XARXES NEURONALS BASADES EN GRAFS PER A PROCESSOS DE SCREENING MOLECULAR

Guiu Oms Font

Resum– El treball es planteja en el context de cerca de molècules candidates mitjançant metodologies de deep learning, per tal de trobar-ne alguna amb certa acció biològica requerida. Per això, s'han explorat xarxes neuronals basades en grafs (GNN) que prediguin les energies lliures d'interacció entre una molècula i una proteïna en impactar sobre la base de l'estructura química d'ambdues. S'han investigat diferents estratègies que optimitzin la precisió per obtenir el rànquing de candidats el més fiable possible. En particular, el problema ha estat aplicat a la recerca de medicaments (principis actius) que estan disponibles al mercat i que interaccionen amb alguna de les proteïnes seleccionades del SARS-CoV-2 com un exercici de drug repurposing aplicat a la Covid-19. La idea final és poder proveir al model una proteïna i un conjunt de molècules (medicaments), per seleccionar les que possiblement interaccionaran en funció de les prediccions fetes.

Paraules clau– GNN, intel·ligència artificial, aprenentatge profund, xarxes neuronals, drug repurposing, screening molecular, Covid-19, interacció molecular, binding pocket, binding affinity.

Abstract– This work has been developed in the context of the research of candidate molecules using deep learning methodologies, in order to find some with some biologically required action. Therefore, Graph-based Neural Networks (GNN) have been explored in order to predict the free energies of an interaction between a molecule and a protein when they impact on the basis of the chemical structure of both. Different strategies have been studied to optimize the accuracy in order to obtain the ranking of candidates as reliable as possible. In particular, the problem has been applied to the search of drugs (active principle) that are available on the market and interact with some of the selected proteins of SARS-CoV-2 as a drug repurposing exercise applied to Covid- 19. The final aim is to provide the model with a protein and a set of molecules (drugs), to select those that will possibly interact based on the predictions made.

Keywords– GNN, artificial intelligence, deep learning, neural networks, drug repurposing, molecular screening, Covid-19, molecular interaction, binding pocket, binding affinity.

1 INTRODUCCIÓ - CONTEXT DEL TREBALL

La creació de nous medicaments és un procés complex, llarg i costós, que moltes vegades comporta una gran quantitat de proves, estudis, requisits legals i fracassos que deriven en el desistiment de les investigacions. Tot i això, també és cert que hi ha una gran quantitat de principis actius aprovats al mercat que tenen un ús específic, però que podrien emprar-se per a altres finalitats, sense haver de passar totes les corresponents proves, inversions i dedicació de recursos, ja que aquestes haurien estat passades en recerques anteriors.

L'acció de reaprofitar aquests medicaments per a finalitats que no són les seves es coneix com a drug repurposing i és el principal objectiu d'aquest treball, on s'ha intentat, a través de les tecnologies més punteres del sector de la intel·ligència artificial, com podrien ser les GNN, estudiar d'una manera ràpida i eficaç les interaccions entre la proteïna corresponent a una malaltia i la totalitat de principis actius del mercat, obtenint una llista de candidats a interaccionar (anul·lant l'efecte de la proteïna) més reduïda que la inicial, per tal de procedir a dur a terme les proves

• E-mail de contacte: guiuomsfont@gmail.com
• Menció realitzada: Computació
• Tutor: Josep Lladós Canet (Ciències de la computació)
• Tutor a l'empresa: Guillermo Rodriguez Lázaro
• Curs 2021/22

experimentals als laboratoris i comprovar si alguna de les molècules estudiades, i que ja forma part d'un medicament, podria ser útil per a una altra finalitat, agilitzant i embaratint el procés de solució d'una malaltia al públic general.
Aquest TFG ha estat realitzat a través de l'empresa Grupo AIA, una empresa d'intel·ligència artificial que treballa en diversos temes del món científic i tecnològic actual, i que, entre aquests, es troba la bioquímica. Per això, des del departament de I+D, s'ha impulsat la realització d'una investigació en aquest tema rellevant per poder obrir nous horitzons i aportar valor empresarial a l'organització.
La resta del document està estructurat de la següent manera: a la secció 2 es descriuen els objectius, a la 3 el marc teòric per contextualitzar el treball i l'estat de l'art, a la 4 la metodologia que s'ha emprat, a la 5 el desenvolupament de les tasques, a la 6 els resultats obtinguts i, finalment, a la 7 les conclusions extretes.

2 OBJECTIUS

Els principals objectius d'aquest treball han estat:

- Realització d'un estudi profund de l'estat de l'art en aquest camp per a poder aplicar els coneixements obtinguts al projecte.
- Obtenció d'una o diverses bases de dades útils i confiables per a l'entrenament, validació i test del model.
- Disseny d'un model funcional optimitzant al màxim les mètriques de qualitat dels resultats.
- Millora del model aconseguit a través de diversos mètodes com la modificació d'arquitectura o d'entrenament.
- Aplicació a un cas real com el SARS-CoV-2 i anàlisis dels resultats.
- Creació d'un prototip final útil per a la seva aplicació en l'àmbit empresarial.

3 MARC TEÒRIC

Aquesta secció segueix un fil d'estudi similar al dut a terme al llarg de la investigació, on primerament, s'ha fet una introducció a la part bioquímica i informàtica, seguida del plantejament del treball i culminada en la recerca de l'estat de l'art, comprensible gràcies a tots els passos realitzats anteriorment.

3.1 INTRODUCCIÓ A LES GNN

Al llarg dels últims anys han agafat rellevància al món de la computació unes xarxes neuronals que es basen en dades estructurades en forma de grafs, i que coneixem com a GNN o Graph Neural Networks. Aquestes obren la porta a un nou món, ja que infinitat de coses tenen la capacitat de ser representades per aquesta estructura, i, com a conseqüència, les tecnologies d'aprenentatge basades en xarxes neuronals (deep learning) poden ser aplicades a elles.
Els grafs o $G = (V, A)$ són un conjunt de nodes o vèrtexs units entre ells pel que coneixem com a arestes. Durant el treball, s'utilitzaran aquestes terminologies de la següent manera:

- Vèrtex, Node o V: unitat fonamental del graf, representada a la figura 1 amb cercles i un número o valor a dins. $V = \{v_i\}$ és el conjunt de vèrtexs o nodes.
- Aresta o A: relació entre dos vèrtexs. Pot ser dirigida o no en funció de si la relació entre els nodes és simètrica o unidireccional. $A = \{a_i\}$ és el conjunt d'arestes.

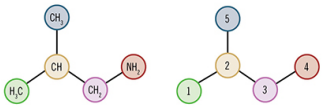

Fig. 1: Exemple d'un graf

A més de la informació topològica, aquesta estructura pot aportar més informació sobre propietats del que representa a partir de l'ordre (nombre de vèrtexs), la mida (nombre d'arestes), tipus de connectivitat, etc. Per altra banda, un graf etiquetat o d'atributs ofereix dades com el context, les relacions entre les diferents parts o informació dels components, entre d'altres, en forma de vectors de característiques associats als vèrtexs i a les arestes.
Per representar-lo s'acostuma a utilitzar una matriu d'adjacència, que és una manera de representar un graf a través d'una taula o matriu quadrada on cada una de les posicions indica si hi ha connexió o no entre Vx i Vy. Per tant, si es té un graf sense dirigir, aquesta matriu serà simètrica.

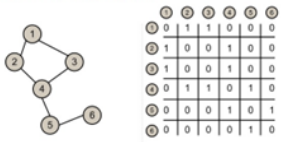

Fig. 2: Exemple d'una matriu d'adjacència

Si es busquen exemples a la vida real, es trobaran en infinitat llocs, ja sigui des d'un simple text on les paraules estan connectades entre elles, una imatge on els píxels adjacents estan units i el seu valor RGB és la informació que contenen, representacions de relacions entre persones com arbres genealògics o xarxes socials, les carreteres d'una regió, etc. Fins i tot, si es mira d'una manera microscòpica, les molècules són grafs, i aquí és on se centra aquest treball. Les molècules estan construïdes per àtoms i electrons a l'espai, on totes les partícules estan interactuant, però quan un parell d'àtoms s'uneixen a certa distància d'una forma estable, es diu que comparteixen un enllaç covalent i interaccionen. Els diferents parells d'àtoms i enllaços tenen diferents distàncies i això es pot representar en forma de graf.
Tornant a les GNN, poden tenir diferents objectius. Aquests, estan dividits en tres, però al final, tots busquen fer una transformació de tots els atributs del graf (nodes, arestes i context global) conservant les seves simetries (invariàncies de permutació). La primera, a escala de graf, on es busca predir propietats o característiques generals d'aquests, com per exemple, un etiquetatge o una nota global. La segona a escala de vèrtex o node on es busca predir les anteriors característiques mencionades d'aquests. La tercera i última, a escala d'aresta, on es busca predir característiques de les relacions entre aquests o si existeix o no una connexió entre vèrtexs.

Per tant, l'input serà un graf d'entrada amb característiques com les mencionades anteriorment, i l'output serà un graf amb aquestes informacions o embeddings modificats. Per dur a terme aquest procés de transformació, es realitza una seqüència de passos igual per a cada un dels components del graf que interessi a la vegada. Aquest procés agafarà les característiques dels veïns en funció de com estigui dissenyada la xarxa, siguin els nodes, les arestes o la informació global i les transformarà amb una funció A, per posteriorment empaquetar els resultats de totes aquestes amb una funció B (Agregació). El resultat d'aquest primer procés es coneix com a missatge. Una vegada fet això, el missatge s'ajunta amb l'estat actual del node i a través d'una funció de transformació C (Actualització), s'aconsegueix el següent estat.

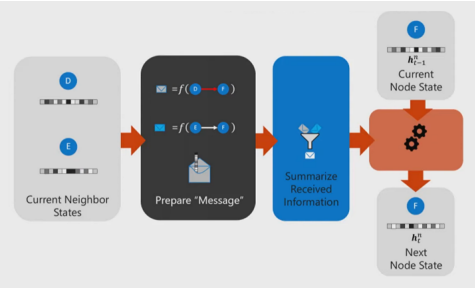

Fig. 3: Transmissió de missatges i actualització de valors

$$M_{N(u)}^{(k)} = AGGREGATE^{(k)}(\{h_v^k, \forall v \in N(u)\}) \quad (1)$$

$$h_u^{(k+1)} = UPDATE^{(k)}(h_u^{(k)}, M_{N(u)}^{(k)}) \quad (2)$$

On $UPDATE$ i $AGGREGATE$ són funcions diferenciables arbitràries (funcions B i C mencionades anteriorment) i el resultat d'$AGGREGATE$ és el missatge que s'agrega a partir dels veïns del graf de u, anomenats $N(u)$. S'utilitzen superíndexs per distingir les diferents iteracions del pas de missatges.

Però aquesta transformació feta, només propaga la informació de les coses més properes i potser l'objectiu de la GNN requereix més profunditat. Per fer-ho, cal entendre què són els cicles. La idea principal d'aquesta part és que cada característica del node o aresta s'actualitza amb les característiques dels seus veïns, tal com s'ha dit anteriorment. Les característiques veïnes es passen al node objectiu com a missatges a través de les connexions. Com a conseqüència, la nova representació del node o aresta codifica i representa l'estructura local del graf. Per tant, aquests cicles són la quantitat de períodes de temps en els quals es fa cada una de les actualitzacions dels valors mencionats anteriorment i es coneixen com a nombre de capes de la GNN. En conseqüència, si es tenen k-capes, la informació final del graf es propagarà k-passos. Més senzillament, si s'està actualitzant la informació dels nodes agafant la informació dels seus veïns, el valor del i-node acumularà la informació dels veïns que estiguin a distància k o k-capes d'ell (figura 4).

3.2 Molecular docking i drug repurposing

L'anàlisi de l'acoblament molecular o molecular docking ha estat una de les estratègies més bàsiques i importants per

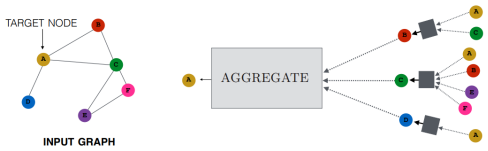

Fig. 4: Procés d'update del node A en la capa/iteració 2

al descobriment de fàrmacs. Permet predir les interaccions moleculars que mantenen units en estat lligat una proteïna i un lligand o molècula. Gràcies a aquesta tècnica es pot predir l'afinitat que tenen o no tenen dues molècules i la força amb la qual s'atrauen a través de les energies lliures o binding affinity resultants de la interacció. Com més negativa és l'energia resultant, més acoblament tenen i com més positiva, més es repel·leixen. En l'àmbit de la medicina es fa servir per a crear medicaments, ja que si s'aconsegueix buscar una molècula (lligand) que sigui capaç d'ajuntar-se amb una proteïna de la malaltia target, aquesta proteïna modificarà la seva composició molecular i possiblement deixarà de fer l'acció perjudicial per al cos, a causa que tindrà una capacitat d'interacció amb altres molècules de l'organisme diferent de la que tenia abans del molecular docking. És a dir, si el receptor i el lligand s'ajunten, es crea un nou paradigma on la proteïna ja no té la capacitat de dur a terme l'efecte negatiu que ha estat realitzant fins al moment.

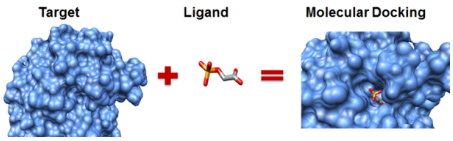

Fig. 5: Exemple de molecular docking

Una vegada explicat això, també s'ha d'entendre què és el drug repurposing, que és l'ús de medicaments ja creats per a funcionalitats o malalties pels quals no es van dissenyar. La seva motivació resideix en el gran cost que pot arribar a tenir la creació d'un medicament, ja sigui d'investigació, de controls sanitaris, de proves, etc. i que aquest s'estalvia si se'n pot reaprofitar un que ja està creat, evitant malgastar recursos innecessaris i totes les conseqüències que la creació d'un fàrmac comporta. Per tant, si es coneixen les molècules dels medicaments que actuaran com a lligands, i la proteïna objectiva, que no és per la qual està creat el medicament, a través del molecular docking es pot predir les energies d'interacció entre ells per a buscar nous medicaments candidats a ser bons pel tractament d'una malaltia, com podria ser el SARS-CoV-2 i la seva proteïna S.

3.3 Enfocament del projecte

Una vegada entès el molecular docking, el drug repurposing i les GNN, sorgeixen dos enfocaments. El primer és la creació d'una GNN capaç de poder emular una proteïna i un medicament en detall amb totes les seves característiques i elements que el conformen, de tal forma que es puguin predir les energies lliures o binding affinity de la interacció entre ells i classificar aquell medicament com a possible candidat per a la realització de drug repurposing. El segon enfocament és la creació d'un interactoma humà, que és el mapa de les interaccions moleculars entre les pro-

teïnes humanes, sense obtenir tant detall sobre les proteïnes i molècules, però aconseguint una visió molt més general i possibles efectes d'altres proteïnes veïnes que no es tenen en compte a l'hora de representar únicament la proteïna i el lligand. D'aquesta manera, l'objectiu no seria estimar un valor sinó fer un link prediction entre el target i el lligand, tenint en compte més factors externs però amb menys detall sobre ells. Les dues opcions tenen els seus pros i els seus contra, ja que mentre una té molt detall, però no es fixa en el context i possibles factors externs, l'altre perd detall, tanmateix, es fixa en el conjunt. Per tant, la situació ideal seria poder extreure les característiques de la primera opció per tal de traslladar-les a la segona. Tot i això, al llarg del projecte s'enfocarà des del primer punt de vista, i la projecció de futur serà que, una vegada acabat aquest, s'ampliïn nous horitzons fora del TFG a través del segon plantejament.

3.4 ESTAT DE L'ART

Per a poder conèixer l'estat de l'art i saber com s'ha d'enfocar el treball, s'ha fet un inventari d'articles que solucionen el problema i dels quals se'n seleccionarà un en el qual basar-se per a la creació de la primera versió de la GNN del treball. A més, s'han consultat altres publicacions útils per a la comprensió total del que s'està fent i per a obtenir idees per a la millora del codi i l'assoliment dels objectius.
Actualment, no hi ha uns estàndards definits per a solucionar aquest problema, ja que és un plantejament molt recent i els diferents investigadors treballen en diversos camins per a millorar constantment els resultats obtinguts. Per exemple, es poden observar diversos articles que utilitzen una CNN per a l'extracció de les característiques de les interaccions, adaptant els grafs perquè puguin ser utilitzables per a xarxes convolucionals, com per exemple l'article [3]. D'altres, com el que ha servit de base per aquest treball, parteixen de grafs i dissenyen GNN de diferents tipus per a veure les diferències entre les interaccions i les no-interaccions. Un punt recurrent en aquestes és la paral·lelització de les capes, duent a terme diverses execucions en una mateixa capa amb diferents representacions i característiques per al mateix graf, i observant les diferències obtingudes entre aquestes. Alguns exemples són els articles [1], [2], que es basen en Novel GNN, també coneguts com capes GAT (Graph Attention Network), l'article [4], que busca donar una solució a través de la unió d'una GNN i una CNN per als lligands i les proteïnes respectivament, o, finalment, l'article [5] que utilitza un IGT o Intermolecular Graph Transformer. També hi ha altres apropaments al problema amb plantejaments com xarxes GoG (Graph of Graphs) [6] que intenten representar un iteractoma i les seves molècules o Deep Neural Networks [7] que se centren més en les propietats físiques del receptor.
En resum, s'ha fet un estudi exhaustiu explicant el plantejament, la solució donada i els resultats d'articles que treballen en aquest problema. Veure apèndix 1. A més, s'han llegit altres textos com articles que donen solució a l'altre plantejament de projecte proposat dels interactomes humans, que busca donar context a la xarxa neuronal proporcionant-li l'interactoma com a graf, articles de bioquímica i deep learning útils per a l'enteniment de com es comporten i es comprenen les estructures de diferents molècules, textos que ensenyen a extreure el màxim rendiment al programari d'Autodocking Vina que s'utilitza en aquest camp o altres publicacions que introdueixen i donen diferents punts de vistes i plantejaments del tema.
Una vegada valorades totes les opcions, s'ha escollit un article com a base per posar en pràctica els coneixements sobre GNN adquirits, creant-ne una i entrenant-la des de zero. La publicació seleccionada és la referència [1]. Els motius principals de la tria han estat la seva gran quantitat d'explicacions i detalls que moltes altres publicacions amagaven, l'ús de dades en 3D i no en 2D com les publicacions més antigues, la qualitat dels resultats, el tipus de preprocessat de dades utilitzat i la procedència (base de dades) d'aquestes, el repositori públic que s'ofereix on hi ha tot el codi que pot ajudar en algun moment que s'estigui encallat per a continuar avançant, les potencials millores que s'hi han detectat i la semblança que té amb un altre article [2] que pot servir per extreure les idees que es creguin millors i plasmar-les a la pròpia xarxa, combinant el millor de cada publicació.
A mesura que s'ha fet l'inventari d'articles, s'ha anat explorant les diferents bases de dades que es mencionaven per tal d'escollir les millors per a l'entrenament de la GNN. A més, aquest pas pren importància si es té en compte la possibilitat que la xarxa aprengui característiques de la base de dades amb la qual és entrenada i després no sigui capaç d'extrapolar-ho a altres interaccions d'altres datasets, que és un dels objectius d'aquest treball. Per tant, s'han estudiat 8 bases de dades diferents que contenen diverses informacions sobre molècules i proteïnes i les seves conseqüents interaccions. Després de l'estudi i la valoració de quins datasets posseïen les millors característiques per a les finalitats del treball, s'ha arribat a la conclusió que tant la DUD-E database [9] com la PDBbind [10] es poden fer servir per a l'entrenament i el testing, respectivament. La primera conté una gran quantitat de mostres utilitzables i el percentatge de receptors-lligands que no interaccionen és molt elevat, fet important per a fer-ho el més realista possible. La segona és interessant perquè les mostres proporcionades són d'alta qualitat i estan totes extretes de procediments experimentals, tot i això, no és molt extensa, i, com és lògic, ja que només interessen les interaccions, només hi ha mostres positives. Es pot veure a la taula 1.

TAULA 1: TAULA DE LES BASES DE DADES

	DUD-E	PDBbind
Interaccions	1311636	13873
Proteïnes	98	13873
Inter. Actives	36859	13873
Inter. Inactives	1274777	0

Independentment d'aquestes, també s'ha sol·licitat accés a Drugbank [11], una plataforma amb totes les estructures moleculars de components actius del mercat. D'aquesta manera, en fer l'aplicació real, es tindran totes les molècules que tenen possibilitats de ser escollida pel drug repurposing. Aquesta base de dades és privada i s'ha de demanar un accés acadèmic o pagar la llicència corresponent per utilitzar-la.

4 METODOLOGIA

La metodologia de treball seguida ha tingut una estructura com la següent:

Estudi – Rèplica – Millora – Aplicació

És a dir, primer s'ha començat fent un estudi exhaustiu de l'estat actual del sector, per introduir-se i apropar-se a l'estat de l'art, veien els recursos i models que s'han fet servir els últims anys. Una vegada fet això, s'ha fet un inventari de bases de dades i arquitectures de GNN estudiades, i se n'ha escollit una per a replicar-la. Quan aquesta ha estat en funcionament, gràcies als coneixements adquirits al llarg de l'estudi, i a través de la comparació entre diferents solucions al problema, s'ha millorat el model creat per a intentar obtenir millors resultats. Finalment, s'ha aplicat a casos reals per tal de comprovar-ne el funcionament i veure el seu potencial.

Les eines emprades s'han anat determinant durant l'avenç de la investigació. Per a fer els models, com s'ha mencionat anteriorment i com es veu a l'estat de l'art, s'ha fet servir les GNN programades amb Python, concretament amb la llibreria PyTorch, i accelerades a través de GPU, tot posat en pràctica en un entorn Linux amb JupyterLab. A part, per la realització del que es coneix com a molecular docking, s'ha fet servir AutoDock Vina [12], i pel tractament de molècules, Pymol [13] i Open Babel.

5 DESENVOLUPAMENT

El desenvolupament del treball ha seguit tres parts clarament diferenciades. La primera és el preprocessat de les dades per tal de preparar-les d'una forma ideal per entrar a la GNN. La segona ha estat la implementació de la GNN. La tercera ha estat el redisseny i la millora de la GNN replicada. Aquestes parts, tot i tenir un ordre cronològic en el temps, s'han anat sobreposant i intercalant per tal de millorar el pipeline d'execució d'una forma unànime i òptima. Tot el codi produït s'ha formatat amb Black i Flake8 seguint l'estàndard PEP 8, s'ha empaquetat amb l'eina PyPI per tal de facilitar la instal·lació a qualsevol persona que volgués replicar el procés i s'ha penjat a un repositori Github al següent enllaç públic: https://github.com/guiuomsfont/GNN-Molecular-Screening.

5.1 PREPROCESSAT

El preprocessat de les dades consta de diverses parts que s'han anat optimitzant, modificant i millorant al llarg del procés. Aquestes, principalment parteixen de la cerca d'un encaix entre la proteïna i la molècula, la millora d'aquest encaix i la selecció de les característiques i dades rellevants per tal d'evitar introduir dades innecessàries o que afegeixin soroll al model.

La primera part del preprocessat es coneix com a molecular docking inicial. Aquest es basa a col·locar la molècula de la interacció en una posició sobre la proteïna que sigui favorable perquè ambdues interaccionin. Cal tenir en compte, que es poden tractar prèviament els agents implicats per tal de millorar-lo. En el cas d'aquest treball s'ha fet a través de l'aigua, ja que en la majoria dels casos, les molècules HOH (H_2O) no estan implicades en la unió, i, per tant, s'eliminen per facilitar els càlculs i netejar el binding pocket de possibles molècules que distorsionarien la cerca de la postura. És a dir, a través de l'eina Pymol [13] s'han eliminat aquestes molècules, i, com es pot observar en els resultats de la secció 6, al final s'ha fet una comparació amb ambdues opcions per veure la veracitat d'aquestes afirmacions i comprovar si aquest canvi respecte a l'article original ha estat beneficiós. Llavors, a través de l'eina Smina [12], una branca del programari Autodock Vina, es realitzen aquests càlculs de docking i s'obté un llistat de les 9 millors posicions o binding poses en les quals el programa ha detectat possibilitat d'interacció, factor que no determina que n'hi hagi, ja que això ho realitzarà la GNN, el que fa és determinar si hi ha possibilitats. En el cas que no se n'hagin trobat 9 o més, es retornen les que s'han aconseguit.

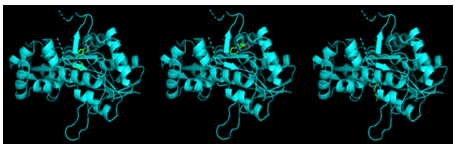

Fig. 6: Diferents posicions trobades d'una molècula (verda) dins d'una proteïna (blava)

Tanmateix, aquest programa no és el millor per a l'ordenament d'aquestes posicions oferides i s'ha buscat la forma de quedar-se amb la millor a través d'altres mètodes de rescoring que intenten donar una energia lliure aproximada de la interacció, però que no és especialment precisa (per això la realització d'aquest estudi). Després de valorar diverses opcions s'ha acabat utilitzant Rf-score, obtingut d'un article [8] públic i que s'adapta a l'objectiu plantejat. Finalment, una vegada ordenades les diverses posicions amb l'energia lliure "temporal", s'ha seleccionat la millor posició, producte de la nova puntuació.

En aquest punt, la proteïna i la molècula ja estan preparades per entrar a la xarxa neuronal, però per a millorar l'eficiència i descontaminar les mostres de soroll, es fa un retall del receptor amb la finalitat de deixar únicament els àtoms propers al lligand, és a dir, el binding pocket. Per fer aquesta selecció, es crea una matriu de distàncies entre els àtoms de la proteïna i els de la molècula i es mantenen tots els que tenen com a mínim una distància inferior a X. Aquesta X és variable, pot donar més o menys context a la mostra, i, per tant, més o menys informació, sempre sabent trobar l'equilibri entre la manca de dades i l'excés de soroll. Per això, el que s'ha fet ha estat utilitzar distàncies de 6, 8, 10 i 12 Angstroms o Å i comparar els resultats dels entrenaments per a veure quina d'elles és més efectiva i obté millors mètriques, procediment explicat a l'apartat 5.3 de desenvolupament. A la figura 7 s'aprecien les diferències entre els possibles retalls.

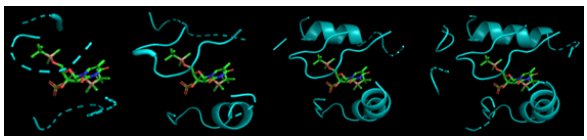

Fig. 7: Retall de la mateixa molècula i proteïna per 6, 8, 10 i 12 Å

El conjunt d'aquest preprocessat ha tardat al voltant de dos mesos d'execució, i, per tant, ha estat crucial paral·lelitzar la feina i realitzar els llançaments sense error, ja que una errada important podia determinar l'èxit o el fracàs del projecte a causa de les restriccions temporals.

5.2 GNN

En paral·lel a la tasca de preprocessat, s'ha programat la GNN partint de l'article que es menciona anteriorment [1]. L'arquitectura seguida es pot observar a la figura 8, però principalment consta de dues parts.
La primera se centra en la duplicació del graf, donant-ne a la xarxa un de desconnectat (només les connexions dins de la mateixa molècula i proteïna), i un altre de connectat seguint un criteri de distància definit posteriorment. Cada un d'ells passa per una GAT i es resten els vectors de característiques resultants. D'aquesta manera es troba la diferència entre una postura d'unió i l'estructura separada. Aquesta seqüència es realitza 4 vegades. La segona part consta d'un conjunt de capes fully connected (multi-layer perceptrons) regularitzades amb un dropout i connectades amb una funció ReLU, excepte l'última en què no hi ha dropout i s'utilitza un sigmoide final per saber si hi haurà interacció o no.

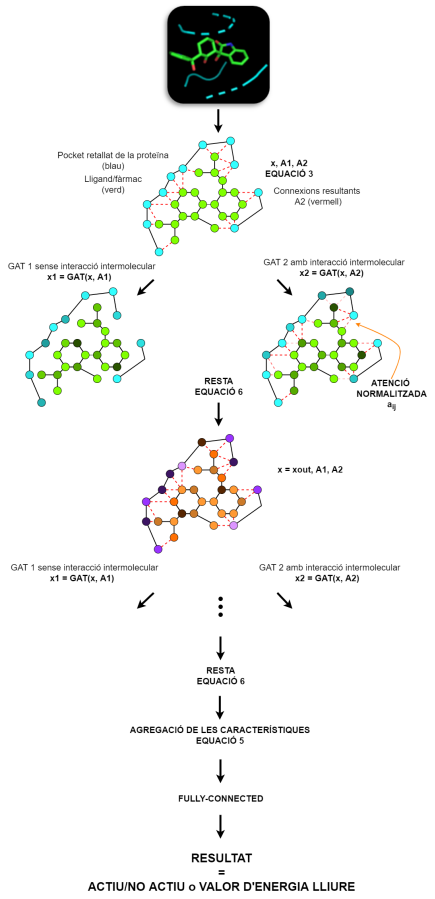

Fig. 8: Arquitectura general de la GNN proposada

Més en detall, es rep d'entrada una molècula o lligand amb una representació 3D dels seus àtoms (que estan col·locats en la millor posició que s'ha trobat després de realitzar el docking i rescoring), i una proteïna situada a l'espai i retallada on només es mantenen els àtoms més propers a les coordenades 3D del lligand. D'aquest input, se n'extreu el vector de característiques x dels nodes (àtoms), que tindrà mida 56. Les primeres 28 posicions representaran els àtoms del lligand de la interacció i les 28 últimes els àtoms de la proteïna, totes elles codificades en one-hot. D'aquestes, les primeres 10 estaran dedicades al tipus d'àtom que s'està representant en el vector de característiques (C, N, O, S, F, P, Cl, Br, B, H), les següents 6 seran pel grau de 0 a 5, les següents 5 pel nombre d'hidrògens adjunts (0 a 4), les altres 6 pel nombre d'electrons de valència implícita (0 a 5) i finalment l'últim bit que queda per un booleà representant l'aromaticitat.
Una vegada extretes aquestes característiques, es creen dues matrius d'adjacència que s'anomenaran A_1 i A_2. La primera, tal com s'ha mencionat anteriorment, només representa els enllaços covalents amb 0 i 1. És a dir, en cas que hi hagi un enllaç ja definit, dins de la mateixa proteïna o dins de la pròpia molècula, la matriu contindrà 1, en cas contrari contindrà 0. La segona matriu A_2, també contindrà la informació de la matriu anterior, però a més, independentment d'aquesta, s'aplicarà la fórmula 3 entre els àtoms de la proteïna i els de la molècula. Aquesta fórmula tindrà unes constants definides a l'inici de l'entrenament que per defecte seran 4 per μ i 1 per σ, i a mesura que la distància entre els àtoms sigui major, més tendirà a 0.

$$e^{-(dist_{ij}-\mu)^2/\sigma} \tag{3}$$

Una vegada les dues matrius estan codificades i el vector de característiques x_i per a cada un dels àtoms creat, s'entrarà a una GAT1 la matriu A_1 i a una GAT2 la matriu A_2 amb els corresponents vectors, de tal manera que se centrin en propietats diferents i siguin capaces d'extreure diferents característiques de la interacció. Dins d'aquestes GAT, els diferents nodes transformaran els seus vectors a través d'una matriu de pesos W apresa (fórmula 4) i els coeficients d'atenció normalitzats a_{ij}, també apresos, que definiran la importància del node-j sobre el node-i (fórmula 5).

$$x_i^{'} = W x_i \tag{4}$$

$$x_i^{''} = \sum_{j \in N_i} a_{ij} x_j^{'} \tag{5}$$

El vector resultant es passarà a les següents capes mitjançant la resta entre aquest i el de l'altre GAT, de tal manera que s'aprengui la diferència entre l'estructura acoblada amb la binding pose i l'estructura separada (fórmula 6). Aquest procés es repeteix 3 vegades més, és a dir, 4 capes en total.

$$x^{out} = x^{outGAT2} - x^{outGAT1} \tag{6}$$

Els vectors d'output de l'última capa GAT (la quarta) se sumen seguint la fórmula 7, i s'introdueixen a un perceptró multicapa, MLP o fully-connected de 4 capes que classificarà si la posició entre el lligand i la proteïna és activa o no. Per acabar utilitzant una funció d'activació ReLU que proporcionarà el valor final de la classificació, donant un 1 si s'ha determinat que la interacció és activa, i un 0 si s'ha determinat que no hi ha interacció.

$$x^{graf} = \sum_{i \in lligand} x_i \tag{7}$$

Al final, per a mesurar els errors obtinguts i entrenar la xarxa s'ha utilitzat un BCE o Binary Cross Entropy com a funció de loss.

5.3 MILLORA DELS PARÀMETRES I LA DISTÀNCIA DE RETALL

Una vegada implementada la GNN adaptada al cas d'ús personal, s'han buscat millores al procés d'entrenament per tal d'optimitzar-ne els resultats i perfeccionar el producte final. La primera de les millores ha estat l'ajust de paràmetres per a millorar el funcionament. Aquests s'han tocat manualment i s'han anat modificant per comprovar els efectes que poden tenir sobre l'entrenament de la xarxa. Finalment, s'ha observat que el millor learning rate es troba en 0,0001 i el nombre de capes òptim es troba en 4 GAT i 4 MLP amb unes dimensions de 140 i 128 respectivament. El batch size emprat ha estat de 128 i s'han realitzat 100 iteracions en tots els entrenaments per assolir la convergència.
També s'ha millorat la distància de retall del binding pocket, explicada anteriorment, i s'han dut a terme entrenaments exactament iguals als que estan explicats a l'apartat 6 amb distàncies de retall del binding pocket de 6, 8, 10, i 12, obtenint els resultats de les taules 2.1 i 2.2.

TAULA 2: TAULA DE RESULTATS I MÈTRIQUES DELS ENTRENAMENTS EN FUNCIÓ DE LA DISTÀNCIA DE RETALL

Distància de retall	ROC	Bal Acc
6	0,936	0,812
8	0,947	0,849
10	0,949	0,843
12	0,936	0,849

Distància de retall	TN	FP	FN	TP
6	253730	7140	2759	5170
8	248115	12755	2014	5915
10	249386	11484	2148	5781
12	243104	17766	1860	6069

Tal com es pot observar, s'obtenen moltes mostres classificades com a positives que realment no ho són. Això es deu al fet que el percentatge de positius i de negatius que hi ha en el test està molt descompensat, i tal com revela l'AUCROC o l'accuracy balancejada, els resultats són molt bons. Tot i això, s'ha de tenir en compte que el límit per a separar una mostra activa d'una inactiva està en una probabilitat de 0,5, aspecte modificable i que s'ha pres com una nova millora. Si mirem els valors que es tenen, es veu com, amb no molta diferència, el millor entrenament és el de 8 Å, ja que aconsegueix la millor accuracy balancejada i està dins de les millors mètriques en la resta de camps, i, per tant, aquest és el que s'ha fet servir posteriorment per a trobar la separació de la probabilitat d'actiu vs. inactiu i per entrenar un model adaptat a una regressió.

5.4 MILLORA DEL LLINDAR

Amb la millor distància de retall ja definida, s'ha prosseguit a determinar quin era el millor límit per a separar les mostres positives i negatives. Tot i això, per a comprendre aquesta part cal recordar que un dels objectius finals que busca el treball és la recerca de candidats per a fer drug repurposing. Per tant, és bastant indiferent que surtin més o menys positius en el test, perquè sempre s'agafaran les n millors mostres, encara que se n'hagin etiquetat moltes més de positives. És a dir, en aquest cas s'intentarà buscar el millor percentatge de TP respecte TP+FP (precision), o la millor F1 score, per veure realment el bo que pot arribar a ser el model, però en les aplicacions finals, aquest threshold que trobarem no servirà, perquè el que es farà serà agafar les puntuacions més altes de les prediccions. En la figura 9 es pot veure l'evolució de la precisió en funció d'on se situï el threshold per a fer la classificació.

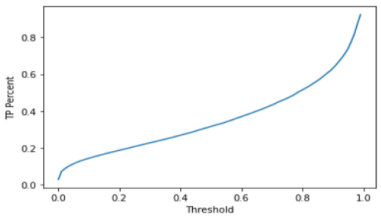

Fig. 9: Gràfica de l'evolució de la precisió en funció del threshold

Veient la gràfica, la millor separació es troba en la probabilitat 0,99. En aquesta, tot i perdre una gran quantitat de positius reals i convertir-los en falsos negatius, el volum de falsos positius disminueix dràsticament, fent que el percentatge de positius que realment són positius, s'elevi bastant. Això és un bon senyal, ja que indica que les mostres classificades amb més probabilitat gairebé sempre seran TP, i, per tant, serviran per a la creació d'un top. Si mirem una altra mètrica que fa una valoració més global en conjunts de dades desbalancejats, com podria ser l'F1 score, veiem que el millor threshold se situa en 0,91 a la figura 10.

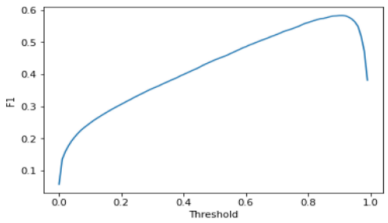

Fig. 10: Gràfica de l'evolució de la F1 score en funció del threshold

Per acabar d'entendre el perquè, es pot observar l'histograma que mostra la quantitat de positius reals que hi ha dins de cada tram de probabilitat predita a la figura 11. D'aquesta forma, s'aprecia visualment com l'objectiu del model s'està assolint, ja que s'agrupen la gran majoria de mostres positives en el rang de probabilitat de 0,95-1.

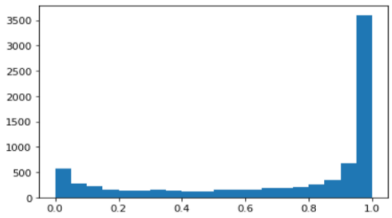

Fig. 11: Histograma que mostra la quantitat de positius reals que hi ha dins de cada tram de probabilitat predita

Per tant, les conclusions de la millora han estat que el límit ideal per a classificar i obtenir bones mètriques és de 0,91.

5.5 MILLORA PER A LA PREDICCIÓ D'UN VALOR CONTINU

Finalment, l'última millora que s'ha implementat ha tingut la finalitat de predir l'energia lliure producte de la interacció entre el receptor i el lligand, per tant, passar de regressió a classificació. Aquest canvi pren sentit si es mira des d'un aspecte científic, ja que a part de determinar si hi ha interacció entre un fàrmac i la proteïna objectiu, permet veure el valor de la binding affinity d'aquesta, que com més petita sigui, més interacció es tindrà.
Per dur-ho a terme s'ha agafat la GNN amb totes les millores explicades fins al moment, i s'hi han realitzat modificacions donant-li la capacitat de predir valors continus. Els resultats d'aquests canvis han estat estudiats i analitzats conjuntament amb els de la classificació a la secció 6.
La primera modificació ha estat l'obtenció dels targets continus, que fins al moment eren binaris de 0 o 1. Per fer-ho, s'ha accedit a la base de dades de DUD-E [9] i manualment s'han descarregat els valors a predir per a cada una de les molècules. En el cas de les molècules sense interacció que no disposen d'una energia lliure, se'ls hi ha donat un valor molt diferent que les que sí que interaccionen per diferenciar-les clarament. A més, com que d'una molècula amb interacció a una sense interacció hi pot haver grans diferències d'escala, amb valors de 10^6 de diferència, s'ha aplicat el logaritme natural a tots els targets, i quan es realitzin les prediccions s'haurà d'aplicar l'exponencial per a desfer aquest canvi.
El segon pas ha estat la modificació de la funció de sortida de la GNN, que ha passat de ser probabilística a ser contínua. En concret, s'ha passat d'una sigmoide a una funció lineal simple que fa que la sortida deixi de ser una probabilitat i passi a ser un valor de regressió.
El tercer i últim pas ha estat modificar la funció de loss, ja que la que hi havia fins al moment no servia a causa que estava adaptada a les probabilitats, passant d'un BCE (Binary cross entropy) a un MSE (Mean squared error).

6 RESULTATS

En aquest apartat es mostren els resultats obtinguts, i el marc experimental emprat per arribar a ells. Principalment, hi ha hagut tres tipus de proves. Les primeres per comprovar la qualitat i funcionament del model de classificació i regressió amb una part de la base de dades d'entrenament que no se li ha mostrat i s'ha reservat pel test, les segones per dur a terme una validació i comparació amb una altra base de dades, comprovant la capacitat d'extrapolar que té la GNN, i finalment les últimes han estat una aplicació real al virus SARS-CoV-2, comparant els resultats aconseguits amb investigacions reals per veure si s'ha demostrat experimentalment l'eficàcia d'alguna de les prediccions.

6.1 MARC EXPERIMENTAL

Tal com ja s'ha mencionat, els següents experiments han estat realitzats amb les bases de dades de DUD-E [9], PDBbind [10] i DrugBank [11] respectivament. Totes elles han passat per un preprocessat explicat a la secció 5.1 i que consta de l'encaix entre el receptor o proteïna i la molècula o lligand per a cada una de les interaccions, procés anomenat docking i dut a terme amb Smina [12], seguit de la selecció dels millors encaixos obtinguts amb Rf-score [8], que serveix per aconseguir una estimació de l'energia lliure d'interacció, i el retall del binding pocket, és a dir, la zona propera al lloc on encaixen la molècula i la proteïna per evitar donar excessiva informació a la GNN que produeix distorsió i augmenta el còmput. A més, en els casos on se cita que hi ha mostres sense aigua, seccions 6.3 i 6.4, s'ha passat per un pas previ a tots els anteriors amb Pymol [13], on s'han eliminat les molècules H_2O, aspecte que millora el docking. Cal destacar que en les dades de DUD-E [9], i, per tant, en la informació utilitzada per a l'entrenament, no s'ha fet aquest pas, ja que disposa d'un tractament previ similar, que ja està realitzat quan es descarrega.
A més, els models i les dades emprats per a totes les proves ja disposen de les millores explicades, és a dir, amb els paràmetres de la xarxa, el batch size i les iteracions òptimes trobades a l'apartat 5.3, la distància de retall realitzada a 8 Å, el threshold de 0,5 i el de 0,91 definit, i, quan es mencioni, l'ús de la GNN adaptada a la regressió que retorna el valor d'energia lliure (que com més baix, millor interacció), en comptes d'una probabilitat d'interaccionar.

6.2 ENTRENAMENT AMB DUD-E

Per a l'entrenament i testeig dels dos models s'ha dividit la base de dades de DUD-E [9], i s'han separat les proteïnes en dos grups, un per train i l'altre per test. És a dir, es disposa de 98 proteïnes, les quals tenen associades a elles n interaccions, és a dir, n lligands etiquetats amb si hi ha interacció o no. D'aquesta forma, totes les proteïnes que es mostren en el test, el model no les haurà vist mai al llarg de l'entrenament, i, per tant, no podrà estar condicionat ni tindrà experiència sobre elles, com hauria de passar en un cas real. La partició ha estat feta amb 74 proteïnes per entrenar i 24 per testejar. A més, com que la proporció de positius envers negatius és extremadament baixa, s'ha fet un sampleig amb la proporció fixa d'1:1 quan es realitza la preparació d'un batch d'entrenament. En el test s'han mantingut les proporcions reals per a comprovar l'eficàcia dels models.
Finalment, per a comprovar la validesa de les mètriques obtingudes, s'ha dut a terme un Cross Validation mantenint la proporció de train-test mencionada (0,75 - 0,25 aprox.), és a dir, amb un k-fold on k = 4 (k1 = 73 - 25, k2 = 73 - 25, k3 = 74 - 24 i k = 74 - 24), obtenint uns resultats molt similars als entrenaments anteriors i que es mostra la mitjana resultant a les taules 3.1 i 3.2 per la classificació i a la taula 4 per la regressió.

TAULA 3: TAULES DE MÈTRIQUES DEL CV AMB EL MODEL DE CLASSIFICACIÓ

	ROC	Bal Acc
Train	0,994	0,968
Test	0,946	0,822

	TN	FP	FN	TP
Train	474239.7	17713,0	13623,5	478150,7
Test	310412,5	8281,7	2995,2	6219,5

També, per veure l'eficàcia del model obtingut, s'ha comparat la millora amb el mètode que actualment està més àmpliament estès, i que, en aquest treball, ja s'ha utilitzat prèviament, Rf-score [8]. Aquest model és un random fo-

TAULA 4: TAULA DE MÈTRIQUES DEL CV AMB EL MODEL DE REGRESSIÓ

	MSE	MAE
Train	6,410	6,438
Test	1,621	1,136

rest que, amb el docking fet, repuntua les posicions trobades intentant donar una energia lliure. El que s'ha fet ha estat l'adquisició d'un dels conjunts de test del CV, i el recompte de X positius reals totals que té. Tenint aquest número, s'han ordenat totes les mostres predites i s'han agafat les X primeres, comprovant quantes d'elles són predites com a positives. Els resultats es poden observar a la taula 5.

TAULA 5: TAULA DE MILLORES RESPECTE A RF-SCORE

Model	Fracció Positius	Millora respecte a Rf-score
Rf-score	1019/7929	-
Classificació	7228/7929	709,323 %
Regressió	6087/7929	597,35 %

Analitzant els resultats obtinguts, es pot veure que s'ha aconseguit crear un model completament funcional per a la finalitat que es busca, ja que les mètriques dels entrenaments i el CV han estat d'alta qualitat amb una gran ROC i accuracy en el cas de la classificació, i amb bon MSE i MAE en el cas de la regressió. A més, a l'hora de realitzar la comparació amb un dels mètodes més usat actualment, es pot veure com les millores són molt bones, assolint grans percentatges de millora i una taxa de positius predita molt més elevada.
Si s'observa la regressió envers la classificació, es podria determinar que la regressió no ha estat una millora, però aquesta afirmació no és del tot certa, ja que gràcies a la regressió s'aconsegueix una energia lliure que pot servir per a molts càlculs, en canvi, a la classificació únicament s'obté una probabilitat. A més, si es miren els resultats més en detall, i tal com es pot veure en seccions posteriors, la regressió és pitjor a l'hora de comparar les mètriques, però la confiança en els seus resultats, és a dir, la probabilitat que les millors mostres reals amb una energia lliure o d'interacció baixa (mostra positiva) es trobin en posicions altes del rànquing de possibles candidats a fer drug repurposing, és més alta que en la classificació.

6.3 VALIDACIÓ AMB PDBBIND

Per a poder validar la capacitat de generalitzar del model, s'ha buscat un dataset extern, amb diferents proteïnes i lligands de naturalesa diversa, per descartar overfitting i veure com degrada el model fora del dataset d'entrenament, ja que, al llarg de l'estudi de l'estat de l'art, s'ha vist que les mètriques baixaven en picat en modificar l'estil de les dades. Això s'ha fet a través de PDBbind [10], una base de dades on la totalitat de les mostres interaccionen entre elles. És a dir, és una petita base de dades amb 13873 mostres que s'ha comprovat experimentalment que poden interaccionar. Tot i això, cal destacar que algunes interaccions són molt lleugeres o ínfimes, i la GNN podria detectar-les com a negatives. Les proves realitzades han estat sense i extraient l'aigua, de tal forma que també es pugui veure l'eficàcia d'aquesta millora. Es poden observar a taula 6.

TAULA 6: PERCENTATGE D'ENCERTS (TP)

	H_2O	NOH_2O
Class - Thresh = 0,5	0,345	0,605
Class - Thresh = 0,91	0,215	0,382
Regr - Thresh = 5000 nM	0,499	0,596

Tal com es pot veure, els resultats cauen notablement, baixant fins a 0,60 en el percentatge d'encert màxim. Tanmateix, no és una cosa especialment negativa, ja que, primer de tot, cal recalcar que l'objectiu del treball és determinar un conjunt de candidats, no obtenir una mètrica, i, en aquest cas, s'ha vist que les mostres classificades com a positives han estat les que més possibilitats d'interaccionar posseeixen, encara que segons la base de dades totes interaccionin en més o menys mesura, sobretot en el cas de la regressió. A més, d'aquesta forma es mostra com la millora de l'extracció de l'aigua ha estat un encert, trobant diferències notòries entre les prediccions fetes amb H_2O i sense. També, es pot observar una altra propietat bona de la regressió, i és que és capaç de tenir un millor percentatge d'encert en les mostres amb aigua, permetent, en casos extrems, no haver de fer un tractament previ tan exhaustiu de les dades. Per tant, la conclusió d'aquests resultats és que, tot i haver disminuït considerablement les mètriques, el model està entrenat i funciona remarcablement bé per a la finalitat amb la qual ha estat dissenyat, si es té en compte que només es volen les millors mostres de totes les prediccions fetes, pot estar-se enfrontant amb dades completament diferents de les que ha vist, la base de dades conté mostres de dubtosa interacció positiva i que les prediccions en el cas de la regressió o les classificacions en el cas de la classificació que han estat positives, concorden destacablement amb les mostres de PDBbind [10] que més interacció tenen.

6.4 DRUG REPURPOSING APLICAT A LA COVID-19

Per acabar les proves i fer una exemplificació en el món real, s'ha estudiat el virus del SARS-CoV-2 i s'han agafat 4 de les seves proteïnes més importants que, després de llegir diversos articles, es determina que poden interaccionar amb medicaments i que aquests facin efecte. Les seleccionades han estat la 6LU7, 6M71, 6VSB i 6VXX sense treure i traient l'aigua. Aquestes, s'han creuat amb la base de dades de Drugbank [11], que com s'ha explicat en seccions anteriors, conté tots els principis actius de medicaments actualment al mercat. Conté 8349 mostres, per tant, s'han avaluat un total de 33396 interaccions amb aigua i 33396 interaccions sense aigua. Els resultats complets de les prediccions ordenades per energia lliure o probabilitat es troben a l'apèndix 2.
Realitzant diverses cerques a través de la web, amb els principis actius que la GNN ha predit que tenen més opcions per ser candidats al reaprofitament de medicaments per les proteïnes mencionades, s'ha vist que bastants ja s'estan estudiant o ja s'estan aplicant per la cura de la malaltia. Alguns dels exemples i les seves respectives proves es mostren a continuació, mentre que altres resultats positius es poden veure a l'apèndix 2. Es considera interacció a partir de $<$ 5000 nM. Enllaç de l'estudi sobre el nom del medicament.

- 6M71_121304016 - Remdisvir - Predicció: 2,691 nM

- 6LU7_1549008 - Saquinavir - Predicció: 12,568 nM
- 6LU7_45375808 - Sofosbuvir - Predicció: 35,661 nM

Per tant, es pot comprovar que el model està funcionant correctament pel fet que un percentatge notori de les interaccions que ha predit s'estan estudiant o es fan servir per a combatre la Covid-19, mentre que una altra part, no ha coincidit, aspecte que no significa que no tingui interacció, sinó que pot passar, que, per ser un tema tan present, no s'hagi descobert la relació, i, amb una eina com la creada, es puguin trobar nous medicaments per fer drug repurposing.

6.5 Discussió dels resultats

Una vegada analitzats els resultats, es pot veure que aquests han estat de bastant qualitat. En el primer apartat, on sempre s'han emprat dades tractades de la mateixa manera i provinents del mateix lloc, les mètriques obtingudes han estat extremadament bones. Tanmateix, a l'hora de comprovar amb dades de procedències diferents el funcionament del model, aquestes han baixat notablement. Tot i això, no és completament vàlid fixar-se únicament amb els valors resultants en general, ja que no s'ha de perdre de vista l'objectiu final del treball, que és trobar una llista dels millors principis actius candidats a ser eficaços envers una malaltia, i, per tant, a la realitat deixa de ser rellevant on es determina el threshold per definir si una mostra és positiva o negativa, o la quantitat de positius i negatius que hi ha, ja únicament es quedaran els n millors per a procedir amb investigacions experimentals. A més, la conversió de la GNN de classificació a regressió ha fet que, no únicament es tinguin els que més probabilitat hi ha que siguin actius o interaccionin, sinó que s'aconsegueix un valor aproximat sobre l'energia lliure estimada que tindran. Finalment, tal com es pot observar a l'últim apartat, el propòsit del treball ha estat assolit satisfactòriament, ja que un nombre important de mostres han coincidit amb medicaments reals que s'estan usant o que estan en trams finals d'investigació per la gran quantitat de probabilitats que tenen de ser útils per a la malaltia escollida.

7 Conclusions

Una vegada acabat el treball, els objectius que es platejaven inicialment s'han assolit de forma satisfactòria. S'ha realitzat un estudi exhaustiu i profund de l'estat de l'art que ha servit per a contextualitzar i poder implementar el projecte, obtenint en el mateix procés la informació sobre les bases de dades necessàries, i veient la importància d'aquestes per l'èxit del projecte. A més, s'ha pogut replicar un model d'un dels articles llegits aplicant-hi considerables millores tant en el preprocessat de les dades, com en les mètriques o en l'arquitectura, acabant amb una validació rigorosa d'aquest i amb una aplicació real que actualment s'està començant a implementar en un projecte de l'empresa Grupo AIA per al tractament de malalties diverses amb fàrmacs de procedència natural, i que s'espera que pugui servir per a més persones en un futur no molt llunyà.
Personalment, s'està satisfet per haver cursat el grau d'Enginyeria Informàtica i haver pogut emprar molts dels coneixements apresos al llarg de la carrera, que han estat útils tant per la part més formal del treball, com per l'estructuració de codi, i, sobretot, la implementació de models d'IA, l'anàlisi posterior dels resultats i l'aplicació real feta.
Per concloure, d'aquest projecte se n'ha extret una gran experiència plena de nous coneixements i vivències que et fan créixer com a professional, però també com a persona, i s'espera poder continuar treballant en ell per aconseguir aportar nou valor a la societat, facilitant i agilitzant el reaprofitament de medicaments, per tal de poder tractar malalties d'una forma més ràpida i segura gràcies a la tecnologia.

Agraïments

M'agradaria donar les gràcies a la meva família pel suport rebut al llarg del procés i per la formació que m'han proporcionat al llarg dels anys. També a l'empresa Grupo AIA per l'interès en el projecte i les facilitats oferides, ja que sense ells no hi hauria treball, en especial a en José Luis i en Guillermo per tota la seva ajuda i suport constant. I finalment, al meu tutor, en Josep, que ha estat al meu costat en tot moment, fent-me gaudir de l'experiència i fent del procés una cosa enriquidora.

Referències

[1] Lim, J., et al. (2019). Predicting Drug–Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. JCIM, 59(9), 3981–3988. DOI.

[2] Wang, X., et al. (2021). Protein Docking Model Evaluation by Graph Neural Networks. FIMB, 8. DOI.

[3] Hu, S., et al. (2019). Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks. BMC, 20(S25). DOI.

[4] Tsubaki, M., et al. (2018). Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. BMC, 35(2), 309–318. DOI.

[5] Liu, S., et al. (2021). Improved Drug-target Interaction Prediction with Intermolecular Graph Transformer (2). DOI.

[6] Harada, S., et al. (2020). Dual graph convolutional neural network for predicting chemical networks. BMC, 21(S3). DOI.

[7] Verma, N., et al. (2021). SSnet: A Deep Learning Approach for Protein-Ligand Interaction Prediction. IJMS, 22(3), 1392. DOI.

[8] Hongjian L., et al. (2016). Improving AutoDock Vina Using Random Forest. Verlag GmbH, 10.11. DOI.

[9] Database of Useful (Docking) Decoys - Enhanced. (2021). DUD-E. Enllaç.

[10] Database P. (2022). PDBbind-CN database. PDBbind. Enllaç.

[11] DrugBank Online (2022). DrugBank. Enllaç.

[12] AutoDock Vina. (2021). CCSB. Enllaç.

[13] PyMOL - pymol.org. (2022). Schrödinger, Inc. Enllaç.

APÈNDIX

A.1 INVENTARI D'ARTICLES

En aquest apèndix es proporciona una relació d'articles d'interès relacionats amb el treball, que per motius d'espai no s'han inclòs a la revisió de l'estat de l'art.

DOVE (enllaç)
Arquitectura:
DOVE = DOcking decoy eValuation scorE. La representació es fa a través de dos grafs per capturar les interaccions intermoleculars de proteïna-lligand. Ambdós grafs seran processats per una xarxa neuronal (GNN) per produir una puntuació, que és una probabilitat que la interacció tingui una qualitat CAPRI acceptable. CAPRI = Critical Assessment of PRediction of Interactions. El primer graf descriu els àtoms de la regió de la interfície i només considera els vincles covalents entre àtoms de residus d'interfície dins de cada subunitat com a arestes. El segon graf connecta la interacció de residus covalents (per tant, els inclou) i no covalents com a arestes, on un parell d'àtoms no covalents es defineix com aquells que estan més a prop de 10,0 Å entre si. Es representaran amb una matriu d'adjacència i vectors de característiques pels nodes.
L'agregació es fa amb un mecanisme de GNN a partir del mecanisme d'atenció, i les noves característiques dels nodes s'actualitzen tenint en compte els seus nodes veïns, a través d'una combinació lineal de les característiques del node veí amb el coeficient d'atenció final. Es generen quatre capes de GNN-DOVE per processar la informació d'incrustació dels nodes dels veïns i per generar la incrustació de nodes actualitzada. Es generen 4 capes FC per classificar si el model de complex proteic és correcte o incorrecte amb les funcions d'activació RELU excepte l'última que es fa amb Sigmoide. S'obté un valor de la interacció intermolecular, no una probabilitat. Per evitar l'overfitting es realitza un drop-out del 0,3 a cada capa excepte l'última FC.
DB:
Dockground dataset 1.0, ZDOCK dataset
Resultats:
S'ha avaluat el rendiment de GNN-DOVE al conjunt de dades Dockground. Es va comparar GNN-DOVE amb DOVE i altres cinc mètodes de puntuació de models d'estructura existents, com ara GOAP (Zhou i Skolnick, 2011), ITScore (Huang i Zou, 2008), ZRANK (Pierce i Weng, 2007), ZRANK2 (Pierce i Weng, 2007). , 2008) i IRAD (Vreven et al., 2011). Es forma GNN-DOVE en un conjunt de dades més gran i s'avaluen dos conjunts de dades més, inclòs el CAPRI Scoreset, que confirma un rendiment superior de GNN-DOVE als mètodes existents. Per avaluar la qualitat dels models d'estructura, tenint en compte les interaccions de diversos cossos (àtoms o residus) s'ha demostrat que és un model efectiu. Les GNN consideren patrons d'interaccions multiàtoms.

DUAL GNN (enllaç)
Arquitectura:
Es planteja un problema de predicció de la xarxa química com a problema de predicció d'enllaços en un graf de grafs (GoG). Hi ha dos tipus de capes: capes de convolució de grafs interns i capes de convolució de grafs externs. Es fa un graf on hi ha nodes i cada un dels nodes és un altre graf, referint-se al general com extern i a l'interior com intern. En el graf exterior, es representen compostos químics en els nodes, i si tenen interacció a les arestes. En els grafs interiors es representen els àtoms dels compostos i els lligams entre aquests àtoms. La intenció és obtenir una representació de característiques dels compostos amb el graf intern per tal de poder predir si hi haurà interacció entre els compostos en el graf extern.
Es proposa una xarxa neuronal convolucional de graf dual per a un GoG que consta de tres components: la capa de convolució de grafs interna, la capa de convolució de grafs externs i la capa de predicció d'enllaços. Les mides de les dues capes a la funció de predicció d'enllaç s'estableixen a 128 i 64, respectivament. S'implementa la xarxa convolucional de graf dual suggerit utilitzant Chainer i utilitzem ADAM com a optimitzador. La taxa d'aprenentatge s'estableix en 0,001.
S'utilitzen conjunts de dades de desenvolupament distribuïts per triar el nombre de dimensions de les representacions de grafs interns, de 32, 62, 128 i el nombre de passos de convolució T i L de 1, 3, 5. En xarxes externes denses, el nombre de convolucions externes sembla més important que el de les convolucions internes. S'utilitza un dropout de 0,2.
DB:
DrugBank database, SIDER2, version 4.1, KEGG LIGAND database, Release 62.0.
Resultats:
Tots els conjunts de dades estan desequilibrades pel que fa al nombre d'enllaços positius i negatius; per tant, es mesura el rendiment predictiu de cada mètode mitjançant (i) ROC-AUC que no es veu afectada pel desequilibri de l'etiqueta i (ii) PR-AUC que pot avaluar adequadament el rendiment en conjunts de dades desequilibrats. L'enfocament de doble convolució aconsegueix un alt rendiment de predicció tot i que les característiques eren de dimensions inferiors en comparació amb les característiques comercials en xarxes relativament denses. Mentre que el rendiment esdevé inferior en xarxes externes extremadament escasses a causa de la dificultat d'explotar la informació sobre les xarxes externes.

GNN-CNN (enllaç)
Arquitectura:
El disseny es basa en la tècnica d'aprenentatge de la representació d'extrem a extrem que consta de tres passos: (i) incrustar símbols d'entrada discrets, com ara paraules, en un espai vectorial de valor real de dimensions baixes, (ii) dissenyar diverses xarxes neuronals tenint en compte les estructures de dades (per exemple, seqüències i grafs) i (iii) l'aprenentatge de tots els paràmetres de la xarxa per backpropagation, inclosos els vectors d'inserció de símbols d'entrada discrets. S'utilitzen les representacions de compostos i proteïnes resultants d'una GNN i una CNN, que tenen la mateixa dimensionalitat, com a entrada per a un classificador per predir si interactuen.
La GNN utilitzada per a les molècules rep SMILES i els transforma en uns embeddings basats en subgrafs r-radis que són induïts pels vèrtexs veïns i arestes dins del radi r des d'un vèrtex. Es fan servir diferents funcions per a la transició de vèrtex i arestes i s'obté un vector de

representació molecular. CNN utilitzada per a les proteïnes que rep d'input embeddings basats en n-gram aminoàcids i ens proporciona un conjunt de vectors que acaben resumits en un vector de representació proteica. El CPI es calcula en una probabilitat (no en un valor) i obtenim a través d'un softmax classifier si interactuaran o no. Entrenada a través d'hiperparàmetres i gridsearch entre aquests.
DB:
DUD-E, DrugBank 4.1, Matador DB
Resultats:
Els experiments amb tres conjunts de dades de CPI van demostrar que l'enfocament d'extrem a extrem proposat aconsegueix un rendiment competitiu o superior en comparació amb diversos mètodes de predicció del CPI existents. Les representacions basades en dades de compostos i proteïnes obtingudes per GNN i CNN d'extrem a extrem són més robustes que les característiques químiques i biològiques tradicionals obtingudes a partir de bases de dades.

Graph-CNN (enllaç)
Arquitectura:
S'utilitza un graf-convolucional de dos passos. Al pas I, s'entrena un autocodificador de graf de pocket no supervisat en un conjunt de pocket drogable representatiu per conèixer les característiques generals de la pocket i incrustar butxaques de proteïnes en un espai latent de mida fixa. Al pas II, es construeix un pocket Graph-CNN i un lligand Graph-CNN per extreure característiques dels grafs de pocket i dels grafs de lligands 2D, respectivament. Per permetre que la xarxa reconegui diverses característiques de pocket, la pocket Graph-CNN s'inicia amb els pesos apresos del pas I. A continuació, la capa d'interacció integra característiques apreses del pocket i lligand Graph-CNN. Finalment, el classificador incorpora les interaccions apreses per realitzar prediccions vinculants. Al pas II, l'entrenament del model està impulsat per les etiquetes de classificació vinculants. Per tant, el model extreu automàticament les característiques específiques de la tasca que caracteritzen les interaccions entre l'objectiu i els lligands. A més, com que el model incorpora grafs de pocket i lligands per separat, el model no requereix complexos proteïna-lligand com a entrada.
DB:
No es menciona clarament.
Resultats:
Conjunt de resultats bons però no són millors que els anteriors ni hi ha res destacable.

IGT (enllaç)
Arquitectura:
L'IGT consta de tres mòduls, és a dir, un mòdul d'extracció de funcions, un mòdul de pas de missatges i un mòdul de lectura. El mòdul d'extracció de característiques extreu les característiques d'àtom i enllaç del lligand, el receptor i l'estructura complexa, respectivament. A continuació, les característiques extretes s'introdueixen als grafs corresponents al mòdul de pas de missatges, que consta de blocs IGT repetits en tàndem. A cada bloc IGT, adoptem una atenció al producte de punts conscient del gràfic per a cada gràfic i una atenció intermolecular per agregar tota la informació per actualitzar el gràfic complex. Les característiques del node de tres gràfics del bloc final s'alimenten al mòdul de lectura. Tots els missatges s'agreguen per l'operació d'agregació i es preveu la puntuació de l'activitat d'enllaç o la puntuació de la posició d'enllaç.
DB:
DUD-E, LIT-PCBA.
Resultats:
Han estat els resultats més bons obtinguts fins al moment i molt semblants al Novel-DTI. Hi ha un apartat aplicat al SARS-CoV-2 de manera exitosa.

Novel-DTI (enllaç)
Explicada i escollida al llarg de l'article.

ParaVS (enllaç)
Arquitectura:
En aquest article, es proposa un mètode basat en acoblament (ParaVS-Dock) i un mètode no basat en acoblament (ParaVS-ND) per a tasques SBVS, i s'estableix un marc que els contingui tots dos, tal com es mostra a la figura 1. S'avalua ambdós mètodes en dos grans conjunts de dades. L'objectiu és desenvolupar un mètode de cribratge virtual (VS) basat en l'acoblament i no basat en l'acoblament. Tenint en compte aquest objectiu, en la implementació, s'utilitza la GNN desenvolupada internament, HagNet.
DB:
DUD-E, NoDecoy, BindingDB.
Resultats:
A DUD-E s'aconsegueix un AUC de 0,981 i un factor d'enriquiment al 2% de 36,2; a NoDecoy s'aconsegueix un AUC de 0,974. Són resultats bons, però no són els millors i no es veu molt clar el procediment seguit per obtenir-los.

SAG-DTA (enllaç)
Arquitectura:
En aquest estudi es consideren dos tipus d'arquitectura pel que fa a l'estratègia d'agrupació, és a dir, l'arquitectura d'agrupació global i l'arquitectura d'agrupació jeràrquica. L'arquitectura d'agrupació global consta de tres capes convolucionals de grafs, i les sortides d'aquestes tres capes es concatenen abans d'introduir-se a una capa de SAG-Pooling, és a dir, agrupar-se de manera global. Els nodes restants passen per la capa de lectura i finalment es passen a capes completament connectades per a representacions de molècules de fàrmacs. L'arquitectura de pooling jeràrquica es compon de tres blocs, i cadascun d'ells conté una capa convolucional de gràfics i una capa de SAGPooling. Així, els resultats convolucionals de cada capa s'agrupen i es llegeixen jeràrquicament. Aquestes sortides se sumen abans de passar a les capes completament connectades per obtenir les representacions finals del fàrmac.
DB:
DAVIS, KIBA.
Resultats:
Els resultats experimentals mostren que SAG-DTA és el més precís entre els mètodes avaluats. En detall, el SAG-DTA global aconsegueix un MSE de 0,130 i un CI de 0,892, i el SAG-DTA jeràrquic aconsegueix un MSE de 0,131 i un CI de 0,893. Aquests resultats demostren l'eficàcia i la bona capacitat de generalització del nostre model en predicció de DTA. El model aconsegueix un rendiment superior al de diversos mètodes de predicció de DTA existents, cosa que suggereix l'eficàcia de l'enfoca-

ment proposat per predir l'afinitat dels parells de fàrmacs i proteïnes. A més, el bon rendiment de SAG-CPI, que és la versió CPI de SAG-DTA, demostra la bona capacitat de generalització del mètode proposat així com l'eficàcia dels mecanismes d'autoatenció.

SSNet (enllaç)
Arquitectura:
En aquesta xarxa es denota l'entrada com una matriu 2D (vector 1D amb curvatura i torsió remodelades per contenir curvatura en una fila i torsió en l'altra), X(0), on cada columna representa un residu únic i les files corresponents a la curvatura i la torsió. La primera capa és una convolució de branques amb diferents mides de finestra. És a dir, cada branca és una circumvolució amb un filtre de longitud diferent. Es fa aquesta operació perquè la xarxa neuronal pugui reconèixer patrons de longituds variables a la trama de descomposició. A continuació, cada branca s'alimenta de més circumvolucions de la mateixa mida de finestra. Això permet que la xarxa reconegui patrons més complexos en X(0) que poden ser més difícils de reconèixer amb una sola convolució. La sortida d'aquestes branques convolucionals es concatena, s'agrupa al llarg de la seqüència i s'alimenta a una capa densa totalment connectada. La branca superior més dreta de la mostra un vector lligand que es genera i s'alimenta a una capa densa completament connectada. La sortida d'aquesta capa es coneix normalment com a incrustació. Intuïtivament, aquesta incrustació és una representació de dimensionalitat reduïda de la proteïna i el lligand. Les sortides de la incrustació de proteïnes i la incrustació de lligands es concatenen i s'alimenten a capes més denses per predir el PLI.
DB:
DUD-E, DrugBank 4.1, Matador.
Resultats:
Al conjunt de proves DUD-E, SSnet:DUD-E té el millor rendiment amb un AUCROC mitjà de 0,97, seguit de prop per GNN-CNN amb 0,96 (taula S6). No obstant això, hi ha hagut crítiques contra els models ML entrenats en el conjunt de dades DUD-E pel que fa a l'overfitting al conjunt de dades.

A.2 Resultats de les prediccions del SARS-CoV-2 i anàlisi d'aquests

En aquest apèndix es mostren els resultats ampliats de la utilització del model creat pel SARS-CoV-2. A més, es veuen alguns dels exemples més importants de prediccions fetes que coincideixen amb principis actius estudiats per a la malaltia a la realitat, donant èmfasis al Remdisvir, l'únic medicament amb ús universal per a la proteïna 6M71 del virus.

RESULTATS DEL TOP 25 PER A CLASSIFICACIÓ I REGRESSIÓ SENSE I AMB AIGUA DEL SARS-COV-2

TOP 25 CLASSIFICATION HOH:

Id	Name	Pred
30464	6M71_33675	1,000000
21319	6VXX_5288615	1,000000
31605	6VXX_6914621	1,000000
2823	6VXX_4369	1,000000
7781	6VSB_445018	1,000000
4025	6M71_4368	1,000000
9060	6M71_5289389	1,000000
31836	6VXX_2120	1,000000
78	6VXX_5111	1,000000
3671	6M71_5288615	1,000000
13810	6M71_4369	1,000000
16451	6VSB_6914621	1,000000
29723	6VXX_16129579	1,000000
28975	6VXX_16117309	1,000000
22268	6VXX_11749858	1,000000
906	6VXX_4848	1,000000
15881	6VXX_6102708	1,000000
19382	6VXX_3342298	1,000000
20832	6VSB_5111	1,000000
5929	6M71_24768528	1,000000
30343	6VSB_466151	1,000000
21928	6M71_3342298	1,000000
22511	6VXX_179337	1,000000
6591	6VXX_11291932	0,999999
31903	6M71_5111	0,999999

TOP 25 CLASSIFICATION NO HOH:

Id	Name	Pred
20850	6VSB_206044	0,999999
4153	6LU7_206044	0,999999
29199	6M71_206044	0,999998
12502	6VXX_206044	0,999997
23160	6VSB_10913	0,999996
6463	6LU7_10913	0,999994
31509	6M71_10913	0,999993
14812	6VXX_10913	0,999991
11126	6VXX_449124	0,999986
27823	6M71_449124	0,999982
24039	6VSB_9888484	0,999982
28842	6M71_154000	0,999982
3796	6LU7_154000	0,999974
15691	6VXX_9888484	0,999972
25026	6VSB_5464097	0,999971
29637	6M71_6914621	0,999966
17214	6VSB_5362119	0,999965
30356	6M71_5288615	0,999960
30423	6M71_3815	0,999958
20493	6VSB_154000	0,999952
18558	6VSB_447522	0,999949
21288	6VSB_6914621	0,999947
7342	6LU7_9888484	0,999940
26907	6M71_447522	0,999938
9005	6VXX_656629	0,999935

TOP 25 REGRESSION HOH:

Id	Name	Pred
26097	6VSB_36294	0,031975
24421	6VSB_6032	0,034128
16479	6VSB_3037209	0,045046
10392	6VXX_446724	0,055385
29950	6VSB_6131	0,058564

33170	6VXX_440483	0,060164
22445	6LU7_68682	0,070776
7501	6LU7_9444	0,077661
26639	6VSB_657041	0,080403
29501	6VXX_439318	0,081639
25067	6LU7_6032	0,084980
30992	6VSB_5289086	0,085724
30872	6VXX_447657	0,117607
24025	6VXX_68682	0,119587
32344	6M71_129856752	0,119924
17938	6VXX_448726	0,141589
16182	6M71_31264	0,149456
15059	6VSB_65309	0,171162
10507	6VXX_60855	0,175056
21191	6M71_65309	0,176845
12690	6M71_445534	0,178723
18039	6VXX_72392	0,181849
16980	6VSB_68682	0,183409
4055	6VXX_5288994	0,185591
26988	6VSB_12803287	0,186071

TOP 25 REGRESSION NO HOH:

Id	**Name**	**Pred**
17264	6VSB_119031	0,692481
8916	6VXX_119031	0,729945
22867	6VSB_20055267	0,787674
25047	6M71_91443	0,833199
4023	6LU7_130731	0,888012
10789	6VXX_447271	1,055850
25613	6M71_119031	1,184784
28424	6M71_119055	1,284245
8350	6VXX_91443	1,545305
20507	6VSB_9831652	1,631738
3810	6LU7_9831652	1,666033
1528	6LU7_98792	1,677029
29361	6M71_3037981	1,690422
216	6LU7_44257	1,698163
20720	6VSB_130731	1,703516
14519	6VXX_20055267	1,710270
23832	6VSB_447268	1,933415
21012	6VSB_3037981	2,005775
23077	6VSB_5611	2,106167
27024	6M71_5288251	2,170672
21646	6VSB_6323250	2,192747
23815	6VSB_11376392	2,254511
27005	6M71_656989	2,280820
1959	6LU7_656989	2,281667
11745	6VXX_65074	2,302720

LLISTAT DE PREDICCIONS SARS-COV-2 COMPROVADES

Llistat d'algunes prediccions positives del top realitzades amb algun o varis dels models i que el seu funcionament està demostrat experimentalment o amb les investigacions molt avançades. Enllaç de l'estudi sobre el nom del medicament. N'hi ha bastants més, però únicament s'han destacat les més rellevants.

- 6VSB_36294 - Tobramycin - Predicció: 0,031 nM
- 6LU7_68682 - Arbekacin - Predicció: 0,070 nM
- 6VSB_20055267 - Enviomicin - Predicció: 0,787 nM
- 6M71_91443 - Tetrahydrofolic - Predicció: 0,833 nM
- 6M71_3037981 - Viomycin - Predicció: 1,690 nM
- 6VXX_439318 - Bekanamycin - Predicció: 0,082 nM

Mostra de tres medicaments que funcionament experimental i estan en ús, i que algun o varis dels models ha predit com a positius tot i no trobar-se en el top. Enllaç de l'estudi sobre el nom del medicament.

- 6M71_121304016 - Remdisvir - Predicció: 2,691 nM
- 6LU7_1549008 - Saquinavir - Predicció: 12,568 nM
- 6LU7_45375808 - Sofosbuvir - Predicció: 35,661 nM

Actualment, l'únic medicament completament aprovat i en ús regular pel SARS-CoV-2 és el Remdisvir, que se situa en el top 35 de la regressió sense aigua i que, si únicament agafem la proteïna amb la qual està demostrat que interacciona, es troba en el top 10. Aquest medicament va ser dissenyat durant més d'un any per experts en bioquímica i ha sortit com un dels candidats principals de gairebé 11500 medicaments, demostrant el potencial d'aquest model per a l'screening molecular. Per tant, si en un passat s'hagués disposat d'aquesta GNN, es podria haver descobert el reaprofitament del principi actiu molt abans.