Evaluación de un modelo neuronal para la estimación de similaridad entre compuestos a partir de representaciones one-hot

Eugenio Borzone, Leandro Ezequiel Di Persia, Matias Gerard

Instituto de investigación en señales, sistemas e inteligencia computacional (sınc(i)), FICH–UNL/CONICET, Ciudad Universitaria UNL, (S3000) Santa Fe, Argentina.
eborzone@sinc.unl.edu.ar
www.sinc.unl.edu.ar

Resumen El estudio de la similaridad entre elementos de un conjunto es un problema común en áreas tan diversas como la bioinformática, la informática química y la medicina. En el caso de compuestos químicos, para calcularla, se utilizan descriptores moleculares, como es el caso de las fingerprints, que son representaciones vectoriales de cada compuesto. En este trabajo se estudian diferentes fingerprints ampliamente utilizadas en la literatura, para identificar la más adecuada para el cálculo de similaridad. Además, se busca determinar si es posible predecir ésta similaridad a través de un modelo neuronal. Para esto se caracterizan diferentes fingerprints por su desempeño en términos de predicción de similaridad, distribución de resultados en el intervalo [0,1] y frecuencia de uso en el ámbito científico. Posteriormente, se evalúa la capacidad de un Perceptrón Multicapa (MLP) para predecir la similaridad entre compuestos representados mediante vectores one-hot.
Los resultados muestran que las claves MACC proporcionan una buena distribución en los valores de similaridad. El MLP es capaz de inferir con un bajo error (aproximadamente 10 % en términos absolutos) la similaridad entre compuestos empleando una representación one-hot.

Keywords: Perceptrón multi capa · Fingerprints · Similaridad molecular · Indice de Tanimoto

1. Introducción

La similaridad molecular [19,12] es uno de los conceptos más explotados de la informática química y áreas relacionadas como la bioinformática, la química medicinal y el descubrimiento de fármacos. Se aplica en múltiples tareas, tales como la predicción de propiedades [2], el cribado virtual [12] que es un conjunto de técnicas que permiten la selección de potenciales candidatos de fármacos, el análisis de diversidad molecular [9], la búsqueda por similaridades [19] como es el caso de la búsqueda y diseño de vías metabólicas a través del uso de la estructura molecular de los compuestos para guiar el proceso de búsqueda [11,14] maximizando la similaridad entre el producto de cada reacción de la secuencia y el producto final de interés.

La forma más intuitiva de calcular similaridad entre compuestos es identificar el solapamiento entre los grafos que representan a un par de moléculas [10], donde los nodos representan átomos y los arcos enlaces. Dado que estos métodos son muy lentos, habitualmente para el cálculo de similaridad se utilizan descriptores moleculares, que son características extraídas a partir de las estructuras de los compuestos. Existen descriptores 1D como el peso molecular, el logaritmo del coeficiente de reparto o la superficie polar [20]. Otros descriptores son bidimensionales, como los índices topológicos [4] y las fingerprints moleculares [19], que codifican en un vector las representaciones estructurales 2D de las moléculas. También existen descriptores en tres dimensiones como la forma molecular en el espacio, donde se especifican las coordenadas de cada átomo.

Existen diferentes enfoques para la construcción de fingerprints moleculares, como el de tipo estructural o a través de funciones de hash. En todos los casos es necesario contar con la estructura de ambos compuestos para calcular la similaridad. Una vez obtenida la fingerprint que caracteriza cada compuesto, se puede calcular un índice de similaridad, por ejemplo el de Tanimoto [1]. Sin embargo, cuando la estructura de alguno se desconoce resulta imposible calcular la similaridad, es por esto que resulta de interés explorar formas de predecir la similaridad de compuestos sin necesidad de calcular las fingerprints.

Las redes neuronales han proporcionado importantes desarrollos en las últimas décadas. Han demostrado ser muy útiles para aprender relaciones complejas en los datos. En particular, en el campo de la química computacional se han empleado para una amplia gama de tareas, como Mol2vec [8] que genera representaciones moleculares a partir de las estructuras de compuestos. Es un enfoque de aprendizaje automático no supervisado para aprender representaciones vectoriales de subestructuras moleculares. La información producida por la conjunción entre la química y el aprendizaje computacional, a través de los análisis basados en datos y las predicciones de las redes neuronales, permiten impulsar la capacidad de comprender la complejidad de los datos químicos, racionalizar y diseñar experimentos [5]. Se pueden citar como ejemplos de este uso, optimizar o predecir las relaciones estructura-propiedad [3], o producir moléculas estables a partir de las propiedades deseadas [16].

Este estudio tiene como objetivo determinar si es posible predecir la similaridad entre compuestos a partir de redes neuronales artificiales, sin que sea necesario calcular fingerprints para ellos. La organización del trabajo es la siguiente: dentro de la Sección 2 se detallan los materiales y métodos utilizados: la subsección 2.1 se describen las fingerprints a estudiar. La subsección 2.2 presenta el modelo y sus componentes. En la subsección 2.3 se describen los datos empleados y cómo se construye el dataset. En la Sección 3 se presentan los resultados donde se analiza el desempeño del modelo. Finalmente, en la Sección 4 se presentan las conclusiones del trabajo.

2. Materiales y métodos

En esta Sección se describen los conceptos teóricos y los materiales empleados en la realización de este trabajo. Primero se presentan los métodos para cálculo de fingerprint que son considerados durante la etapa experimental. Luego se describe el modelo neuronal utilizado para estimar similaridad a partir de las representaciones one-hot de cada compuesto. Finalmente, se detalla la construcción del dataset utilizado, explicitando las fuentes de datos y metodología seguida.

2.1. Descripción de fingerprints

En este trabajo se seleccionaron 6 métodos de construcción de fingerprints, de las cuales 2 corresponden a la familia de tipo estructural, y 4 corresponden a la familia basadas en códigos de hash. La elección se realizó de acuerdo a la frecuencia de uso en la literatura.

En el caso de las fingerprints estructurales diferentes aspectos de los compuestos llamadas claves, son codificados en un vector binario: cada bit corresponde a una subestructura predefinida (por ejemplo un anillo aromático o un grupo hidroxilo). Como se observa en la Figura 1a si la molécula tiene presente en su composición la estructura, el bit de la posición correspondiente será seteado en 1, en otro caso 0. De este tipo se consideraron para el análisis las Molecular ACCess System (MACC) keys [6] de versión publica con 166 claves y PubChem fingerprint (PCFP) [1] de 880 claves.

Las fingerprint basadas en códigos de hash no requieren la definición y listado previo de claves. En cambio, se generan enumerando subestructuras de un cierto tamaño definido y aplicándoles una función de hash de forma iterativa. Dentro de las más populares se encuentran las Extended-connectivity fingerprints (ECFP) [15] representadas en la Figura 1b, son figerprints de estructura radial, que se definen como el conjunto de identificadores resultantes de un procedimiento hash. Éste mapea las características subestructurales centradas en los átomos (no hidrógeno) y registradas en múltiples capas circulares hasta un diámetro determinado de forma sistemática. También se analizará un caso particular de las ECFP, llamadas MORGAN con capas circulares de radio 3. Otras fingerprints a tratar son las MinHash fingerprint (MHFP) [13] que codifican subestructuras detalladas utilizando el principio de conectividad extendida de ECFP de una manera fundamentalmente diferente, aumentando el rendimiento de las búsquedas de vecinos más cercanos exactos en los estudios de evaluación comparativa y permitiendo la aplicación de hashing sensible a la localidad y las SMILES Extended Connectifity Fingerprint (SECFP) que es una variante del esquema de hashing de subestructura circular basado en SMILES de MHFP, doblado por la misma operación de módulo n que utiliza ECFP.

[1] `https://web.cse.ohio-state.edu/~zhang.10631/bak/drugreposition/list_fingerprints.pdf`

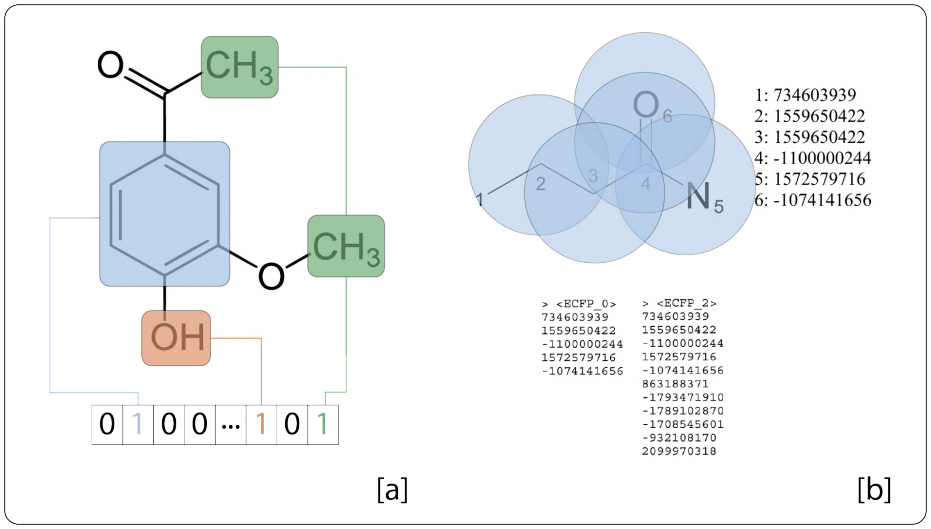

Figura 1: Esquema de construcción de fingerprints. [a] Fingerprint de tipo estructural. [b] Fingerprints basadas en código de hash (ECFP).

2.2. Modelo de red neuronal

Para realizar la predicción de la similaridad entre compuestos se empleó un modelo neuronal basado en la arquitectura básica del perceptrón multicapa [7]: capa de entrada, capas ocultas y capa de salida. En este modelo, la función de activación empleada en las capas ocultas es la ReLU, definida como:

$$ReLU(x) = max(0, x), \tag{1}$$

donde x es la salida lineal al final de cada capa. En la capa de salida se emplea la función sigmoidea, definida como y capa de salida con una función de activación sigmoidea

$$Sigmoidea(x) = \frac{1}{1 + e^{-x}}, \tag{2}$$

donde x es la salida lineal de la capa. Se emplea esta función dado que permite acotar el rango de salida al intervalo $[0, 1]$. Este modelo toma como entrada las representaciones de los compuestos entre los que se desea estimar la similaridad, y devuelve como salida el valor de la misma. La Figura 2 muestra un esquema del modelo. Dado que cada compuesto se encuentra representado mediante un vector one-hot de N elementos (la posición en la que se encuentra el 1 indica el compuesto del que se trata) y se busca comparar la similaridad entre pares de compuestos, el vector de entrada a la red posee $2N$ características. Puede observarse así que la entrada a la red se compone de $2N$ neuronas, donde cada una procesa individualmente cada una de las características de este vector. Se utilizó Dropout como regularizador [17] con una probabilidad p única para todas

las capas ocultas. Para representar la entrada (es decir, los dos compuestos a los que se le quiere medir la similaridad), se utilizó codificación one-hot como se explicará en la próxima sección.

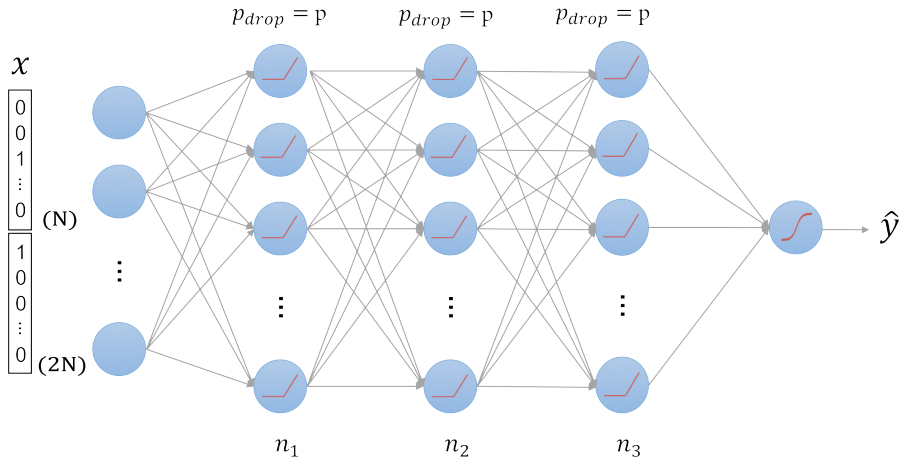

Figura 2: Arquitectura neuronal. El vector de características x contiene las representaciones one-hot de los compuestos a comparar $[0, N]$ y $[N + 1, 2N]$ respectivamente.

2.3. Construcción del dataset

Se utilizó la totalidad de la vía metabólica de la Glucólisis [2]. Esta vía está compuesta por 52 reacciones, que involucran a un total de 60 compuestos. De estos sólo 47 tienen estructura conocida, lo que se necesita para calcular las fingerprints, por lo tanto, ese será el conjunto de compuestos a usar en este trabajo. Los datos fueron extraídos de la base de datos KEGG[3] (versión 95.2). Para cada compuesto se descargó la estructura molecular en formato SMILES [18], cuando estaba disponible de la base de datos de PubChem[4]. Con base en estas estructuras, se generaron las diferentes fingerprints empleando la librería RDKit[5], se eligió esta representación debido a su sencillez para representar la información de los compuestos. Luego se calculó la similaridad entre pares de compuestos empleando el índice de Tanimoto, tal y como se describe en la literatura [1],

[2] `https://www.genome.jp/pathway/map00010`
[3] `https://www.genome.jp/kegg/`
[4] `https://pubchem.ncbi.nlm.nih.gov/`
[5] `https://www.rdkit.org`

definido de la siguiente forma:

$$T(\mathbf{p}_i, \mathbf{p}_j) = \frac{\sum_k (\mathbf{p}_i^k \wedge \mathbf{p}_j^k)}{\sum_k (\mathbf{p}_i^k \vee \mathbf{p}_j^k)} \tag{3}$$

donde $\wedge$ y $\vee$ son los operadores binarios *and* y *or*, respectivamente, y $\mathbf{p}_i$ y $\mathbf{p}_j$ son representaciones binarias de las estructuras de los compuestos i y j. La sumatoria se realiza sobre las k características de cada compuesto. El coeficiente de Tanimoto toma valores en el intervalo $[0, 1]$ y calcula la proporción de características compartidas entre ambas estructuras.

En total se definieron 1081 patrones que resultan de la combinatoria de a pares de los 47 compuestos con estructura conocida, a los cuales se los codificó con vectores one-hot: se enumeran los 47 compuestos de la vía metabólica de la Glucólisis que poseen estructura conocida, y se codifica el compuesto con un vector de ceros de dimensión 47 con un uno en la posición del compuesto correspondiente y se los concatena de a pares para formar los datos de entrada x como se explico en la sección anterior.

3. Resultados

En esta sección se presentan los resultados obtenidos. En primer lugar, se realiza la selección de la fingerprint mas apropiada según los criterios ya detallados. Una vez definido este aspecto, se calcula la salida deseada para la similaridad de todos los pares de compuestos. Luego se presenta la optimización de los hiperparámetros del modelo y construcción del modelo definitivo.

3.1. Selección de fingerprint

Para determinar cuál de las 6 fingerprints presentadas en la Sección (2.1) resulta más adecuada para generar los valores a predecir con el modelo neuronal, se realizó un análisis basado en 3 criterios: predicción de similaridad, separabilidad de resultados (resolución) y frecuencia de uso en el ámbito científico. Con el objetivo de determinar cuán bien se distribuyen los valores de similaridad a lo largo del rango de posibles valores (intervalo [0, 1]), se construyeron los histogramas que se muestran en la Figura 3 a-f. Cada gráfico presenta la distribución del índice de Tanimoto para cada una de las fingerprints estudiadas. Como puede apreciarse en las Figuras 3a, 3b, 3c, 3e y 3f las distribuciones están concentrados valores bajos. Se observa que las fingerprints MACC (Figura 3d) presentan una distribución de coeficientes de Tanimoto mas uniforme a lo largo del intervalo, esto es deseable para que el modelo entrenado produzca predicciones con similar calidad en todo el rango de variación del índice.

Por otro lado, se realizó un nuevo estudio para profundizar en la comparación de las fingerprints. Para esto se seleccionó de un subconjunto de compuestos de la Glucólisis, sobre el que se realizó un análisis subjetivo que determinó la bondad de cada similaridad calculada. Se construyó una matriz de similaridad entre los

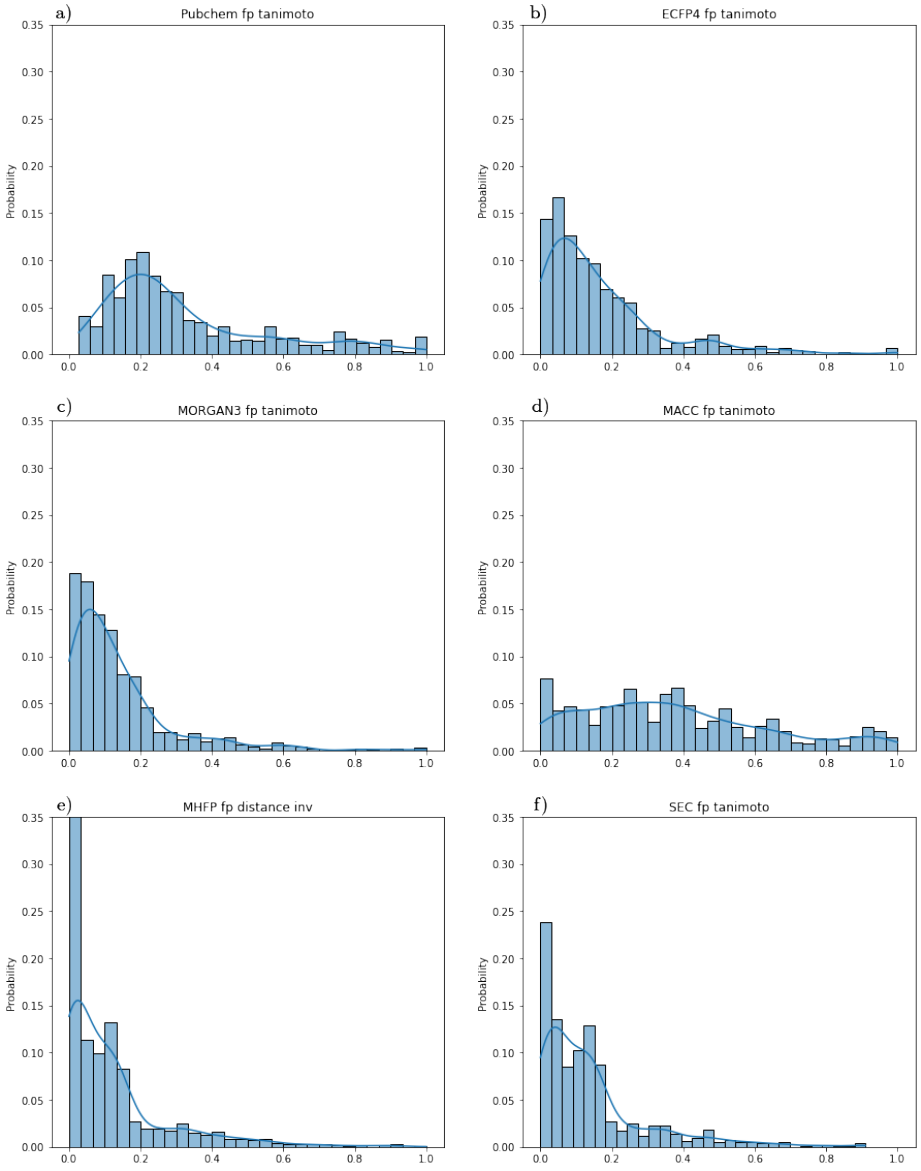

Figura 3: Histograma de la distribución de los valores de similaridad calculado con el índice de Tanimoto para las diferentes fingerprints consideradas sobre el camino metabólico de la Glucólisis: a) Pubchem, b) ECFP de radio 4, c) MORGAN (radio 3), d) MACC, e) MHFP, f) SEC

compuestos cuyas KEGG ids son: C00103, C00668, C01172, C05345, C05378, C00118, C00197, C00074, C00068, C00024, C05125, C00022, C00033, C00084, C00469. Los mismos fueron seleccionados por tener estructuras de diferentes complejidades, y con distinto grado de parecido al compararlas de a pares.

En la Figura 4 se presenta la matriz de similaridad (matriz triangular inferior) para un subconjunto de los compuestos seleccionados para la prueba. Dentro de cada cuadro se muestra la similaridad calculada con las fingerprints PCFP, ECFP4 (ECFP de radio 4), MOR3 (MORGAN de radio 3), MACC, MHFP la SEC respectivamente. Como resulta esperable, todas las medidas devuelven similaridad 1 cuando la estructura de un compuesto se compara consigo mismo. Además se ve claramente como las fingerprints ECFP, MORGAN, MHFP y SEC, tienden a otorgar valores bajos de similaridad, incluso cuando las estructuras presentes son muy parecidas, como es el caso de C00668 y C00103. Esto es consistente con el estudio anterior. Hay que señalar que las MACC keys y las PCFP fueron las que mejor desempeño tuvieron al generar los índices. Se observó que las fingerprint de Pubchem tienden a otorgar valores altos de similaridad al comparar compuestos pequeños.

Dadas las comparativas realizadas, se determinó la utilización de las MACC keys como fingerprint para para construir el dataset para entrenar el modelo neuronal.

3.2. Búsqueda de hiperparámetros y resultados de la red neuronal

En esta Sección se evalúa la capacidad de un modelo neuronal basado en MLP para inferir la similaridad. Para calcular los valores targets y se utilizó el coeficiente de Tanimoto calculado a partir de las las MACC keys de los compuestos. Se emplearon como valores de entrada x representaciones one-hot de los mismos, como ya se explicó. Para esto, se realizó el entrenamiento del modelo neuronal utilizando cross-validation con 5 folds con un porcentaje de validación del 10 %. El modelo fue implementado utilizando Pytorch 1.9. El entrenamiento se realizó empleando el optimizador Adam y el error cuadrático medio (MSE) como función de costo. Se consideraron dos arquitecturas diferentes: una con tres capas ocultas y otra con cuatro, todas ellas con funciones de activación ReLU. Para la capa de salida se utilizó una sigmoidea. Los experimentos realizados se llevaron a cabo empleando 6000 como número máximo de épocas de entrenamiento y 1000 épocas de paciencia para la detención temprana, evitando que la red continúe el entrenamiento cuando no hay mejora en el error de validación. En los casos donde se supera el umbral de paciencia, el algoritmo es detenido y se recupera el modelo con el menor error de validación hasta el momento.

Dado que existen múltiples hiperparámetros que pueden afectar el desempeño del modelo, como primer paso se realizó una exploración de los mismos. En particular, se realizó una búsqueda en grilla detallada en la Tabla 1. Esto implica que se entrenaron 16384 modelos, uno para cada combinacion de los hiperparámetros descriptos

La Tabla 2 presenta los 5 mejores conjuntos de hiperparámetros obtenidos de esta exploración. El modelo fue entrenado con una función de costo del error

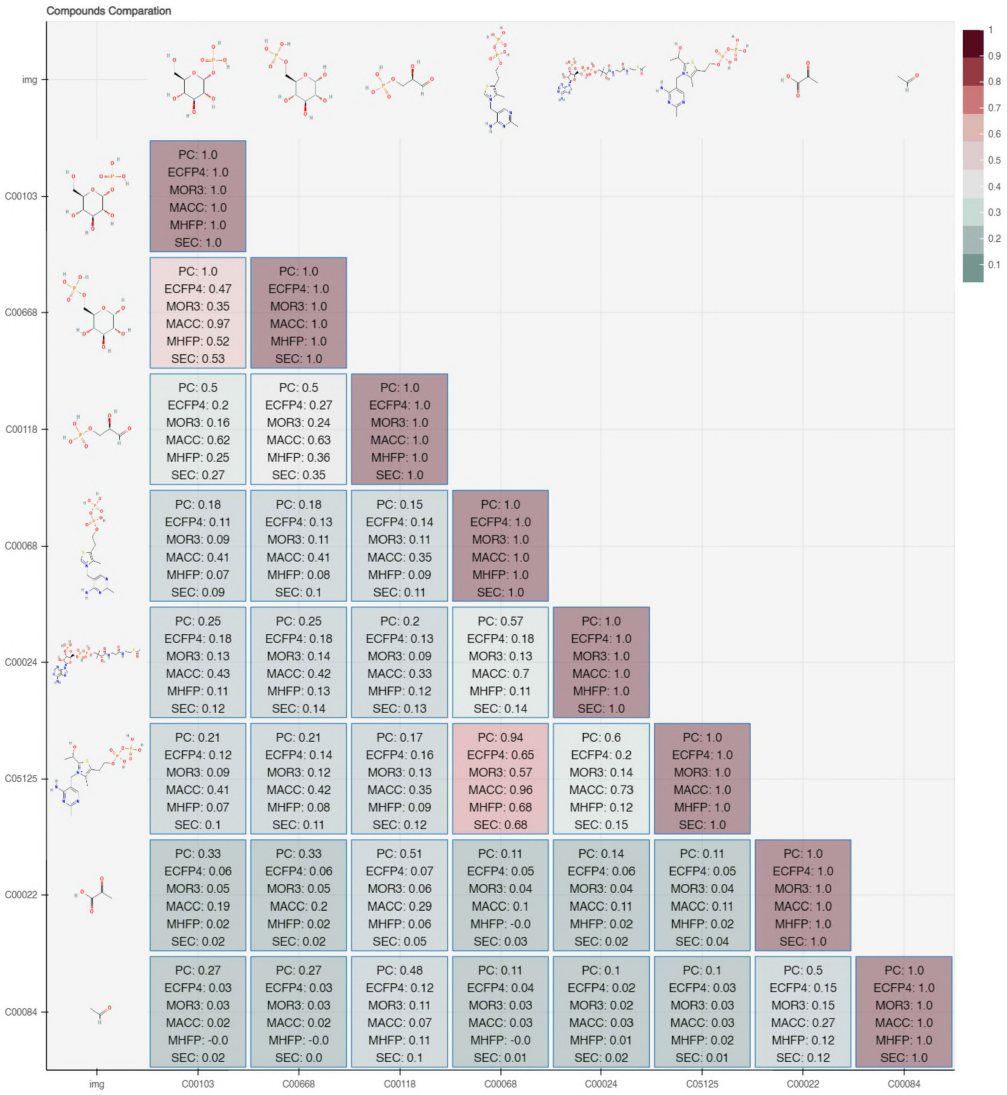

Figura 4: Parte de la comparativa de fingerprints: en verde los compuestos que poseen similaridad promedio baja, en rojo alta.

Hiperparametro	Rango	Paso
Tasa de aprendizaje	$5 \cdot 10^{-4}$, $3 \cdot 10^{-4}$	
Tamaño de batch	10, 100	
P dropout	0.3, 0.5	
Tamaño capa 2	16 - 256	16
Tamaño capa 3	16 - 256	16
Tamaño capa 4	16 - 256	16

Tabla 1: Grilla de búsqueda para el modelo base

cuadrático medio, sin embargo, se presenta la raíz cuadrada del mismo porque ésta se encuentra en el mismo orden de magnitud del índice a predecir. Como puede apreciarse se obtuvo el mejor conjunto de hiperparámetros: tasa de aprendizaje de $5 \cdot 10^{-4}$, tamaño de batch de **10**, la probabilidad de dropout de **0.3** y un número de neuronas de [**94,256,16,16,1**] en cada capa respectivamente. En cuanto al error, la raíz cuadrada del error cuadrático medio $RMSE$ de validación cruzada obtenido es de **0.1019** y un coeficiente de determinación R^2 de test de **0.9**, determinado a partir de los valores reales y los predichos por la red. Este resultado se condice con lo observado al comparar los valores predichos vs los reales. En la Figura 5 se representan los índices de Tanimoto targets (y) comparados con los predichos por la red (yhat) para uno de los fold, con resultados similares para los demás folds. Adicionalmente, se agregaron líneas de tendencia de cada set. En azul se muestran los valores de entrenamiento, en naranja los de test y en verde los de validación. Lo primero que se observa es que la mayoría de los valores de similaridad se ubican sobre la diagonal, lo que indica una buena predicción de los valores. A su vez, si se observan las rectas de regresión para cada partición se aprecia que están muy cercanas a la recta de 45°. Por último son presentadas las densidades de distribución del índice, que son similares tanto para los valores reales como los predichos.

RMSE	Batch	Lr	Capa 2	Capa 3	Capa 4
0.101974	**10**	**0.0005**	**256**	**16**	**16**
0.102195	10	0.0005	256	16	16
0.102852	10	0.0005	256	256	16
0.102863	10	0.0005	128	16	16
0.103010	10	0.0003	256	16	16

Tabla 2: Mejores resultados de la búsqueda de grilla. Batch se refiere al tamaño del batch, Lr es la taza de aprendizaje y Capa se refiere al numero de neuronas de esa capa

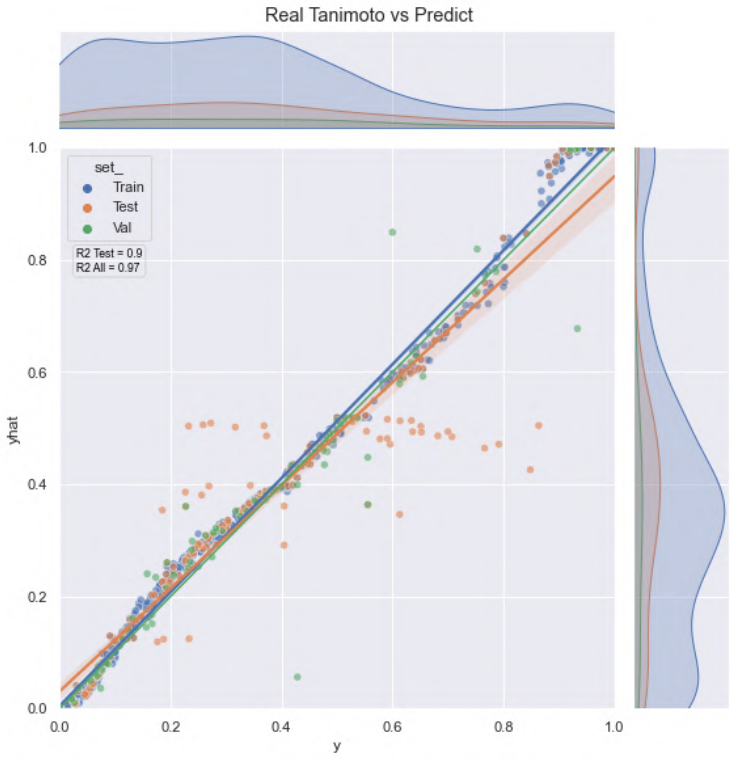

Figura 5: Indice de Tanimoto real vs Indice de Tanimoto predicho por la red

4. Conclusiones

En este trabajo se abordó el problema de cálculo de similaridad entre compuestos. Se compararon métodos para generar fingerprints y se determinó que las más adecuadas para la creación de un dataset son las MACC keys, tanto por su eficacia a la hora del cálculo de la similaridad, como por la distribución pareja del índice de Tanimoto en todo el intervalo. Con base en estos resultados se evaluó la capacidad de un MLP para predecir la similaridad entre compuestos, empleando como entrada representaciones one-hot de los mismos. Los resultados muestran que los modelos neuronales son prometedores para predecir similaridad entre compuestos, aún utilizando modelos sencillos y con poca información para el entrenamiento. Cabe destacar que debido a la naturaleza de la información de entrada, es imposible establecer relaciones entre los compuestos que no poseen estructura definida con los que si la tienen. Con base en estos resultados, el siguiente paso será explorar diferentes estrategias para generar embeddings que permitan representar en espacios n-dimensionales los compuestos a partir del procesamiento de la información disponible, como el grafo de reacciones de un camino metabólico y las propiedades fisicoquímicas propias de cada uno.

Referencias

1. Bajusz, D., Rácz, A., Héberger, K.: Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal of Cheminformatics **7**(1), 20 (2015). https://doi.org/10.1186/s13321-015-0069-3
2. Brown, R.D., Martin, Y.C.: Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. J. Chem. Inf. Comput. Sci. **36**, 572–584 (1996)
3. Chandrasekaran, A., Kamal, D., Batra, R., Kim, C., Chen, L., Ramprasad, R.: Solving the electronic structure problem with machine learning. npj Computational Materials **5**(1), 1–7 (Feb 2019). https://doi.org/10.1038/s41524-019-0162-7, number: 1 Publisher: Nature Publishing Group
4. Chaques, C.S.: Influencia del empleo de diferentes indices topologicos en la prediccion de distintas propiedades fisico-quimicas y farmacologicas de un grupo de farmacos antihistaminicos y antiinflamatorios. http://purl.org/dc/dcmitype/Text, Universitat de València (1988)
5. Cova, T.F.G.G., Pais, A.A.C.C.: Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns. Frontiers in Chemistry **7** (2019)
6. Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G.: Reoptimization of mdl keys for use in drug discovery. Journal of Chemical Information and Computer Sciences **42**, 1273–1280 (11 2002). https://doi.org/10.1021/ci010132r
7. Haykin, S.: Neural networks: a comprehensive foundation. Prentice Hall PTR (1994)
8. Jaeger, S., Fulle, S., Turk, S.: Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. Journal of Chemical Information and Modeling **58**(1), 27–35 (Jan 2018). https://doi.org/10.1021/acs.jcim.7b00616, `https://doi.org/10.1021/acs.jcim.7b00616`, publisher: American Chemical Society
9. Koutsoukas, A., Paricharak, S., Galloway, W.R., Spring, D.R., Ijzerman, A.P., Glen, R.C., Marcus, D., Bender, A.: How diverse are diversity assessment methods? a comparative analysis and benchmarking of molecular descriptor space. Journal of Chemical Information and Modeling **54**, 230–242 (1 2014). https://doi.org/10.1021/ci400469u
10. Kumar, A., Zhang, K.Y.: Advances in the development of shape similarity methods and their application in drug discovery. Frontiers in Chemistry **6** (2018). https://doi.org/10.3389/fchem.2018.00315
11. McShan, D.C., Rao, S., Shah, I.: PathMiner: predicting metabolic pathways by heuristic search. Bioinformatics **19**(13), 1692–1698 (1 Sep 2003)
12. Muegge, I., Mukherjee, P.: An overview of molecular fingerprint similarity search in virtual screening. Expert Opinion on Drug Discovery **11**, 137–148 (2 2016). https://doi.org/10.1517/17460441.2016.1117070
13. Probst, D., Reymond, J.L.: A probabilistic molecular fingerprint for big data settings. Journal of Cheminformatics **10** (12 2018). https://doi.org/10.1186/s13321-018-0321-8
14. Rahman, S.A., Advani, P., Schunk, R., Schrader, R., Schomburg, D.: Metabolic pathway analysis web service (pathway hunter tool at CUBIC). Bioinformatics **21**(7), 1189–1193 (2005)
15. Rogers, D., Hahn, M.: Extended-Connectivity Fingerprints. J. Chem. Inf. and Model. **50**(5), 742–754 (2010). https://doi.org/10.1021/ci100050t
16. Schneider, G.: Generative models for artificially-intelligent molecular design. Molecular Informatics **37**(1-2), 1880131 (2018). https://doi.org/https://doi.org/10.1002/minf.201880131

17. Wager, S., Wang, S., Liang, P.S.: Dropout training as adaptive regularization. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013)
18. Weininger, D.: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules (May 2002). https://doi.org/10.1021/ci00057a005, archive Location: world Publisher: American Chemical Society
19. Willett, P., Barnard, J.M., Downs, G.M.: Chemical similarity searching. Journal of Chemical Information and Computer Sciences **38**, 983–996 (1998). https://doi.org/10.1021/ci9800211
20. Xue, L., Bajorath, J.: Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening. Combinatorial Chemistry & High Throughput Screening **3**(5), 363–372 (Oct 2000). https://doi.org/10.2174/1386207003331454