Algoritmo K-NN para la identificación de posibles fármacos contra la COVID-19

The K-NN algorithm for identifying potential COVID-19 drugs

Raúl Isea [1]

Fundación Instituto de Estudios Avanzados, Miranda, Venezuela[1]

raul.isea@gmail.com[1]

Fecha de recepción: 05/05/2025

Fecha de aceptación: 26/08/2025

Pág: 2 – 12

DOI: 10.5281/zenodo.17466373

Resumen

El objetivo de la investigación explorar y validar la aplicación del algoritmo K-NN para la identificación de grupos de compuestos que pueden ser empleadas contra la COVID-19 mediante métodos de quimioinformática. Para lograrlo, se analizaron los componentes de la base de datos ChEMBL empleados en estudios experimentales sobre el SARS-CoV-2. Esta información fue analizada de forma manual y, finalmente, se obtuvieron 1904 biomoléculas categorizadas como "Activas" o "Inactivas" en función de su actividad inhibitoria frente a dicho virus. Después, se empleó un algoritmo de K-vecinos más cercano (K-NN) para agrupar las biomoléculas en función de su similitud fisicoquímica. Finalmente, el estudio evidenció que este tipo de algoritmos es una herramienta valiosa para identificar posibles compuestos iniciales para posteriores investigaciones que ayuden a combatir la COVID-19, estableciendo de esta manera una base metodológica para futuros trabajos en el presente tema.

Palabras clave: ChEMBL, clústers, K-NN, quimioinformática, SARS-CoV-2.

Abstract

The goal of the study is to use cheminformatics techniques to identify possible medications that combat COVID-19. This was carried out by analyzing the ChEMBL database components used in SARS-CoV-2 experimental investigations. Following a manual analysis of this data, 1904 biomolecules were classified as "Active" or "Inactive" according to their ability to inhibit the virus.

Esta obra está bajo licencia CC BY-NC-SA 4.0.

After that, a K-nearest neighbors (K-NN) algorithm was used to help classify the biomolecules according to how similar they were physicochemically. Lastly, the study showed that this kind of algorithm is a useful tool for finding possible compounds for further investigation to aid in the fight against COVID-19.

Keywords: ChEMBL, clusters, K-NN, chemioinformatics, SARS-CoV-2.

Introducción

Frank Brown definió el concepto de quimioinformática (o informática química) en 1998 al utilizar varias aplicaciones informáticas para comprender los mecanismos de acción de los medicamentos, así como identificar nuevas moléculas con potencial terapéutico (Brown, 1998). De hecho, las técnicas computacionales se han erigido cada vez más como una opción viable para acelerar la identificación de posibles terapias, tras los progresos de metodologías como el cribado virtual y el docking molecular (Hernández et al., 2007), la dinámica molecular (Isea et al., 2013), la genómica inversa (Isea et al., 2016), y los modelos cuantitativos de relación estructura-actividad (Tian et al., 2024).

La ventaja principal de estos estudios es poder detectar compuestos efectivos contra dicha enfermedad antes de su elaboración, lo que implicaría una disminución de costos y tiempo de estudios. El fundamento de estas investigaciones radica en que los compuestos con estructuras moleculares similares tienden a tener acciones biológicas parecidas, un principio conocido como principio de similitud molecular (Maggiora et al., 2014).

Sin embargo, el problema es que mínimas modificaciones estructurales también pueden provocar alteraciones significativas en la actividad biológica, a menudo conocidas como *activity cliffs* (Hu et al., 2018). El mejor caso conocido es la morfina y el fentanilo (Woodhouse et al., 1996). Aunque ambas moléculas tienen estructuras químicas diferentes, ambas trabajan como agonistas opioides al ser responsables de activar los receptores opioides mu en el sistema nervioso central (Woodhouse et al., 1996).

La utilización de estas herramientas para la búsqueda de fármacos contra la COVID-19 no es un concepto innovador. Citemos el estudio conducido por De Clercq (2020) cuando revisó los compuestos antivirales que se estaban desarrollando contra este flagelo, destacando el papel que pueden jugar las técnicas computacionales para identificar nuevos compuestos contra dicho virus. Además, Khan et al. (2021) subrayaron la creación de modelos QSAR específicos para inhibidores de dianas virales como la proteasa principal (Mpro) y la ARN polimerasa dependiente de ARN (RdRp), con el objetivo de agilizar el proceso de identificación de medicamentos.

En paralelo, Liu et al. (2020) desarrollaron métodos tanto de QSAR como los utilizados en cribado virtual para poder detectar posibles inhibidores de la proteasa (Mpro del SARS-CoV-2, y así poder combatir dicha pandemia. Por su parte, Ghasemi et al. (2021) también identificaron medicamentos cuya acción inhibitoria estuviera centrada en el punto de entrada del SARS-CoV-2, es decir, la enzima angiotensina 2 (ACE2).

Asimismo, la labor de Dias et al. (2021) emplearon métodos fundamentados en aprendizaje profundo para anticipar la actividad antiviral de una serie de compuestos que deberían ser útiles para combatir el SARS-CoV-2. El uso de técnicas de aprendizaje profundo en QSAR constituye un progreso importante en la habilidad para modelar vínculos complejos entre la estructura molecular y la actividad biológica Ojha et al. (2021).

Por todo ello, el presente trabajo identifica agrupaciones o conjuntos de compuestos que muestran una actividad antiviral agrupados en clúster para combatir la COVID-19, empleando para ello el algoritmo de K-vecinos más cercanos (Zhang, 2016).

La selección de esta clase de algoritmo no se realizó de forma aleatoria. En realidad, se ha utilizado en otras investigaciones relacionadas al COVID-19, como se señala a continuación (Alie et al., 2024; Rabie et al., 2023; Sejuti y Islam, 2023):

- Detección de personas infectadas con COVID-19 basándose en la información clínica, síntomas, historial médico y resultados de exámenes de laboratorio (Alie et al., 2024; Rabie et al., 2023).
- Detección por radiografías, tomografías computarizadas o cualquier otro tipo de imagen para un rápido diagnóstico clínico (Sejuti y Islam, 2023).
- Detectar el grado de fatalidad en pacientes con infección de COVID (Alie et al., 2024).

Sin pasar por alto que estos algoritmos K-NN ha sido validados en otras áreas como por ejemplo, la identificación de patrones y la categorización de imágenes (Ávila et al., 2021), evaluación del riesgo crediticio, así como detectar fraude (Rjoub et al., 2023), la agrupación de muestras biológicas (Cottrell et al., 2023), entre otros.

Por lo tanto, la aplicación del algoritmo K-NN debería ser la técnica más apropiada para agrupar compuestos con similitudes fisicoquímicas anteriormente identificadas como inhibidores antivirales frente al SARS-CoV-2. Este método permitió la detección de aspirantes comprometidos para futuros estudios experimentales. Por todo ello, el presente trabajo concibe las bases metodológicas para la búsqueda de tratamientos para la COVID-19.

Metodología

Los datos para este estudio se obtuvieron de la base de datos ChEMBL (disponible gratuitamente en www.ebi.ac.uk/chembl) (Bento et al., 2020; Zdrazil et al., 2024) de donde

se extrajeron aquellas moléculas bioactivas empleadas en estudios experimentales contra el coronavirus. Recordemos que la base de datos ChEMBL posee información referente a la química estructural, así como de la bioactividad experimental y genómica que han sido procesados manualmente por el Instituto Europeo de Bioinformática (EBI), siendo una herramienta imprescindible para el descubrimiento y desarrollo de medicamentos (Zdrazil, 2025).

Los datos obtenidos fueron analizados con diferentes algoritmos escritos en Python (versión 3.11, disponible gratuitamente en www.python.org), donde se emplearon, básicamente, las bibliotecas desarrolladas en Python llamadas `chembl_webresource_client` (Davies et al., 2015) así como RDKit (Bento et al., 2020).

Seguidamente, se eliminaron las entradas repetidas utilizando la notación InChI (*International Chemical Identifier*), especialmente diseñada para ser sensible a la conformación molecular (Cornell et al., 2024). Es importante tener en cuenta que las moléculas con la misma conexión atómica, pero con distintos estados de carga, producen distintas cadenas de InChI facilitando el proceso de depuración manual.

El próximo paso consistió en calcular los descriptores moleculares correspondientes a cada una de las entradas obtenidas de ChEMBL basadas tanto en sus propiedades estructurales como fisicoquímicas empleando la librería RDKit (Bento et al., 2020). Entre las que destacaremos el logaritmo del coeficiente de partición octanol/agua (logP) que indica la hidrofobicidad de una molécula, el peso molecular (MW), la cantidad de donantes y receptores de puentes de hidrógeno, y el área de superficie polar topográfica (TPSA) que indica la polaridad de la molécula. Además, se vincularon los datos numéricos de bioactividad con cada compuesto, en particular los valores de concentración inhibitoria media (IC50) cuando se encontraban disponibles (Nowotka et al., 2017).

Se logró establecer la similitud entre los compuestos, tanto en términos estructurales como en sus características fisicoquímicas, utilizaremos el criterio de distancia de Tanimoto (Bajusz et al., 2015). Para ello, se ha desarrollado e implementado un algoritmo basado en los K-vecinos más cercanos (K-NN) (Zhang, 2016). Es importante tener en cuenta que el algoritmo K-NN es una técnica de aprendizaje automático supervisado que se emplea tanto en tareas de clasificación como de regresión. Se categoriza como no paramétrico, dado que no presupone una distribución específica de los datos en su base, lo que le otorga versatilidad para ajustarse a diferentes estructuras de datos (Zhang, 2016). Su fundamento esencial se basa en el aprendizaje fundamentado en ejemplos.

El algoritmo K-NN llevará a cabo lo siguiente (detalles en Bajusz et al., 2015):

- Determina la distancia (es decir, la separación) entre el nuevo punto de datos y todos los puntos del conjunto de datos vinculados a la etapa de entrenamiento, empleando para ello

el coeficiente de Tanimoto.

- Identificará los K vecinos basándose en las distancias calculadas.

Es vital especificar algunos elementos teóricos de la metodología en relación con los valores de K, es decir, un valor bajo de K puede provocar que el modelo sea extremadamente susceptible al ruido presente en los datos, mientras que un valor alto puede suavizar excesivamente los límites de decisión, lo que podría resultar en un subajuste del modelo (conocido en inglés como *underfitting*) (Zhang, 2016). Frecuentemente, la elección ideal de K se lleva a cabo a través de métodos de validación cruzada, pero en este trabajo inicial no se ha tomado en cuenta.

Finalmente, mencionar que hay otros indicadores para calcular la distancia, tales como la euclídea, la distancia a partir de Manhattan y la distancia de Minkowski (Ehsani y Drabløs, 2020). En este contexto, la selección de la métrica puede afectar el diseño de las regiones de decisión y el desempeño del algoritmo para un conjunto específico.

Resultados

El estudio se centró en la detección de moléculas bioactivas estudiadas para combatir la COVID-19, utilizando la librería `chembl_webresource_client` en Python para la descarga eficaz de datos desde la base de datos ChEMBL. Esta metodología de descarga a nivel local ofrece el beneficio de reducir la necesidad de la disponibilidad y rapidez de la conexión a internet durante el análisis. No obstante, un posible inconveniente reside en la obligación de administrar y renovar la base de datos local para garantizar el acceso a la información más reciente.

En un principio, se consiguió un grupo de 3060 moléculas identificadas como bioactivas frente a la COVID-19. En esta base de datos, cada molécula se caracteriza por un identificador único que comienza con la palabra “CHEMBL” y continúa con un número.

Para valorar la actividad inhibitoria de estas moléculas, se realizó el cálculo del valor pIC50 utilizando los datos de IC50 disponibles de acuerdo a datos experimentales (detalles en la sección de Metodología). Sin embargo, se descartaron 134 compuestos del estudio por carecer de dicha información. Esta exclusión, pese a ser imprescindible para asegurar la integridad del cálculo de la actividad inhibitoria, implica una pérdida de información que podría ser de relevancia. La elección de no especificar estos compuestos en la investigación se basa en su insuficiencia para el análisis subsiguiente, aunque podría ser útil señalar la cantidad de esta pérdida en relación al porcentaje del conjunto de datos inicial.

Luego, el estudio se concentró en determinar la actividad inhibitoria. Se estableció un criterio de exclusión para los compuestos de actividad moderada, caracterizados por un valor de pIC50 que oscila entre 5 y 6. El beneficio de esta elección reside en centrar el estudio en

moléculas con una actividad más marcada, lo que podría simplificar la detección de patrones y propiedades pertinentes para la creación de medicamentos más eficaces. No obstante, la eliminación de compuestos de actividad intermedia podría resultar en la pérdida de información relevante acerca de posibles puntos de partida para la optimización de medicamentos con una actividad más baja, pero potencialmente con un perfil farmacocinético más favorable o menor toxicidad.

Para garantizar la unicidad de las entidades químicas analizadas, se descartaron las moléculas duplicadas. Un ejemplo ilustrativo es el compuesto "CHEMBL4495583" que estaba presente más de cinco veces en la lista inicial de datos. Esta fase es crucial para prevenir sesgos en el análisis y asegurar que cada molécula aporte de manera independiente a los resultados. Todo ello redujo los datos a tan solo 2301 moléculas.

Para finalizar el proceso de curación mencionado en la sección previa, se concluyó el filtrado de los datos utilizando la notación InChi, con el objetivo de eliminar aquellas moléculas que poseen la misma estructura química sin importar su carga.

El motivo de eliminar todos esos duplicados es para prevenir que una misma sustancia química tenga un impacto desmedido en la elaboración de modelos predictivos dándole un factor de peso adicional con respecto a otras moléculas. Es fundamental mantener un conjunto de datos únicos de moléculas para conseguir resultados consistentes y representativos.

Así se logró un conjunto definitivo de 1904 compuestos exclusivos utilizados contra el coronavirus. Este procedimiento final garantiza la uniformidad química del conjunto de datos para el estudio de descripciones moleculares.

La Figura 1 muestra una representación tridimensional que evidencia la diversidad de valores obtenidos empleando algunos de los descriptores moleculares utilizados en el estudio. En sus ejes se grafican los valores obtenidos del peso molecular ($MolWt$), el coeficiente de partición octanol-agua (LogP) y el área de superficie polar topográfica (TPSA). Se empleó una escala de colores para ilustrar la cantidad de aceptores de puentes de hidrógeno ($NumHAcceptors$). Esta figura permite visualizar la inexistencia de una tendencia en sus propiedades fisicoquímicas.

La Figura 2 describe el resultado obtenido después de aplicar el algoritmo K-NN. Aunque los datos esenciales se encuentran en las matrices de distancia generadas por el uso de la métrica de Tanimoto, esta figura proporciona una representación visual más fácil de comprender en la distribución espacial de los compuestos. Se han identificado cuatro clúster, los cuales se han señalado a través de distintos colores para simplificar su identificación. Esta selección de K=4 clústeres se basó en una separación evidente en los resultados acorde en esta etapa exploratoria de estudio. El próximo paso sería determinar el número de clústeres, pero la misma sería para trabajos a futuro.

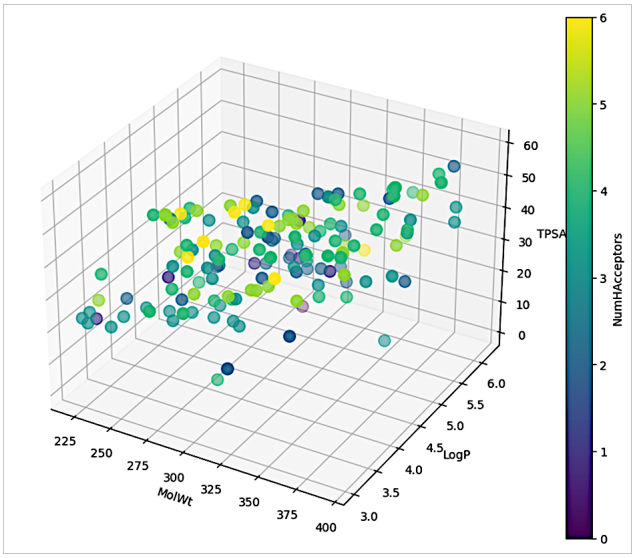

Figura 1: Representación tridimensional de los 1904 moléculas empleados en el trabajo dónde se representa el peso molecular (*MolWt*), el coeficiente de partición octanol-agua (LogP) y el área de superficie polar topográfica (TPSA), así como la cantidad de aceptores de puentes de hidrógeno (*NumHAcceptors*) en la escala de colores.

Fuente: Elaboración propia (2025).

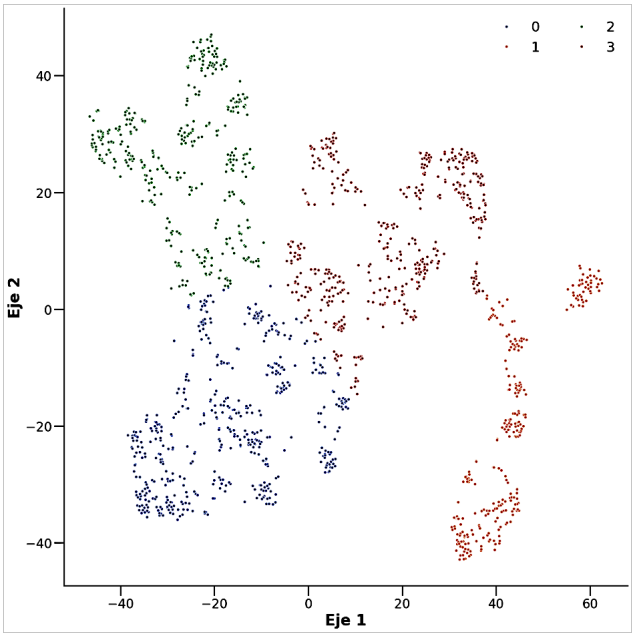

Figura 2: Representación en dos dimensiones de las distancias de Tanimoto, en la que se ubican cuatro grupos representados con cuatro colores distintos.

Fuente: Elaboración propia (2025).

Por último, la Figura 3 muestra una sección de las moléculas que forman el clúster identificado con el número 1, destacando la significativa similitud estructural entre las moléculas que lo conforman. Este hallazgo inicial corrobora la factibilidad de la metodología de trabajo. Sin embargo, es vital subrayar que este análisis no ha abordado la optimización del número de clúster debido a la ausencia de uniformidad en la distribución de los datos. Por esta razón, no se muestran las diversas tablas de moléculas que constituyen cada uno de dichos clústers (se presentará en un próximo estudio).

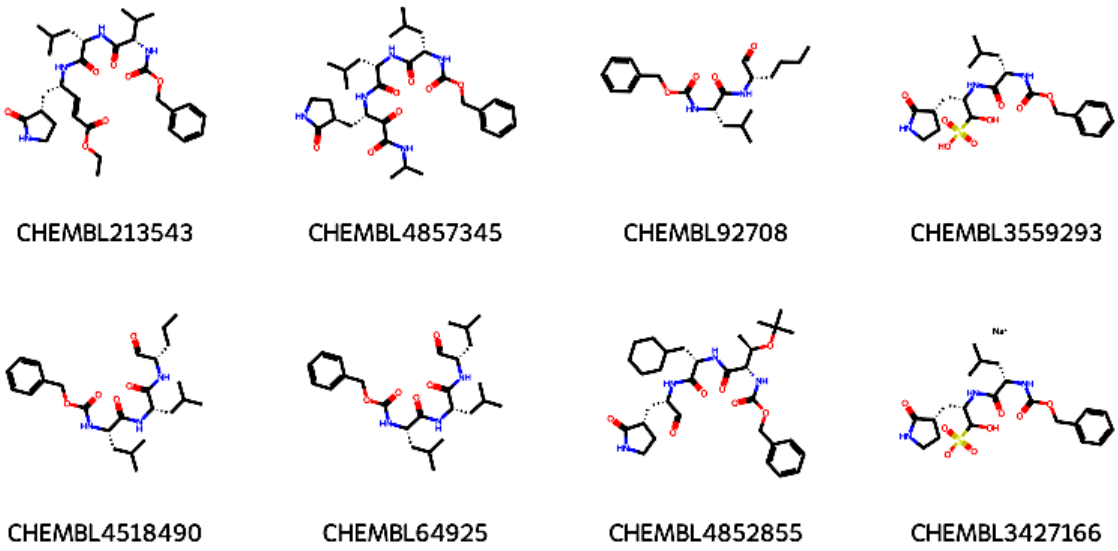

Figura 3: Representación en dos dimensiones de las distancias de Tanimoto, en la que se ubican cuatro grupos representados con cuatro colores distintos.
Fuente: Elaboración propia (2025).

Conclusiones

El presente trabajo identificó un conjunto de moléculas utilizadas en investigaciones experimentales contra la COVID-19 obtenidas en la base de datos ChEMBL. Una vez obtenida la información, fue necesario verificar que los datos no estén incompletos así como eliminar moléculas duplicadas. Dicho análisis identificó 1904 compuestos únicos basados en su actividad inhibitoria contra la COVID-19. Luego, se determinaron los descriptores moleculares de todas esas moléculas que fue la base para determinar la matriz de similaridad con ayuda del algoritmo K-NN. Se obtuvieron cuatro clústers. Dicho resultado quizás refleje la falta de optimización del número de clústeres ('K') por lo que se está trabajando actualmente en ello.

Referencias

Alie, M., Negesse, Y., Kindie, K., y Merawi, D. (2024). Machine learning algorithms for predicting COVID-19 mortality in Ethiopia. *BMC Public Health*, *24*(1), 1728. https://doi.org/10.1186/s12889-024-19196-0

Ávila, J., Mayer, M., y Quesada, V. (2021). La inteligencia artificial y sus aplicaciones en medicina II: importancia actual y aplicaciones prácticas [Artificial intelligence and its applications in medicine II: Current importance and practical applications]. *Atención Primaria*, *53*(1), 81-88. https://doi.org/10.1016/j.aprim.2020.04.014

Bajusz, D., Rácz, A., y Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, *7*, 20. https://doi.org/10.1186/s13321-015-0069-3

Bento, A., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., Bellis, L., De Veij, M., y Leach, A. (2020). An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics*, *12*(1), 51. https://doi.org/10.1186/s13321-020-00456-1

Brown, F. (1998). Chapter 35. Chemoinformatics: What is it and How does it Impact Drug Discovery. *Annual Reports in Medicinal Chemistry*, *33*, 375-384. https://doi.org/10.1016/S0065-7743(08)61100-8

Cornell, A., Kim, S., Cuadros, J., Bucholtz, E., Pence, H., Potenzone, R., y Belford, R. (2024). IUPAC International Chemical Identifier (InChI)-related education and training materials through InChI Open Education Resource (OER). *Chemistry Teacher International*, *6*(1), 77-91. https://doi.org/10.1515/cti-2023-0009

Cottrell, S., Hozumi, Y., y Wei, G. (2023). K-Nearest-Neighbors Induced Topological PCA for Single Cell RNA-Sequence Data Analysis. *ArXiv [Preprint]*. https://doi.org/10.48550/arXiv.2310.14521

Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., y Overington, J. (2015). ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, *43*(W1), W612-W620. https://doi.org/10.1093/nar/gkv352

De Clercq, E. (2020). Antiviral drugs in development for the treatment of COVID-19. *Biochemical Pharmacology*, *176*, 113747.

Dias, D., Viana, W., De Azevedo, W., y Andricopulo, A. (2021). Deep learning applied to QSAR for the identification of potential anti-SARS-CoV-2 compounds. *European Journal of Medicinal Chemistry*, *212*, 113175.

Ehsani, R., y Drabløs, F. (2020). Robust Distance Measures for kNN Classification of Cancer Data. *Cancer Informatics*, *19*, 1176935120965542. https://doi.org/10.1177/1176935120965542

Ghasemi, S., Saadati, S., Ebrahimiasl, S., y Fassihi, A. (2021). QSAR study of angiotensin-converting enzyme 2 (ACE2) inhibitors as potential therapeutic agents for COVID-19. *Journal of Molecular Liquids*, *323*, 114582.

Hernández, V., Blanquer, I., Aparicio, G., Isea, R., Chaves, J., Hernández, A., Mora, H., Fernández, M., Acero, A., Montes, E., y Mayo, R. (2007). Advances in the biomedical applications of the EELA Project. Stud Health Technol Inform. *Studies in Health Technology and Informatics*, *126*, 31-36. https://ebooks.iospress.nl/publication/10828

Hu, H., Stumpfe, D., y Bajorath, J. (2018). Rationalizing the Formation of Activity Cliffs in Different Compound Data Sets. *ACS Omega*, *3*(7), 7736-7744.6. https://doi.org/10.1021/acsomega.8b01188

Isea, R., Hoebeke, J., y Mayo, R. (2013). Designing a peptide-dendrimer for use as a synthetic vaccine against Plasmodium falciparum 3D7. *American Journal of Bioinformatics and Computational Biology*, *1*(1), 1.

Isea, R., Mayo, R., y Restrepo, S. (2016). Reverse Vaccinology in Plasmodium falciparum 3D7. *Journal of Immunological Techniques & Infectious Diseases*, *5*(3), 1. https://doi.org/10.4172/2329-9541.1000145

Khan, M., Shahid, M., Ali, S., Asif, H., y Ashraf, M. (2021). Quantitative structure-activity relationship (QSAR) studies on potential inhibitors of SARS-CoV-2 main protease. *Journal of Biomolecular Structure and Dynamics*, *39*(16), 5949-5963.

Liu, X., Zhang, R., Jin, M., Zhao, M., Li, J., Wei, S., y Liu, H. (2020). Identification of potential inhibitors against SARS-CoV-2 main protease by QSAR modeling and virtual screening. *European Journal of Pharmaceutical Sciences*, *152*, 105454.

Maggiora, G., Vogt, M., Stumpfe, D., y Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, *57*(8), 3186-3204. https://doi.org/10.1021/jm401411z

Nowotka, M., Gaulton, A., Mendez, D., Bento, A., Hersey, A., y Leach, A. (2017). Using ChEMBL web services for building applications and data processing workflows relevant to drug discovery. *Expert Opinion on Drug Discovery*, *12*(8), 757-767. https://doi.org/10.1080/17460441.2017.1339032

Ojha, S., Roy, K., y Mitra, I. (2021). Machine learning-based QSAR modeling for the prediction of SARS-CoV-2 main protease inhibitors. *Journal of Molecular Graphics and Modelling*, *107*, 107939.

Rabie, A., Mohamed, A., Abo-Elsoud, M., y Saleh, A. (2023). A new Covid-19 diagnosis strategy using a modified KNN classifier. *Neural Computing and Applications*, *35*(27), 1-25. https://doi.org/10.1007/s00521-023-08588-9

Rjoub, H., Adebayo, T., y Kirikkaleli, D. (2023). Blockchain technology-based FinTech banking sector involvement using adaptive neuro-fuzzy-based K-nearest neighbors algorithm. *Financial Innovation*, *9*(1), 65. https://doi.org/10.1186/s40854-023-00469-3

Sejuti, Z., y Islam, M. (2023). A hybrid CNN-KNN approach for identification of COVID-19 with 5-fold cross validation. *Sensors International*, *10*(4), 100229. https://doi.org/10.1186/s12889-024-19196-0

Tian, Y., Tong, J., Liu, Y., y Tian, Y. (2024). QSAR Study, Molecular Docking and Molecular Dynamic Simulation of Aurora Kinase Inhibitors Derived from Imidazo[4,5-b]pyridine Derivatives. *Molecules*, *29*(8), 1772. https://doi.org/10.3390/molecules29081772

Woodhouse, A., Hobbes, A., Mather, L., y Gibson, M. (1996). A comparison of morphine, pethidine and fentanyl in the postsurgical patient-controlled analgesia environment. *Pain*, *64*(1), 115-121. https://doi.org/10.1016/0304-3959(95)00082-8

Zdrazil, B. (2025). Fifteen years of ChEMBL and its role in cheminformatics and drug discovery. *Journal of Cheminformatics*, *17*(1), 32. https://doi.org/10.1186/s13321-025-00963-z

Zdrazil, B., Felix, E., Hunter, F., Manners, E., Blackshaw, J., Corbett, S., De Veij, M., Ioannidis, H., Lopez, D., Mosquera, J., Magarinos, M., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A., Adasme, M., Monecke, P., Landrum, G., y Leach, A. (2024). The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, *52*(D1), D1180-D1192. https://doi.org/10.1093/nar/gkad1004

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicinal*, *4*(11), 218. https://doi.org/10.21037/atm.2016.03.37