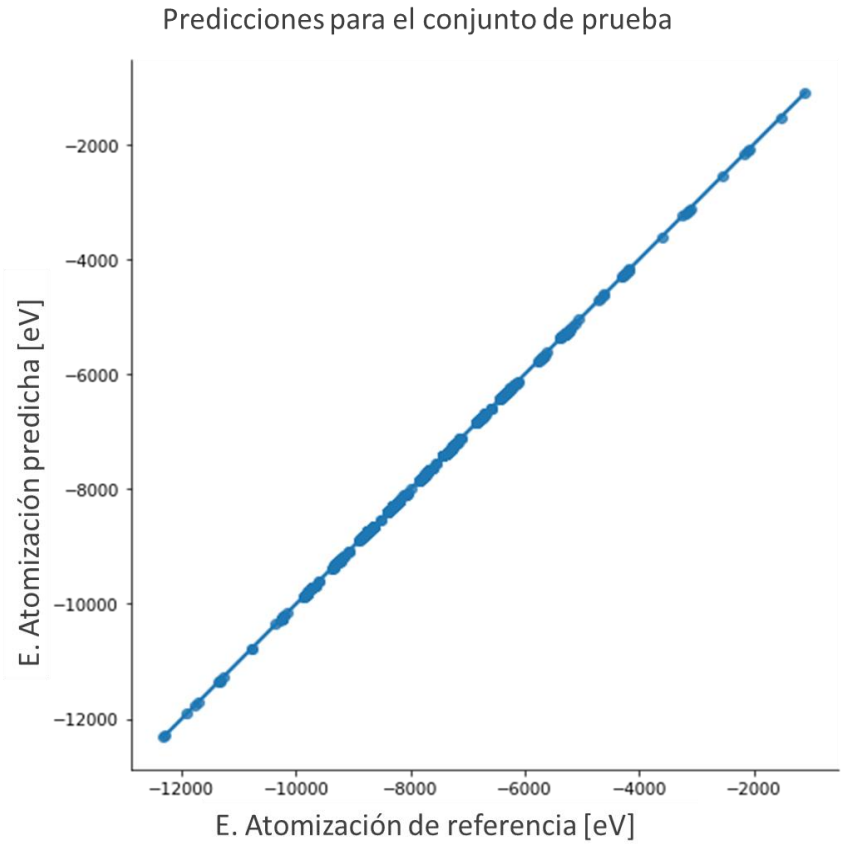


Figura 20. Gráfico de valores de energías calculados y predichos por el modelo.

En la figura 21 se muestran las predicciones del modelo para la energía de orbitales HOMO, sobre el conjunto de prueba. En este caso se aprecian algunos puntos de datos ubicados fuera de la diagonal del gráfico; sin embargo, la mayoría de los valores están en correspondencia con los datos de energías de referencia. Este comportamiento se refleja en un valor de Error Absoluto Medio ligeramente superior al obtenido para el caso anterior, siendo igual a 0.05 eV.

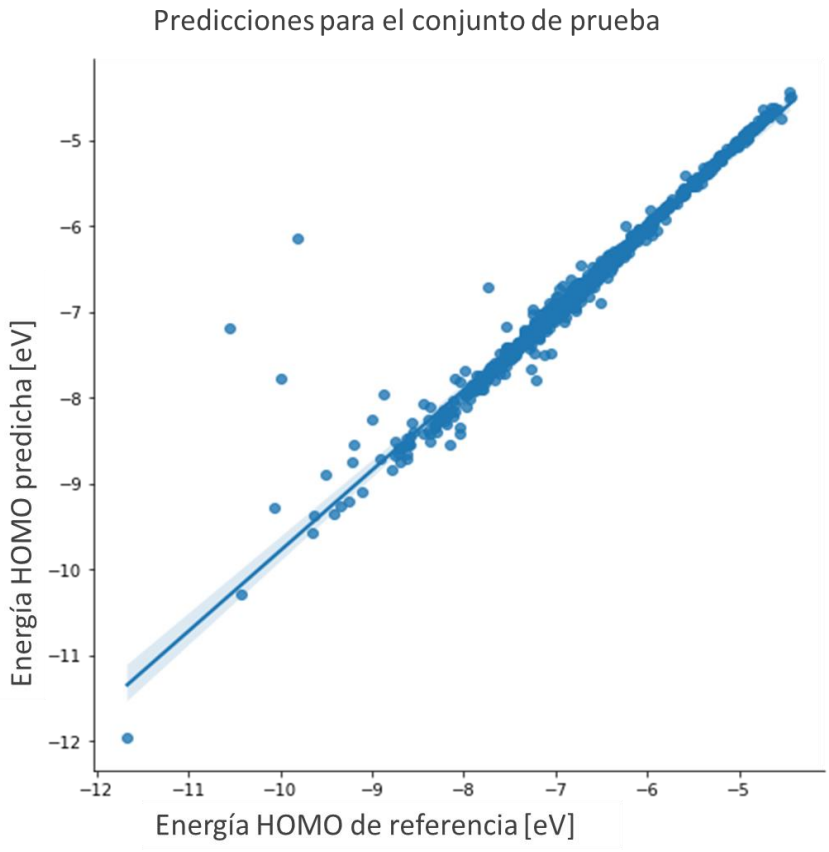


Figura 21. Energía del orbital HOMO para los compuestos en el conjunto de prueba.

La figura 22 muestra los resultados de las predicciones realizadas por el modelo, para la energía de orbitales LUMO, durante la etapa de prueba. En este caso, se obtienen valores de Error Absoluto Medio de 0.04 eV y en el gráfico se aprecia el correcto desempeño del modelo, con una coincidencia de los cálculos realizados y las etiquetas del conjunto, para la mayoría de los casos.

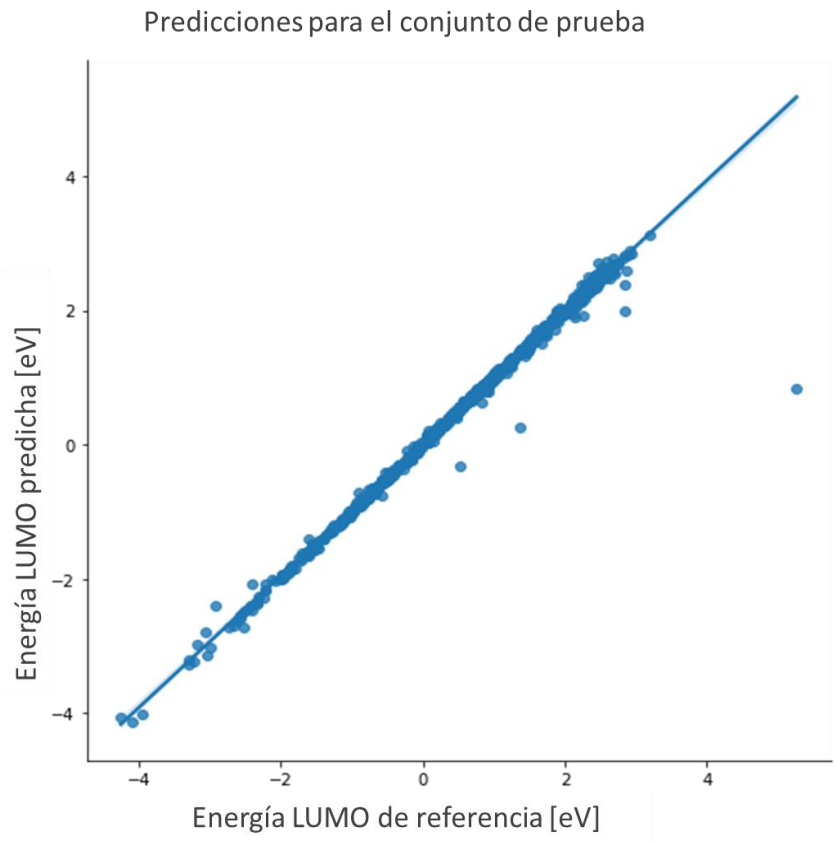


Figura 22. Energía del orbital LUMO para los compuestos en el conjunto de prueba.

La figura 23 muestra el gráfico que relaciona los valores de brecha energética entre orbitales HOMO-LUMO y las predicciones generadas por el modelo, para los compuestos del conjunto de prueba. En esta figura, se aprecia que los puntos mantienen una tendencia hacia la diagonal, con ejemplos en los que valores calculados por el modelo difieren de los de referencia. El Error Absoluto Medio de las predicciones para este caso, es de 0.08 eV; que a pesar de no ser un valor elevado muestra ser superior a los errores obtenidos para el cálculo de las propiedades anteriores.

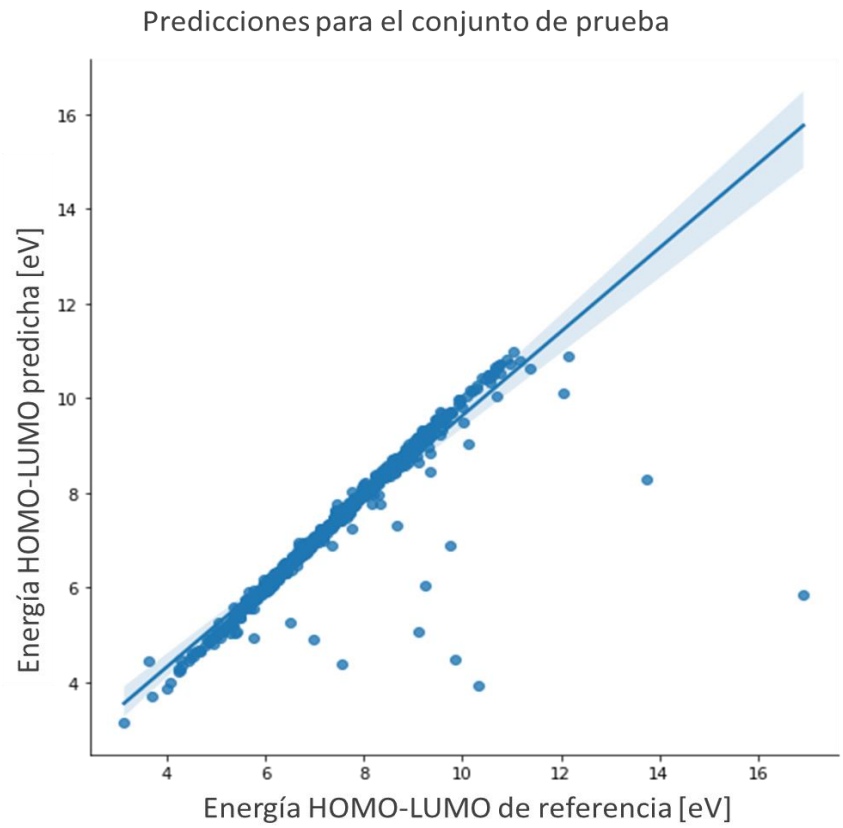


Figura 23. Brecha energética HOMO-LUMO para los compuestos del conjunto de prueba.

Para todos los casos de las propiedades predichas, se obtienen valores de Error Absoluto Medio similares a los reportados en la literatura consultada: 0.014 eV para las Energías de Atomización, 0.041 eV para energías de orbitales HOMO, 0.034 eV para energías de orbitales LUMO y 0.063 eV para la brecha energética entre orbitales HOMO-LUMO. En ese estudio, se construyen 64 características atómicas, empleando 6 bloques de interacción y 110000 elementos para entrenar el modelo. (Schütt et al. 2018). En el presente trabajo, solo se usan 50000 compuestos para la etapa de entrenamiento, debido a limitaciones de recursos computacionales; sin embargo, la similitud de los resultados obtenidos con los reportados, indican que el desempeño del modelo, de modo general es aceptable. Este desempeño pudiese mejorarse, aumentando el número de épocas y compuestos a utilizar durante la etapa de entrenamiento; así como efectuando variaciones en el proceso de optimización, que

incluyen disminuir el factor en que cambia la tasa de aprendizaje y aumentar el valor de patience empleado.

IX. CONCLUSIONES

A continuación, se muestran las conclusiones obtenidas durante el desarrollo de este estudio:

- Se ha construido un conjunto de datos de moléculas colorantes, que incluye un total de 1153 compuestos y que puede ser ampliado posteriormente. Este conjunto, se encuentra disponible para ser utilizado en futuros trabajos que requieran de un repositorio con este tipo de compuestos.

- Se obtuvo 171 compuestos colorantes, que presentan grupos característicos de las moléculas de índigo y dehidroíndigo; los cuales constituyen candidatos a formar pigmentos híbridos estables a base de paligorskita.

- Se cuenta con un modelo de Regresión de Kernel Laplaciano, que predice energías de atomización en compuestos de hasta 7 átomos de C, N, O, S.

- El empleo de un modelo de Redes Neuronales Convolucionales, ajustado sobre el conjunto QM9, muestra ser útil para el cálculo de propiedades como Energía de orbitales HOMO, LUMO y gap existente entre estos.

Como recomendación para trabajos futuros queda:

- Aumentar el número de compuestos a utilizar durante la etapa de entrenamiento del modelo de redes neuronales.
- Realizar estudios de Dinámica Molecular para analizar las interacciones que pudiesen producirse entre las moléculas candidatas y la paligorskita.

X. BIBLIOGRAFÍA

Alsuliman, Tamim, Dania Humaidan, and Layth Sliman. 2020. “Machine Learning and Artificial Intelligence in the Service of Medicine: Necessity or Potentiality?” *Current Research in Translational Medicine* 68(4):245–51. doi: https://doi.org/10.1016/j.retram.2020.01.002.

Antony John Williams. 2007. “ChemSpider.” *Royal Society of Chemistry*. Retrieved (http://www.chemspider.com/).

AS Christensen, FA Faber, B Huang, LA Bratholm, A Tkatchenko, KR Muller, OA von Lilienfeld. 2017. “QML: A Python Toolkit for Quantum Machine Learning.”

AS Christensen, FA Faber, B. Huang, LA Bratholm, A. Tkatchenko, KR Muller, and OA von Lilienfeld. 2017. “‘QML: A Python Toolkit for Quantum Machine Learning.’”

Bender, L. (Bru. Conti). 2013. “Architectural Paint.” Pp. 250–56 in, edited by J. A. Siegel, P. J. Saukko, and M. M. B. T.-E. of F. S. (Second E. Houck. Waltham: Academic Press.

Bernardino, N. D., V. R. L. Constantino, and D. L. A. De Faria. 2018. “Probing the Indigo Molecule in Maya Blue Simulants with Resonance Raman Spectroscopy.” *Journal of Physical Chemistry C* 122(21):11505–15. doi: 10.1021/acs.jpcc.8b01406.

Blum, Lorenz C., and Jean-Louis Reymond. 2009. “970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13.” *Journal of the American Chemical Society* 131(25):8732–33. doi: 10.1021/ja902302h.

Caliandro, Rocco, Valentina Toson, Luca Palin, Eleonora Conterosito, Maurizio Aceto, Valentina Gianotti, Enrico Boccaleri, Eric Dooryhee, and Marco Milanesio. 2019. “New Hints on the Maya Blue Formation Process by PCA-Assisted In Situ XRPD/PDF and Optical Spectroscopy.” *Chemistry – A European Journal* 25(49):11503–11. doi: https://doi.org/10.1002/chem.201901932.

Capecchi, Alice, Daniel Probst, and Jean-Louis Reymond. 2020. “One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome.” *Journal*

of Cheminformatics 12(1):43. doi: 10.1186/s13321-020-00445-4.

Cavalla, David. 2015. “Chapter 35 - Web Alert: Using the Internet for Medicinal Chemistry.” Pp. 825–42 in, edited by C. G. Wermuth, D. Aldous, P. Raboisson, and D. B. T.-T. P. of M. C. (Fourth E. Rognan. San Diego: Academic Press.

Cereto-Massagué, Adrià, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. 2015. “Molecular Fingerprint Similarity Search in Virtual Screening.” *Methods* 71:58–63. doi: https://doi.org/10.1016/j.ymeth.2014.08.005.

Chen, Huiwen, Zepeng Zhang, Guanzheng Zhuang, and Rui Jiang. 2019. “A New Method to Prepare ‘Maya Red’ Pigment from Sepiolite and Basic Red 46.” *Applied Clay Science* 174(March):38–46. doi: 10.1016/j.clay.2019.03.023.

Collins, Christopher R., Geoffrey J. Gordon, O. Anatole von Lilienfeld, and David J. Yaron. 2018. “Constant Size Descriptors for Accurate Machine Learning Models of Molecular Properties.” *The Journal of Chemical Physics* 148(24):241718. doi: 10.1063/1.5020441.

Dejoie, Catherine, Pauline Martinetto, Eric Dooryhée, Elsavan Elslande, Sylvie Blanc, Patrice Bordat, Ross Brown, Florence Porcher, and Michel Anne. 2010. “Association of Indigo with Zeolites for Improved Color Stabilization.” *Applied Spectroscopy* 64(10):1131–38. doi: 10.1366/000370210792973622.

Doménech-Carbó, Antonio, María Teresa Doménech-Carbó, Francisco Manuel Valle-Algarra, Marcelo E. Domine, and Laura Osete-Cortina. 2013. “On the Dehydroindigo Contribution to Maya Blue.” *Journal of Materials Science* 48(20):7171–83. doi: 10.1007/s10853-013-7534-z.

Doménech, Antonio, María Teresa Doménech-Carbó, Laura Osete-Cortina, and Noemí Montoya. 2013. “Application of Solid-State Electrochemistry Techniques to Polyfunctional Organic-Inorganic Hybrid Materials: The Maya Blue Problem.” *Microporous and Mesoporous Materials* 166:123–30. doi: 10.1016/j.micromeso.2012.04.031.

Dong, Jie, and Junping Zhang. 2019. “13 - Maya Blue Pigments Derived From Clay Minerals.” Pp. 627–61 in *Micro and Nano Technologies*, edited by A. Wang and W. B. T.-N. from C. M. Wang. Elsevier.

Faber, Felix A., and O. Anatole von Lilienfeld. 2019. “Modeling Materials Quantum Properties with Machine Learning.” *Materials Informatics* 171–79.

Gao, Peng, Jie Zhang, Yuzhu Sun, and Jianguo Yu. 2020. “Toward Accurate Predictions of Atomic Properties via Quantum Mechanics Descriptors Augmented Graph Convolutional Neural Network: Application of This Novel Approach in NMR Chemical Shifts Predictions.” *The Journal of Physical Chemistry Letters* 11(22):9812–18. doi: 10.1021/acs.jpclett.0c02654.

Giustetto, R., O. Wahyudi, I. Corazzari, and F. Turci. 2011. “Chemical Stability and Dehydration Behavior of a Sepiolite/Indigo Maya Blue Pigment.” *Applied Clay Science* 52(1–2):41–50. doi: 10.1016/j.clay.2011.01.027.

Hamilton, William L. 2020. “Graph Representation Learning.” *Synthesis Lectures on Artificial Intelligence and Machine Learning* 14(9781681739632):1–159. doi: 10.2200/S01045ED1V01Y202009AIM046.

He, Jian-Bo, Guang-He Ma, Jun-Cong Chen, Ye Yao, and Yan Wang. 2010. “Voltammetry and Spectroelectrochemistry of Solid Indigo Dispersed in Carbon Paste.” *Electrochimica Acta* 55(17):4845–50. doi: https://doi.org/10.1016/j.electacta.2010.03.064.

Jeon, Woosung, and Dongsup Kim. 2019. “FP2VEC: A New Molecular Featurizer for Learning Molecular Properties.” *Bioinformatics* 35(23):4979–85. doi: 10.1093/bioinformatics/btz307.

Jiang, Tammy, Jaimie L. Gradus, and Anthony J. Rosellini. 2020. “Supervised Machine Learning: A Brief Primer.” *Behavior Therapy* 51(5):675–87. doi: https://doi.org/10.1016/j.beth.2020.05.002.

Ju, Zixin, Jie Sun, and Yanping Liu. 2019. “Molecular Structures and Spectral Properties of Natural Indigo and Indirubin: Experimental and DFT Studies.”

Molecules (Basel, Switzerland) 24(21):3831. doi: 10.3390/molecules24213831.

Kim, Sunghwan, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. 2016. "PubChem Substance and Compound Databases." *Nucleic Acids Research* 44(D1):D1202–13. doi: 10.1093/nar/gkv951.

Kuenemann, Melaine A., Malgorzata Szymczyk, Yufei Chen, Nadia Sultana, David Hinks, Harold S. Freeman, Antony J. Williams, Denis Fourches, and Nelson R. Vinueza. 2017. "Weaver's Historic Accessible Collection of Synthetic Dyes: A Cheminformatics Analysis." *Chemical Science* 8(6):4334–39. doi: 10.1039/C7SC00567A.

Landrum, Greg. 2019. *RDKit Documentation. Release 2019.09.1*.

Li, Qingliang, Tiejun Cheng, Yanli Wang, and Stephen H. Bryant. 2010. "PubChem as a Public Resource for Drug Discovery." *Drug Discovery Today* 15(23):1052–57. doi: https://doi.org/10.1016/j.drudis.2010.10.003.

Li, Yi, and Jihong Yu. 2014. "New Stories of Zeolite Structures: Their Descriptions, Determinations, Predictions, and Evaluations." *Chemical Reviews* 114(14):7268–7316. doi: 10.1021/cr500010r.

Li, Yichuan, Guofu Zhu, Yu Wang, Yongming Chai, and Chenguang Liu. 2020. "Preparation, Mechanism and Applications of Oriented MFI Zeolite Membranes: A Review." *Microporous and Mesoporous Materials* 110790. doi: https://doi.org/10.1016/j.micromeso.2020.110790.

Muegge, Ingo, and Prasenjit Mukherjee. 2016. "An Overview of Molecular Fingerprint Similarity Search in Virtual Screening." *Expert Opinion on Drug Discovery* 11(2):137–48. doi: 10.1517/17460441.2016.1117070.

Müller, A. C., and S. Guido. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media.

Nápoles, José Manuel and Quintana, Lissete. 2021a. "51 Moléculas." Retrieved (https://figshare.com/articles/dataset/54_Mol_culas/17192972).

Nápoles, José Manuel, and Quintana, Lissete. 2021b. “Compuestos Candidatos.” Retrieved (https://figshare.com/articles/dataset/Colorantes_obtenidos/17192966).

Nápoles, José Manuel, and Quintana, Lissete. 2021c. “Conjunto de Compuestos Colorantes.” Retrieved (https://figshare.com/articles/dataset/Conjunto_de_compuestos_colorantes/17257148).

Nápoles, José Manuel, and Quintana, Lissete. 2021d. “Conjunto de Datos Inicial (No Colorantes).” Retrieved (https://figshare.com/articles/dataset/Conjunto_de_datos_inicial_no_colorantes_/17256278).

Nápoles, José Manuel, and Quintana, Lissete. 2021e. “Conjuntos de Datos Iniciales de Colorantes.” Retrieved (https://figshare.com/articles/dataset/Dye_datasets/17185799).

Nápoles, José Manuel, and Quintana, Lissete. 2021f. “Coordenadas de Las 51 Moleculas.” Retrieved (https://figshare.com/articles/dataset/Coordenadas_de_las_51_moleculas/17256731).

Pascual, María Luisa Vázquez de Ágredos, María Teresa Doménech Carbó, and Antonio Doménech Carbó. 2011. “Characterization of Maya Blue Pigment in Pre-Classic and Classic Monumental Architecture of the Ancient Pre-Columbian City of Calakmul (Campeche, Mexico).” *Journal of Cultural Heritage* 12(2):140–48. doi: https://doi.org/10.1016/j.culher.2009.12.002.

Patle, A., and D. S. Chouhan. 2013. “SVM Kernel Functions for Classification.” Pp. 1–9 in *2013 International Conference on Advances in Technology and Engineering (ICATE)*.

Pattanaik, Lagnajit, and Connor W. Coley. 2020. “Molecular Representation: Going Long on Fingerprints.” *Chem* 6(6):1204–7. doi: https://doi.org/10.1016/j.chempr.2020.05.002.

Pfaff, G. 2017. *Inorganic Pigments*. De Gruyter.

Prashanth, B., Mruthyunjaya Mendu, and Ravikumar Thallapalli. 2021. “Cloud Based Machine Learning with Advanced Predictive Analytics Using Google Colaboratory.” *Materials Today: Proceedings*. doi: https://doi.org/10.1016/j.matpr.2021.01.800.

R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C., and J. Chem. Inf. Model. Steinbeck, J. Wegner, E. L. Willighagen. 2006. “OpenBabel.” 46, 991.

Ramakrishnan, Raghunathan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. 2014. “Quantum Chemistry Structures and Properties of 134 Kilo Molecules.” *Scientific Data* 1(1):140022. doi: 10.1038/sdata.2014.22.

Rupp, Matthias. 2015. “Machine Learning for Quantum Mechanics in a Nutshell.” *International Journal of Quantum Chemistry* 115(16):1058–73. doi: https://doi.org/10.1002/qua.24954.

Rupp, Matthias, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. 2012. “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning.” *Physical Review Letters* 108(5):58301. doi: 10.1103/PhysRevLett.108.058301.

Russo, Daniel P., Xiliang Yan, Sunil Shende, Heng Huang, Bing Yan, and Hao Zhu. 2020. “Virtual Molecular Projections and Convolutional Neural Networks for the End-to-End Modeling of Nanoparticle Activities and Properties.” *Analytical Chemistry* 92(20):13971–79. doi: 10.1021/acs.analchem.0c02878.

Sánchez-Ochoa, F., Gregorio H. Cocoletzi, and G. Canto. 2017. “Trapping and Diffusion of Organic Dyes inside of Palygorskite Clay: The Ancient Maya Blue Pigment.” *Microporous and Mesoporous Materials* 249:111–17. doi: 10.1016/j.micromeso.2017.04.060.

Schütt, K. T., P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K. R. Müller. 2019. “SchNetPack: A Deep Learning Toolbox For Atomistic Systems.” *Journal of Chemical Theory and Computation* 15(1):448–55. doi: 10.1021/acs.jctc.8b00908.

Schütt, K. T., P. J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K. R. Müller. 2017. “SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions.” Pp. 992–1002 in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, *NIPS’17*. Red Hook, NY, USA: Curran Associates Inc.

Schütt, K. T., H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K. R. Müller. 2018. “SchNet – A Deep Learning Architecture for Molecules and Materials.” *The Journal of Chemical Physics* 148(24):241722. doi: 10.1063/1.5019779.

Sharafi, Hassan, Isa Ebtehaj, Hossein Bonakdari, and Amir Hossein Zaji. 2016. “Design of a Support Vector Machine with Different Kernel Functions to Predict Scour Depth around Bridge Piers.” *Natural Hazards* 84(3):2145–62. doi: 10.1007/s11069-016-2540-5.

Sharma, Neha, Reecha Sharma, and Neeru Jindal. 2021. “Machine Learning and Deep Learning Applications-A Vision.” *Global Transitions Proceedings* 2(1):24–28. doi: https://doi.org/10.1016/j.gltp.2021.01.004.

Silakari, Om, and Pankaj Kumar Singh. 2021. “Chapter 3 - Small Molecule Databases: A Collection of Promising Bioactive Molecules.” Pp. 65–88 in, edited by O. Silakari and P. K. B. T.-C. and E. P. of M. and I. in D. D. Singh. Academic Press.

Stratmann, Heidi, Maria Hellmund, Ulrich Veith, Nicole End, and Wera Teubner. 2020. “Indicators for Lack of Systemic Availability of Organic Pigments.” *Regulatory Toxicology and Pharmacology* 115:104719. doi: https://doi.org/10.1016/j.yrtph.2020.104719.

Stulp, Freek, and Olivier Sigaud. 2015. “Many Regression Algorithms, One Unified Model: A Review.” *Neural Networks* 69:60–79. doi: https://doi.org/10.1016/j.neunet.2015.05.005.

Swain, Matt. 2014. “PubChemPy Documentation.” Retrieved (https://pubchempy.readthedocs.io/en/latest/).

Theodoridis, Sergios. 2020. “Chapter 11 - Learning in Reproducing Kernel Hilbert

Spaces." Pp. 531–94 in, edited by S. B. T.-M. L. (Second E. Theodoridis. Academic Press.

Torng, Wen, and Russ B. Altman. 2019. "Graph Convolutional Neural Networks for Predicting Drug-Target Interactions." *Journal of Chemical Information and Modeling* 59(10):4131–49. doi: 10.1021/acs.jcim.9b00628.

Volkov, V. V, R. Chelli, R. Righini, and C. C. Perry. 2020. "Indigo Chromophores and Pigments: Structure and Dynamics." *Dyes and Pigments* 172:107761. doi: https://doi.org/10.1016/j.dyepig.2019.107761.

Williams, Antony. 2017. "150 Analog Max Weaver Dye Library Subset." Retrieved (https://figshare.com/articles/dataset/150_Analog_Max_Weaver_Dye_Library_Subset/4590250).

Williams, Tova N., Melaine A. Kuenemann, George A. Van Den Driessche, Antony Williams, Denis Fourches, and Harold S. Freeman. 2017. "Hair Dye Substance Database (HDSD)." Retrieved (https://figshare.com/articles/dataset/Hair_Dye_Substance_Database_HDSD_/5505856).

Williams, Tova N., Melaine A. Kuenemann, George A. Van Den Driessche, Antony J. Williams, Denis Fourches, and Harold S. Freeman. 2018. "Toward the Rational Design of Sustainable Hair Dyes Using Cheminformatics Approaches: Step 1. Database Development and Analysis." *ACS Sustainable Chemistry & Engineering* 6(2):2344–52. doi: 10.1021/acssuschemeng.7b03795.

Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2017. "Chapter 7 - Extending Instance-Based and Linear Models." Pp. 243–84 in, edited by I. H. Witten, E. Frank, M. A. Hall, and C. J. B. T.-D. M. (Fourth E. Pal. Morgan Kaufmann.

Woodtli, Pascal, Sandro Giger, Patrick Müller, Lucie Sägesser, Nicola Zucchetto, Michael J. Reber, Achim Ecker, and Dominik Brühwiler. 2018. "Indigo in the Nanochannels of Zeolite L: Towards a New Type of Colorant." *Dyes and Pigments*

149:456–61. doi: https://doi.org/10.1016/j.dyepig.2017.10.029.

Yamamoto, Yui, Takahiro Okazaki, Yasuhiro Sakai, Shun Iwasaki, and Nobuyoshi Koga. 2020. “Kinetic Analysis of the Multistep Thermal Decomposition of Maya Blue-Type Pigments to Evaluate Thermal Stability.” *Journal of Thermal Analysis and Calorimetry* 142(2):1073–85. doi: 10.1007/s10973-020-09278-7.

Yang, Chihae, Susan M. Barlow, Kristi L. Muldoon Jacobs, Vessela Vitcheva, Alan R. Boobis, Susan P. Felter, Kirk B. Arvidson, Detlef Keller, Mark T. D. Cronin, Steven Enoch, Andrew Worth, and Heli M. Hollnagel. 2017. “Thresholds of Toxicological Concern for Cosmetics-Related Substances: New Database, Thresholds, and Enrichment of Chemical Space.” *Food and Chemical Toxicology* 109:170–93. doi: https://doi.org/10.1016/j.fct.2017.08.043.

Yuan, Qing, Zhiqiang Wei, Xu Guan, Mingjian Jiang, Shuang Wang, Shugang Zhang, and Zhen Li. 2019. “Toxicity Prediction Method Based on Multi-Channel Convolutional Neural Network.” *Molecules* 24(18).

Zhang, Yujie, Ling Fan, Hao Chen, Junping Zhang, Yuan Zhang, and Aiqin Wang. 2015. “Learning from Ancient Maya: Preparation of Stable Palygorskite/Methylene Blue@SiO2 Maya Blue-like Pigment.” *Microporous and Mesoporous Materials* 211:124–33. doi: 10.1016/j.micromeso.2015.03.002.

Zhou, Wei, Hong Liu, Tingting Xu, Yeling Jin, Shijie Ding, and Jing Chen. 2014. “Insertion of Isatin Molecules into the Nanostructure of Palygorskite.” *RSC Advances* 4(94):51978–83. doi: 10.1039/C4RA06299J.

XI. APÉNDICES

Apéndice I. Conjunto de moléculas candidatas (171 compuestos).

Fórmula molecular	Peso molecular [g/mol]	Número de átomos	Número átomos C,N,O,S
$C_{12}H_{12}N_2OS_2$	264.3665	29	17
$C_{17}H_{12}N_2OS_2$	324.4200	34	22
$C_{18}H_{12}N_4O_4S$	380.3773	39	27
$C_{21}H_{18}N_2O_3$	346.3792	44	26
$C_{20}H_{14}N_2O_5$	362.3356	41	27
$C_{22}H_{18}N_6O_3S$	446.4817	50	32
$C_{25}H_{27}N_{11}O_4S_2$	609.6832	69	42
$C_{13}H_{13}N_3O_4S$	307.3250	34	21
$C_{29}H_{26}N_6O_6S$	586.6183	68	42
$C_{12}H_9N_3O_3$	243.2182	27	18
$C_{28}H_{17}NO_2S$	431.5051	49	32
$C_{31}H_{28}N_6O$	500.5936	66	38
$C_{23}H_{15}NO_3S$	385.4351	43	28
$C_9H_7N_5O_3S_2$	297.3136	26	19
$C_{14}H_{14}N_6O_3$	314.2994	37	23
$C_{17}H_{13}N_3O_3S_2$	371.4334	38	25
$C_{17}H_{11}N_3O_2S$	321.3531	34	23
$C_{22}H_{14}N_6O_3S$	442.4500	46	32
$C_{23}H_{19}N_3O_4$	401.4147	49	30
$C_{19}H_{19}N_5O_2S$	381.4515	46	27
$C_{13}H_9N_3O_2S$	271.2945	28	19
$C_{29}H_{24}N_6O_4S$	552.6037	64	40
$C_{18}H_{22}N_4O$	310.3935	45	23
$C_{17}H_{14}N_2O_3S$	326.3697	37	23
$C_{28}H_{22}N_2O_8S$	546.5479	61	39
$C_{24}H_{27}N_9O_3$	489.5297	63	36
$C_{14}H_{14}N_6O_3S_2$	378.4294	39	25
$C_{14}H_9ClN_4O_3$	316.6993	31	21
$C_{14}H_{14}N_4O_4$	302.2854	36	22
$C_{20}H_{16}F_6N_4O_5S$	538.4203	52	30
$C_{34}H_{22}N_4O_4$	550.5629	64	42
$C_{33}H_{32}N_4O_3S_4$	660.8922	76	44
$C_{28}H_{28}N_4O_5S_2$	564.6757	67	39
$C_{21}H_{20}Cl_2N_6O_2S$	491.3935	52	30

$C_{29}H_{29}N_7O_6$	571.5839	71	42
$C_{33}H_{37}N_7O_4$	595.6914	81	44
$C_{35}H_{19}Cl_4N_5O_4$	715.3685	67	44
$C_{20}H_{19}N_3O_2$	333.3838	44	25
$C_{20}H_{18}N_2O_2S$	350.4341	43	25
$C_{12}H_{11}NO_4$	233.2200	28	17
$C_{16}H_{14}N_2O_6$	330.2921	38	24
$C_{33}H_{39}N_3O_5S$	589.7449	81	42
$C_{26}H_{26}N_2OS_2$	446.6274	57	31
$C_{25}H_{26}N_2O_2S$	418.5511	56	30
$C_{25}H_{26}N_2OS$	402.5517	55	29
$C_{50}H_{42}N_4O$	714.8947	97	55
$C_{33}H_{34}N_4O_3$	534.6481	74	40
$C_{49}H_{40}N_4O$	700.8681	94	54
$C_{23}H_{26}N_2O$	346.4653	52	26
$C_{47}H_{38}N_4O$	674.8308	90	52
$C_{35}H_{32}N_4O_5$	588.6524	76	44
$C_{28}H_{30}N_2OS$	442.6156	62	32
$C_{25}H_{24}N_2OS_2$	432.6009	54	30
$C_{23}H_{24}N_2OS$	376.5145	51	27
$C_{23}H_{26}N_2O_2S$	394.5297	54	28
$C_{32}H_{30}N_2O_4S_2$	570.7216	70	40
$C_{32}H_{36}N_2O$	464.6410	71	35
$C_{31}H_{34}N_2O$	450.6145	68	34
$C_{23}H_{23}ClN_2OS_3$	475.0895	53	29
$C_{22}H_{24}N_2O_3S$	396.5026	52	28
$C_{21}H_{21}ClN_2O_2S$	400.9216	48	26
$C_{24}H_{27}N_3O_2S_2$	453.6201	58	31
$C_{21}H_{22}N_2O_2S$	366.4766	48	26
$C_{26}H_{25}N_3OS$	427.5612	56	31
$C_{28}H_{27}N_3O$	421.5335	59	32
$C_{21}H_{22}N_2OS_3$	414.6072	49	27
$C_{22}H_{23}N_3O_2S_3$	457.6319	53	30
$C_{27}H_{26}N_4O_2S$	470.5859	60	34
$C_{22}H_{23}N_3O_3S$	409.5013	52	29
$C_{41}H_{44}N_4O_4S_2$	720.9425	95	51
$C_{31}H_{34}N_2O_4S_2$	562.7427	73	39
$C_{28}H_{31}N_3O_2S_2$	505.6946	66	35

$C_{23}H_{22}N_2OS_2$	406.5636	50	28
$C_{27}H_{30}N_2O$	398.5399	60	30
$C_{26}H_{24}N_2O$	380.4816	53	29
$C_{39}H_{33}N_3O_3S_3$	687.8926	81	48
$C_{44}H_{43}N_3O_3$	661.8305	93	50
$C_{39}H_{38}N_2O$	550.7318	80	42
$C_{37}H_{32}N_2O_2S_2$	600.7922	75	43
$C_{41}H_{40}N_2O_2$	592.7685	85	45
$C_{40}H_{38}N_2O_2$	578.7419	82	44
$C_{27}H_{26}N_2OS_2$	458.6381	58	32
$C_{29}H_{32}N_2O$	424.5772	64	32
$C_{30}H_{29}N_3OS$	479.6358	64	35
$C_{25}H_{26}N_2OS_3$	466.6817	57	31
$C_{18}H_{15}NOS$	293.3828	36	21
$C_{26}H_{26}N_2OS_2$	446.6274	57	31
$C_{21}H_{21}NO$	303.3975	44	23
$C_{19}H_{19}NO_5$	341.3579	44	25
$C_{23}H_{18}N_2O$	338.4018	44	26
$C_{18}H_{16}N_2O_8$	388.3282	44	28
$C_{34}H_{34}N_4O_4$	562.6582	76	42
$C_{13}H_{12}N_2O_3$	244.2460	30	18
$C_{15}H_{13}NO_4$	271.2680	33	20
$C_{49}H_{57}N_9O_{12}$	964.0300	127	70
$C_{18}H_{11}NO_2$	273.2854	32	21
$C_{28}H_{30}N_2O_3$	442.5494	63	33
$C_{30}H_{26}N_2O_{13}$	622.5330	71	45
$C_{16}H_{10}N_2O_2$	262.2628	30	20
$C_{12}H_{12}N_2OS_2$	264.3665	29	17
$C_{16}H_{10}N_2O_8S_2$	422.3892	38	28
$C_{12}H_{11}NO_7P^+$	312.1900	32	20
$C_{12}H_{11}NO_7P^+$	312.1900	32	20
$C_{12}H_{11}NO_7P^+$	312.1900	32	20
$C_{42}H_{35}N_3O_4S_3$	741.9000	87	52
$C_{42}H_{35}N_3O_4S_3$	741.9000	87	52
$C_{35}H_{35}N_3O_8S$	657.7000	82	47
$C_{45}H_{52}N_4O_{14}S_2$	937.0000	117	65
$C_{23}H_{19}NO_4$	373.4000	47	28
$C_{22}H_{20}N_4O$	356.4000	47	27

$C_{29}H_{32}N_2O_{12}S_2$	664.7000	77	45
$C_{32}H_{40}N_2O_5S$	564.7000	80	40
$C_{23}H_{26}N_2O_4S_3$	490.7000	58	32
$C_{40}H_{48}N_4O_6S$	712.9000	99	51
$C_{48}H_{47}N_3O_4S_3$	826.1000	105	58
$C_{20}H_{22}N_4O_3$	366.4000	49	27
$C_{41}H_{40}N_2O_7S$	704.8000	91	51
$C_{43}H_{50}N_4O_{14}S_2$	911.0000	113	63
$C_{31}H_{34}N_2O_8S_2$	626.7000	77	43
$C_{37}H_{42}N_3O_{10}S^{2-}$	752.9000	94	52
$C_{35}H_{40}N_3O_{10}S^{2-}$	726.8000	90	50
$C_{58}H_{70}N_2O_2S_4$	955.5000	136	66
$C_{43}H_{54}N_4O_{12}S_2$	883.0000	115	61
$C_{25}H_{24}N_4O$	396.5000	54	30
$C_{18}H_{25}BF_2N_4O$	362.2000	51	23
$C_{26}H_{21}F_5N_4O_4S$	580.5000	61	35
$C_{16}H_{24}N_6O_4$	364.4000	50	26
$C_{43}H_{49}N_4O_{14}S^{2-}$	910.0000	112	63
$C_{48}H_{47}N_3O_4S_3$	826.1000	105	58
$C_{27}H_{21}N_3O$	403.5000	52	31
$C_{38}H_{46}N_4O_6S$	686.9000	95	49
$C_{17}H_{21}N_4O_2S^{+}$	345.4000	45	24
$C_{20}H_{24}BF_2N_5O$	399.2000	53	26
$C_{34}H_{42}N_2O_5S$	590.8000	84	42
$C_{42}H_{35}N_3O_4S_3$	741.9000	87	52
$C_{16}H_{10}N_2O_2$	262.2600	30	20
$C_{37}H_{30}N_2O_3S_2$	614.8000	74	44
$C_{31}H_{28}N_2O_3$	476.6000	64	36
$C_{20}H_{22}N_4O$	334.4000	47	25
$C_{35}H_{28}N_2O_2$	508.6000	67	39
$C_{26}H_{29}N_3O_2$	415.5000	60	31
$C_{38}H_{45}N_3O_7S$	687.8000	94	49
$C_{48}H_{30}N_4O_8$	790.8000	90	60
$C_{46}H_{40}N_4O_2S_2$	745.0000	94	54
$C_{19}H_{18}N_4O_4$	366.4000	45	27
$C_{42}H_{35}N_3O_4S_3$	741.9000	87	52
$C_{21}H_{22}B_2N_2O_7$	436.0000	54	30
$C_{42}H_{35}N_3O_4S_3$	741.9000	87	52

$C_{54}H_{58}N_2O_6S$	863.1000	121	63
$C_{52}H_{53}N_3O_6S_3$	912.2000	117	64
$C_{38}H_{26}N_4O_2$	570.6000	70	44
$C_{31}H_{28}N_2O_4S_2$	556.7000	67	39
$C_{32}H_{31}N_3O_{12}$	649.6000	78	47
$C_{48}H_{47}N_3O_4S_3$	826.1000	105	58
$C_{21}H_{21}N_3O_2$	347.4000	47	26
$C_{16}H_9NO_2S$	279.3000	29	20
$C_{21}H_{17}NOS$	331.4000	41	24
$C_{10}H_{10}N_2O$	174.0793	23	13
$C_8H_5NO_2$	147.0320	16	11
$C_{25}H_{15}N_3O_2$	389.1164	45	30
$C_{20}H_{19}N_3O_5S$	413.1045	48	29
$C_{18}H_{18}N_4O$	306.1480	41	23
$C_{26}H_{12}N_4O_2$	412.0960	44	32
$C_{16}H_{10}N_2O_2$	262.0742	30	20
$C_{20}H_{10}N_2O_9$	422.0386	41	31
$C_{16}H_{14}N_4O$	278.1167	35	21
$C_5H_5N_5O$	151.0494	16	11
$C_{24}H_{26}N_2O_{13}$	550.1434	65	39
$C_{16}H_{12}N_4O_9S_2$	468.0045	43	31
$C_{20}H_{20}N_3O^{2+}$	334.1550	45	25
$C_{19}H_{22}N_5O^{+}$	336.1818	47	25

Apéndice II. Modelo de Red Neuronal SchNet.

A continuación, se explica acerca de la arquitectura de los modelos de Redes Neuronales SchNet y su funcionamiento (Schütt et al. 2017; Schütt et al. 2018).

Una molécula puede ser descrita por un conjunto de n sitios atómicos, con cargas nucleares $Z = (Z_1, \ldots Z_n)$ y posiciones $R = (r_1, \ldots, r_n)$. En el algoritmo de SchNet, los átomos se describen por un arreglo de atributos $X^l = (x^l{}_1, \ldots, x^l{}_n)$, con $x^l{}_i \in R^F$; siendo F el número de atributos, n el número de átomos y l el número de la capa.

La representación del sitio i se inicializa usando una representación embebida, dependiente de la especie atómica (ecuación 13):

$$x^0{}_i = aZ_i \tag{13}$$

Los vectores aZ_i son inicializados de forma aleatoria y posteriormente son optimizados durante el entrenamiento. En esta etapa, estos vectores representan los átomos del sistema, sin involucrar información sobre el entorno; pero posteriormente dicha información se va incorporando.

Capas “Atom-wise”

Estas capas son en realidad capas densas, que se aplican de forma separada a las representaciones $x^l{}_i$ del i-ésimo átomo (ecuación 14):

$$x^{l+1}{}_i = W^l x^1{}_i + b^l \tag{14}$$

Dado que los pesos W^l y sesgos b^l son compartidos entre los átomos, es independiente del tamaño del sistema.

Bloques de interacción

Los bloques de interacción de SchNet contribuyen al refinamiento de la representación atómica, mediante la incorporación de “interacciones a pares” de funciones de base radial, con los átomos del entorno cercano (alrededor de 5 Angstronms). Para una representación “atom-wise”, X^l en las posiciones R, se obtiene la interacción del átomo i con el entorno, mediante una capa convolucional de filtro continuo (cfconv) como se muestra a continuación (ecuación 15):

$$x^{l+1}{}_i = \sum_{j=0}^{n_{átomos}} x^l{}_j \circ W^l(r_j - r_i) \quad (15)$$

donde ∘ representa un producto Hadamard.

La figura 24 muestra la arquitectura completa de SchNet, donde se pueden observar las diferentes capas y bloques de interacción:

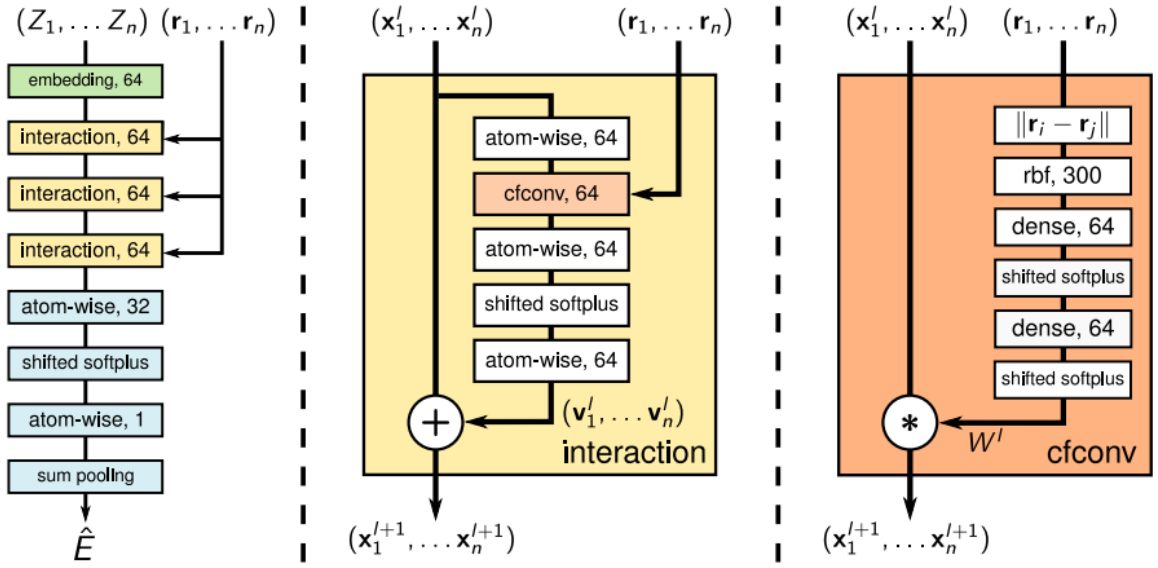


Figura 24. Arquitectura SchNet (izquierda), bloque de interacción (medio) y filtro continuo de convolución (derecha). Tomado de (Schütt et al. 2017).

Apéndice III. Código correspondiente a la selección de compuestos colorantes.

Cargar los conjuntos de datos y añadir columna de estructura molecular.

- Biblioteca de tintes Max Weaver.

```
mweaver_sdf = os.path.join(RDConfig.RDDataDir,'/content/150_max weaver
   library.sdf')
frame_mweaver_sdf = PandasTools.LoadSDF(mweaver_sdf, molColName='Molecule'
   , removeHs=False)
print(frame_mweaver_sdf.shape)   #(Contiene 150 compuestos)
```

```
(150, 13)
```

- Base de datos ChemSpider.

```
lambdamax_sdf = os.path.join(RDConfig.RDDataDir
   ,'/content/198_Lambda_max 400-780.sdf')
frame_lambdamax_sdf = PandasTools.LoadSDF(lambdamax_sdf
  , molColName='Molecule', removeHs=False)

dye_sdf = os.path.join(RDConfig.RDDataDir,'/content/232_Dye_Data.sdf')
frame_dye_sdf = PandasTools.LoadSDF(dye_sdf, molColName='Molecule'
   , removeHs=False)

#Concatenar los dos dataframes
frame_chemspider_sdf=pd.concat([frame_lambdamax_sdf, frame_dye_sdf])
print(frame_chemspider_sdf.shape)   #(Contiene 430 compuestos)
```

```
(430, 13)
```

- Base de datos PubChem.

```
frame_colour_csv=pd.read_csv
  ('/content/PubChem_compound_text_colour_summary.csv')
```

```
frame_color_csv=pd.read_csv
  ('/content/PubChem_compound_text_color_summary.csv')
frame_dyecomp_csv=pd.read_csv
  ('/content/PubChem_compound_text_dye_summary.csv')

#Concatenar los dos dataframes
frame_pubchem_csv=pd.concat([frame_colour_csv, frame_color_csv
  , frame_dyecomp_csv])

#Agregar columna de moléculas
smiles_list_pubchem=frame_pubchem_csv['isosmiles']
mol_list_pubchem=[]
for smiles in smiles_list_pubchem:
  RDLogger.DisableLog('rdApp.*')
  mol=Chem.MolFromSmiles(smiles)
  mol_list_pubchem.append(mol)
frame_pubchem_csv['Molecule']=mol_list_pubchem
print(frame_pubchem_csv.shape) #(Contiene 660 compuestos)
```

```
(660, 25)
```

- Base de datos COSMOS.

```
cosmosinv_sdf = os.path.join(RDConfig.RDDataDir
    ,'/content/COSMOS_CosmeticsInventoryV220180822.sdf')
frame_cosmosinv_sdf = PandasTools.LoadSDF(cosmosinv_sdf
    , molColName='Molecule', removeHs=False)

mol_list_cosmosinv=frame_cosmosinv_sdf['Molecule']
MW_list=[]
MF_list=[]
for mol in mol_list_cosmosinv:
  MW=Descriptors.ExactMolWt(mol)
  MF=CalcMolFormula(mol)
  MW_list.append(MW)
  MF_list.append(MF)
frame_cosmosinv_sdf['MW']=MW_list
frame_cosmosinv_sdf['MF']=MF_list


#Seleccionar colorantes del dataframe
frame_cosmosinv_sdf1=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
    ['COSING Chemical Function'] == 'HAIR DYEING')]
```

```
frame_cosmosinv_sdf2=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
    ['COSING Chemical Function'] == 'COSMETIC COLORANT | HAIR DYEING')]
frame_cosmosinv_sdf3=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
    ['COSING Chemical Function'] == 'COSMETIC COLORANT')]
frame_cosmosinv_sdf4=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
    ['COSING Chemical Function'] == 'HAIR DYEING | MASKING')]
frame_cosmosinv_sdf5=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
    ['COSING Chemical Function'] == 'COSMETIC COLORANT | HAIR DYEING
    | SKIN CONDITIONING')]
frame_cosmosinv_sdf6=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
    ['COSING Chemical Function'] == 'DENATURANT | HAIR DYEING | MASKING')]
frame_cosmosinv_sdf7=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
    ['COSING Chemical Function'] == 'COSMETIC COLORANT | OPACIFYING
    | UV ABSORBER | UV FILTER')]
frame_cosmosinv_sdf8=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
    ['COSING Chemical Function'] == 'ANTIMICROBIAL | COSMETIC COLORANT')]
frame_cosmosinv_sdf9=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'ABRASIVE | ABSORBENT | ANTICAKING
   | BULKING | COSMETIC COLORANT | EMULSION STABILISING | OPACIFYING
   | VISCOSITY CONTROLLING')]
frame_cosmosinv_sdf10=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'ANTIMICROBIAL | COSMETIC COLORANT
   | SKIN CONDITIONING')]
frame_cosmosinv_sdf11=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'COSMETIC COLORANT | DEODORANT')]
frame_cosmosinv_sdf12=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'ABSORBENT | BINDING | BULKING
   | COSMETIC COLORANT | OPACIFYING')]
frame_cosmosinv_sdf13=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'COSMETIC COLORANT | OPACIFYING')]
frame_cosmosinv_sdf14=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'ABRASIVE | ABSORBENT | ANTICAKING
   | BULKING | COSMETIC COLORANT | OPACIFYING')]
frame_cosmosinv_sdf15=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'ANTICAKING | COSMETIC COLORANT
   | VISCOSITY CONTROLLING')]
frame_cosmosinv_sdf16=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'COSMETIC COLORANT | SKIN CONDITIONING')]
frame_cosmosinv_sdf17=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'ABRASIVE | BUFFERING | BULKING
   | COSMETIC COLORANT | OPACIFYING | ORAL CARE')]
frame_cosmosinv_sdf18=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'BULKING | COSMETIC COLORANT
   | SKIN PROTECTING | UV ABSORBER')]
```

```
frame_cosmosinv_sdf19=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'ABRASIVE | BULKING | COSMETIC COLORANT
   | OPACIFYING | PEARLESCENT')]
frame_cosmosinv_sdf20=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'ANTICAKING | COSMETIC COLORANT
   | EMULSION STABILISING | VISCOSITY CONTROLLING')]
frame_cosmosinv_sdf21=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'BINDING | HAIR CONDITIONING
   | HAIR DYEING | SKIN CONDITIONING | SKIN PROTECTING')]
frame_cosmosinv_sdf22=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'HAIR CONDITIONING | HAIR DYEING
   | MASKING | SKIN CONDITIONING')]
frame_cosmosinv_sdf23=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'COSMETIC COLORANT | HAIR DYEING
   | SKIN CONDITIONING')]
frame_cosmosinv_sdf24=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'ANTIMICROBIAL | BULKING | HAIR CONDITION
   | HAIR DYEING | MASKING | SKIN CONDITIONING')]
frame_cosmosinv_sdf25=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'BINDING | FILM FORMING | HAIR DYEING
   | HAIR FIXING | VISCOSITY CONTROLLING')]
frame_cosmosinv_sdf26=frame_cosmosinv_sdf[(frame_cosmosinv_sdf
   ['COSING Chemical Function'] == 'BINDING | COSMETIC COLORANT
   | HAIR CONDITIONING | HAIR DYEING | SKIN CONDITIONING | SKIN PROTECTING')

frame_cosmosinv_sdftotal=pd.concat([frame_cosmosinv_sdf1
   ,frame_cosmosinv_sdf2, frame_cosmosinv_sdf3, frame_cosmosinv_sdf4
   , frame_cosmosinv_sdf5, frame_cosmosinv_sdf6
   , frame_cosmosinv_sdf7, frame_cosmosinv_sdf8
   , frame_cosmosinv_sdf9, frame_cosmosinv_sdf10
   , frame_cosmosinv_sdf11, frame_cosmosinv_sdf12
   , frame_cosmosinv_sdf13, frame_cosmosinv_sdf14
   , frame_cosmosinv_sdf15, frame_cosmosinv_sdf16
   , frame_cosmosinv_sdf17, frame_cosmosinv_sdf18
   , frame_cosmosinv_sdf19, frame_cosmosinv_sdf20
   , frame_cosmosinv_sdf21, frame_cosmosinv_sdf22
   , frame_cosmosinv_sdf23, frame_cosmosinv_sdf24
   , frame_cosmosinv_sdf25, frame_cosmosinv_sdf26])
print(frame_cosmosinv_sdftotal.shape)  #(Contiene 323 compuestos)

     (323, 14)
```

- Hair Dye Substance Database (HDSD).

```
hairdye_sdf = os.path.join(RDConfig.RDDataDir,'/content/HDSD_(SDF).sdf')
frame_hairdye_sdf = PandasTools.LoadSDF(hairdye_sdf, molColName='Molecule'
   , removeHs=False)

mol_list_hairdye=frame_hairdye_sdf['Molecule']
MF_list=[]
for mol in mol_list_hairdye:
  MF=CalcMolFormula(mol)
  MF_list.append(MF)
frame_hairdye_sdf['MF']=MF_list

#Seleccionar colorantes del dataframe
frame_hairdye_sdf=frame_hairdye_sdf
   [(frame_hairdye_sdf['Substance Type'] == 'Dye')]
print(frame_hairdye_sdf.shape)
```

```
(196, 137)
```

Reducir los conjuntos de datos.

```
frame_mweaver_sdf=frame_mweaver_sdf
   [['MF','SMILES', 'MW', 'InChI', 'Molecule']]


frame_chemspider_sdf=frame_chemspider_sdf
   [['MF','SMILES', 'MW', 'InChI', 'Molecule']]


frame_pubchem_csv=frame_pubchem_csv.rename
   (columns={'cid':'cid','cmpdname':'cmpdname','cmpdsynonym':'cmpdsynonym'
   ,'mw':'MW','mf':'MF','polararea':'polararea','complexity':'complexity'
   ,'xlogp':'xlogp','heavycnt':'heavycnt','hbonddonor':'hbonddonor'
   ,'hbondacc':'hbondacc','rotbonds':'rotbonds','inchi':'InChI'
   ,'isosmiles':'SMILES','inchikey':'inchikey','iupacname':'iupacname'
   ,'meshheadings':'meshheadings','annothits':'annothits'
   ,'annothitcnt':'annothitcnt','aids':'aids','cidcdate':'cidcdate'
   ,'sidsrcname':'sidsrcname','depcatg':'depcatg','annotation':'annotation'
   ,'Molecule':'Molecule','sid':'sid','sidextid':'sidextid'
   ,'subssynonym':'subssynonym','sidmdate':'sidmdate','depdate':'depdate'})


frame_pubchem_csv=frame_pubchem_csv[['MF','SMILES', 'MW', 'InChI'
   , 'Molecule']]
```

```
frame_cosmosinv_sdftotal=frame_cosmosinv_sdftotal[['MF','SMILES', 'MW'
   , 'InChI', 'Molecule']]


frame_hairdye_sdf=frame_hairdye_sdf[['MF', 'Canonical SMILES'
   , 'ExactMW', 'InChI Strings', 'Molecule']]
frame_hairdye_sdf=frame_hairdye_sdf.rename(columns={'MF':'MF'
   ,    'Canonical SMILES':'SMILES',    'ExactMW':'MW'
   , 'InChI Strings':'InChI', 'Molecule':'Molecule'})
```

Concatenar los datasets

```
dataset=pd.concat([frame_mweaver_sdf, frame_chemspider_sdf
   , frame_pubchem_csv, frame_cosmosinv_sdftotal
   , frame_hairdye_sdf])
print(dataset.shape)
```

```
(1759, 5)
```

Agregar columna de número de átomos y número de átomos de C, N, O, S.

```
#Añadir número de átomos incuyendo átomos de Hs
mol_list=dataset['Molecule']
num_atoms_list=[]
for mol in mol_list:
  if mol is None: continue
  H_atoms=Chem.AddHs(mol)
  num_atoms=H_atoms.GetNumAtoms()
  num_atoms_list.append(num_atoms)
dataset['Num_atoms']=num_atoms_list
print(dataset.shape)  #(1759 compuestos)
```

```
(1759, 6)
```

```
# Átomos de C, N, O, S.
```

```
mol_list=dataset['Molecule']
C_atoms_list=[]
N_atoms_list=[]
O_atoms_list=[]
S_atoms_list=[]
for mol in mol_list:
  C_atoms = rdqueries.AtomNumEqualsQueryAtom(6)
  N_atoms = rdqueries.AtomNumEqualsQueryAtom(7)
  O_atoms = rdqueries.AtomNumEqualsQueryAtom(8)
  S_atoms = rdqueries.AtomNumEqualsQueryAtom(16)
  num_C_atoms=len(mol.GetAtomsMatchingQuery(C_atoms))
  num_N_atoms=len(mol.GetAtomsMatchingQuery(N_atoms))
  num_O_atoms=len(mol.GetAtomsMatchingQuery(O_atoms))
  num_S_atoms=len(mol.GetAtomsMatchingQuery(S_atoms))
  C_atoms_list.append(num_C_atoms)
  N_atoms_list.append(num_N_atoms)
  O_atoms_list.append(num_O_atoms)
  S_atoms_list.append(num_S_atoms)
CNOS_count=[]
for i in range(len(C_atoms_list)):
  CNOS_count.append(C_atoms_list[i] + N_atoms_list[i]
                + O_atoms_list[i] + S_atoms_list[i])
dataset['CNOS_count']=CNOS_count
```

Organizar el conjunto de datos.

```
dataset=dataset.rename(columns={'MF':'Formula','SMILES':'SMILES'
   , 'Molecule':'Molecule', 'MW':'Molecular_weight'
   , 'Num_atoms':'Num_atoms', 'CNOS_count':'CNOS_count'
   , 'InChI':'InChI'})
print(dataset.shape)  #(1759 compuestos)
```

```
(1759, 7)
```

Curación de los datos.

Eliminar compuestos repetidos según el código InChI

```
dataset=dataset.drop_duplicates('InChI')
```

```
print(dataset.shape) #(Quedan 1572 compuestos)
```

```
(1572, 7)
```

Número de datos faltantes en cada columna del Dataframe

```
print(dataset.isnull().sum())
```

```
Formula             0
SMILES              0
Molecular_weight    0
InChI               0
Molecule            0
Num_atoms           0
CNOS_count          0
dtype: int64
```

Eliminar compuetos que no contienen carbono: Inorgánicos

```
mol_list=dataset['Molecule']
mol_c=[]
for mol in mol_list:
  c_atom = rdqueries.AtomNumEqualsQueryAtom(6)
  num_c_atom=len(mol.GetAtomsMatchingQuery(c_atom))
  if num_c_atom != 0:
    mol_c.append(mol)
dataset=dataset.loc[dataset['Molecule'].isin(mol_c)]
print(dataset.shape)  #(Se reduce el conjunto
                      # de 1572 a 1539 compuestos)
```

```
(1539, 7)
```

Remover isótopos

```
def MolWithIsotopes(mol):
  mol_isot=[]
  atom_data = [(atom, atom.GetIsotope()) for atom in mol.GetAtoms()]
```

```
  for atom, isotope in atom_data:
      if isotope:
          mol_isot.append(mol)
  return mol_isot #Lista de las moléculas que tienen isótopos

mol_list=dataset['Molecule']
for mol in mol_list:
  isot=MolWithIsotopes(mol)
dataset=dataset.loc[~dataset['Molecule'].isin(isot)]
print(dataset.shape)
```

```
(1539, 7)
```

Eliminar moléculas que contengan metales: Organometálicos

```
mol_list=dataset['Molecule']
mol_met=[]
for mol in mol_list:
  met_atom = rdqueries.MAtomQueryAtom()
  num_met_atom=len(mol.GetAtomsMatchingQuery(met_atom))
  if num_met_atom != 0:
    mol_met.append(mol)
dataset=dataset.loc[~dataset['Molecule'].isin(mol_met)]
print(dataset.shape) #(Se reduce el conjunto a 1153 compuestos)
```

```
(1153, 7)
```

Exportar conjunto de datos final (formato CSV)

```
dataset.to_csv('Conjunto_compuestos_colorantes.csv'
   , header=True, index=False)
```

Seleccionar moléculas con grupos característicos del índigo y dehidroíndigo.

Moléculas que contengan grupos C=O

```
mol_list=dataset['Molecule']
co_group=Chem.MolFromSmiles('C(=O)')
matches_co=[]
for mol in mol_list:
  if mol.HasSubstructMatch(co_group):
    matches_co.append(mol)
dataset=dataset.loc[dataset['Molecule'].isin(matches_co)]
dataset.shape         #(Se reduce el conjunto a 567 compuestos)
```

```
(567, 7)
```

Moléculas que contengan grupos N-H

```
mol_list=dataset['Molecule']
nh_group=Chem.MolFromSmiles('N[H]')
matches_nh=[]
for mol in mol_list:
  if mol.HasSubstructMatch(nh_group):
    matches_nh.append(mol)
dataset=dataset.loc[dataset['Molecule'].isin(matches_nh)]
dataset.shape          #(Se reduce el conjunto a 392 compuestos)
```

```
(392, 7)
```

Moléculas que contengan anillos de 5 miembros

```
mol_list=dataset['Molecule']
r5_group=Chem.MolFromSmarts('[r5]')
matches_r5=[]
for mol in mol_list:
  if mol.HasSubstructMatch(r5_group):
    matches_r5.append(mol)
dataset=dataset.loc[dataset['Molecule'].isin(matches_r5)]
dataset.shape          #(Se reduce el conjunto a 183 compuestos)
```

```
(183, 7)
```

Inspección de los compuestos obtenidos

```
dataset=dataset.reset_index(drop=True)
dataset=dataset.drop(index=[19,20,26,31,133,135,146,147,153,159,169,173])
dataset.shape  #(Se proponen 171 compuestos candidatos)
```

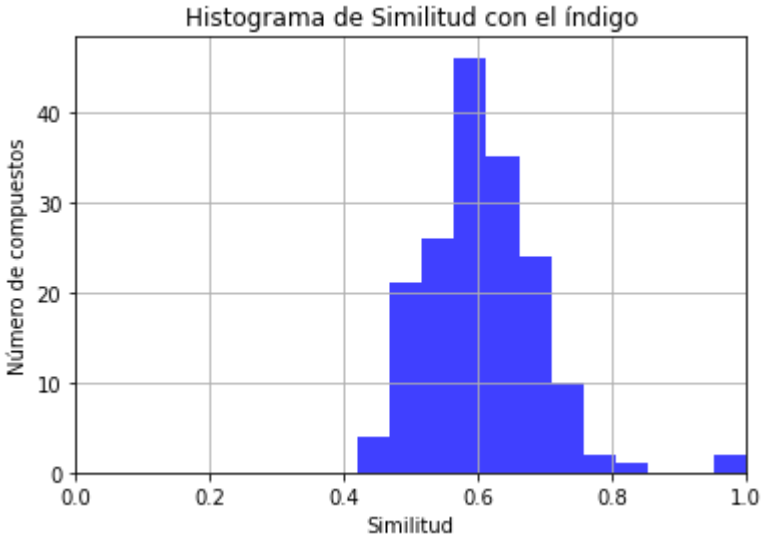
```
(171, 7)
```

Similitud molecular

```
indigo=Chem.MolFromSmiles('O=C1/C(=C2Nc3ccccc3C2=O)Nc2ccccc21')
```

```
mol_list=dataset['Molecule']
bi={}
fp_indigo=MACCSkeys.GenMACCSKeys(indigo)
simil_ind_list=[]
for mol in mol_list:
  fp_mol=MACCSkeys.GenMACCSKeys(mol)
  simil_ind=DataStructs.DiceSimilarity(fp_indigo, fp_mol)
  simil_ind_list.append(simil_ind)
```

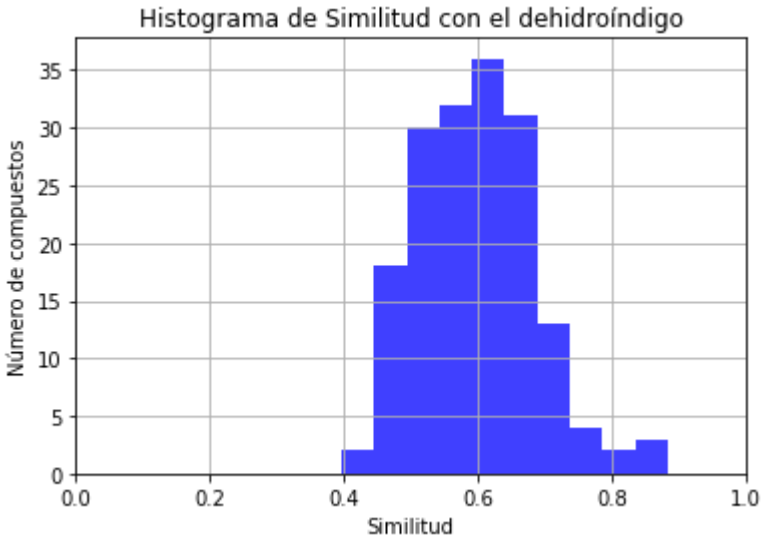
```
n, bins, patches = plt.hist(simil_ind_list, 12, facecolor='blue', alpha=0.75
plt.xlabel('Similitud')
plt.ylabel('Número de compuestos')
plt.xlim(0, 1.0)
plt.title(r'Histograma de Similitud con el índigo')
plt.grid(True)
plt.show()
```




```
dehidroindigo=Chem.MolFromSmiles('O=C1C(=Nc2ccccc12)C4=Nc3ccccc3C4=O')


mol_list=dataset['Molecule']
bi={}
fp_dehidroindigo=MACCSkeys.GenMACCSKeys(dehidroindigo)
simil_dhind_list=[]
for mol in mol_list:
  fp_mol=MACCSkeys.GenMACCSKeys(mol)
  simil_dhind=DataStructs.DiceSimilarity(fp_dehidroindigo, fp_mol)
  simil_dhind_list.append(simil_dhind)


n, bins, patches = plt.hist(simil_dhind_list, 10, facecolor='blue', alpha=0.
plt.xlabel('Similitud')
plt.ylabel('Número de compuestos')
plt.xlim(0, 1.0)
plt.title(r'Histograma de Similitud con el dehidroíndigo')
plt.grid(True)
plt.show()
```



Exportar conjunto de datos final (formato CSV)

```
dataset.to_csv('Colorantes.csv', header=True, index=False)
```