Nova Scientia
Multidisciplinary scientific journal
ISSN 2007-0705
Universidad La Salle Bajío

Castro Vázquez, A., García Guaderrama, R., y Gallo, M. (2024). Predicción de la toxicidad de líquidos iónicos utilizando los descriptores moleculares ECFP y ACSF en conjunto con algoritmos de aprendizaje máquina. Nova Scientia, 16(32). https://doi.org/10.21640/ns.v16i32.3433

Predicción de la toxicidad de líquidos iónicos utilizando los descriptores moleculares ECFP y ACSF en conjunto con algoritmos de aprendizaje máquina

Toxicity prediction of ionic liquids using ECFP and ACSF molecular descriptors in conjunction with machine learning algorithms

Arnulfo Castro-Vázquez[1-2] / Reyna García-Guaderrama[3] / Marco Tulio Gallo Estrada[1] ✉

[1]Tecnológico Nacional de México, Instituto Tecnológico de Ciudad Juárez. Chihuahua, México
[2]Universidad Autónoma de Ciudad Juárez, Instituto de Ingeniería y Tecnología. Chihuahua, México
[3]Tecnológico Nacional de México, Instituto Tecnológico de Ciudad Cuauhtémoc. Chihuahua, México
✉Correspondencia: mgallo@itcj.edu.mx

Recepción: 01-08-2023 / Aceptación: 25-01-2024 / Publicación: 10-04-2024
 https://doi.org/10.21640/ns.v16i32.3433

Resumen: se describe el proceso de predicción de toxicidad de los líquidos iónicos, en particular con respecto a la línea celular en ratas IPC-81. Se estudiaron 355 estructuras moleculares de líquidos iónicos, cuya geometría tridimensional está codificada mediante cadenas de símbolos en lenguaje *Simplified Molecular Input Line Entry System* (SMILES). La alimentación de los datos de entrada a los diferentes modelos de aprendizaje máquina requiere que la información geométrica y de contactos atómicos cercanos de cada liquido iónico sea mapeada o trasformada a notación vectorial numérica (x_i), utilizando los siguientes descriptores moleculares: funciones de simetría centradas en cada átomo *Atom-Centered Symmetry Functions* (ACSF), y huellas digitales de conectividad extendida *Extended Connectivity Fingerprints* (ECFP). Se usaron tres algoritmos de aprendizaje máquina: *Extreme Gradient Boosting* (XGBoost), *Support Vector Regression* (SVR) y *Kernel Ridge Regression* (KRR) para construir el modelo matemático de regresión predictivo que relacione los valores de entrada x_i con el valor de respuesta, representado por el logaritmo de la concentración media efectiva $(y_i = logEC_{50})$ en la evaluación de toxicidad, usando como métrica del grado de ajuste, el coeficiente de determinación (r^2). Los resultados indican que la combinación ECFP, con una distancia radial de 6 vecinos atómicos, en conjunto con el algoritmo KRR, proporciona el mejor ajuste promedio con $r^2 = 0.8602 \pm 0.032$, y con respecto al descriptor molecular ACSF, el mejor ajuste promedio se obtuvo con el algoritmo XGBoost con $r^2 = 0.8029 \pm 0.055$.
Palabras clave: líquidos iónicos; algoritmos de aprendizaje máquina; toxicidad; estructuras moleculares

Abstract: this work describes the process of toxicity prediction of ionic liquids, specifically toxicity with respect to the IPC-81 rat cell line. We studied 355 molecular structures of ionic liquids, whose three-dimensional geometry is encoded by means of symbol strings such as Simplified Molecular Input Line Entry System (SMILES) language. The feeding of the input data to the different machine learning models requires that the geometrical and near atomic neighbor information of each ionic liquid be mapped or transformed to numerical vector notation (x_i) using the following two molecular descriptors: Atom-Centered Symmetry Functions (ACSF), and Extended Connectivity Fingerprints (ECFP). Three machine learning algorithms: Extreme Gradient Boosting (XGBoost), Support Vector Regression (SVR) and Kernel Ridge Regression (KRR) were used to build the predictive mathematical regression model relating the input values x_i to the response value represented by the logarithm of the mean effective concentration $(y_i = \log EC_{50})$ in the toxicity assessment, using the coefficient of determination (r^2) as a metric of the degree of fitness. The results obtained indicated that the ECFP combination with a radial distance of 6 atomic neighbors in conjunction with the KRR algorithm provides the best average fit with $r^2 = 0.8602 \pm 0.032$, and with respect to the ACSF molecular descriptor the best average fit was obtained with the XGBoost algorithm with $r^2 = 0.8029 \pm 0.055$.
Keywords: ionic liquids; machine learning algorithms; toxicity; molecular structures

1. Introducción

Los líquidos iónicos (ILs) se forman mediante la combinación de aniones y cationes, y se encuentran en fase líquida a temperaturas inferiores a 100°C. Hay muchos aniones y cationes que pueden combinarse para formar compuestos con diversas propiedades, así como diferentes aplicaciones en la industria. Los ILs pueden utilizarse como lubricantes, resinas, insecticidas ecológicos, catalizadores, etc. (Chipofya et al., 2022; Welton, 2018; Zhou y Qu, 2017).

El estado del arte muestra cómo varias herramientas de aprendizaje máquina han sido utilizadas con éxito para determinar propiedades fisicoquímicas de ILs líquidos iónicos. En una investigación reciente, Acar y su equipo (2022) calcularon la temperatura de fusión de 1253 ILs a partir una relación cuantitativa estructura-actividad (QSPR) empleando inicialmente 5272 descriptores moleculares. Los descriptores utilizados en esta investigación están relacionados con la estructura geométrica, grado de ramificaciones, peso molecular, tamaño, cargas etc. Se eliminaron los descriptores de alta y baja correlación quedando únicamente 137 descriptores que fueron usados para alimentar una red neuronal consistente en una capa de entrada, cinco capas intermedias y una capa de salida. Las dimensiones de la red fueron 137 neuronas en la capa de entrada, 512 neuronas en las primeras tres capas intermedias, 256 neuronas en la cuarta capa, 64 neuronas en la quinta capa y una sola neurona en la capa de salida que corresponde al valor de la temperatura de fusión. Los datos se dividieron en 80% como conjunto de entrenamiento y 20% para prueba y validación. El ajuste con los datos del conjunto de prueba proporcionó un coeficiente de determinación de $r^2 = 0.90$. Además, subsecuentemente al utilizar el algoritmo de permutación importante de características ELI5, se obtuvieron los diez descriptores más significativos para el modelo a base de red neuronal.

Otra aplicación científica relacionada con el aprendizaje máquina es el diseño de solventes a base de ILs. Por ejemplo, Hutchinson y Kobayashi (2019) calcularon la energía libre de solvatación de 643 moléculas orgánicas neutras, con un rango de polaridad de 0 a 7.14 debyes, empleando los descriptores moleculares *Function Class Fingerprints* (FCFP) y *Extended Connectivity Fingerprints4* (ECFP) para obtener los valores de entrada para el algoritmo de aprendizaje máquina. El algoritmo XGBoost se utilizó en la regresión, se dividió el conjunto de datos alimentados en 80% para entrenamiento, 10% para prueba y 10% para validación. Como resultado del ajuste se obtuvieron los siguientes valores r^2 =0.78 para validación, y $r^2 = 0.81$ de prueba para FCFP, en contraste con $r^2 = 0.74$ para validación, y $r^2 = 0.78$ de prueba con ECFP.

Los ILs también han sido empleados con éxito como solventes para la remoción de sustancias toxicas como el sulfuro de hidrógeno (H2S). Abdi et al. (2022) desarrollaron un modelo de predicción inteligente de absorción de H2S en 792 ILs, aplicando los siguientes seis algoritmos de aprendizaje máquina: sistema adaptivo de inferencia neuro-difusa, máquina de vectores de soporte de mínimos cuadrados, función de base radial, cascada, perceptrón multicapa y red neuronal de regresión generalizada. Las variables de entrada al modelo fueron: temperatura, presión, factor acéntrico, presión crítica, temperatura crítica y el valor de salida fue la solubilidad del sulfuro de hidrógeno; los resultados indican que la mejor predicción de la solubilidad se obtuvo con el modelo máquina de vectores soporte de mínimos cuadrados (LS-SVM), a través de un coeficiente de determinación de $r^2 = 0.990$.

Bouarab et al. (2021), a través de un artículo de revisión de literatura, presentaron diferentes tipos de modelos que se utilizan para la predicción de la viscosidad en líquidos iónicos, como ecuaciones termodinámicas, reglas de mezclado y métodos de aprendizaje máquina. Para la predicción de la viscosidad por medio de redes neuronales utilizaron relaciones cuantitativas estructura-actividad o la relación cuantitativa estructura-propiedad (QSAR/QSPR), empleando diferentes tipos de descriptores como constitucionales, geométricos, topológicos y derivados de la mecánica cuántica como el momento dipolar y la energía del último orbital ocupado. Estos modelos de aprendizaje máquina solo son capaces de predecir la viscosidad de líquidos iónicos de la misma familia, y la capacidad de predicción depende de la calidad de los datos suministrados a la red neuronal.

En su tesis de maestría, Sakloth et al. (2018) investigó la capacidad de modelos basados en redes neuronales artificiales (ANN) para predecir de forma simultánea las siguientes propiedades: densidad, viscosidad y capacidad calorífica. Los valores de entrada a la red utilizados incluyen la temperatura, la presión y los descriptores geométricos de conectividad, basados en propiedades moleculares, cargas, área superficial, volumen etc. La base de datos utilizada incluyó 23 000 estructuras moleculares obtenidas del repositorio ILThermo (Ionic Liquids Database - ILThermo, n.d.). La arquitectura de la ANN evaluadas contiene 2, 3, 4 y 5 capas interconectadas con 16, 32, 64, 128, 256 y 512 neuronas por capa, con funciones de activación ReLU

«Rectified Linear Unit». Los datos de entrada se dividieron en 75% entrenamiento y un 25% para prueba. Los resultados obtenidos muestran coeficientes de determinación r^2 en el rango de 0.87 a 0.97.

Por otro lado, Petkovic et al. (2011) realizaron un análisis sobre las aplicaciones de los ILs respecto de la toxicidad y su impacto ambiental. La participación de Ranke et al. (2004) destacó al ser los primeros en proponer el uso de líneas celulares de rata, concretamente de leucemia IPC-81 para evaluar la citotoxicidad en los siguientes líquidos iónicos [$C_n$mim][BF4] (n=4,6,8), los cuales presentaron mayor toxicidad que las cadenas alquílicas de mayor tamaño.

Un trabajo importante sobre la toxicidad en ILs es el trabajo desarrollado por Yan et al. (2012) que implementaron un modelo de predicción QSAR en la línea celular de una rata leucémica $(\log EC_{50}\ \mathrm{IPC}-81)$, con 173 estructuras moleculares de ILs. El modelo emplea varios índices topológicos (IT) que incluyen características atómicas como radio, electronegatividad y posición del átomo, y la interrelación estructural del anión y del catión, como resultado se obtiene un coeficiente de ajuste r^2 =0.938 y un error absoluto promedio de 0.226.

Sosnowska et al. (2017) desarrollaron modelos locales y globales de regresión múltiple QSAR, basado en la línea celular de una rata leucémica $(\log EC_{50}\ \mathrm{IPC}-81)$, en el que utilizaron 304 estructuras moleculares de ILs. Los modelos locales se aplican solo a un tipo específico de compuestos, por ejemplo, líquidos iónicos del tipo imidazolio, mientras que los globales se aplican a todos los compuestos. Los descriptores moleculares utilizados son invariantes holísticas, que se apoyan en los anillos topológicos, constitucionales y basados en el número de grupos funcionales presentes. Los líquidos iónicos estudiados correspondieron a seis grupos funcionales: amonio, imidazolio, morfolinio, piperidonio, piridonio y pirrolidonio. Los resultados muestran una r^2 de alrededor de 0.7 a 0.8 para el modelo global, con excepción de los líquidos iónicos del tipo morfolinio que presentan una r^2 de alrededor de 0.50.

Wu et al. (2020) implementaron un modelo de predicción log EC50 IPC-81, en el que emplearon 304 estructuras moleculares de ILs, basado en una relación cuantitativa estructura-toxicidad (QSTR) por medio de 33 descriptores. Algunos de estos incluyen el tamaño de la cadena de carbono en el catión, el número de átomos de oxígeno en el catión, el número de sustituyentes en el catión. Los resultados de la regresión múltiple muestran un coeficiente de $r^2 = 0.90$ para el conjunto de prueba.

En este trabajo de investigación se desarrollaron modelos de aprendizaje máquina de regresión predictivos como el de *Refuerzo de gradientes extremo* (XGBoost), *Regresión ridge de kernel* (KRR) y *Soporte de vectores de regresión* (SVR) para la determinación de la toxicidad en la línea celular de la rata leucémica $(\log EC_{50}\ \mathrm{IPC}-81)$ en 355 ILs. Los descriptores utilizados se basan en las *Funciones de simetría centrada en el átomo* (ACSF), y *Huella dactilar de conectividad extendida* (ECFP) para transformar o mapear la información geométrica tridimensional en vectores numéricos.

2. Métodos, técnicas e instrumentos

La metodología utilizada en este trabajo para el modelo de regresión de predicción de la toxicidad de ILs, que utiliza descriptores moleculares, en conjunción con modelos de aprendizaje máquina comprende las siguientes cinco etapas, como se muestran en la figura 1.

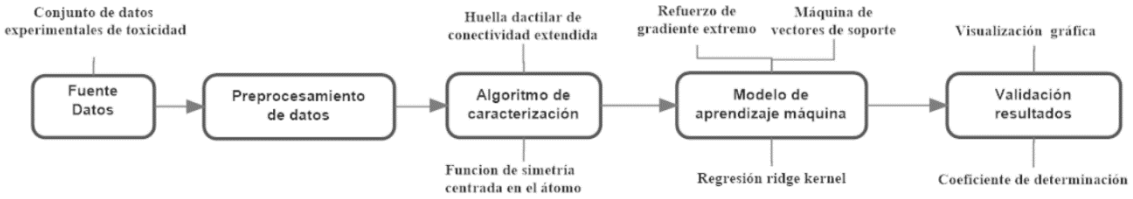

Figura 1. Metodología utilizada para la predicción de toxicidad celular de líquidos iónicos.
Figure 1. Methodology used for the cell toxicity prediction of ionic liquids.

La primera etapa, en la de obtención de datos, incluye la búsqueda de valores de toxicidad en las bases de datos de la comunidad científica, de las cuales se obtuvieron valores experimentales del logaritmo de la concentración media máxima efectiva de toxicidad $(logEC_{50})$ para 355 líquidos iónicos en células de la rata del tipo IPC81 (Sosnowska et al., 2017; Wang et al., 2020; Wu et al., 2020). En esta base de datos, se encuentran

líquidos iónicos que contienen cationes del tipo amonio, pirrolidinio, piridonio e imidazolio, y aniones como sulfatos, tetraborofluorato, metanosulfanato, benzoato, Bis(trifluorometilsulfonil)imida, acetato, bromuro, cloruro, ioduro, dicianamida, nitrato, hexanoato, tetracianidoboranuida, tiocianato, hexafluorofosfato, cianoboratos, tetracarbonilcobalato, nitrato, trifluoroacetato, etc.

La segunda etapa de preprocesamiento consiste en convertir las cadenas de texto del formato SMILES (Zheng et al., 2019) a coordenadas xyz para cada uno de los 355 líquidos iónicos utilizando el software Open babel (O'Boyle et al., 2011).

La tercera, la de caracterización, comprende la codificación de la información contenida en la geometría e identidad atómica a un arreglo numérico o vector, es decir, el cálculo de las variables x_i para la ecuación $Y = f(x)$, en donde las y_i son el valor de respuesta, en nuestro caso, el valor la toxicidad $(\log EC_{50})$ para cada liquido iónico, mediante el uso de descriptores moleculares locales invariantes a la rotación, translación y permutación atómica. En este trabajo, se utilizan dos tipos de descriptores moleculares: el ECFP y el ACSF para el cálculo de las variables x_i. Antes de alimentar los valores de x_i, y_i al modelo de regresión, es necesario llevar a cabo la normalización de los datos, así como la eliminación de los valores atípicos o extremos.

La cuarta etapa engloba el entrenamiento de los algoritmos de aprendizaje máquina XGBoost, SVM y KRR para la obtención de los coeficientes del modelo de regresión no-lineal, la cual involucra la optimización de los hiperparámetros, por ejemplo, el parámetro de regularización para evitar el sobreajuste, tipo de kernel, etc.

La quinta etapa abarca la evaluación del modelo de regresión de aprendizaje máquina, se usa el conjunto de prueba para determinar el grado de predicción a través de datos distintos a los utilizados en el conjunto de entrenamiento.

A continuación, se describe a detalle los descriptores moleculares ACSF y ECFP empleados en esta investigación.

El objetivo de los descriptores moleculares es convertir la información geométrica y atómica (número atómico, masa, carga atómica, número de átomos vecinos, tipos de enlace entre los átomos) en expresiones matemáticas que puedan suministrarse a algoritmos de aprendizaje máquina como se muestra en la figura 2 (Dong et al., 2015).

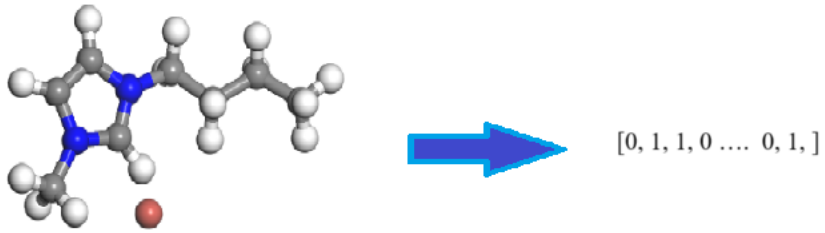

Figura 2. Representación matemática de la geometría molecular de un líquido iónico a través de un mapeo utilizando descriptores moleculares.
Figure 2. Mathematical representation of the molecular geometry of an ionic liquid through a molecular descriptors mapping.

El ACSF es un descriptor cuya finalidad es transformar las coordenadas atómicas cartesianas en funciones de simetría en un entorno local basado en un radio de corte para cada átomo (Behler, 2011; Behler, 2015; Gastegger et al., 2018; Scaomath, 2019), la función de corte (f_c) utiliza la siguiente expresión matemática:

$$f_c(R_{ij}) = \begin{cases} 0.5\left[cos\left(\pi \frac{R_{ij}}{R_c}\right) + 1\right], & R_{ij} \leq R_c \\ 0, & R_{ij} > R_c \end{cases} \quad (1)$$

En donde R_{ij} es la distancia entre los átomos i y j, si esta distancia es mayor que el radio de corte R_c, su valor es igual a cero. Las funciones de simetría más utilizadas son del tipo radial y angular, las funciones de simetría radiales se construyen como sumas de términos de dos cuerpos, las angulares involucran sumas de términos de tres cuerpos. En este trabajo de investigación se utilizaron tres funciones de simetría para describir el entorno del átomo: i: G_i^1, G_i^2, G_i^4 (Behler, 2011; Behler, 2015; Gastegger et al., 2018; Scaomath, 2019).

La función de simetría G_i^1 describe el entorno alrededor del átomo *i* mediante la suma de funciones de corte respecto a los átomos vecinos *j* alrededor de una esfera en un radio de corte, y su interpretación física se relaciona con el número de coordinación alrededor de un átomo central (Behler, 2011; Behler, 2015; Gastegger et al., 2018; Scaomath, 2019).

$$G_i^1 = \sum_{j=1}^{N_{atom}} f_c(R_{ij}) \quad (2)$$

La función de simetría G_i^2 consiste en la suma de gaussianas multiplicadas por la función de corte.

$$G_i^2 = \sum_j e^{-\eta(R_{ij}-R_s)^2} \cdot f_c(R_{ij}) \quad (3)$$

El ancho de las funciones gaussianas se encuentra definido por el parámetro η y el centro de las gaussianas puede desplazarse a cierta distancia radial mediante el parámetro R_s. Al utilizar valores pequeños de η y $R_s = 0$, la función G_i^2 se convierte en G_i^1, la interpretación física de G_i^2 se relaciona con el número de interacciones entre pares atómicos (Behler, 2011; Behler, 2015; Gastegger et al., 2018; Scaomath, 2019).

La función G_i^4 comprende la parte angular, en donde θ_{ijk} es el ángulo con respecto al átomo central i delimitado por las distancias interatómicas R_{ij} y R_{ik}.

$$G_i^4 = 2^{1-\xi} \sum_{j,k\neq i}^{all} (1 + \lambda cos\theta_{ijk})^{\xi} e^{-\eta\left(R_{ij}^2+R_{ik}^2+R_{jk}^2\right)} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \quad (4)$$

La parte angular es simétrica con respecto a $\theta_{ijk} = 180^o$. El parámetro λ puede valer entre $+1$ y -1, desplazando el máximo hacia valores de $\theta_{ijk} = 0$ y $\theta_{ijk} = 180^o$, respectivamente. El ancho de la función angular se define mediante el parámetro ζ. Además, las funciones de distribución angular se pueden calcular a ciertas distancias del átomo central empleando diversos valores de R_c y η, en la parte radial. La interpretación física de la función G_i^4, se relaciona con el número de ángulos existentes entre los átomos triples en la molécula (Behler, 2011; Behler, 2015; Gastegger et al., 2018; Scaomath, 2019).

Los valores numéricos de ACSF, para cada una de las coordenadas cartesianas de los 355 ILs estudiados, se obtuvieron empelando la librería computacional Dscribe (Himanen et al., 2020). La siguiente tabla muestra el conjunto de parámetros y sus respectivos valores utilizados en el cálculo de este descriptor molecular.

Tabla 1. Parámetros utilizados para el cálculo del descriptor ACSF.
Table 1. Parameters employed for the ACSF calculation.

Parámetro		Datos
Especies atómicas		'H', 'B', 'C', 'N', 'O','F', 'Al', 'Si', 'P', 'S', 'Cl', 'Co', 'As', 'Br', 'Sb', 'I'
Radio de corte R_c		12 ángstrom
Conjunto 1 C1	G_i^2 (η, R_s)	[5, 1], [2, 1], [0.5, 1], [5, 3], [2, 3], [0.5, 3]
	G_i^4 (η, ζ, λ)	[0.5, 2, 1], [0.5, 6, 1], [0.5, 16, 1], [0.05, 2, 1], [0.05, 6, -1], [0.05, 16, 1], [0.5, 2, -1], [0.5, 6, -1], [0.5, 16, 1], [0.05, 2, -1], [0.05, 6, -1], [0.05, 16, 1]
Conjunto 2 C2	G_i^2 (η, R_s)	[0.032, 0], [0.3, 0], [0.7, 0]
	G_i^4 (η, ζ, λ)	[0.032, 1, 1], [0.3, 1, -1], [0.7, 1, 1], [0.032, 1, -1], [0.3, 1, -1], [0.7, 1, -1]

Para cada uno de los 355 líquidos iónicos, se obtuvieron 1744 funciones de simetría (G_i^1, G_i^2, G_i^4), correspondiendo a los valores de x_i en el modelo de regresión, en conjunto con su respectivo valor de toxicidad $y_i = logEC_{50}$.

El descriptor molecular ECFP, denominado huellas de conectividad extendida, codifica de forma circular las características moleculares responsables de cierta funcionalidad física, química o biológica. Por ejemplo, puede ser usado para el grado de inhibición de enzimas para el desarrollo de nuevos fármacos, el

punto de ebullición, toxicidad celular, coeficientes de partición, etc. (Apodaca, 2019; Kumar, 2021; Rogers y Hahn, 2010).

Este descriptor realiza la codificación de las moléculas de la siguiente forma. Inicialmente, en la primera iteración para cada átomo, se le asigna un identificador numérico, resultado de aplicar un algoritmo de criptografía del tipo hash a un arreglo unidimensional de valores que comprenden el número de átomos vecinos distintos al átomo de hidrógeno, el número atómico, la masa atómica, la carga atómica, el número de átomos de hidrógeno enlazados, además, indica si el átomo forma parte de un anillo (Apodaca, 2019; Kumar, 2021; Rogers y Hahn, 2010). Subsecuentemente, en la segunda iteración, se crea un nuevo arreglo para cada átomo de la molécula, que incluye el identificador de la primera iteración para cada átomo y, adicionalmente, se agregan los identificadores de sus átomos vecinos inmediatos enlazados y se incluye un valor numérico del uno al cuatro para indicar si es enlace sencillo, doble, triple o aromático. A este nuevo arreglo se le aplica también el algoritmo hash para crear un nuevo identificador para cada átomo.

En la tercera iteración, se incluyen los identificadores de los átomos vecinos inmediatos enlazados, y se adicionan los identificadores de los vecinos enlazados a los átomos vecinos previos, creando un arreglo numérico concatenado con un mayor número de términos al cual se le aplica de nuevo el algoritmo hash para crear un nuevo identificador para cada átomo en la molécula. La siguiente etapa consiste en la remoción de los identificadores repetidos y la recolección de todos los identificadores atómicos de todas las iteraciones. Este arreglo final se convierte mediante una nueva codificación matemática en un arreglo binario (bits) ceros y unos, dando como resultado una longitud de 2048 elementos (Apodaca, 2019; Kumar, 2021; Rogers y Hahn, 2010).

A los descriptores ECFP se le adiciona un número que indica el diámetro del fragmento de mayor tamaño y equivale al doble de las iteraciones efectuadas, por ejemplo, si el fragmento de mayor tamaño tiene un radio de 4 enlaces, el descriptor molecular se denominara ECFP4 (Apodaca, 2019; Kumar, 2021; Rogers y Hahn, 2010).

El descriptor ECFP se implementó en este trabajo utilizando la librería computacional DeepChem (Bharath Ramsundar et al., 2019), con diferentes longitudes radiales (r=2, 4, 6 enlaces) para cada uno de los 355 líquidos iónicos, de los cuales se obtuvieron vectores de longitud de 2048 bits para cada sustancia.

En este trabajo de investigación se utilizaron tres algoritmos supervisados de aprendizaje máquina en la determinación de modelos de regresión predictivos para calcular la toxicidad de cada uno de los 355 líquidos iónicos como se muestra en la figura 3. Para cada uno los tres algoritmos de aprendizaje máquina, el conjunto de datos se dividió de forma aleatoria en 80% para entrenamiento y el 20% de prueba, para los cuales se aplicaron subrutinas computacionales en lenguaje Python, dentro del paquete computacional Scikit-learn (Pedregosa F., 2011).

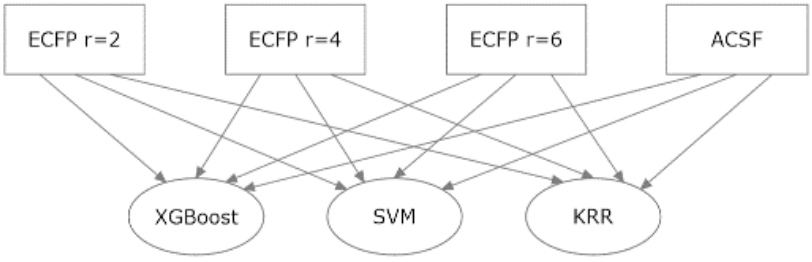

Figura 3. Descriptores moleculares y algoritmos de aprendizaje máquina empleados en la predicción de toxicidad de líquidos iónicos.
Figure 3. Molecular descriptors and machine learning algorithms used in the toxicity prediction of ionic liquids.

El algoritmo XGBoost, denominado *Refuerzo de gradiente extremo* (*Extreme Gradient Boosting*), es un algoritmo de aprendizaje máquina que se basa en la combinación de un conjunto o ensamble de modelos «f» multiplicados por sus respectivos pesos w, como se muestra en la ecuación 5, en donde cada modelo involucra árboles de decisión. Los modelos se añaden de uno en uno en forma secuencial con la finalidad de mejorar los errores de predicción, una especie de refuerzo o *boosting*, en donde se busca una nueva función en cada

iteración que maximice el producto interno con el gradiente negativo de la función de perdida evaluada en la iteración previa $f = F_{m-1}$ (Brownlee J., 2021; Chen y Guestrin, 2016).

$$F_m(x) = w_1 f_1(x) + w_2 f_2(x) + \cdots + w_m f_m(x) \qquad (5)$$

El algoritmo XGBoost se encuentra disponible a través de un paquete computacional de acceso libre (Chen y Guestrin, 2016), y los hiperparámetros empleados se muestran en la siguiente tabla:

Tabla 2. Hiperparámetros empleados en el algoritmo de aprendizaje máquina XGBoost.
Table 2. Hyperparameters used within the XGBoost machine learning algorithm.

Hiperparámetro	Valor
n_estimators	500
Max_depth	7
Eta	0.1
Subsample	0.7
Colsample_bytree	0.8

El hiperparámetro n_estimators representa el número de árboles que se utilizan en el modelo matemático, max_depth representa la profundidad máxima de cada árbol; en tanto, Eta, la velocidad de cambio utilizada por el algoritmo de optimización para la actualización de los pesos aplicados a las variables x_i; subsample controla la fracción de observaciones empleadas por cada árbol de decisión; colsample_bytree controla el número de variables x_i empleadas por cada árbol, además de los parámetros adicionales de regularización, como lambda con sus valores predeterminados (Brownlee, 2021; Chen y Guestrin, 2016; Jiang, 2021).

El algoritmo de los vectores de soporte para la regresión *Support vector regression* (SVR) se encuentra incluido dentro del paquete computacional Scikit-learn (Pedregosa, 2011). Dicho algoritmo de aprendizaje máquina supervisado se basa en determinar una curva de mejor ajuste «hiperplano», que contenga el número máximo de puntos, utilizando una función de mapeo no-lineal (kernel) a las variables de entrada, de tal forma que, se maximice la distancia (margen) entre el hiperplano y los datos más cercanos, minimizando también el error de predicción entre los valores de salida experimentales y los de respuesta obtenidos mediante el algoritmo SVR (Brunton y Kutz, 2019; Jiang, 2021; Pedregosa, 2011).

Tabla 3. Hiperparámetros empleados en el algoritmo de aprendizaje máquina SVR.
Table 3. Hyperparameters employed within the SVR machine learning algorithm.

Hiperparámetro	Valor
Kernel	Radial Basis Function (RBF)
C	[0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000]
Gamma	[0.01, 0.1, 1, 10, 100]

Los hiperparámetros para el algoritmo SVR se muestran en la tabla 3, en donde kernel especifica el tipo de función matemática el algoritmo de aprendizaje para el mapeo de las variables de entrada. En nuestro caso es del tipo *Funciones de base radial* (RBF) o gaussianas. Gamma es el parámetro relacionado con el ancho de las gaussianas en el kernel y C es el valor inverso del parámetro de regularización para mantener los coeficientes de ajuste en un rango predeterminado en la optimización (Pedregosa, 2011).

El algoritmo de aprendizaje máquina del tipo *Regresión ridge del kernel* consiste en la aplicación de gaussianas a los valores x_i para crear funciones matemáticas suaves ponderadas por medio de pesos w_j, lográndose un ajuste preciso entre los datos experimentales y el modelo o función matemática $f^{ML}(x)$, una especie de regresión no-lineal (Pedregosa, 2011; Zhu, 2022; Rupp, 2015; Langer et al., 2022).

$$f^{ML}(x) = \sum_{j=1}^{N_T} w_j k(x, x_j) \qquad (6)$$

En donde las w_j son los pesos y k es el kernel «gaussianas» $k(x, x') = \exp(-\gamma(x - x')^2)$, el hiperparámetro gamma se relaciona con el ancho de las gaussianas. Como ejemplo, en la figura 4, se ajusta una función real (curva con líneas roja) mediante la suma de gaussianas o kernels (curvas grises) centradas en cada punto de color rojo (x_i, y_i), mediante la ecuación 6, para construir la $f^{ML}(x)$ representada por la curva de color negro; en este ajuste, se utilizó el siguiente conjunto de valores para w_j: $[0.58, 0.65, 0.70, 0.30, 0.15]$ y el valor de gamma en el kernel, $\gamma = 12.5$ (Langer et al., 2022; Langer, 2023).

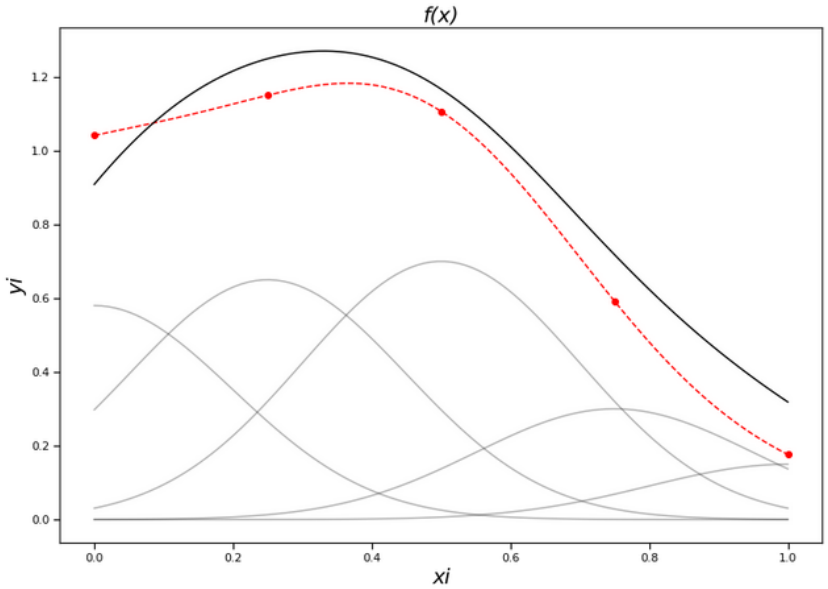

Figura 4. Aproximación de una función original o real (curva de color rojo) al utilizar el algoritmo KRR (curva de color negro); Fuente: $(Langer\ F., 2023)$.
Figure 4. Function approximation (original curve shown in red color) using the KRR algorithm, obtained curve shown in black color; Source. $(Langer\ F., 2023)$.

El proceso de implementación de este algoritmo se realizó por medio de la librería computacional Scikit-Learn (Pedregosa F., 2011). Los parámetros utilizados para el algoritmo se muestran en la tabla 4.

Tabla 4. Hiperparámetros empleados en el algoritmo de aprendizaje maquina KRR.
Table 4. Hyperparameters employed within the KRR machine learning algorithm.

Hiperparámetro	Valor
Alpha	[1.0, 1.5, 0.095, 0.01, 0.015, 0.001]
Gamma	[0.001, 0.01, 1 , 10, 100]
Kernel	Radial Basis Function (RBF)

Kernel representa la función matemática utilizada para el mapeo de las variables de entrada, mientras que alpha el parámetro de regularización para mantener los valores de w_j en un rango predeterminado, y gamma es el parámetro asociado al kernel del tipo RBF (Pedregosa, 2011; Zhu, 2022; Rupp, 2015; Haga clic o pulse aquí para escribir texto.Langer et al., 2022; Langer, 2023).

Al comenzar la etapa de entrenamiento de los modelos de aprendizaje máquina, se determinan los valores de los hiperparámetros que proporcionen el menor error y la menor desviación estándar, mediante el algoritmo computacional GridsearchCV, contenido en el paquete Scikit-Learn (Pedregosa F., 2011) y, subsecuentemente, se emplean los hiperparámetros dentro del modelo de regresión en el conjunto de prueba. El algoritmo GridSearchCV prueba todas las posibles combinaciones de los valores de los hiperparámetros mediante una cuadricula multidimensional y calcula el error para cada modelo de aprendizaje máquina, en conjunto con el método de validación cruzada para obtener los valores óptimos.

La validación cruzada de K iteraciones $k = 5$ consiste en dividir los datos de entramiento (x_i, y_i) en k subconjuntos de datos. En la primera iteración, se utilizan $(k-1)$ subconjuntos para entrenamiento y un subconjunto de datos como prueba para determinar las métricas relacionadas con el error de predicción. Las iteraciones subsecuentes separan un subconjunto diferente al de la iteración previa como subconjunto de prueba para calcular el error. El error obtenido representa el promedio de las k iteraciones, y el conjunto de hiperparámetros con menor error se utilizan para el modelo final de aprendizaje máquina para evaluar su capacidad predictiva empleando un conjunto de datos frescos o nuevos (Pedregosa, 2011; Brunton y Kutz, 2019; Jiang, 2021).

3. Resultados y discusión

En la tabla 5, se presenta el cálculo del coeficiente de determinación (r^2) para cada modelo de regresión para el que se emplean los dos descriptores moleculares (ACSF y ECFP), en conjunto con los tres diferentes algoritmos de aprendizaje máquina (XGBOOST, SVR y KRR). Para la determinación de valores estadísticos (media y desviación estándar), los cálculos de regresión se repitieron 20 veces, en donde el conjunto de datos se dividió para cada cálculo de forma aleatoria en los subconjuntos de entrenamiento (80%) y de prueba (20%).

Tabla 5. Resultados obtenidos para cada descriptor molecular en conjunto con el algoritmo de aprendizaje máquina en el conjunto de prueba.
Table 5. Results obtained for each molecular descriptor in conjunction with machine learning algorithm using the test set.

Descriptor molecular	Modelo de aprendizaje máquina	Valor promedio del coeficiente de determinación	Desviación estándar del coeficiente de determinación
ECFP2	XGBoost	0.7250	0.0570
ECFP2	SVR	0.6967	0.0410
ECFP2	KRR	0.8309	0.0344
ECFP4	XGBoost	0.7814	0.0674
ECFP4	SVR	0.7092	0.0551
ECFP4	KRR	0.8392	0.0374
ECFP6	XGBoost	0.8211	0.0411
ECFP6	SVR	0.6896	0.0765
ECFP6	KRR	0.8602	0.0320
ACSF_C1	XGBoost	0.8029	0.0559
ACSF_C1	SVR	0.7148	0.0629
ACSF_C1	KRR	0.7696	0.0518
ACSF_C2	XGBoost	0.7851	0.0562
ACSF_C2	SVR	0.6948	0.0584
ACSF_C2	KRR	0.7180	0.0594

Se puede apreciar que el mejor desempeño en el proceso de predicción se obtuvo con el caracterizador ECFP6 y el algoritmo de aprendizaje máquina KRR, los cuales presentaron una desviación estándar $\sigma = 0.0320$ y un valor promedio de $r^2 = 0.8602$. La figura 5 muestra la gráfica de dispersión para los valores de predicción de la toxicidad contra los valores experimentales en el conjunto de prueba.

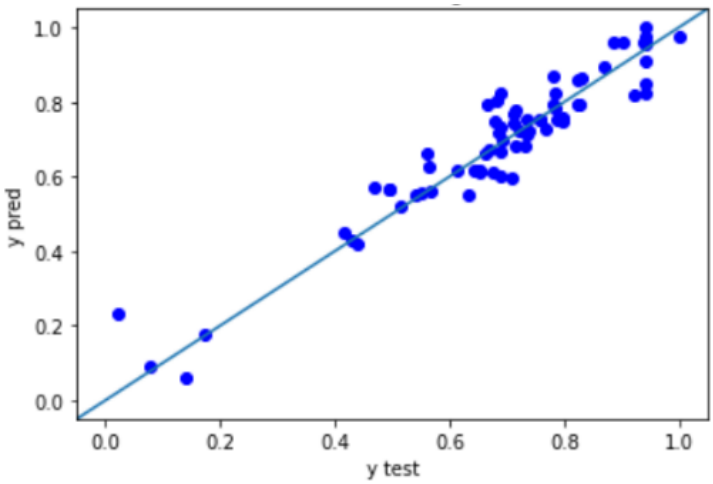

Figura 5. Comparación entre el valor calculado de la toxicidad (normalizada) vs. el valor experimental (normalizado) para el descriptor molecular ECFPr6 al emplear el algoritmo KRR, la curva y=x se utiliza como guía para visualizar el nivel de error.
Figure 5. Comparison between the calculated toxicity value (normalized) vs. the experimental value (normalized) for the molecular descriptor ECFPr6 using the KRR algorithm, the y=x curve is used as a guide to visualize the error level.

En contraste, para el descriptor molecular ECFP6, en conjunto con el algoritmo SVR, se obtuvo un peor ajuste en el nivel de predicción, el valor del coeficiente de determinación en promedio fue de $r^2 = 0.6896$, con una desviación estándar $\sigma = 0.0765$. La figura 6 muestra la gráfica de dispersión para los valores de predicción de la toxicidad contra los valores experimentales en el conjunto de prueba para este resultado.

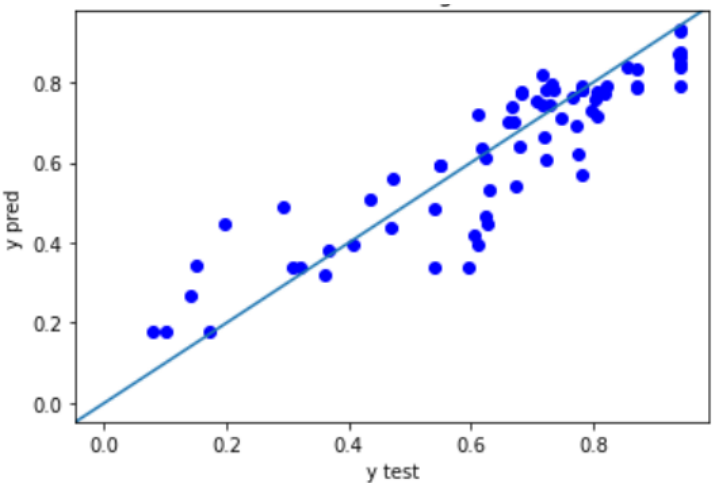

Figura 6. Comparación entre el valor calculado de la toxicidad (normalizada) vs. el valor experimental (normalizado) para el descriptor molecular ECFP6 al emplear el algoritmo SVR, la curva y=x se utiliza como guía para visualizar el nivel de error.
Figure 6. Comparison between the calculated toxicity value (normalized) vs. the experimental value (normalized) for the molecular descriptor ECFP6 using the SVR algorithm, the y=x curve is used as a guide to visualize the error level.

Respecto al descriptor molecular ACSF, se utilizaron dos conjuntos de valores de parámetros denominados $C1$ y $C2$ para el cálculo de las funciones G_i^2 , G_i^4. El mejor modelo de regresión se obtuvo utilizando el algoritmo XGBoost, en conjunto con los parámetros C1, lo que da como resultado una desviación estándar de $\sigma = 0.0559$, y un valor promedio de $r^2 = 0.8029$. La figura 7 muestra la gráfica de dispersión para los valores de predicción de la toxicidad contra los valores experimentales en el conjunto de prueba.

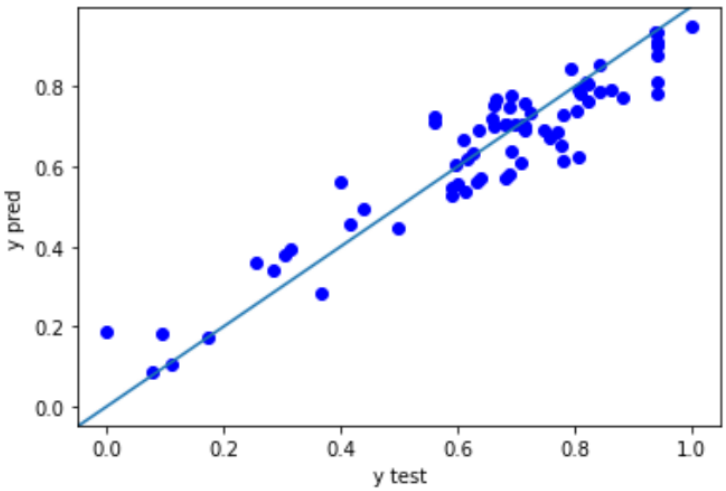

Figura 7. Comparación entre el valor calculado de la toxicidad (normalizada) vs. el valor experimental (normalizado) para el descriptor molecular ACSF al emplear el algoritmo XGBOOST, la curva y=x se utiliza como guía para visualizar el nivel de error.
Figure 7. Comparison between the calculated toxicity value (normalized) vs. the experimental value (normalized) for the molecular descriptor ACSF using the XGBOOST algorithm, the y=x curve is used as a guide to visualize the error level.

Es importante observar que para el descriptor molecular ACSF, el conjunto $C1$ permite una mayor resolución angular ζ entre tres átomos, lo cual aporta una mayor flexibilidad en comparación con el conjunto $C2$, en donde se utiliza un valor fijo de $\zeta = 1$, lo que da como resultado un mejor ajuste en el modelo de regresión. Además, la resolución radial η para el conjunto $C1$ que involucra dos átomos se encuentra más restringida en comparación con los valores del conjunto $C2$.

Respecto al descriptor molecular ECFP, el efecto de la distancia radial para la selección de los fragmentos atómicos indica que para los algoritmos KRR y XGBOOST, a mayor valor de la distancia radial (ECFP6), mejor es el ajuste en el modelo de regresión.

4. Conclusiones

En este trabajo, se utilizaron dos descriptores moleculares diferentes, el primero denominado ACSF, se basa en funciones de simetría radiales y angulares, y el segundo descriptor molecular ECFP que codifica de forma circular las siguientes características moleculares: número de átomos vecinos distintos al átomo de hidrogeno, número atómico, masa atómica, carga atómica, número de átomos de hidrogeno enlazados y aromaticidad. Estos descriptores, en conjunto con tres diferentes algoritmos de regresión de aprendizaje máquina (XGBOOST, SVR, KRR), se utilizaron para desarrollar modelos matemáticos que permitan determinar la toxicidad celular $(\log EC_{50}\ \mathrm{IPC} - 81)$ de líquidos iónicos compuestos de diferentes cationes y aniones.

Los resultados obtenidos muestran que los modelos matemáticos con un mejor ajuste de regresión, valor más cercano a uno del coeficiente de determinación se presentaron para el descriptor molecular ECFP6-KRR, con un valor promedio de $r^2 = 0.8602$ y una desviación estándar $\sigma = 0.0320$, mientras que el descriptor molecular ACSF-XGBOOST proporcionó un valor promedio de $r^2 = 0.8029$, con una desviación estándar de $\sigma = 0.0559$.

Como futura línea de investigación es importante diseñar nuevos descriptores moleculares que proporcionen un menor error de predicción en las propiedades fisicoquímicas de solventes. Sin embargo, estos modelos de aprendizaje máquina, en conjunto con los descriptores ECFP y ACSF, constituyen una alternativa robusta de bajo costo para evaluar de forma rápida la toxicidad celular de nuevos líquidos iónicos para aplicaciones como solventes en las diferentes actividades industriales.

5. Información adicional

No hay información adicional.

6. Agradecimientos

Arnulfo Castro agradece al Conahcyt por la beca otorgada para la realización de estudios de Doctorado; Marco Gallo agradece al Laboratorio Nacional de Supercómputo del Sureste de México (LNS), perteneciente al padrón de laboratorios nacionales Conahcyt , por los recursos computacionales, el apoyo y la asistencia técnica brindados a través del proyecto 202301015N.

Información de los autores

Arnulfo Castro Vázquez[1,2] 0000-0002-2345-2953
Reyna García-Guaderrama[3] 0000-0001-8641-5394
Marco Tulio Gallo Estrada[1] 0000-0001-7400-9233

Contribución de los autores en el desarrollo del trabajo

Los autores declaramos haber participado en la elaboración del documento de acuerdo con lo siguiente: Arnulfo Castro-Vázquez: desarrollo, implementación, escritura, validación; Reyna García-Guaderrama: Redacción y revisión del documento final; Marco Tulio Gallo Estrada: Conceptualización, administración y supervisión del proyecto, desarrollo, validación, escritura, revisión y edición del documento final.

Conflicto de interés

Los autores declaran que no existe conflicto de interés.

Referencias

Abdi, J., Hadipoor, M., Esmaeili-Faraj, S. H., y Vaferi, B. (2022). A modeling approach for estimating hydrogen sulfide solubility in fifteen different imidazole-based ionic liquids. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-08304-y

Acar, Z., Nguyen, P., y Lau, K. C. (2022). Machine-Learning Model Prediction of Ionic Liquids Melting Points. *Applied Sciences (Switzerland)*, *12*(5). https://doi.org/10.3390/app12052408

Apodaca, R. L. (2019, 14 de enero). *Computing Extended Connectivity Fingerprints*. Depth-First. https://depth-first.com/articles/2019/01/11/extended-connectivity-fingerprints/

Behler, J. (2011). Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *Journal of Chemical Physics*, *134*(7). https://doi.org/10.1063/1.3553717

Behler, J. (2015). Constructing high-dimensional neural network potentials: A tutorial review. *International Journal of Quantum Chemistry*, *115*(16), 1032–1050. https://doi.org/10.1002/qua.24890

Bharath Ramsundar, Peter Eastman, Patrick Walters, y Vijay Pande. (2019). *Deep Learning for the Life Sciences* (1st ed., Vol. 1). O´REILLY.

Bouarab, A. F., Harvey, J. P., y Robelin, C. (2021). Viscosity models for ionic liquids and their mixtures. *Physical Chemistry Chemical Physics*, *23*(2), 733–752. https://doi.org/10.1039/d0cp05787h

Brownlee J. (2021). *XGBoost for Regression - MachineLearningMastery.com*. https://machinelearningmastery.com/xgboost-for-regression/

Brunton, S., y Kutz I. (2019). *Data-Driven science and engineering: Machine Learning, dinamic systems and control* (2019th ed.). IEEE Control Systems Magazine.

Chen, T., y Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Chipofya, M., Tayara, H., y Chong, K. T. (2022). Deep Probabilistic Learning Model for Prediction of Ionic Liquids Toxicity. *International Journal of Molecular Sciences*, *23*(9), 5258. https://doi.org/10.3390/ijms23095258

Dong, J., Cao, D.-S., Miao, H.-Y., Liu, S., Deng, B.-C., Yun, Y.-H., Wang, N.-N., Lu, A.-P., Zeng, W.-B., y Chen, A. F. (2015). ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of Cheminformatics*, *7*(1), 60. https://doi.org/10.1186/s13321-015-0109-z

Gastegger, M., Schwiedrzik, L., Bittermann, M., Berzsenyi, F., y Marquetand, P. (2018). WACSF - Weighted atom-centered symmetry functions as descriptors in machine learning potentials. *Journal of Chemical Physics*, *148*(24), 1–15. https://doi.org/10.1063/1.5019667

Himanen, L., Jäger, M. O. J., Morooka, E. V., Federici Canova, F., Ranawat, Y. S., Gao, D. Z., Rinke, P., y Foster, A. S. (2020). DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, *247*, 106949. https://doi.org/10.1016/j.cpc.2019.106949

Hutchinson, S. T., y Kobayashi, R. (2019). Solvent-Specific Featurization for Predicting Free Energies of Solvation through Machine Learning. *Journal of Chemical Information and Modeling*, *59*(4), 1338–1346. https://doi.org/10.1021/acs.jcim.8b00901

Jiang, H. (2021). *Machine Learning Fundamentals* (Vol. 1). Cambridge University Press.

Kumar, M. (2021, March 25). *A beginner´s guide for understanding Extended-connectivity fingerprints(ECFPs)*. ChemicBook.

Langer F., M. (2023, June 5). *krr4mat - Jupyter Notebook*. NOMAD. https://nomad-lab.eu/prod/analytics/public/user/c8a96aa1-df79-4055-8fe0-67d43fed0f85/notebooks/tutorials/krr4mat.ipynb

Langer, M. F., Goeßmann, A., y Rupp, M. (2022). Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *Npj Computational Materials 2022 8:1*, *8*(1), 1–14. https://doi.org/10.1038/s41524-022-00721-x

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., y Hutchison, G. R. (2011). Open Babel: An Open chemical toolbox. *Journal of Cheminformatics*, *3*(10), 1–14. https://doi.org/10.1186/1758-2946-3-33/TABLES/2

Pedregosa F. (2011, May 5). *Scikit-Learn:Machine Learning in Python*. Https://Jmlr.Csail.Mit.Edu/Papers/V12/Pedregosa11a.Html.

Petkovic, M., Seddon, K. R., Rebelo, L. P. N., y Pereira, C. S. (2011). Ionic liquids: A pathway to environmental acceptability. *Chemical Society Reviews*, *40*(3), 1383–1403. https://doi.org/10.1039/c004968a

Ranke, J., Mölter, K., Stock, F., Bottin-Weber, U., Poczobutt, J., Hoffmann, J., Ondruschka, B., Filser, J., y Jastorff, B. (2004). Biological effects of imidazolium ionic liquids with varying chain lengths in acute Vibrio fischeri and WST-1 cell viability assays. *Ecotoxicology and Environmental Safety*, *58*(3), 396–404. https://doi.org/10.1016/S0147-6513(03)00105-2

Rogers, D., y Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742–754. https://doi.org/10.1021/ci100050t

Rupp, M. (2015). Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry*, *115*(16), 1058–1073. https://doi.org/10.1002/qua.24954

Sakloth, K., Beckner, W., Pfaendtner, J., y Goh, G. B. (2019). IL-Net: Using Expert Knowledge to Guide the Design of Furcated Neural Networks. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 1465–1473. https://doi.org/10.1109/BigData.2018.8622512

Scaomath, C. (2019). *LB -2.0 features: dscribe ACSF Descriptors | Kaggle*. https://www.kaggle.com/scaomath/lb-2-0-features-dscribe-acsf-descriptors

Sosnowska, A., Grzonkowska, M., y Puzyn, T. (2017). Global versus local QSAR models for predicting ionic liquids toxicity against IPC-81 leukemia rat cell line: The predictive ability. *Journal of Molecular Liquids*, *231*, 333–340. https://doi.org/10.1016/j.molliq.2017.02.025

Wang, Z., Song, Z., y Zhou, T. (2020). Machine Learning for Ionic Liquid Toxicity Prediction. *Processes*, *9*(1), 65. https://doi.org/10.3390/pr9010065

Welton, T. (2018). Ionic liquids: a brief history. *Biophysical Reviews*, *10*(3), 691–706. https://doi.org/10.1007/s12551-018-0419-2

Wu, T., Li, W., Chen, M., Zhou, Y., y Zhang, Q. (2020). Estimation of Ionic Liquids Toxicity against Leukemia Rat Cell Line IPC-81 based on the Empirical-like Models using Intuitive and Explainable Fingerprint Descriptors. *Molecular Informatics*, *39*(10), 2000102. https://doi.org/10.1002/minf.202000102

Yan, F., Xia, S., Wang, Q., y Ma, P. (2012). Predicting the Toxicity of Ionic Liquids in Leukemia Rat Cell Line by the Quantitative Structure–Activity Relationship Method Using Topological Indexes. *Industrial & Engineering Chemistry Research*, *51*(43), 13897–13901. https://doi.org/10.1021/ie301764j

Zheng, S., Yan, X., Yang, Y., y Xu, J. (2019). Identifying Structure-Property Relationships through SMILES Syntax Analysis with Self-Attention Mechanism [Research-article]. *Journal of Chemical Information and Modeling*, *59*(2), 914–923. https://doi.org/10.1021/acs.jcim.8b00803

Zhou, Y., y Qu, J. (2017). Ionic Liquids as Lubricant Additives: A Review. *ACS Applied Materials & Interfaces*, *9*(4), 3209–3222. https://doi.org/10.1021/acsami.6b12489

Zhu, R. (2022). *Statistical Learning and Machine Learning with R*. Ruoqing Zhu. https://teazrq.github.io/SMLR/