DOI: 10.37761/rsqp.v90i2.472

Recibido : 15.10.24
Aceptado : 13.12.24

PREDICCIÓN DEL COEFICIENTE DE PARTICIÓN DE SIDERÓFOROS USANDO REDES NEURONALES ARTIFICIALES

Jesús Alvarado-Huayhuaz[a], Miquéias Amorim Santos Silva[b], Karina dos Santos Machado[b], Ana Cecilia Valderrama Negrón*[a]

RESUMEN

El coeficiente de partición octanol-agua (logP) es un indicador importante en el estudio de lipofilia y permeabilidad celular, por ello, es un descriptor molecular recurrente en las reglas empíricas que evalúan la farmacocinética de una molécula. Los sideróforos son moléculas de interés farmacológico, debido a su potencial efecto caballo de Troya, sin embargo, encontramos dos problemáticas principales, su peso molecular mayor a 500 Dalton y la falta de bases de datos con información de las coordenadas atómicas de la estructura tridimensional y descriptores moleculares. En este trabajo, hemos creado una base de datos con el código SMILES de los sideróforos, nombre, microorganismo asociado, descriptores moleculares, entre otros, que se encuentra disponible en nuestro repositorio https://github.com/inefable12/siderophores_database. También creamos una página web para visualizar las estructuras 2D y 3D (https://sideroforos.streamlit.app). Además, demostramos una manera rápida y eficiente de estimar el logP para los sideróforos, usando redes neuronales artificiales en R. La información que proveemos en este artículo permitirá facilitar el estudio estructural de los sideróforos, el diseño de potenciales metalofármacos, la generación de sus estructuras tridimensionales para simulaciones con docking y dinámica molecular, así como también, el desarrollo de nuevos modelos predictivos de propiedades empleando inteligencia artificial.

Palabras clave: Coeficiente de partición, logP, sideróforo, redes neuronales artificiales.

PREDICTION OF SIDEROPHORES PARTITION COEFFICIENT USING ARTIFICIAL NEURAL NETWORKS

ABSTRACT

The octanol-water partition coefficient (logP) is a crucial indicator in the study of lipophilicity and cell permeability, making it a recurring molecular descriptor in empirical rules for evaluating a molecule's pharmacokinetics. Siderophores are pharmacologically relevant molecules due to their potential Trojan horse effect; however, two major challenges arise: their molecular weight often exceeds 500 Daltons, and there is a lack of

[a] Laboratorio de Investigación en Biopolímeros y Metalofármacos (LIBIPMET), Facultad de Ciencias, Universidad Nacional de Ingeniería, Av. Túpac Amaru 210, Lima, Perú, anitacvn29@yahoo.com.mx
[b] COMBI-Lab, Grupo de Biologia Computacional, Centro de Ciências Computacionais, Universidade Federal do Rio Grande, Rio Grande, RS, Brasil

databases containing atomic coordinates of their three-dimensional structures and molecular descriptors. In this work, we have created a database containing the SMILES codes of siderophores, their names, associated microorganisms, molecular descriptors, among other information, which is available in our repository at https://github.com/inefable12/siderophores_database. We have also developed a web page to visualize the 2D and 3D structures (https://sideroforos.streamlit.app). Additionally, we demonstrate a quick and efficient way to estimate the logP for siderophores using artificial neural networks in R. The information provided in this article aims to facilitate the structural study of siderophores, the design of potential metallodrugs, the generation of their three-dimensional structures for docking and molecular dynamics simulations, as well as the development of new predictive models for properties using artificial intelligence.

Keywords: Partition coefficient, logP, siderophore, artificial neural networks.

INTRODUCCIÓN

Una molécula con potencial actividad farmacológica debe poseer ciertas características químicas que favorezcan su biodisponibilidad, permeabilidad celular, solubilidad, entre otros. Muchas propiedades químicas o descriptores se pueden estimar con modelos computacionales o *in silico* y pueden facilitar su selección en bases de datos moleculares, para luego ser estudiadas *in vitro*. Existen reglas empíricas para evaluar la capacidad de una molécula de convertirse en un fármaco, conocida como drogabilidad o druglikeness, y algunas de las reglas más populares son las de Lipinski (1), Ghose (2), Veber (3), Egan (4) y Muegge (5). Estas consideran un rango de valores para la masa molecular, el área de superficie polar topológica (TPSA), el número de átomos capaces de formar enlaces de hidrógeno, entre otros, como se muestra en la Tabla 1.

Tabla 1. Druglikeness de reglas empíricas clásicas.

Modelo	Druglikeness	Ref.
Lipinski	MW<500, logP<4.15, N ó O <10, NH ó OH <5	(1)
Ghose	160<MW<480, -0.4<logP<5.6, 40<MR<130, 20<átomos<70	(2)
Veber	Enlaces rotables <10, TPSA <140	(3)
Egan	logP<5.88, TPSA<131.6	(4)
Muegge	200<MW<600, -2<logP<5, TPSA<150, anillos<7, C>4, heteroátomos>1, enlaces rotables< 15, HBA<10, HBD<5	(5)

MW: masa molecular, logP: coeficiente de partición agua-octanol, MR: refractividad molar, TPSA: área superficial polar topológica, HBA: aceptor de enlace puente de hidrógeno, HBD: donador de enlace puente de hidrógeno.

Uno de los parámetros más recurrentes es el coeficiente de partición, que consiste en la relación de las concentraciones de la molécula en estudio entre dos disolventes inmiscibles en equilibrio. El coeficiente de partición octanol-agua (logP) se utiliza como una medida de lipofilia y un indicador general de la permeabilidad celular (6), con valores óptimos reportados entre 1 y 3. Se han desarrollado varias herramientas para la predicción

de logP (7), principalmente para ahorrar en reactivos químicos, esfuerzo y tiempo. De este modo, encontramos métodos basados en mecánica cuántica (ab initio, semiempíricos, DFT, etc.), que requieren el cálculo de la energía electrónica de la molécula en octanol, agua y vacío, para estimar el logP mediante la determinación del valor de la Energía Libre de Gibbs. Estos métodos son menos adecuados para su aplicación en grandes bases de datos debido a su alto costo computacional. Métodos menos costosos, y por lo tanto considerados estándar, se basan en la suma de las contribuciones al logP por átomo o grupo funcional (también llamado fragmento) (8–14). Aunque estos métodos presentan un buen grado de predicción, tienen diferentes limitaciones, como valores irreales, sesgo que subestima el logP, número limitado de átomos de metales de transición, entre otros. Con respecto al uso de metales de transición (15), es necesario generar bases de datos de compuestos bioinorgánicos (16) debido al significativo potencial farmacológico representado por complejos o compuestos de coordinación entre biometales y fármacos orgánicos. Estos posibles metalofármacos exhiben una alta versatilidad química en términos de farmacocinética y farmacodinámica.

R es uno de los lenguajes de programación más utilizados en quimioinformática, biología estructural e inteligencia artificial, entre otros (17,18). Algunas bibliotecas en R facilitan el preprocesamiento y la visualización de datos, el desarrollo de redes neuronales o el uso de computación paralela en los procesadores; las que usamos en este trabajo se mencionan específicamente a continuación. Tidyverse es un conjunto de paquetes en R diseñados para facilitar y mejorar el flujo de trabajo en análisis y manipulación de datos (19). Tidyverse proporciona un conjunto coherente y consistente de herramientas para realizar tareas comunes en análisis de datos, como tibble, ggplot2, entre otros. DoParallel es otro paquete en R que se utiliza para facilitar la ejecución paralela de tareas en múltiples núcleos de procesadores (20). La ejecución paralela implica realizar múltiples tareas simultáneamente, distribuyendo la carga de trabajo entre los diferentes núcleos de procesadores de una computadora. Esto puede conducir a mejoras significativas en la velocidad de ejecución y la eficiencia del programa, especialmente cuando se trata de operaciones intensivas computacionalmente. El paquete caret (Classification And REgression Training) en R es una herramienta versátil y poderosa que ayuda eficientemente en la creación, evaluación y ajuste de modelos de aprendizaje automático (21). Su objetivo es simplificar y estandarizar el proceso de desarrollo de modelos, desde la preparación de datos hasta la evaluación y selección del modelo. El paquete VIM (Visualization and Imputation of Missing Values) en R está diseñado específicamente para abordar el problema de los valores faltantes en los conjuntos de datos (22). Los valores faltantes son una ocurrencia común en el análisis de datos y pueden ser problemáticos, potencialmente introduciendo sesgos o afectando la calidad del análisis y los resultados del modelo. El paquete neuralnet en R es una herramienta utilizada para construir y entrenar redes neuronales artificiales (ANNs). Las ANNs son un tipo de modelo de aprendizaje automático inspirado en la estructura y funcionamiento del cerebro humano. Estas redes son capaces de aprender a partir de datos y realizar tareas como clasificación, regresión y reconocimiento de patrones. En este trabajo, proponemos la predicción de logP utilizando redes neuronales en el lenguaje de programación R, basada en una base de datos original compuesta por moléculas con una masa molecular superior a 500 Daltons. Estas moléculas son de interés en la química bioinorgánica medicinal debido a su alta afinidad por el hierro (sideróforos). La predicción del logP para sideróforos nos permitirá filtrar rápidamente moléculas con el potencial de llevar a cabo

la estrategia farmacológica del efecto Caballo de Troya, que involucra la internalización de metales abióticos en microorganismos patógenos resistentes a los antibióticos.

PARTE EXPERIMENTAL

En este trabajo desarrollamos una base de datos de sideróforos y mostramos cómo a través de la generación de sus descriptores moleculares podemos predecir eficientemente su coeficiente de partición (LogP). Primero, hemos creado nuestra base de datos de 232 sideróforos mediante una búsqueda en la literatura científica en repositorios y páginas web. Se extrajeron los códigos SMILES (23) de las estructuras o se generaron si en caso no estaban disponibles. Para esta tarea nos apoyamos de Avogadro y además desarrollamos la página web https://quimicaorganica.streamlit.app. Esta web es de acceso libre y fue desarrollada usando recursos de código abierto. Usando el código SMILES generamos las coordenadas tridimensionales y pre-optimizamos las estructuras usando el algoritmo de descenso de gradiente y el campo de fuerza MMFF94 con Open Babel. Después de generar las 232 estructuras tridimensionales en formato MOL2, utilizamos el programa RFL-Score (24) para extraer 250 descriptores químicos utilizando los programas de quimioinformática Padel (25) y RDKit (26), donde nuestro atributo a predecir es el logP, indicado como MolLogP. Los paquetes de R: Tidyverse, doParallel, caret, VIM, neuralnet, y modelr, fueron empleados para realizar el preprocesamiento de datos, la eliminación recursiva de características (RFE), la separación de los datos (80% para el entrenamiento y 20% para la prueba), el desarrollo del modelo y el cálculo de métricas de evaluación, como ya hemos reportado previamente (27). Este procedimiento se resume en la Figura 1.

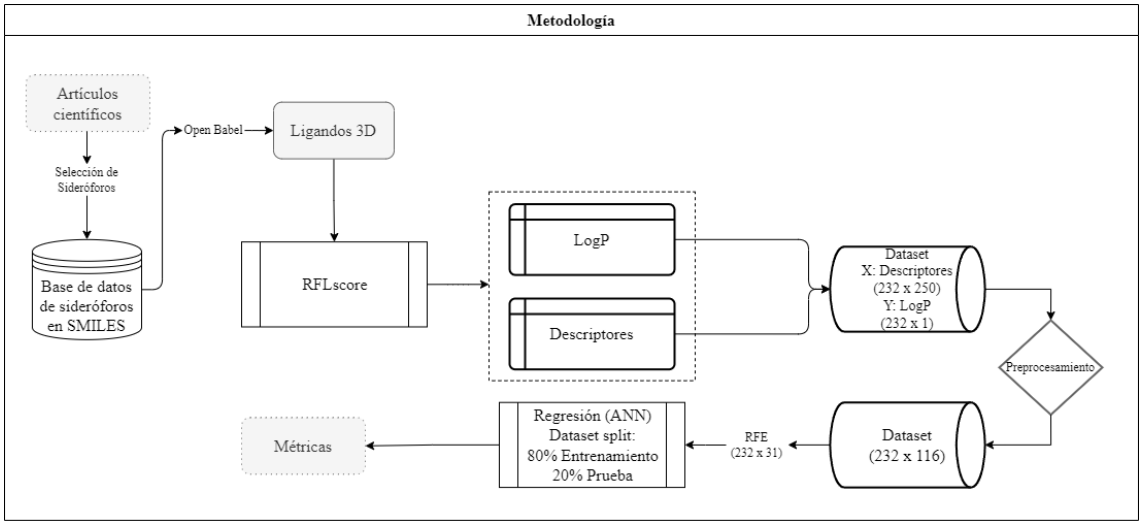

Figura 1. Resumen de la metodología.

RESULTADOS Y DISCUSIÓN

La base de datos moleculares en formato CSV (separado por comas), los descriptores generados, las estructuras tridimensionales, el modelo y los códigos implementados se encuentran disponibles en nuestro repositorio: https://github.com/inefable12/logP_ann. De los 250 descriptores obtenidos con Padel y RDKit (Figura 2), una gran cantidad estaba compuesta de ceros, por lo cual, desarrollamos un script para considerar las columnas con un mayor aporte de datos no nulos. De esta manera, conservamos únicamente 116 columnas de las 250 iniciales, que contenían al menos 150 datos diferentes de cero, lo cual significó una disminución del 46% del total de atributos. También se incorporó una columna con el tipo de sideróforo a la base de datos. Esta información es valiosa para posteriores investigaciones asociadas con el desarrollo de modelos de clasificación basado en descriptores moleculares.

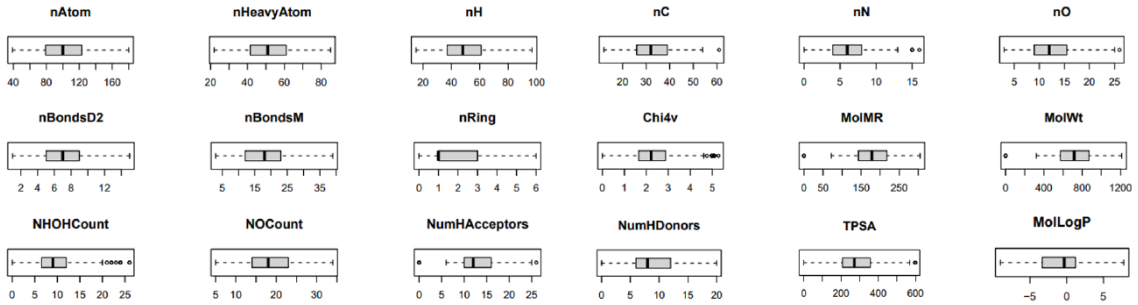

Figura 2. Boxplots de algunos atributos en nuestra base de datos.

Durante la selección de atributos con RFE aplicada a los 116 atributos, se identificó un menor error cuadrático medio (RMSE) en la validación cruzada repetida cuando se utilizaron 31 atributos, como se destaca con un punto azul en la Figura 3. Estos corresponden a los descriptores: TPSA, PEOE_VSAn (n: 1, 6, 7, 8, 9, 10, 12), nN, SlogP_VSA5, SlogP_VSAm (m: 1, 6, 7), SlogP_VSA2, SA_EState7, MQNs_polarity_counts_hba, Chi3v, SMR_VSA5, MQNs_atom_counts_ao, NumHeteroatoms, Kappa3, NOCount, NumHDonors, nC, FractionCSP3, Chi2v, MQNs_atom_counts_c, SMR_VSA3, BalabanJ, SMR_VSA7, NHOHCount. En seguida los datos fueron normalizados y particionados en conjuntos de entrenamiento (80%) y prueba (20%) para la siguiente etapa.

La arquitectura de la ANN consistió en 31 descriptores como entrada (input), 3 capas ocultas (3, 5 y 3 neuronas, respectivamente) y 1 salida, la predicción de LogP (Figura 4). Al representar gráficamente los datos de prueba, "LogP real" frente a los valores predichos o "LogP predicho" observamos la tendencia de estos valores (Figura 5). La regresión lineal nos da 0.99, 0.011, 0.00012, 0.0083 para las métricas R^2, RMSE, MSE y MAE, respectivamente, lo cual es un indicador de la alta precisión del modelo, como se observa en la Figura 5. Empleamos un script para distribuir la ejecución de cálculos en los procesadores en paralelo, lo cual permitió poder explorar diferentes combinaciones de capas ocultas, número de neuronas y tasa de aprendizaje, sin depender del uso de supercomputación que es habitual en estos casos.

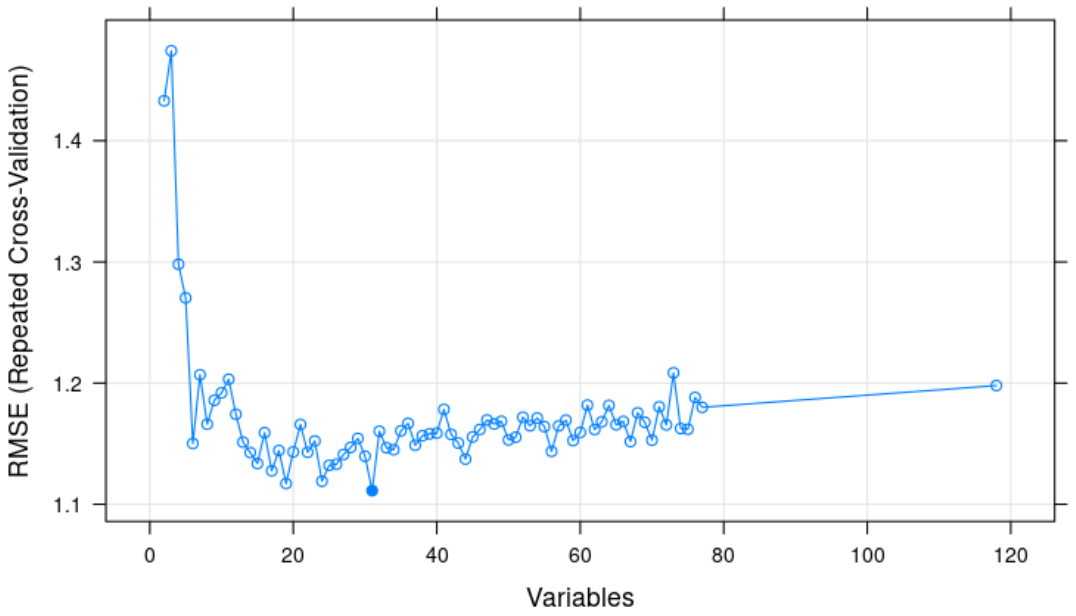

Figura 3. RMSE y validación cruzada respecto al número de variables.

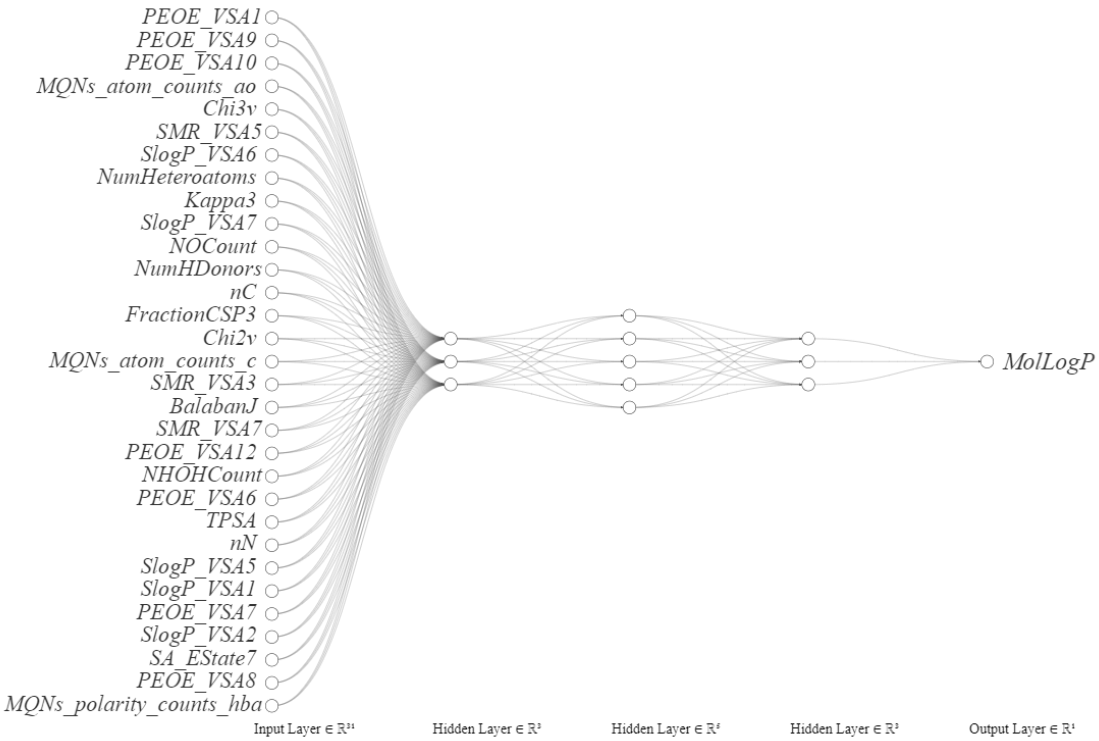

Figura 4. Red neuronal: 31 inputs, 3 capas ocultas (3, 5 y 3 neuronas) y 1 output.

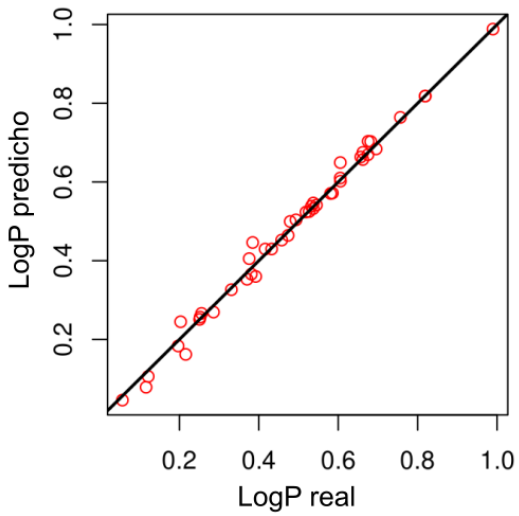

Figura 5. Valores de LogP normalizados del grupo de prueba.

Si bien estamos demostrando la obtención de un modelo con excelentes métricas para la predicción del target, en general es requerido realizar una optimización de hiperparámetros lo cual requiere de computación de alto rendimiento, como se han reportado en otros trabajos. Tetko y colaboradores trabajaron la predicción de logP basado en el método de distribución atómica reportando un R^2 de 0.918, utilizando un generador de descriptores comercial (MOE) y SMARTS, un codificador molecular más sofisticado que SMILES (18), mientras que otra publicación reciente basada en aprendizaje por transferencia predice el logP de pequeñas moléculas de un conjunto de datos de Reaxys y PhysProp con un R2 de 0.988 (13). Por lo tanto, destacamos nuestro procedimiento por tratarse de un modelo eficiente y constituido con herramientas de código abierto. Finalmente, la base de datos es también accesible desde https://sideroforos.streamlit.app, desde dispositivos móviles y computadoras en cualquier sistema operativo (Figura 6).

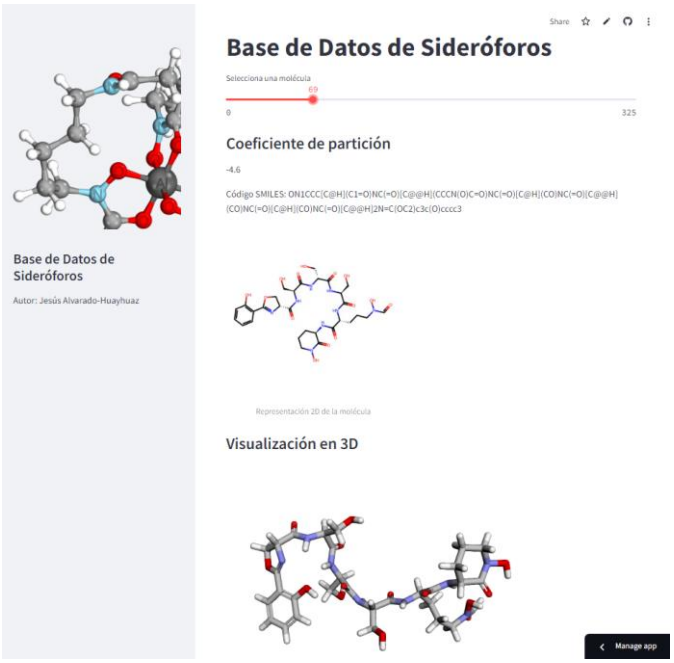

Figura 6. Página web para la visualización de la base de datos de sideñoforos.

CONCLUSIONES

Este trabajo ha generado una base de datos única que incluye el código SMILES, las coordenadas atómicas tridimensionales y 250 descriptores moleculares de 232 sideróforos, lo que permitirá a otros investigadores trabajar más fácilmente en estudios de propiedades fisicoquímicas y farmacológicas de estas moléculas. Se ha demostrado que es posible predecir el coeficiente de partición octanol-agua (logP) de los sideróforos mediante redes neuronales artificiales con un alto grado de precisión ($R^2 = 0.99$). Este método es particularmente relevante, dado que el cálculo experimental del logP es costoso y la predicción computacional tradicional se ve limitada por la masa molecular y el tipo de átomos en las moléculas. A diferencia de otros estudios que dependen de herramientas computacionales costosas y bases de datos comerciales, este trabajo demuestra que es posible obtener modelos predictivos eficientes utilizando recursos computacionales domésticos y software de código abierto, lo que facilita su aplicación en una variedad de entornos de investigación. La base de datos y el modelo desarrollado proporcionan una herramienta valiosa para estudios de fármacos basados en el efecto Caballo de Troya, la simulación molecular y el desarrollo de nuevos modelos predictivos en química medicinal. Aunque se ha logrado una excelente precisión en la predicción del logP, destacamos también la necesidad de optimizar los hiperparámetros del modelo para obtener resultados aún más robustos, particularmente si se emplean técnicas de computación de alto rendimiento. Este estudio sienta las bases para investigaciones futuras en el campo de la quimioinformática y el diseño de fármacos basados en sideróforos, utilizando inteligencia artificial para abordar desafíos clave en la predicción de propiedades moleculares.

AGRADECIMIENTO

JAAH agradece a FONDECYT (Convenio 237- 2015-FONDECYT) por la concesión de la beca de doctorado.

REFERENCIAS BIBLIOGRÁFICAS

1. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1. Adv Drug Deliv Rev. 2001;46(1):3–26.
2. Ghose AK, Viswanadhan VN, Wendoloski JJ. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. J Comb Chem. 1999;1(1):55–68.
3. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem. 2002 Jun 6;45(12):2615–23.
4. Egan WJ, Merz Kenneth M., Baldwin JJ. Prediction of Drug Absorption Using Multivariate Statistics. J Med Chem. 2000;43(21):3867–77.
5. Muegge I, Heald SL, Brittelli D. Simple Selection Criteria for Drug-like Chemical Matter. J Med Chem. 2001;44(12):1841–6.

6. Leo A, Hansch C, Elkins D. Partition coefficients and their uses. Chem Rev. 1971;71(6):525–616.
7. Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep. 2017;7(1):42717.
8. Fu L, Ye F, Feng Y, Yu F, Wang Q, Wu Y, et al. Both Boceprevir and GC376 efficaciously inhibit SARS-CoV-2 by targeting its main protease. Nat Commun. 2020;11(1):4417. doi: 10.1038/s41467-020-18233-x. .
9. Cheng T, Zhao Y, Li X, Lin F, Xu Y, Zhang X, et al. Computation of Octanol−Water Partition Coefficients by Guiding an Additive Model with Knowledge. J Chem Inf Model. 2007;47(6):2140–8.
10. Ghose AK, Crippen GM. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. J Chem Inf Comput Sci. 1987;27(1):21–35.
11. Plante J, Werner S. JPlogP: an improved logP predictor trained using predicted data. J Cheminform. 2018 Dec 14;10(1):61. doi: 10.1186/s13321-018-0316-5.
12. Tetko I V, Tanchuk VY. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. J Chem Inf Comput Sci. 2002;42(5):1136–45.
13. Pedretti A, Villa L, Vistoli G. VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs. J Mol Graph Model. 2002;21(1):47–9.
14. Goss K-U. Predicting the equilibrium partitioning of organic compounds using just one linear solvation energy relationship (LSER). Fluid Phase Equilib. 2005;233(1):19–22.
15. Chen Y-K, Shave S, Auer M. MRlogP: Transfer Learning Enables Accurate logP Prediction Using Small Experimental Training Datasets. Processes. 2021;9(11):2029. doi: 10.3390/pr9112029.
16. Medina-Franco JL, López-López E, Andrade E, Ruiz-Azuara L, Frei A, Guan D, et al. Bridging informatics and medicinal inorganic chemistry: Toward a database of metallodrugs and metallodrug candidates. Drug Discov Today. 2022;27(5):1420–30.
17. Mente S, Kuhn M. The Use of the R Language for Medicinal Chemistry Applications. Curr Top Med Chem. 2012;12(18):1957–64.
18. Rodríguez-Pérez R, Miljković F, Bajorath J. Machine Learning in Chemoinformatics and Medicinal Chemistry. Annu Rev Biomed Data Sci. 2022;5(1):43–65.
19. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4(43):1686. doi: 10.21105/joss.01686.
20. Corporation M, Weston S. doParallel: Foreach Parallel Adaptor for the "parallel" Package. CRAN: Contributed Packages. 2011. [citado 11 jul 2024]. Disponible en: https://cran.r-project.org/package=doParallel
21. Kuhn M. Caret: classification and regression training. Astrophys Source Code Libr. [Internet]. 2015;ascl-1505. [citado 20 set 2024]. Disponible en: https://ascl.net/1505.003
22. Blomquist R, Lell RM, Gelbard EM. VIM: a continuous energy Monte Carlo code at ANL. United States; 1980 p. 31-46

.

23. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–6.
24. Arrua OE, Aderhold A, Werhli A V., Dos Santos Machado K. RFL-Score: Random Forest with Lasso Scoring Function for Protein-Ligand Molecular Docking. In: 2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) [Internet]. IEEE; 2024. p. 1–8. [citado 15 jul 2024]. Disponible en: https://ieeexplore.ieee.org/document/10702128/
25. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–74.
26. Landrum G. RDKit: Open-source cheminformatics. [Internet]. [Citado 13 set 2024]. Disponible en: https://www.rdkit.org
27. Alvarado-Huayhuaz JA, Amorim Santos Silva M, Jimenez Peña EM, Claudio Rengifo Maraví J, Cordova-Muñoz AM, dos Santos Machado K. Artificial Neural Networks for the Rapid Prediction of Possible Ferroptosis Inducers Using the GPx4 Enzyme. In: 2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). Natal, Brazil; 2024. pp. 1–6. doi: 10.119/CIBCB58642.2024.10702169.