ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA
Universitat Rovira i Virgili

PREDICCIÓ DE LA GED I DISTÀNCIES DE DESCRIPTORS MORDRED MITJANÇANT KNN A PARTIR DE GRAFS QUÍMICS

TREBALL FINAL DE GRAU ENGINYERIA INFORMÀTICA

Javier Vega Cuadros

Doble Grau Biotecnologia i Enginyeria Informàtica

Tutora acadèmica: Natàlia Segura Alabart

Departament: Departament d'Enginyeria Informàtica i Matemàtiques

E-Mail: natalia.segura@urv.cat

Tarragona, Juny 2025

1. Agraïments

Aquest treball de fi de grau és molt especial per mi, perquè significa tancar una etapa al voltant de cinc meravelloses persones que pràcticament puc considerar família. El doble grau m'ha fet passar dels millors i pitjors moments a la meva vida, i això va dedicat a ells, que han estat durant tot el camí. Gràcies Montse, Alba, Mario, Diego i Cristian, sempre us guardaré un raconet molt especial al meu cor. També al Josep, que per casualitat la vida va ajuntar els nostres camins i em va haver d'aguantar durant un quadrimestre a Eslovènia. Des de llavors és també de la família.

Ara em cal agrair a la meva família de veritat, gràcies papa per la teva rocambolesca manera de fer que sempre doni el 200% de mi, l'any que ve anirem al Camp Nou a celebrar-ho. Gràcies mama pel teu suport incondicional i creure en mi al llarg d'aquests llargs cinc anys, gran part del mèrit és teu també. Al "tete" per fer-me sortir de polleguera en qualsevol de les situacions. Tot i que hi ha maneres i maneres, jo no seria el mateix sense això. Gràcies "awelo" per cada dia recordar-me que tant tu com la iaia, estigui on estigui, celebrareu els meus èxits amb llàgrimes de felicitat. També als avis, que sé que sempre tindré una segona casa amb ells.

Vull agrair també als de sempre, no sé quants anys porto aguantant-los, però sé que una part de qui soc jo és degut a com són ells. Amb aquests no cal tenir enemics, però sé que puc confiar en ells amb la meva vida que sempre hi estaran. Visca la Jinga.

Aquest any ha estat tota una troballa per a mi el club voleibol Sant Pere i Sant Pau. M'ha permès gaudir d'un esport que vaig descobrir a Eslovènia i del qual ara m'he enamorat. Gràcies al Dani i tots aquests que en formen part, ja que no seria el mateix sense vosaltres.

Gràcies Clàudia per, de manera incondicional, escoltar els meus problemes i motivar-me a seguir endavant. Ets una persona meravellosa que se que, passi el temps que passi, sempre ens podrem fer costat l'un a l'altre.

Gràcies Marina per ser com ets, per aguantar-me durant aquest procés i per estar present dia a dia. En poc temps has fet que tingui plena confiança en tu i em pugui dissuadir dels contratemps escoltant música i parlant amb tu.

Per finalitzar, vull agrair a la Natàlia, la meva tutora. Gràcies a la teva paciència i versatilitat, i gràcies per confiar en mi per aquest projecte i fer-lo encara més enriquidor amb les reunions setmanals i les meves aparicions al laboratori. A part d'una gran tutora, ets una persona amb un gran cor, i sempre t'has preocupat per la meva situació. Sense aquest suport no hauria sigut possible acabar aquesta tesi.

Índex

2. Resum

Aquest treball de final de grau ha analitzat la predicció de la distància entre molècules químiques mitjançant dues metodologies: la Graph Edit Distance (GED) i els descriptors moleculars Mordred, combinades amb l'algoritme de K-Nearest Neighbors (KNN). S'han fet servir dues bases de dades (ESOL i FreeSolv) i dos enfocaments per calcular la GED: un mitjançant CPU (amb la llibreria NetworkX) i un altre mitjançant GPU (amb l'algorisme Fast Bipartite). Els resultats obtinguts mostren que el càlcul amb GPU és molt més eficient en temps (menys d'una setmana vs. tres setmanes amb CPU) tot i que els valors calculats han variat una mica respecte als de CPU. Els models KNN entrenats amb dades de GED calculades per GPU van obtenir en general millors resultats que els de CPU, amb errors quadràtics mitjans (MSE) i errors absoluts mitjans (MAE) més baixos, així com coeficients de determinació (R^2) més elevats. De totes maneres, els millors models han estat els que s'han entrenat amb els valors de GED que majoritàriament coincidien entre els algorismes de CPU i GPU, en dos subdatasets, un per a cada base de dades. També es va identificar una correlació entre les distàncies Mordred i la GED, fet que obre la porta a predir la similitud dels grafs moleculars a partir de descriptors més fàcilment computables. En conclusió, el treball demostra que és viable predir distàncies moleculars de manera eficient amb KNN i empremtes moleculars, i proposa una eina útil per la quimioinformàtica basada en el codi desenvolupat.

3. Paraules Claus

Molècules químiques, grafs, distància d'edició de graf, aprenentatge automàtic supervisat, K-Nearest Neighbors.

4. Abreviatures

- IUPAC → International Union of Pure and Applied Chemistry.
- SMILES → Simplified Molecular Input Line Entry System.
- CPU → Central Processing Unit
- GPU → Graphics Processing Unit
- GED → Graph Edit Distance, distància d'edició de grafs.
- KNN → K-Nearest Neighbors.
- MSE → Eror quadràtic mitjà.
- MAE → Error absolut mitjà.
- R^2 → Coeficient de determinació.

5. Definicions

Graph Edit Distance: Distància d'edició de graf. Donats uns costos associats a la inserció, deleció o substitució de nodes i arestes, es calcula amb els canvis necessaris per convertir un graf en un altre.

K-Nearest Neighbors: Mètode d'aprenentatge automàtic supervisat que permet realitzar regressions o classificacions de les dades en funció del nombre de veïns més propers amb relació a les dades.

Cross Validation: Mètode que serveix per provar el rendiment d'un mètode d'aprenentatge automàtic. Serveix per afinar els paràmetres del model, normalment mitjançant una k-fold, dividint el conjunt de dades en k subconjunts, dels quals un serà de test i l'altre d'entrenament.

Outliers: Dades que no segueix les tendències del dataset i provoca possibles errors en les prediccions per tant en la precisió del model.

Empremta molecular: Vectors que codifiquen la presència de subestructures dins una molècula. S'utilitzen per representar-la de manera compacta i facilitar la comparació entre molècules. Un exemple són les Morgan Fingerprints.

Descriptor molecular: Vectors numèrics que quantifiquen propietats físiques i químiques d'una molècula. Serveixen per representar-la matemàticament en models computacionals. Un exemple són els descriptors Mordred.

6. Introducció

6.1 Context químic

La química i la bioquímica són dues ciències molt presents en el nostre dia a dia. Tot i que passen desapercebudes, les molècules i estructures químiques són la bastida de tot el que coneixem avui en dia. Els éssers humans estem formats per milers de molècules orgàniques, cadascuna amb la seva funció, que ens donen les propietats i característiques que ens donen vida [1]. Les funcions de les molècules químiques en els organismes biològics en són molt variades. Hi ha molècules estructurals, molècules energètiques, de transport... La funció d'aquestes biomolècules ve donada principalment per la seva composició i estructura química [2]. Per exemple, proteïnes semblants acabaran realitzant funcions semblants entre diferents organismes. Això ho sabem gràcies a l'evolució i la informació genètica que tenim, un altre cop, codificada en molècules químiques [3].

6.2 Representació de molècules químiques com grafs

Les molècules estan formades per àtoms units entre si mitjançant enllaços químics. La mida i forma de la molècula depèn de quants àtoms i enllaços conté. Es pot representar l'estructura d'una molècula química com un graf no dirigit i etiquetat (figura 1.1), on els nodes són els àtoms, i les arestes són els enllaços químics. Les molècules poden ser representades de diverses maneres (taula 1). Les més senzilles només proporcionen informació dels àtoms continguts en cada molècula, com per exemple les fórmules empírica i molecular de la primera columna. Després estan les representacions simplificades, com per exemple la nomenclatura SMILES (Simplified Molecular Input Line Entry System)[4]. Aquestes nomenclatures es fan servir en sistemes informàtics per simplificar en caràcters ASCII tota aquesta informació tridimensional de les molècules, com es pot veure a la segona columna. Seguidament, es troben les representacions espaials (tercera columna taula 1) que si contenen que si contenen informació tridimensional i són més semblants a una representació en forma de graf. Aquestes estructures i nomenclatures segueixen la regulació i especificacions de la IUPAC (International Union of Pure and Applied Chemistry)[5], útils per als éssers humans. L'avantatge de la representació SMILES respecte a les altres és que, tot i ser més complicades d'interpretar per a un ésser humà, estan pensades per representar les molècules a programes quimioinformàtics per traduir-les a una estructura de graf bidimensional.

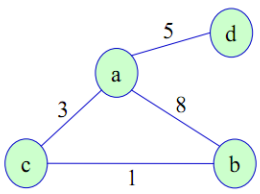
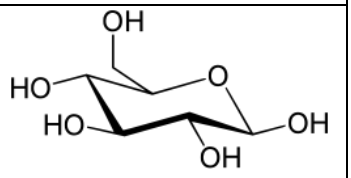
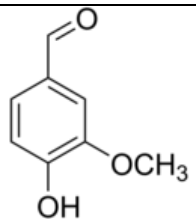
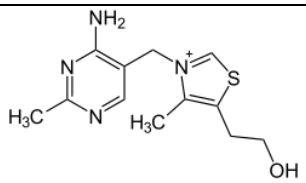


Figura 1.1 Exemple de graf no dirigit i etiquetat. Font: http://wwwestruc-grafosrosy.blogspot.com/2011/06/grafo-etiquetado.html

Taula 1. Conté els tres tipus de representacions de les molècules esmentades per a les molècules de la Glucosa, la Vanil·lina i la vitamina B1. Font: https://es.wikipedia.org/wiki/SMILES

Molècula (Formula Molecular)	SMILES	Estructura (IUPAC)
Glucosa (glucopiranosa) ($C_6H_{12}O_6$)	OC[C@@H](O1)[C@@H](O)[C@H](O)[C@@H](O)[C@@H](O)1	
Vanil·lina ($C_8H_8O_3$)	O=Cc1ccc(O)c(OC)c1	
Tiamina (vitamina B1) ($C_{12}H_{17}N_4OS+$)	OCCc1c(C)[n+](=cs1)Cc2cnc(C)nc(N)2	

Finalment, estan les representacions purament tridimensionals. Aquestes fan servir programes específics, com per exemple en podrien ser Jmol [30] o RasMol [31]. Donada una molècula, sobretot proteïnes, es pot visualitzar la seva conformació en l'espai. Arrossegant el ratolí, podem rotar la molècula, escalar-la, d'entre altres. Amb comandes es pot escollir el tipus de representació, amb esferes, filferro... segons el que interessi visualitzar (Figura 2). De totes maneres, no és d'interès per aquesta tesi aquest tipus de representacions.

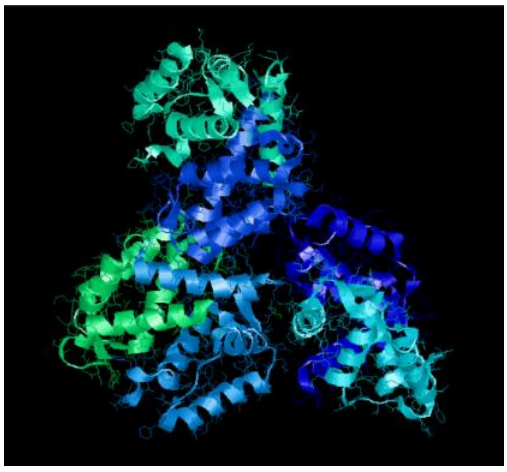


Figura 2. Representació tridimensional des de l'aplicació JMol de la proteïna fosfodiesterasa humana. Font: Autor

Aquestes representacions moleculars es poden representar en el món quimioinformàtic en un graf a través del seu codi SMILES. Un graf, com s'ha esmentat abans, és una estructura de dades especialitzada a desar un conjunt d'objectes (nodes) interrelacionats mitjançant un conjunt de connexions (arestes). Aquesta estructura facilita trobar aquestes connexions i la seva forma en ser representades. Una limitació dels grafs respecte a les molècules químiques és la seva dimensionalitat. Aquesta estructura està normalment representada de manera bidimensional. Si fem servir un graf per representar una molècula química estem perdent informació de cara a la conformació tridimensional de la molècula, molt important per definir segons quines interaccions pot arribar a tenir amb altres molècules [6]. De totes maneres, es pot guardar informació sobre les arestes que facilitin aquesta interpretació tridimensional.

6.3 Mètodes per comparar la distància entre molècules

Els grafs, igual que d'altres estructures de dades, tenen operacions pròpies que permeten extreure informació d'aquests. En el cas de les molècules químiques, podem modelitzar-les com grafs i calcular la seva distància d'edició de graf o graph edit distance (figura 1.2). La GED és un algoritme específic per grafs que permet calcular la distància entre dues estructures modificant els nodes i les arestes per convertir l'una en l'altra. Assignant uns costos d'inserció, substitució i deleció de nodes i arestes podem obtenir un valor numèric que, com més petit, implicarà que la similitud entre ambdues estructures serà major [7]. D'aquesta manera podem calcular com de properes són dues molècules químiques mitjançant la GED dels seus grafs. Aquest serà un dels dos mètodes que seran tractats per trobar la distància entre dues molècules.

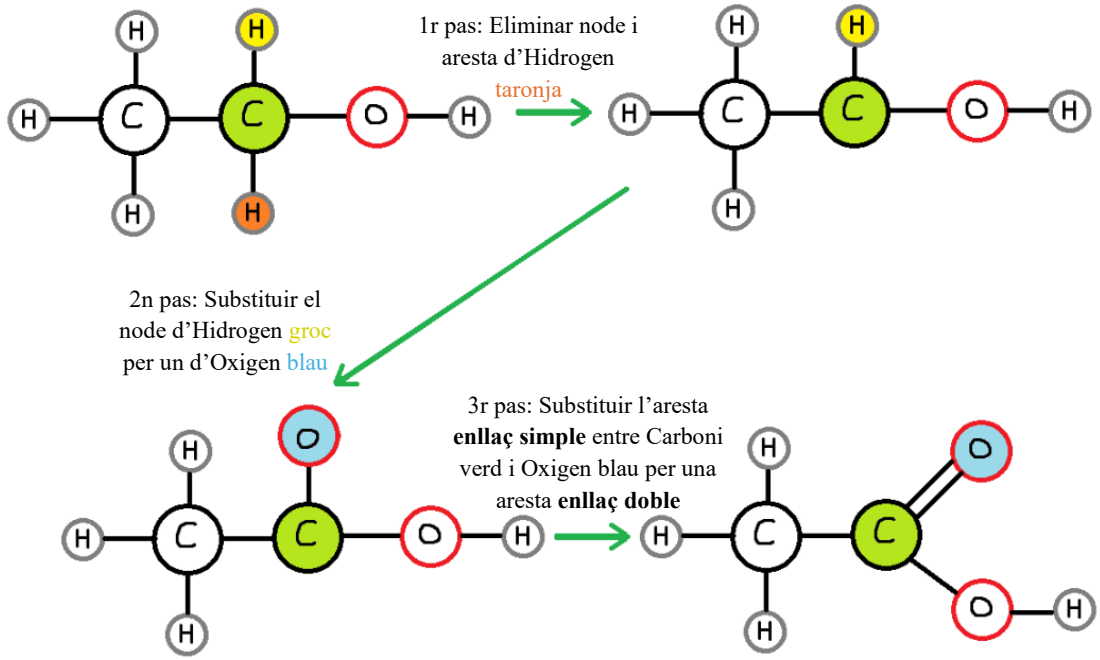


Figura 1.2. Representació del càlcul de la GED amb els passos necessaris per passar dels grafs d'etanol a àcid acètic. Font: Autor

Hi ha diverses implementacions de la GED, però en aquesta tesi no es mencionaran totes, l'enfocament principal serà en la implementació de la llibreria de Python NetworkX [7] i en l'algorisme Fast Bipartite [8]. En el cas general, el problema és NP-Complet [9], ja que no es pot trobar una solució òptima en temps polinomial, per tant, el que es fa és aproximar un resultat local que creiem pot ser subòptim (mínim local). Hi ha casos, però, que converteixen el problema de la GED en un problema que pertany al conjunt de problemes P, resolubles en temps polinomial. Això ocorre quan els grafs d'entrada són arbres ordenats.

Una altra aproximació per trobar la distància entre dues molècules és l'ús de la distància euclidiana entre els vectors de descriptors moleculars. Els descriptors moleculars són el resultat de representar matemàticament la informació química de l'estructura molecular a partir del seu codi SMILES amb una posterior estandardització [10]. Alguns exemples de descriptors moleculars bidimensionals en són el nombre d'àtoms, d'enllaços, energies d'ionització... Les eines actuals per al càlcul de descriptors moleculars més comunes són PaDEL-Descriptors i Mordred. En el cas de Mordred, aquest surt de solucionar alguns errors que presenta la implementació PaDEL, per tant, és el mètode que serà usat en aquesta tesi [11]. Mordred pot calcular més de 1800 descriptors moleculars, generant un vector o matriu que representa numèricament les propietats d'una molècula. Aquesta eina és ràpida, eficient i fàcil d'utilitzar, ja que només cal instal·lar una llibreria a Python. Els descriptors inclouen informació com el pes molecular, la topologia o la polaritat, i poden fer-se servir en models computacionals per comparar molècules. Per exemple, es pot aplicar la distància euclidiana entre vectors de descriptors per mesurar la similitud entre dues molècules de manera quantitativa. Com més petita sigui aquesta distància euclidiana, més properes seran químicament les molècules que estiguem comparant.

$$Distancia\ Euclidiana = \|v_1 - v_2\| = \sqrt{\sum_{i=1}^{n} (v_{1i} - v_{2i})^2}$$

Equació 1: Fórmula de la distància euclidiana per a un vector n-dimensional.

6.3 Mecanismes de predicció de la similitud entre molècules

Donades dues molècules químiques, podem extreure la seva informació respecte a la seva estructura i característiques que les fan úniques. Similar als descriptors moleculars, estan les empremtes moleculars o molecular fingerprints [12].

Les empremtes moleculars són representacions binàries o vectorials que codifiquen la presència de fragments estructurals en una molècula. En concret, les Morgan fingerprints, també conegudes com a ECFP (Extended Connectivity Fingerprints) [13], són un tipus popular de fingerprints que funcionen examinant l'entorn de cada àtom a un determinat radi, generant identificadors únics per a cada subestructura (Figura 3). Aquests identificadors es converteixen en un vector de bits (per exemple, de 1024 o 2048 dimensions) on cada bit indica la presència o absència d'un fragment específic. En cas de

repetir-se la subestructura, s'incrementa el seu pes. Aquesta és una forma compacta i eficient de representar l'estructura d'una molècula, i és especialment útil en tasques de trobar semblança molecular, classificació i agrupació, però pot ser difícil d'entendre per als humans.

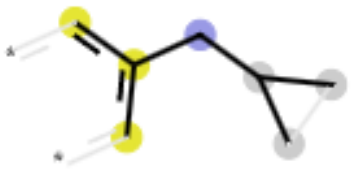


Figura 3. Imatge de la llibreria rdkit que mostra els bits de la empremta Morgan de la molècula ciclopropilmetilbenzè. En color blau es pinta l'àtom central de l'entorn específic, en groc àtoms aromàtics (prop de l'anell benzènic) i en gris àtoms alifàtics. Font: https://www.rdkit.org/docs/_images/mfp2_bit872.svg

Aquests vectors poden fer-se servir perquè un model d'intel·ligència artificial aprengui les subestructures de les molècules. Donades dues molècules, segons com siguin de semblants les seves empremtes moleculars, major similitud hi haurà en l'estructura d'aquestes, per tant menor la distància entre els grafs moleculars.

El propòsit és fer servir aquestes empremtes moleculars per predir, entre dues parelles de molècules químiques, quina serà la seva distància molecular. Per això cal explorar les alternatives que permeten realitzar aquestes prediccions mitjançant algun mecanisme d'intel·ligència artificial. Dins del camp de la intel·ligència artificial, hi ha mètodes que permeten entrenar sistemes per aprendre de les dades proporcionades sense haver de programar-ho a la força. Aquest camp de la intel·ligència artificial s'anomena Machine Learning o aprenentatge automàtic [14]. Dins del camp de l'aprenentatge automàtic, hi ha diverses estratègies, de les quals se'n poden diferenciar de dos tipus, aprenentatge supervisat i no supervisat.

L'aprenentatge automàtic supervisat es diferencia del no supervisat en el fet que el model s'entrena amb dades etiquetades, és a dir, coneixent prèviament la resposta correcta de la predicció. L'aprenentatge no supervisat treballa amb dades sense etiquetar i busca patrons o agrupacions de manera autònoma [15]. Dintre dels mecanismes d'aprenentatge supervisat, hi ha diferents implementacions, algunes més focalitzades en predicció, com és el cas de la regressió lineal o la K-Nearest Neighbors, mentre que les de classificació poden ser les Support Vector Machines (SVM) o regressions logístiques.

En aquesta tesi es fa servir el mètode d'aprenentatge supervisat de predicció KNN. Aquest és el mètode que ha estat escollit com a proposta de treball de fi de grau, ja que és robust per a realitzar aquestes prediccions, tot i que altres alternatives podrien haver sigut estudiades. L'aprenentatge supervisat per KNN consisteix en el càlcul de la distància euclidiana de la nostra nova instància a les dades amb què hem entrenat el model [16][17]. Posteriorment, segons el valor del paràmetre K, ens quedem amb aquelles K dades veïnes que són més properes a la nostra entrada. En cas de voler realitzar una classificació, es

mira de totes les K dades veïnes escollides, quina és la classe majoritària entre elles. En el cas de voler realitzar una regressió, simplement es calcula el valor mitjà de la variable de les instàncies que engloben el cercle de classificació, com es pot observar en la figura 4.

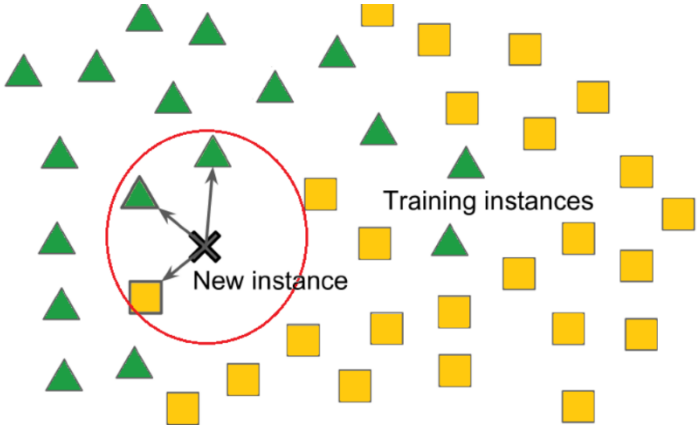


Figura 4. Imatge de com es realitza una predicció d'una instancia nova (conjunt de test) donades les instàncies d'entrenament i el valor del paràmetre k (=3). Font: https://towardsdatascience.com/cross-validation-using-knn-6babb6e619c8/

El valor del paràmetre K no té una manera inequívoca per calcular-ho. No és intuïtiu tampoc esbrinar quin serà el nombre de veïns òptim tot veient les dades representades. El procediment estàndard per trobar el valor d'aquest paràmetre és el de la cross-validation. Aquest algorisme utilitza un mecanisme que divideix el conjunt d'entrenament en subconjunts d'entrenament i de test dins d'aquest. Tot això depèn d'un paràmetre que es coneix com a k-fold [18]. Per exemple, amb un k-fold de 4, es divideix el conjunt d'entrenament en 4 subconjunts idèntics, dels quals s'escull un que farà la funció de test. Aquest subconjunt de test anirà rotant, fins a acabar entrenant i predient totes les possibles combinacions. Per tant, el que estem aconseguint és calcular la precisió del model donat un valor de K, que definirem dins d'un interval que ens interessi, per exemple, d'1 a 20 veïns propers. Posteriorment, s'escull el valor de K que ha obtingut millor precisió durant el cross-validation. Aquest valor de K és el que es farà servir per entrenar el model amb el conjunt total d'entrenament, i posteriorment realitzar les prediccions amb el conjunt total de test.

Un cop obtingut el valor de la predicció, com ja disposem del valor real, es comprova la certesa d'aquesta predicció. Hi ha diferents mètodes per comprovar com de lluny està la resposta del model respecte de la resposta veritable. Les loss functions o les funcions de pèrdua mesuren com de lluny estan les prediccions d'un model d'IA de les respostes correctes reals [19] [20]. Compara la sortida del model amb el valor real i calcula un nombre, anomenat "pèrdua", que mostra la mida de l'error. Quan el model és precís, la pèrdua és baixa. Quan el model fa males prediccions, la pèrdua és alta. Es tracta d'actualitzar els paràmetres que requereixi el model per tal d'ajustar-ho millor a les dades que ens arriben d'input i així minimitzar aquesta funció de pèrdua. Trobar el mínim d'aquesta funció que pot dependre de diverses variables es fa mitjançant el càlcul del gradient:

$$Vector\ Gradient = \nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right)$$

Equació 2: Fórmula del vector que representa el gradient d'una funció mitjançant les derivades parcials, aquest cas en 3 dimensions

On x, y i z serien variables d'aquesta funció.

Cal anar amb cura, perquè si se sintonitza completament el nostre model a les dades d'entrada, patim el risc del que es coneix com a overfitting. Aquest fenomen ocorre quan el model "memoritza" les dades d'entrenament i posteriorment realitza prediccions en base a copiar el que li ha estat ensenyat i això implica que no hi ha hagut cap mena d'aprenentatge [21]. Això és un greu problema, ja que perdem generalització. En aquest cas, el nostre model no seria capaç de realitzar prediccions prou precises o fiables fora del conjunt d'entrenament.

Dins de les possibles funcions de pèrdua per a regressions, hi ha la Mean Squared Error, o MSE, la Mean Absolute Error o MAE i la R^2 Score.

La MSE o error quadràtic mitjà és una funció de pèrdua que calcula la quantitat d'error que hi ha entre les prediccions d'un model d'aprenentatge automàtic i els resultats reals mitjançant la mitjana de les diferències dels quadrats entre ells [22]. Elevar al quadrat les diferències fa que els errors més grans guanyin més pes, cosa que ajuda el model a centrar-se a corregir grans errors. Prenent la mitjana s'assegura que l'error s'ajusta pel nombre de punts de dades, donant una mesura normalitzada de precisió general. La fórmula de la MSE és:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

Equació 3: Fórmula del Mean Squared Error.

On n és el tamany del data set, y_i és el valor predit per l'entrada i, i el valor $\bar{y}$ és el valor real corresponent a aquella entrada.

La MAE o error absolut mitjà, també anomenat L1 Loss, és una funció de pèrdua que mesura la mitjana de les diferències absolutes entre valors predits i valors reals [23]. A diferència de la MSE, no eleva al quadrat els errors, de manera que tots els errors, grans o petits, es tracten amb la mateixa importància. La MAE té com a fórmula:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \bar{y}|}{n}$$

Equació 4: Fórmula del Mean Absolute Error.

On n és el tamany del data set, y_i és el valor predit per l'entrada i, i el valor $\bar{y}$ és el valor real corresponent a aquella entrada.

La R^2 o coeficient de determinació pren un valor que va de 0 a 1. Aquest indica que tan bé un model explica la variabilitat de la predicció [24]. Un valor de 0 significa que el

model no explica res de la variació, mentre que un valor d'1 indica que explica tota la variació. Un R^2 igual a 1 representa un ajust perfecte, la qual cosa implica que el model prediu amb total precisió la variable dependent a partir de les independents. No obstant això, un valor de R^2 alt no garanteix que el model sigui bo. Només mostra que hi ha una forta relació lineal entre les variables independents i la dependent, però no avalua si els supòsits del model es compleixen ni si les variables realment tenen poder predictiu. Per això, és important complementar-ho amb altres mètriques per a tenir una visió completa del rendiment del model. Aquesta funció representa R^2, on el numerador és el sumatori de residuals al quadrat, i el denominador l'error quadràtic mitjà.

$$R^2 = 1 - \frac{\sigma_r^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Equació 5: Fórmula del coeficient de determinació o R^2

7. Hipòtesi de treball i objectius

La hipòtesi del treball consisteix en el fet que podem convertir estructures moleculars a grafs a partir del seu codi SMILES. A partir de bases de dades de molècules químiques es podran obtenir els grafs moleculars corresponents. Un cop hem obtingut els grafs, es podran efectuar operacions característiques d'aquests com insercions, delecions i substitucions. Una de les operacions característiques més complexes també pot ser calcular la GED entre dos grafs distints. Si aparellem dos a dos les molècules de les bases de dades, obtindrem parelles de grafs sobre les quals es podrà calcular la GED.

El principal objectiu d'aquest treball és observar com diferents aproximacions per al càlcul de la GED poden variar en entrenar un model d'intel·ligència artificial d'aprenentatge supervisat com és la KNN. A partir d'aquí, com que calcular la GED pot ser un problema d'elevada complexitat computacional, també es vol facilitar un repositori que contingui els datasets i el codi per facilitar l'entrenament d'aquests models d'intel·ligència artificial per facilitar l'obtenció de prediccions fidels i robustes de GED a partir de dos grafs químics.

També, es vol observar si hi ha alguna relació entre la GED, característica dels grafs, i la distància entre vectors de descriptors Mordred, que són característics de les molècules químiques. En cas que la correlació sigui forta, es calcularà quina regressió ajusta millor els valors de GED donada la distància Mordred.

8. Metodologia

8.1 Planificació del Projecte

8.1.1 Planificació de les tasques

Per tal de demostrar la hipòtesi com a certa i assolir els objectius d'aquest treball de fi de grau, es proposa muntar un Jupyter Notebook. Primerament, cal buscar bases de dades de molècules que continguin la informació de la molècula en format SMILES, i també en algun format comprensible per als humans, com per exemple l'identificador IUPAC. Un cop tinguem prou bases de dades, el primer pas serà un procediment per calcular la GED entre parelles de molècules a la mateixa base de dades. Per a n molècules dins de la base, finalment hi hauria $\frac{n(n-1)}{2}$ parelles de molècules. Com que això incrementa la dimensió de les dades a analitzar, es provaran dues implementacions per al càlcul de la GED.

El primer, implementat sobre CPU, fent servir l'algorisme de la llibreria NetworkX de Python. Com aquest algorisme explora diferents solucions, el temps de càlcul creix exponencialment respecte a la quantitat de molècules. Per evitar temps d'execució excessius, s'ha decidit posar un temporitzador de 5 minuts per cada parella de molècules. La segona implementació serà la de l'algorisme Fast Bipartite, implementat sobre la

tecnologia GPU de la llibreria CuPy de Python, aprofitant la tecnologia CUDA de la que disposa la targeta gràfica de l'ordinador on es realitzaran els càlculs.

Un cop obtinguts els datasets amb el valor de la GED per cada parell de molècules, s'implementarà el mecanisme d'aprenentatge automàtic supervisat de la KNN, fent servir les llibreries de Scikit-learn. Per tal de processar les molècules en format SMILES, es farà servir la llibreria de rdkit, específicament rdkit Chem. D'aquesta es podran extreure les dades de les Morgan Fingerprints a partir del codi SMILES de cada parella de molècules. El vector de bits de cada molècula serà utilitzat per entrenar el model de la KNN. Per decisions de disseny, es farà servir dels propis dataframes obtinguts a partir de la GED el 70% per entrenar el model, i el 30% per realitzar el testing i les prediccions. El valor del millor paràmetre K del model serà escollit mitjançant cross-validation. S'entrenarà un model per cadascuna de les implementacions de la GED, tant per l'algorisme de CPU com per al de GPU. Finalment, s'obtindran les mètriques o funcions de pèrdua per a cadascuna de les prediccions i s'avaluarà la precisió d'ambdós mètodes.

8.1.2 Tecnologia utilitzada

Com s'ha esmentat abans, s'utilitzarà la versió 3.13 de Python. Les llibreries més rellevants utilitzades per implementar la solució, junt amb les que han estat mencionades prèviament, han estat les següents.

NumPy 2.2.5: Processament de càlculs matemàtics i matricials

CuPy 13.4.1: Igual que NumPy, però accelerat per a GPU.

Pandas 2.2.3: Construcció i tractament dels datasets a partir de les bases de dades.

NetworkX 3.4.2: Eines de treball amb grafs i GED.

RDKit 2024.9.6: Obtenció de dades químiques a partir de la nomenclatura SMILES present als datasets, com per exemple el graf molecular o el Morgan Fingerprint.

Scikit-learn 1.6.1 / scipy 1.15.2: Llibreria principal que implementa els models d'aprenentatge automàtic, traint/test split, rutines per al cross-validation i posterior avaluació del model amb els errors esmentats prèviament.

Matplotlib 3.10.1 i Seaborn 0.13.2: Construcció i embelliment dels gràfics per avaluar visualment els models.

Itertools i Multiprocessing: Per incloure paral·lelisme.

El projecte serà desenvolupat en scripts de python. També hi haurà disponible una versió en Jupyter Notebook a un repositori públic al GitHub:

Per als càlculs de la GED i les distàncies Mordred s'utilitzarà un ordinador específic per a realitzar càlculs computacionals costosos, pertanyent al laboratori del grup d'investigació de la tutora. Les característiques de l'ordinador són les següents:

- CPU: AMD Ryzen 9 7950 x3d 16 Cores 32 Threads.
- GPU: NVIDIA RTX A4000 16 GB GDDR6
- RAM: 128 GiB
- Disc: 10 TB SSD
- Sistema Operatiu: Ubuntu 22.04.4 LTS 64bit

8.1.3 Bases de dades de molècules químiques.

Per aquesta tesi s'han escollit les bases de dades de ESOL i FreeSolv.

8.1.3.1 ESOL

La base de dades ESOL conté un recull de molècules químiques d'interès químic per al descobriment de fàrmacs. El fitxer conté un conjunt de dades de predicció de solubilitat en aigua que ocupa uns 95 kB i consta de 1.128 mostres, per tant, un total de 635.628 parelles de molècules [25].

Conté les següents columnes:

"Component ID" - Nom IUPAC del compost.

"ESOL predicted log solubility in mols per litre" – Predicció del logaritme de la solubilitat de la molècula en aigua en mols/L.

"Molecular Weight" – Pes molecular en g/mol.

"Number of H-Bond Donors" – Nombre de donadors d'enllaços d'hidrogen.

"Number of Rings" – Nombre d'anells de la molècula.

"Number of Rotatable Bonds" – Nombre d'enllaços amb rotació lliure.

"Polar Surface Area" – Superfície de la molècula amb propietats polars.

"measured log solubility in mols per litre" – Mesura empírica del logaritme de la solubilitat de la molècula en aigua en mols/L.

"smiles" – Representació SMILES de l'estructura molecular.

8.1.3.2 FreeSolv

La base de dades de solvatació gratuïta, FreeSolv (SAMPL), proporciona energia lliure d'hidratació experimental i calculada de petites molècules en aigua. Aquesta ocupa uns 32 kB, i conté un total de 642 entrades moleculars, el que donarà una combinació de 205.761 parelles de molècules [26].

Els valors calculats es deriven de càlculs d'energia lliure alquímica mitjançant simulacions de dinàmica molecular.

La base de dades conté les següents columnes:

"iupac" – Nom IUPAC del compost.

"smiles" – Representació SMILES de l'estructura molecular.

"expt" – Energia de solvatació mesurada (unitat: kcal/mol) del compost, utilitzada com a etiqueta.

"calc" – Energia de solvatació calculada (unitat: kcal/mol) del compost.

8.1.3.3 Comparativa entre ambdues bases de dades

Realitzant una comparativa d'ambdues bases de dades, tal com es pot observar en la figura 5 es pot observar que aproximadament menys de la meitat de les molècules de FreeSolv estan presents a ESOL. Això és perquè l'article on es menciona ESOL cita a l'article on es menciona FreeSolv, com una de les bases de dades font de molècules per a la creació d'ESOL.

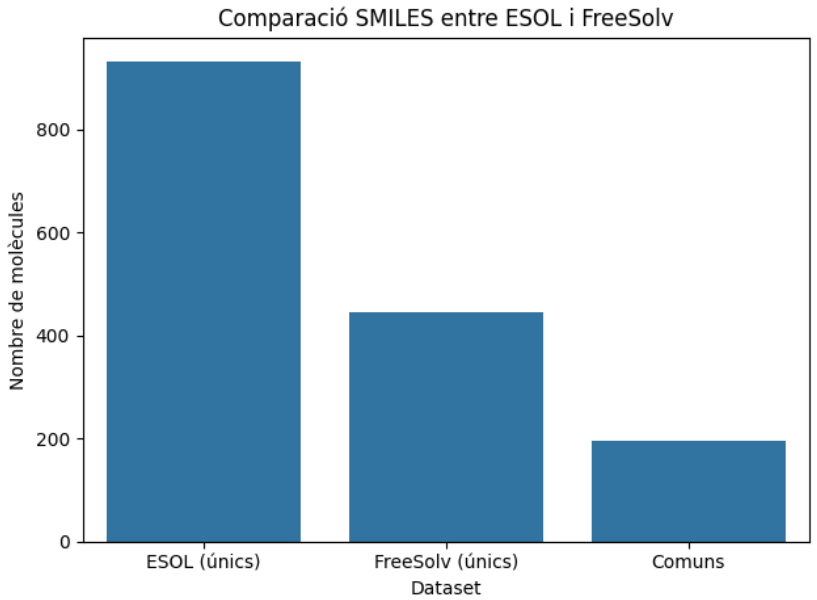


Figura 5. Gràfic de barres comparant els continguts entre ambdues bases de dades.

8.2 Implementació de la proposta

8.2.1 Implementació del càlcul de la GED

Primerament, caldrà generar els datasets de parelles de molècules donada una base de dades. Com s'ha esmentat prèviament, cada molècula de la base de dades se li calcularà la GED respecte a la resta de molècules, per tal de maximitzar la quantitat de dades de què disposem per posteriorment entrenar els nostres models. A partir del codi SMILES d'ambdues molècules, es trobarà el graf corresponent a les seves estructures químiques. Un cop disposem de les parelles de grafs al dataset, caldrà calcular la GED entre elles.

Els costos d'edició, ja que la implementació NetworkX no accepta valors decimals, seran els següents:

- Inserció de nous nodes o arestes: 1
- Deleció de nodes o arestes: 1
- Substitució del tipus d'àtom o enllaç: 2

Computant el nombre de canvis que cal realitzar per passar d'un graf molecular a un altre amb aquests pesos definits ens donarà el valor de la GED.

Tal com s'ha mencionat a la introducció, es faran servir dos dels possibles mecanismes de càlcul de la GED com a manera de calcular la distància entre dues molècules. Aquests són l'algorisme de la llibreria NetworkX, i l'algorisme del Fast Bipartite.

8.2.1.1 Algorisme de NetworkX

El primer algorisme proposat és el de la llibreria NetworkX a Python, que combina el millor de l'algorisme A* i DFS. DF-GED explora l'espai de totes les possibles correspondències entre dos grafs mitjançant un arbre ordenat. Aquest arbre de cerca es construeix dinàmicament en temps d'execució creant iterativament nodes successors enllaçats per arestes al node considerat actualment a l'arbre de cerca. Aquesta solució utilitza una heurística Best-First, per tant, hi ha prou solucions candidates per a saturar la memòria ràpidament. Per evitar aquest problema, fan servir upper i lower bounds que calculen fent una cerca en profunditat prèvia per decidir si seguir explorant una fulla o no mitjançant l'algorisme de Munkres o Hungarian Algorithm, que es pot utilitzar per trobar aparellaments de pes màxim en grafs bipartits, amb un cost $O(n^3)$ [27]. A més a més, també permeten afegir un *timeout*, que para l'execució, retornant la millor GED trobada explorant l'espai de solucions.

$$C_v = \left[\begin{array}{cccc|cc} c_{1,1} & \cdots & \cdots & c_{1,m} & c_{1\leftarrow\varepsilon} & c_{1\rightarrow\varepsilon} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{n,1} & \cdots & \cdots & c_{n,m} & c_{n\leftarrow\varepsilon} & c_{n\rightarrow\varepsilon} \\ \hline c_{\varepsilon\rightarrow 1} & \cdots & \cdots & c_{\varepsilon\rightarrow m} & \infty & \infty \\ c_{\varepsilon\leftarrow 1} & \cdots & \cdots & c_{\varepsilon\leftarrow m} & \infty & \infty \end{array}\right]$$

Figura 6. Matriu de costos del DF-GED

Cada element $c_{i,j}$ de la matriu C_v correspon al cost d'assignar el vèrtex i-èssim del graf g_1 al vèrtex j-èssim del graf g_2. La cantonada superior esquerra de la matriu conté totes les possibles substitucions de nodes, mentre que la cantonada superior dreta representa el cost de totes les possibles insercions i eliminacions de vèrtexs de g_1, respectivament. La cantonada inferior esquerra conté totes les possibles insercions i eliminacions de vèrtexs de g_2, respectivament, mentre que el cost dels elements de la cantonada inferior dreta s'estableix a infinit, cosa que fa referència a la substitució de ε−ε.

8.2.1.2 Algorisme del Fast Bipartite

El segon algorisme proposat és el BP / Fast BP algorithm. Aquests algorismes tenen com a base la idea de trobar assignacions òptimes d'elements del graf G^p al graf G^q minimitzant la suma dels costos. Semblant a l'anterior, fan servir les matrius de costos C i C', respectivament. Ara, en lloc d'assignar costos vèrtex a vèrtex, es realitza amb cliques [28]. Un clique en un graf no dirigit G és un conjunt de vèrtexs V tal que per a tot parell de vèrtexs pertanyents a V, existeix una aresta que les connecta. En altres paraules, un clique és un subgraf en què cada vèrtex està connectat a cada altre vèrtex del graf.

$$C = \left[\begin{array}{cccc|cccc} c_{1,1} & c_{1,2} & \cdots & c_{1,m} & c_{1,\varepsilon} & \infty & \cdots & \infty \\ c_{2,1} & c_{2,2} & \cdots & c_{2,m} & \infty & c_{2,\varepsilon} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \infty \\ c_{n,1} & c_{n,2} & \cdots & c_{n,m} & \infty & \cdots & \infty & c_{n,\varepsilon} \\ \hline c_{\varepsilon,1} & \infty & \cdots & \infty & 0 & 0 & \cdots & 0 \\ \infty & c_{\varepsilon,2} & \ddots & \vdots & 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \infty & \vdots & \ddots & \ddots & 0 \\ \infty & \cdots & \infty & c_{\varepsilon,m} & 0 & \cdots & 0 & 0 \end{array}\right]$$

Figura 7 Matriu de costos del Bipartite Algorithm

La cantonada superior esquerra conté els costos de substitució del clique K^p_a amb el K^q_i. Les cantonades superior dreta i inferior esquerra contenen els costos de deleció dels cliques K^p_a i inserció dels cliques K^q_i, respectivament. L'algorisme de Munkres o Hongarès, que també ho fa servir l'anterior aproximació, es pot executar per trobar el cost mínim per a totes les permutacions. Aquest cost mínim és un valor de distància d'edició subòptim entre els gràfics implicats, ja que les files de la matriu de costos estan relacionades amb les cliques del gràfic G^p i les columnes estan relacionades amb les cliques de G^q.

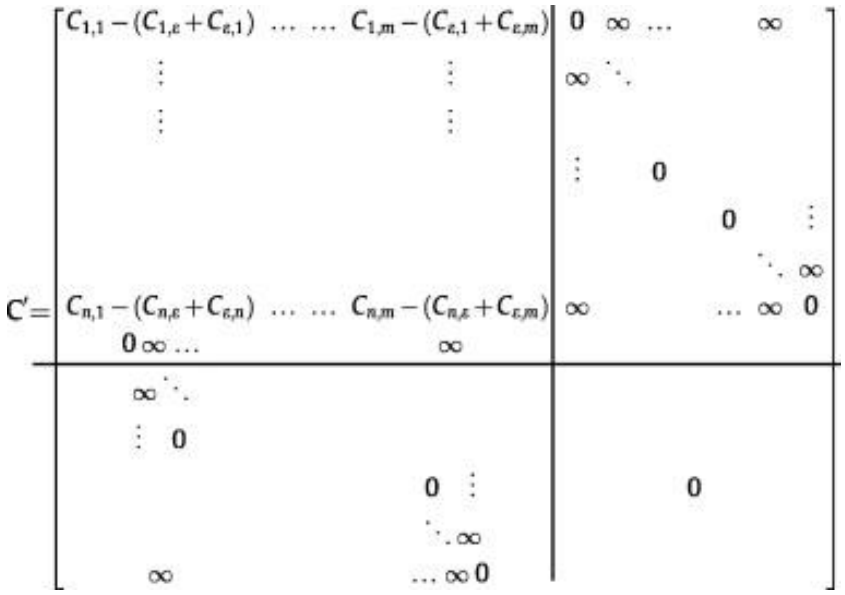


Figura 8 Matriu de costos del Fast Bipartite Algorithm

La metodologia del Fast Bipartite obté un millor speedup assegurant-se que la GED compleix veritablement la definició de distància, seguint els següents atributs:

1) $C_{vs} \leq C_{vd} + C_{vi}\ and\ C_{es} \leq C_{ed} + C_{ei}$
2) $C_{vd} = C_{vi}\ and\ C_{ed} = C_{ei}$
3) $C_{vs} = 0\ and\ C_{es} = 0\ if\ same\ attributes$
4) $All\ costs\ have\ to\ be\ non - negative$

Si es compleix, podem assegurar que, per a dos grafs de mida n i m respectivament, tots els nodes del graf amb ordre inicial inferior o igual se substitueixen si la Distància d'Edició es defineix com una funció de distància. Per tant, no és possible tenir una operació d'inserció i una operació d'eliminació alhora en una bijecció òptima.

També, gràcies al fet que la matriu de costos (Figura 8) està definida seguint aquest lema i que la nova definició de cost d'edició resta del cost d'edició original els costos d'eliminar nodes i arcs de G^p i inserir nodes i arcs de G^q, podem obtenir el cost d'edició a partir de permutacions de la matriu C'. Finalment, només cal escollir el mínim.

8.2.2 Preparació de les dades per la KNN

Un cop s'ha calculat la GED per totes les parelles de molècules de cadascun dels casos, cal desar els nostres propis datasets a un fitxer .csv, un per cada alineament, que contindrà les següents columnes:

id1, smiles1, id2, smiles2, ged.

D'aquesta manera, tenim dades per identificar les molècules llegint els seus identificadors, que tindran el nom acceptat segons la IUPAC. També disposem del codi SMILES, perquè l'ordinador pugui posteriorment extreure les empremtes moleculars a la predicció, i finalment, el resultat calculat de la GED per aquella parella de molècules.

8.2.3 Implementació de la KNN

Per tal d'implementar l'algorisme de predicció mitjançant el mètode de la KNN, primerament cal definir una funció que, a partir de les parelles de molècules del fitxer .csv en format SMILES, calculi les Morgan Fingerprints. Aquestes, com ha estat esmentat abans, són vectors que identifiquen les molècules segons l'entorn molecular de cada àtom dins d'un radi. Aquest vector és el que es farà servir per a entrenar el model.

Un cop el dataframe conté les fingerprints processades, podem definir les variables de X_train X_test, y_train i y_test. Aquestes variables seran escollides de manera pseudoaleatòria dividint el dataframe per tal de separar el conjunt de dades en test i entrenament. El percentatge serà 70% del total de les dades per al conjunt d'entrenament, i 30% per al conjunt de test. La variable X o dependent contindrà la GED. La variable y o independent seran les empremtes moleculars.

Posteriorment, es determinarà el valor del millor paràmetre K que s'ajusta a les dades d'entrenament mitjançant cross-validation, mitjançant la funció predefinida de la llibreria GridSearchCV. El valor del paràmetre K a optimitzar estarà entre 1 i 21 veïns, realitzant 5 validacions creuades del conjunt d'entrenament mitjançant el paràmetre cv=5. La mètrica que servirà per avaluar la qualitat de la predicció serà el Neg-MSE, és a dir, maximitzar l'error quadràtic negatiu, que equival a minimitzar el MSE [29]. El model serà un objecte de la classe KNeighborsRegressor de la llibreria sklearn. Aquest model, un cop obtingut el valor del paràmetre que millor ajusta les dades d'entrenament, serà entrenat amb el valor corresponent de K per tal de realitzar les prediccions sobre el conjunt de test.

Posteriorment, es calcularan les mètriques esmentades prèviament, la MSE, la MAE i la R2. Amb aquests valors validarem la robustesa del nostre model. Per tal de visualitzar els resultats d'una manera més gràfica, es representarà el valor de la GED predita respecte al valor de GED real. També s'inclourà un gràfic de residuals per veure la distribució dels errors.

8.3 Impacte del projecte

El treball posa de manifest un aspecte sovint oblidat en entorns d'alt rendiment computacional: l'empremta de carboni associada a l'execució intensiva de codi. El càlcul de distàncies moleculars com la GED mitjançant CPU pot durar setmanes i consumir molts recursos, però només ha calgut fer-ho una vegada. Ara, gràcies al model de KNN entrenat amb aquestes dades, es poden predir les GED i distàncies Mordred sense haver de repetir el càlcul, cosa que redueix considerablement l'impacte ambiental. Aquesta aproximació contribueix a una recerca més sostenible.

Respecte a l'impacte social, el projecte desenvolupa un model d'intel·ligència artificial aplicable en l'àmbit de la recerca, especialment en quimioinformàtica. Això pot facilitar l'aparició de noves aplicacions científiques i tecnològiques, accelerant processos d'investigació i apropant el coneixement a més persones, especialment en entorns amb menys recursos computacionals.

Finalment, a l'apartat d'igualtat i ètica, aquest projecte no incorpora cap element que impliqui biaixos de gènere, ètnia o cap altra forma de discriminació. A més, com que no inclou una interfície d'usuari amb dades personals, es minimitzen els riscos associats a l'ús indegut de la informació i es manté un enfocament ètic i responsable.

9. Resultats i discussió

En aquesta secció s'exposaran, en primer lloc, els resultats corresponents al càlcul de la GED i les distàncies Mordred per a cada conjunt de dades (secció 9.1), detallant el procediment seguit per obtenir aquestes mesures. Tot seguit, a la secció 9.2, es presentaran els resultats de la predicció de les distàncies moleculars obtingudes utilitzant aquestes distàncies prèviament calculades com a base per als entrenaments, analitzant el seu comportament i rellevància en el context de l'estudi.

9.1 Càlcul de la GED i les distàncies Mordred

A causa del cost computacional elevat associat al càlcul de la GED i tenint en compte que els primers experiments es van dur a terme en un ordinador amb prestacions força limitades, s'ha optat per separar els resultats en funció de com han estat calculats.

Concretament, s'han distingit dos grans grups de càlcul:

1. **Càlcul amb CPU**: Inclou el càlcul de la GED mitjançant el mètode NetworkX i les distàncies Mordred sobre CPU. El càlcul de la GED va requerir la majoria del temps computacional, perquè segons l'enfocament NetworkX, cada parella estava uns 5 minuts aproximant la solució que millor descendís en l'espai de solucions. Per altra banda, el càlcul de les distàncies Mordred és menyspreable, ja que és realitzar una resta de dos mòduls vectorials de cost O(1). Aquests dos procediments van trigar en combinat aproximadament 3 setmanes.
2. **Càlcul amb GPU**: Utilitza l'enfocament de Fast Bipartite per calcular la GED. Aquest mètode, molt més eficient, va permetre obtenir els resultats en menys d'una setmana.

A més, s'han separat els resultats segons la base de dades d'origen de les molècules. Això ha donat lloc a sis conjunts de dades diferenciats, que s'utilitzaran per entrenar sis models KNN i fer prediccions. D'aquesta manera, podrem comparar quin mètode de càlcul i quina base de dades permeten obtenir millors resultats en l'entrenament d'un model predictiu amb KNN per estimar la distància entre dues molècules.

Finalment, a causa de discrepàncies detectades entre els valors de GED calculats amb CPU i amb GPU, s'han creat dos subconjunts addicionals que inclouen només les parelles de molècules per a les quals ambdós mètodes han calculat valors similars de GED. Aquests subdatasets s'utilitzaran també per entrenar dos models per veure si s'obtenen millors prediccions. Aleshores, de 6 conjunts de dades diferenciats, acabarem amb 8 conjunts en total.

Els 8 conjunts de dades utilitzats són els següents:

GED ESOL CPU, GED ESOL GPU, Mordred ESOL, GED FreeSolv CPU, GED FreeSolv GPU i Mordred FreeSolv.

9.1.1 GED a la base de dades ESOL

D'aquest apartat esperem com a resultat dos datasets, els quals contindran per a les mateixes parelles de molècules de la ESOL el valor de GED calculat per cada mètode.

9.1.1.1 Implementació sobre CPU

En aquest cas obtenim un dataset amb un total de 635.629 entrades, que correspon amb el càlcul per a la mida de la base de dades ESOL de $\frac{n(n-1)}{2} = \frac{1128 \cdot 1127}{2} = 635.629$. Això significa que no hi ha hagut cap problema amb els alineaments per parelles o el càlcul de la GED entre elles. La taula 2 mostra els resultats dels 5 primers càlculs de la GED entre parelles de molècules de la base de dades ESOL.

Taula 2. Contingut del dataset que conté els càlculs de la GED per a la base de dades ESOL utilitzant l'enfocament per CPU

ID 1	SMILES Molècula 1	ID 2	SMILES Molècula 2	GED
Amigdalin	OCC3OC(OCC2OC(OC(C#N)…	Fenfuram	Cc1occc1C(=O)Nc2ccccc2	41,0
Amigdalin	OCC3OC(OCC2OC(OC(C#N)…	citral	CC(C)=CCCC(C)=CC(=O)	49,0
Amigdalin	OCC3OC(OCC2OC(OC(C#N)…	Picene	c1ccc2c(c1)ccc3c2ccc4c5ccccc5ccc43	34,0
Amigdalin	OCC3OC(OCC2OC(OC(C#N)…	Thiophene	c1ccsc1	58,0
Amigdalin	OCC3OC(OCC2OC(OC(C#N)…	benzothiazole	c2ccc1scnc1c2	51,0

9.1.1.2 Implementació sobre GPU

Igual que abans, obtenim un altre dataset que conté 635.629 elements, el què indica que el càlcul per GPU tampoc ha fallat en cap parella de molècules. Les 5 primeres entrades, com es poden observar a la taula 3, mostren el mateix que l'anterior, però amb lleugers canvis en els resultats de la GED.

Taula 3. Contingut del dataset que conté els càlculs de la GED per a la base de dades ESOL utilitzant l'enfocament per GPU

ID 1	SMILES Molècula 1	ID 2	SMILES Molècula 2	GED
Amigdalin	OCC3OC(OCC2OC(OC(C#N)...	Fenfuram	Cc1occc1C(=O)Nc2ccccc2	47,0
Amigdalin	OCC3OC(OCC2OC(OC(C#N)...	citral	CC(C)=CCCC(C)=CC(=O)	43,0
Amigdalin	OCC3OC(OCC2OC(OC(C#N)...	Picene	c1ccc2c(c1)ccc3c2ccc4c5ccccc5ccc43	54,0
Amigdalin	OCC3OC(OCC2OC(OC(C#N)...	Thiophene	c1ccsc1	37,0
Amigdalin	OCC3OC(OCC2OC(OC(C#N)...	benzothiazole	c2ccc1scnc1c2	41,0

9.1.2 Mordred a la base de dades ESOL

En aquest cas, també obtenim el dataset prou similar als anteriors dos, però en lloc d'una columna per la GED, conté una columna amb la distància euclidiana entre els vectors de descriptors moleculars Mordred, definida com a distancia_mordred.

Com el cost de calcular aquesta distància és menyspreable, només hi ha una aproximació, per tant, un únic dataset.

Podem observar a la taula 4 els següents resultats de les 5 primeres molècules:

Taula 4. Contingut del dataset que conté els càlculs de la Distància Mordred per a la base de dades ESOL

ID 1	SMILES Molècula 1	ID 2	SMILES Molècula 2	GED
Amigdalin	OCC3OC(OCC2OC(OC(C#N)...	Fenfuram	Cc1occc1C(=O)Nc2ccccc2	71.393,86
Amigdalin	OCC3OC(OCC2OC(OC(C#N)...	citral	CC(C)=CCCC(C)=CC(=O)	73.743,89
Amigdalin	OCC3OC(OCC2OC(OC(C#N)...	Picene	c1ccc2c(c1)ccc3c2ccc4c5cc...	47.923,17
Amigdalin	OCC3OC(OCC2OC(OC(C#N)...	Thiophene	c1ccsc1	96.056,40
Amigdalin	OCC3OC(OCC2OC(OC(C#N)...	benzothiazole	c2ccc1scnc1c2	90.370,00

9.1.3 GED a la base de dades FreeSolv

Igual que per a l'altra base de dades, esperem dos datasets de la mateixa mida, cadascun amb els seus càlculs de la GED segons l'enfocament per a cadascuna de les parelles de molècules.

9.1.3.1 Implementació sobre CPU

En aquest cas obtenim un dataset amb un total de 205.762 entrades. En aquest cas, també correspon amb el càlcul de la mida del nostre dataset a partir de les dades de la base de dades FreeSolv segons: $\frac{n(n-1)}{2} = \frac{642 \cdot 641}{2} = 205.762$. Això és indicatiu de què tampoc hi ha hagut cap tipus de problema en l'aparellament de les molècules o del càlcul de la GED. La taula 5 mostra els resultats dels 5 primers càlculs de la GED entre parelles de molècules de la base de dades FreeSolv.

Taula 5. Contingut del dataset que conté els càlculs de la GED per a la base de dades FreeSolv utilitzant l'enfocament per CPU

ID 1	SMILES Molècula 1	ID 2	SMILES Molècula 2	GED
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	methanesulfonyl chloride	CS(=O)(=O)Cl	19,0
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	3-methylbut-1-ene	CC(C)C=C	17,0
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	2-ethylpyrazine	CCc1cnccn1	10,0
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	heptan-1-ol	CCCCCCCO	11,0
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	3,5-dimethylphenol	Cc1cc(cc(c1)O)C	10,0

9.1.3.2 Implementació sobre GPU

Igual que en el cas anterior, s'obté un altre conjunt de dades que conté 205.762 elements, la qual cosa indica que el càlcul mitjançant GPU tampoc ha fallat en cap parella de molècules. Les cinc primeres entrades, mostrades a la taula 6, presenten resultats similars als de l'anàlisi anterior, amb lleugeres variacions en els valors de la GED.

Taula 6. Contingut del dataset que conté els càlculs de la GED per a la base de dades FreeSolv utilitzant l'enfocament per GPU

ID 1	SMILES Molècula 1	ID 2	SMILES Molècula 2	GED
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	methanesulfonyl chloride	CS(=O)(=O)Cl	18,0
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	3-methylbut-1-ene	CC(C)C=C	18,0
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	2-ethylpyrazine	CCc1cnccn1	21,0
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	heptan-1-ol	CCCCCCCO	21,0
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	3,5-dimethylphenol	Cc1cc(cc(c1)O)C	22,0

9.1.4 Mordred a la base de dades FreeSolv

Igual que per a la base de dades de la ESOL, podem esperar un dataset amb el mateix nombre d'elements que prèviament, amb la columna de distàncies Mordred entre les parelles de molècules, com es pot observar a la taula 7:

Taula 7. Contingut del dataset que conté els càlculs de la Distància Mordred per a la base de dades FreeSolv

ID 1	SMILES Molècula 1	ID 2	SMILES Molècula 2	Distància Mordred
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	methanesulfonyl chloride	CS(=O)(=O)Cl	24.011,40
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	3-methylbut-1-ene	CC(C)C=C	19.093,37
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	2-ethylpyrazine	CCc1cnccn1	16.038,17
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	heptan-1-ol	CCCCCCCO	9.164,41
4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	3,5-dimethylphenol	Cc1cc(cc(c1)O)C	11.890,91

9.1.5 Discrepàncies entre el càlcul de la GED per CPU respecte GPU

Com s'ha pogut veure a les taules 2 i 3 i a les taules 5 i 6, observant manualment els resultats ja es pot veure petites diferències entre el valor de la GED calculat per l'enfocament de la CPU respecte al de la GPU. Per tenir una visió més àmplia i global d'aquestes diferències, i per veure també si poden afectar posteriorment a la precisió del model a entrenar, s'ha creat un script que trobarà la distribució de les diferències de la GED al llarg dels dos conjunts de dades.

La visualització, representada a la figura 9 i figura 10, mostra la freqüència del valor de la diferència centrat al 0. Aquest punt serà 0 quan ambdós mètodes han calculat el mateix valor de GED per a la parella de molècules. A partir d'aquí, els punts negatius (a l'esquerra del 0) indicaran que el valor de la GED calculat per la GPU ha estat superior al calculat per la CPU, i viceversa.

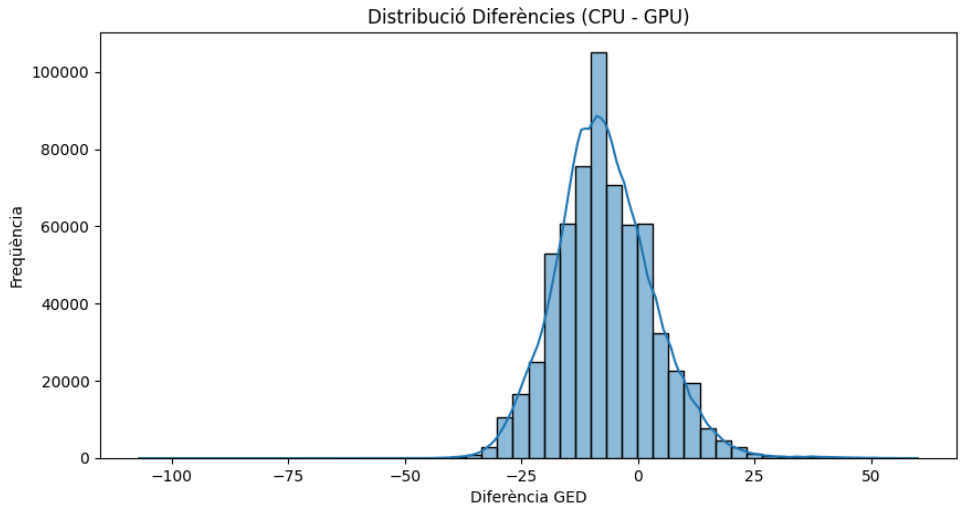


Figura 9 Distribució de les diferències del càlcul de la GED per a la base de dades ESOL

En aquest cas, per a la base de dades ESOL (figura 9), es pot observar una distribució semblant a la gaussiana, però lleugerament desplaçada cap a l'esquerra. Això indica que l'algorisme del Fast Bipartite implementat sobre GPU aproxima una GED superior que l'algorisme de NetworkX.

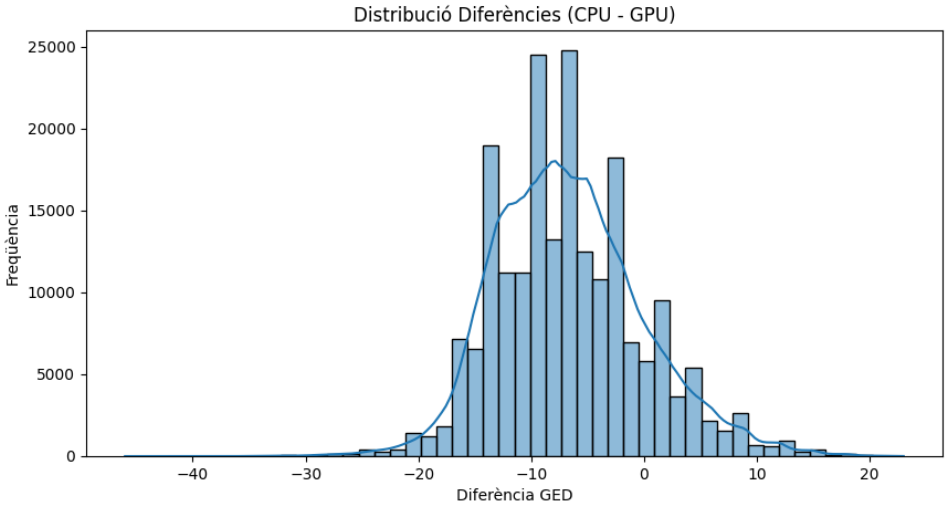


Figura 10 Distribució de les diferències del càlcul de la GED per a la base de dades FreeSolv

Per aquest cas (figura 10), també es veu una distribució semblant a l'anterior, considerant que té menys dades. També està lleugerament desplaçada cap a l'esquerra, aquesta una mica més, per tant, la GED de la GPU serà superior per la majoria de casos.

Tal com s'ha esmentat a l'inici de l'apartat, aquesta observació ha sigut motiu per extreure'n dos subdatasets addicionals on aquesta diferència de GED ha estat per sota d'un valor llindar, per tant, indicant que ambdós algorismes han calculat un valor semblant de GED. El valor llindar escollit com a filtre ha estat de 5. Això vol dir que es crearà un subdataset per a cada base de dades que contindrà les parelles de molècules on la seva diferència de GED ha estat ≤ 5.

9.1.5.1 Característiques del subdataset d'ESOL

Aquest conté un total de 189.581 parelles de molècules del total de 635.629 que hi havia. Això representa un 30% del dataset original, aproximadament, del qual ambdós mètodes han calculat una GED semblant. Observant els primers continguts del nou dataset, podem veure alguns exemples de parelles de molècules a la taula 8.

Taula 8. Subdataset per la base de dades ESOL que conté els parells de molècules amb una valor de GED calculat de manera semblant pels enfocaments CPU i GPU. També conté les columnes per a l'error absolut i relatiu.

ID 1	ID 2	GED calculat a CPU	GED calculat a GPU	Diferència de GED entre els 2	Diferència relativa de GED
Amigdalin	2-Chloronapthalene	47,0	43,0	4,0	0,0851
Amigdalin	2-Undecanol	43,0	44,0	-1,0	0,0227
Amigdalin	Phorate	45,0	45,0	0,0	0,0000
Amigdalin	Phenacetin	44,0	45,0	-1,0	0,0222
Amigdalin	Theophylline	47,0	45,0	2,0	0,0426

En aquesta taula s'han omès els codis SMILES i el valor d'índex de la parella en el dataset original per afavorir la visualització. Es poden observar les molècules amb els valors de GED calculats per als dos casos, amb dues columnes que contenen la diferència absoluta i relativa d'aquest valor.

9.1.5.2 Característiques del subdataset de FreeSolv

En aquest cas, el subconjunt conté un total de 72.875 parelles de molècules, del total de 205.762 inicials al dataset de FreeSolv. Això representa un 35,4% del dataset original. A la taula 9 es poden observar les primeres entrades amb una diferència de GED menor o igual a 5.

Taula 9. Subdataset per la base de dades FreeSolv que conté els parells de molècules amb una valor de GED calculat de manera semblant per l'enfocament CPU i GPU. També conté les columnes per a l'error absolut i relatiu.

ID 1	ID 2	GED calculat a CPU	GED calculat a GPU	Diferència de GED entre els 2	Diferència relativa de GED
4-methoxy-N,N-dimethyl-benzamide	methanesulfonyl chloride	19,0	18,0	1,0	0,0526
4-methoxy-N,N-dimethyl-benzamide	3-methylbut-1-ene	17,0	18,0	-1,0	0,0556
4-methoxy-N,N-dimethyl-benzamide	2,3-dimethylbutane	15,0	19,0	-4,0	0,2105
4-methoxy-N,N-dimethyl-benzamide	2-methylpentan-2-ol	15,0	20,0	-5,0	0,2500
4-methoxy-N,N-dimethyl-benzamide	butan-2-ol	17,0	18,0	-1,0	0,0556

9.2 Predicció de les distàncies moleculars

Ara que es disposa de la informació de la GED i les distàncies Mordred als datasets corresponents, s'entrenaran els models d'aprenentatge automàtic per KNN i s'avaluarà la seva precisió i robustesa en comparativa.

Primerament, cal llegir els codis SMILES de cada parella de molècules i obtenir les seves Morgan Fingerprints, que com ja s'ha esmentat prèviament seran les dades de la X per als conjunts d'entrenament i de test. En cas que alguna molècula donés error, ens avisaria i aquella parella quedaria descartada, però no ha estat el cas per cap aproximació.

Tot i voler realitzar una comparativa final, cada model serà entrenat amb el valor del paràmetre K obtingut com a idoni per separat mitjançant validació creuada. D'aquesta manera, podem obtenir les millors prediccions en qualsevol dels casos independentment de les dades d'entrada. En el cas del train test split, aquest sí que serà el mateix per a l'estudi entre CPU i GPU dins d'una mateixa base de dades. Això s'aconseguirà amb una llavor prefixada per als nombres pseudoaleatoris que fa servir per separar el dataset en els dos subconjunts, que serà la 42.

9.2.1 Model GED ESOL CPU

Aquest primer model serà entrenat amb el dataset obtingut a l'apartat 9.1.1.1.

Primerament, es trobarà per cross-validation el millor valor de K per a entrenar la KNN amb aquestes dades. En aquest cas, representant l'error del MSE per a cadascun dels models entrenats amb el paràmetre de K ∈ [1, 20], tal com es pot apreciar a la figura 11.

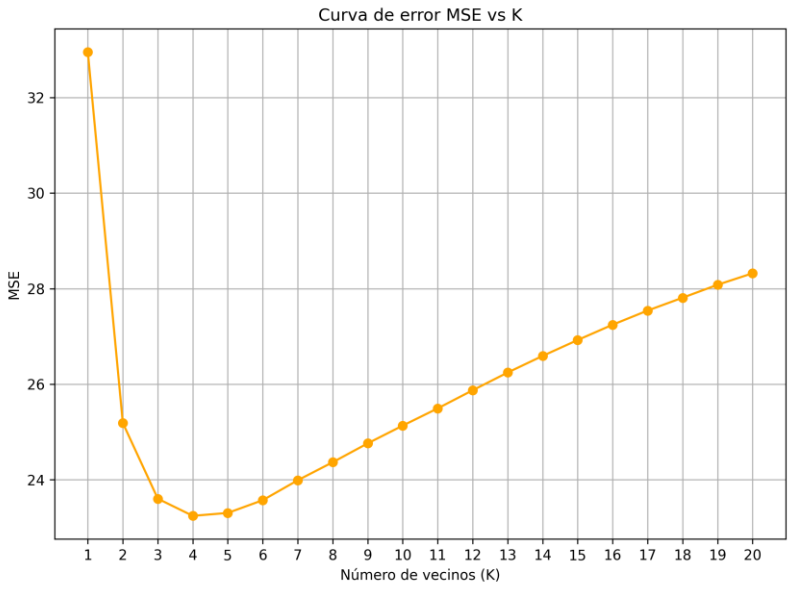


Figura 11 Representació del Mean Squared Error per avaluar els diferents valors del paràmetre K (ESOL CPU)

El que ens interessa és aquell valor de K que minimitzi el error al llarg de les validacions creuades. En aquest cas s'ha trobat que el millor valor ha estat K = 4.

Posteriorment, es tornarà a entrenar el model amb aquest valor de K = 4, i aquí és on es realitzaran les prediccions del conjunt de test.

Les mètriques d'error obtingudes per avaluar el model han estat les següents:

- MSE = 23,24
- MAE = 3,36
- $R^2 = 0,867$

Per aquest cas, s'ha trigat aproximadament 16 hores i mitja per a la validació creuada, mitja hora per tornar a entrenar el model i realitzar les prediccions, i 18 hores per obtenir els gràfics els resultats.

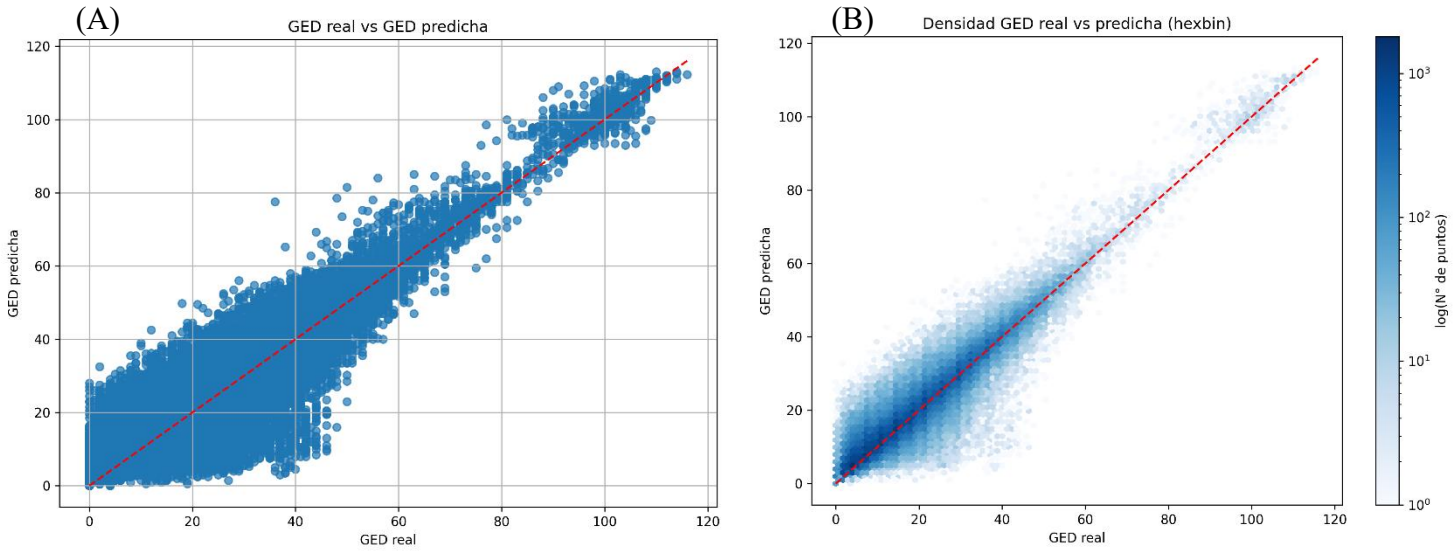


Figura 12. Representació dels valors i la densitat (logaritme del nº de punts) de punts de GED reals (continguts al dataset de test) respecte dels valors de GED predits pel mètode KNN. (ESOL CPU)

Com es pot apreciar, hi ha molta densitat de punts, degut al gran tamany del dataset provinent de la base de dades ESOL. La majoria de punts segueixen la tendència de la línia vermella, que indica una predicció perfecta. Hi ha alguns punts que no s'ajusten tant, això és degut a que la predicció no ha estat tan precisa. Això es veurà reflectit al gràfic de residuals i a l'error calculat prèviament.

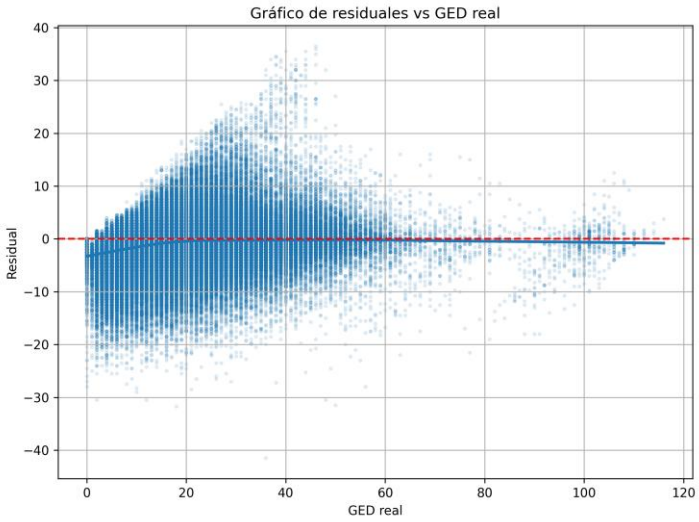


Figura 13. Representació dels valors de GED reals (continguts al dataset de test) respecte dels valors dels residuals $y_i - \hat{y}_i$. (ESOL CPU)

Observant el gràfic de residuals podem observar que la línia de tendència en blau clar està prou ajustada a la línia vermella, que un altre cop indica la predicció perfecta. A molècules prou semblants, el predictor estima una major GED de la que els hi correspon a cada parella.

9.2.2 Model GED ESOL GPU

Aquest model serà entrenat amb el dataset obtingut a l'apartat 9.1.1.2.

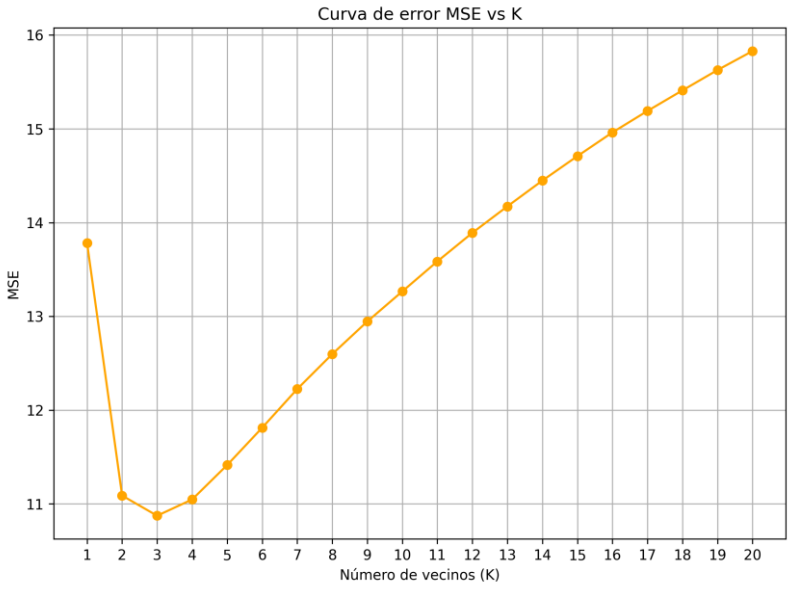


Figura 14 Representació del Mean Squared Error per avaluar els diferents valors del paràmetre K (ESOL GPU)

En el cas de la GED del dataset d'ESOL calculada mitjançant GPU, la K escollida ha estat 3.

Posteriorment, es tornarà a entrenar el model amb aquest valor de K = 3, i aquí és on es realitzaran les prediccions del conjunt de test.

Les mètriques d'error obtingudes per avaluar el model han estat les següents:

- MSE = 10,87
- MAE = 2,16
- R^2 = 0,885

Per aquest cas, s'ha trigat aproximadament 28 hores i mitja per a la validació creuada, 40 minuts per tornar a entrenar el model i realitzar les prediccions, i 13 hores per obtenir els gràfics els resultats.

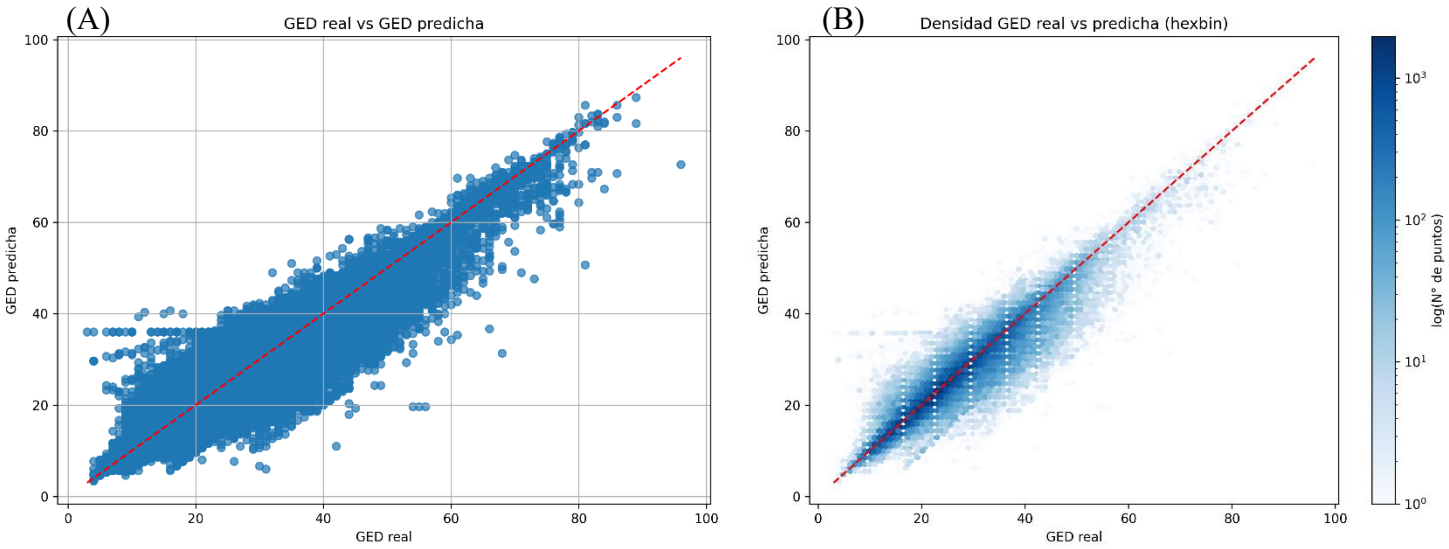


Figura 15. Representació dels valors i la densitat (logaritme del nº de punts) de punts de GED reals (continguts al dataset de test) respecte dels valors de GED predits pel mètode KNN. (ESOL GPU)

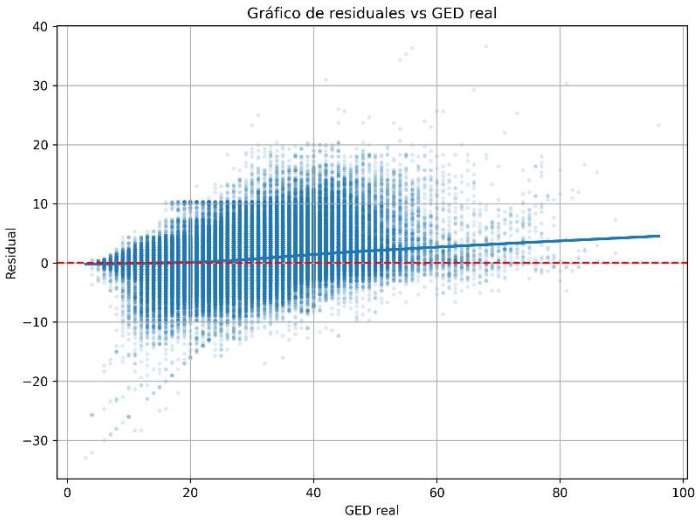


Figura 16. Representació dels valors de GED reals (continguts al dataset de test) respecte dels valors dels residuals $y_i - \hat{y}_i$. (ESOL GPU)

Comentant les figures en conjunt, es pot observar que la densitat de valors de GED reals respecte predits és prou bona. En aquest cas, el gràfic de residuals ens indica, que per a parelles de molècules on la GED és més gran, la tendència que ha seguit el model és predir valors de GED inferiors. Això és possiblement degut a la falta de parelles de molècules amb GED tan grans.

9.2.3 Model GED ESOL Reduït

Aquest model serà entrenat amb el dataset obtingut a l'apartat 9.1.5.1.

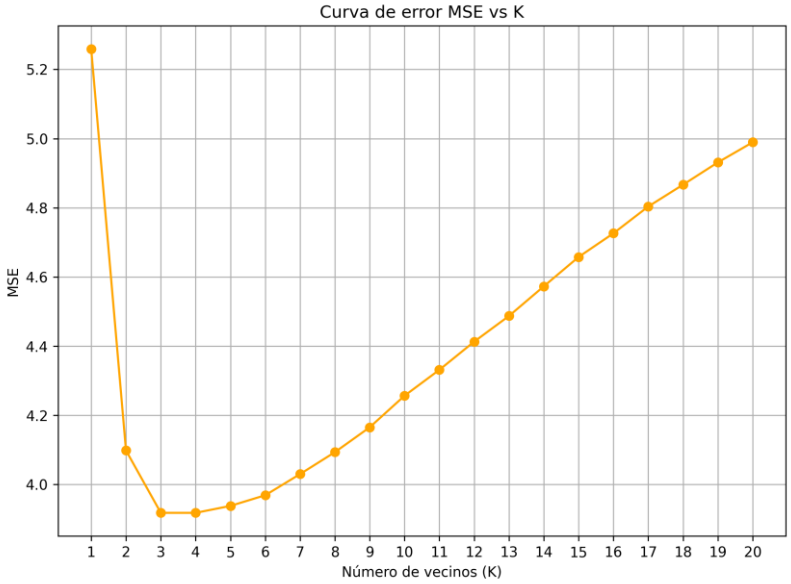


Figura 17 Representació del Mean Squared Error per avaluar els diferents valors del paràmetre K (ESOL Subdataset)

En el cas de la GED del subdataset d'ESOL reduït, la K escollida ha estat 4.

Posteriorment, es tornarà a entrenar el model amb aquest valor de K = 4, i aquí és on es realitzaran les prediccions del conjunt de test.

Les mètriques d'error obtingudes per avaluar el model han estat les següents:

- MSE = 3,92
- MAE = 0,945
- R^2 = 0,952

Per aquest cas, s'ha trigat aproximadament 2 hores i quart per a la validació creuada, 4 minuts per tornar a entrenar el model i realitzar les prediccions, i 1 hora i 10 minuts per obtenir els gràfics els resultats.

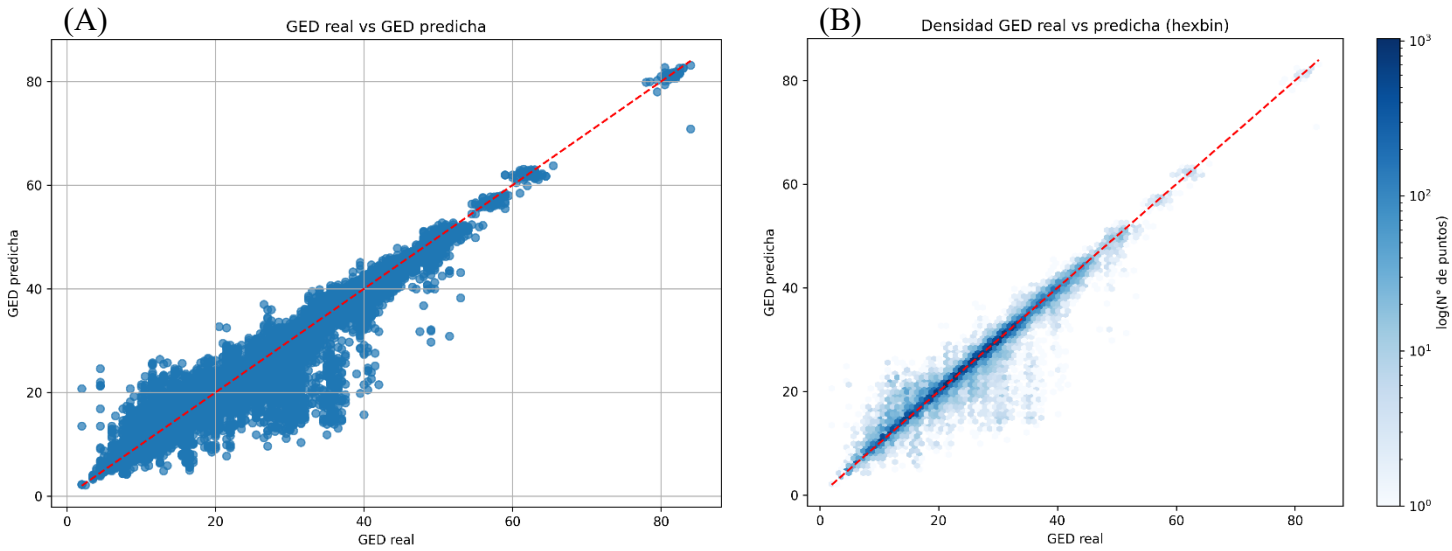


Figura 18. Representació dels valors i la densitat (logaritme del nº de punts) de punts de GED reals (continguts al dataset de test) respecte dels valors de GED predits pel mètode KNN. (ESOL Reduït)

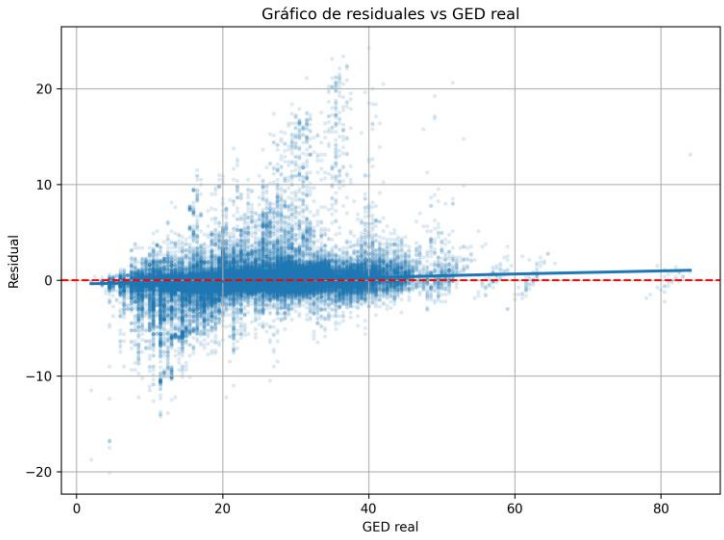


Figura 19. Representació dels valors de GED reals (continguts al dataset de test) respecte dels valors dels residuals $y_i - \hat{y}_i$. (ESOL Reduït)

En aquest cas, la predicció ha estat molt més precisa, es pot observar una falta de densitat de punts entre una GED de 60 i 80, però revisant els subdatasets, això és degut a l'absència de parelles de molècules entre aquestes distàncies, no a altres possibles errors. El gràfic de residuals està pràcticament ajustat a la línia central, realitzant alguna predicció més optimista a les parelles amb GED entre 20 i 40.

9.2.4 Model Mordred ESOL

Aquest model serà entrenat amb el dataset obtingut a l'apartat 9.1.2.

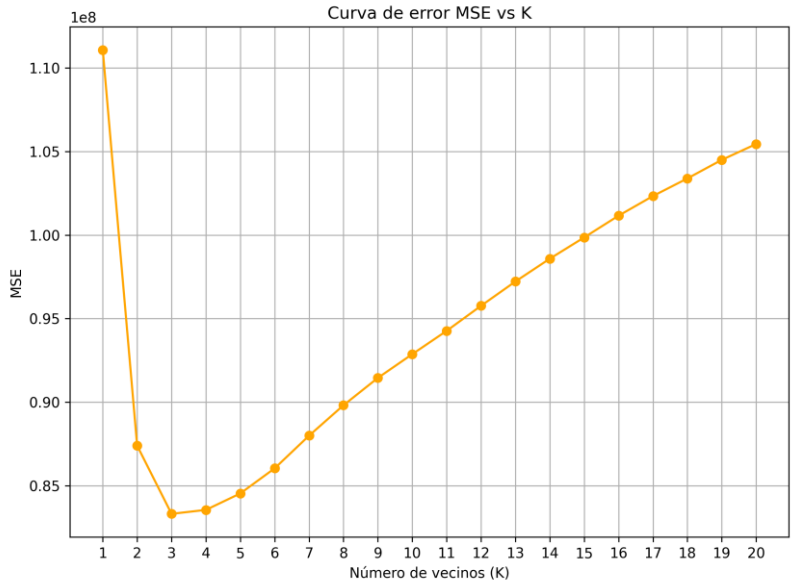


Figura 20 Representació del Mean Squared Error per avaluar els diferents valors del paràmetre K (ESOL Mordred)

En el cas de la distància Mordred per a les parelles del dataset d'ESOL, la K escollida ha estat 3.

Posteriorment, es tornarà a entrenar el model amb aquest valor de K = 3, i aquí és on es realitzaran les prediccions del conjunt de test.

Les mètriques d'error obtingudes per avaluar el model han estat les següents:

- MSE = 83.563.847,82
- MAE = 5881,77
- R^2 = 0,906

Per aquest cas, s'ha trigat aproximadament 15 hores per a la validació creuada, 25 minuts per tornar a entrenar el model i realitzar les prediccions, i 8 hores i 40 minuts per obtenir els gràfics els resultats.

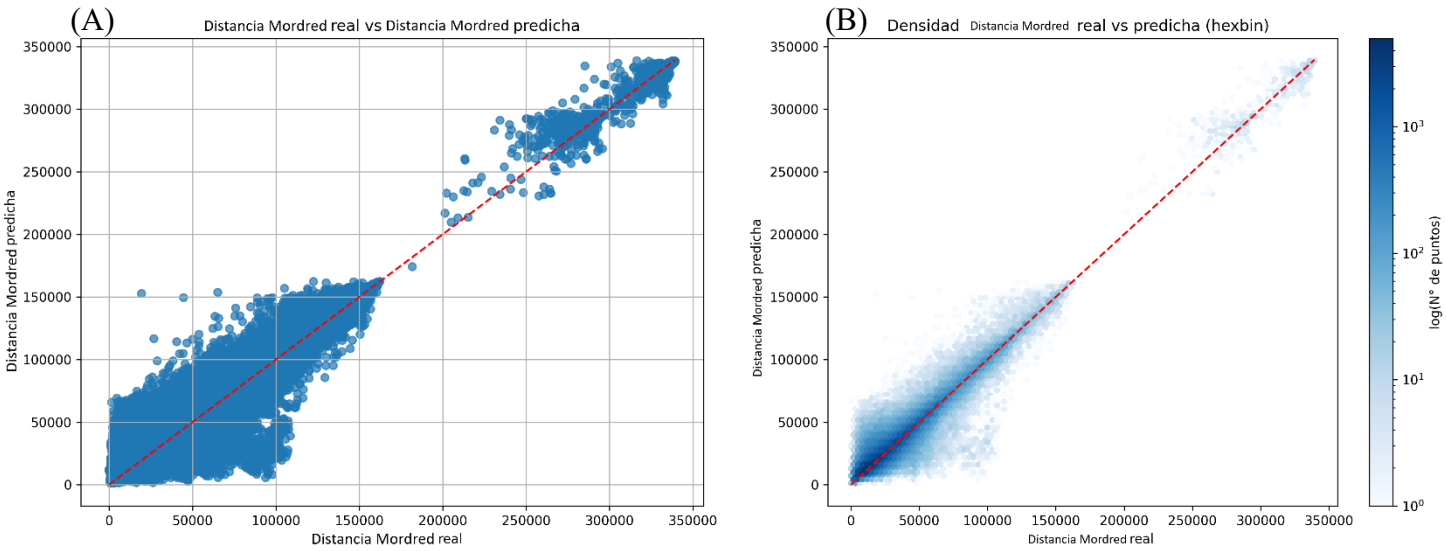


Figura 21 Representació dels valors i la densitat (logaritme del nº de punts) de punts de distància Mordred reals (continguts al dataset de test) respecte dels valors de distància Mordred predits pel mètode KNN. (ESOL)

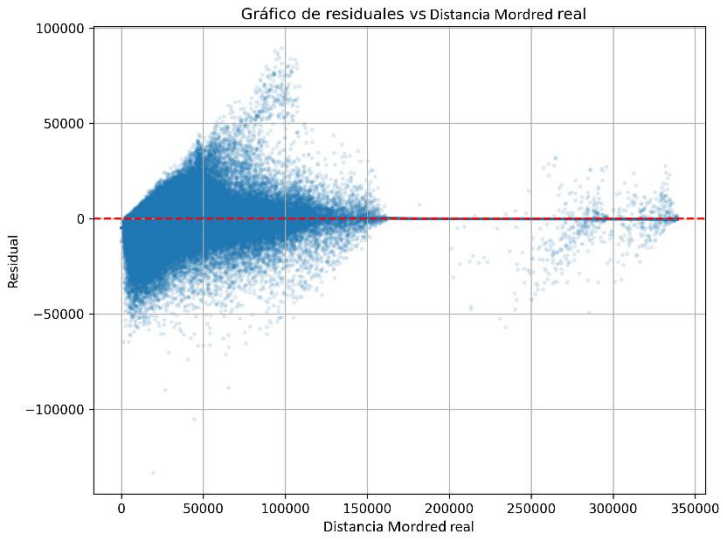


Figura 22. Representació dels valors de distància Mordred reals (continguts al dataset de test) respecte dels valors dels residuals $y_i - \hat{y}_i$. (ESOL)

Aquí es pot observar també una falta de dades al voltant dels 170.000 - 200.000 de distància dels vectors Mordred. En aquest cas, els errors de les prediccions, com l'escala de la distància Mordred és més elevada, es pot veure una gran densitat de punts fora de la línia de tendència sobretot a l'inici.

9.2.5 Model GED FreeSolv CPU

Aquest model serà entrenat amb el dataset obtingut a l'apartat 9.1.3.1.

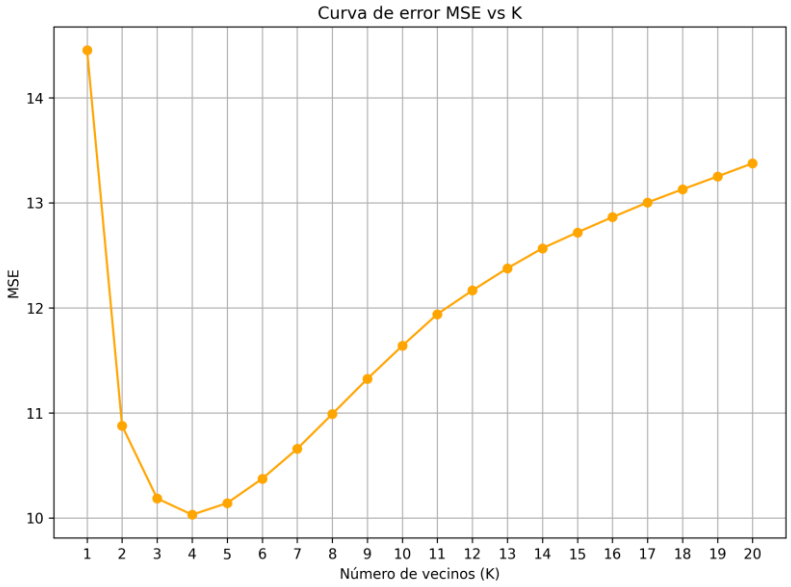


Figura 23 Representació del Mean Squared Error per avaluar els diferents valors del paràmetre K (FreeSolv CPU)

En el cas de la GED del dataset de FreeSolv calculada mitjançant CPU, la K escollida ha estat 4.

Posteriorment, es tornarà a entrenar el model amb aquest valor de K = 4, i aquí és on es realitzaran les prediccions del conjunt de test.

Les mètriques d'error obtingudes per avaluar el model han estat les següents:

- MSE = 10,03
- MAE = 2,28
- R^2 = 0,849

Per aquest cas, s'ha trigat aproximadament 3 hores i 10 minuts per a la validació creuada, 5 minuts per tornar a entrenar el model i realitzar les prediccions, i 1 hora i 20 minuts per obtenir els gràfics els resultats.

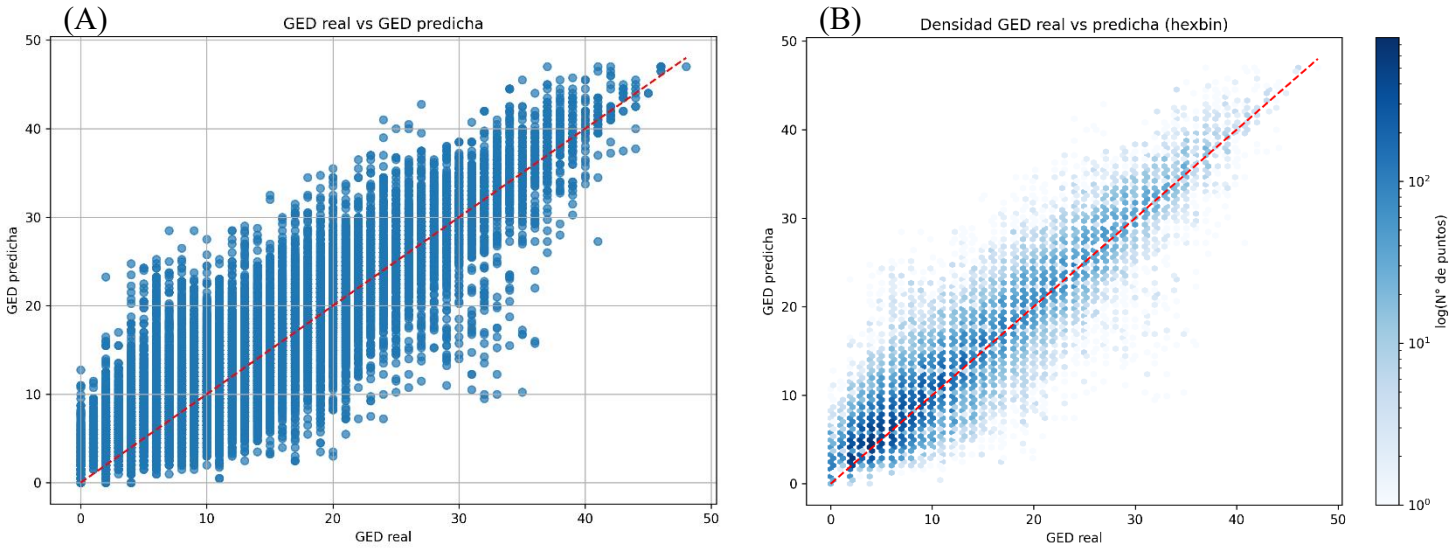


Figura 24. Representació dels valors i la densitat (logaritme del nº de punts) de punts de GED reals (continguts al dataset de test) respecte dels valors de GED predits pel mètode KNN. (FreeSolv CPU)

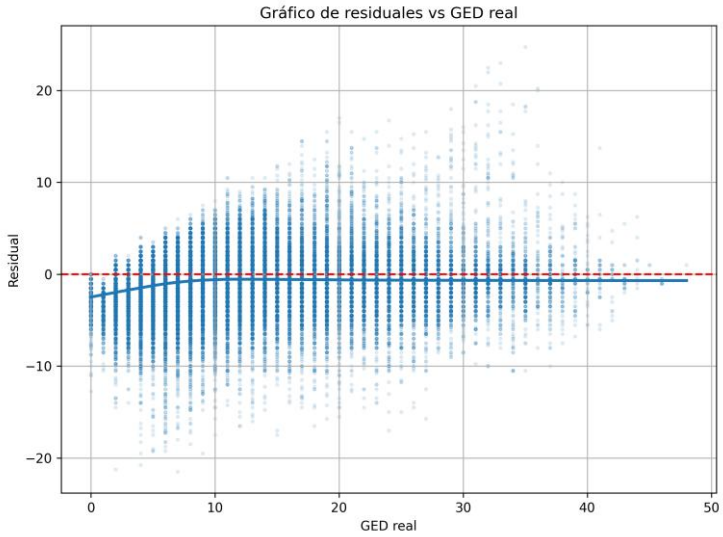


Figura 25. Representació dels valors de GED reals (continguts al dataset de test) respecte dels valors dels residuals $y_i - \hat{y}_i$ (FreeSolv CPU)

En aquest cas, es pot observar tant a les figures dels valors i densitat com en la figura dels residuals que els punts s'han representat malament i hi ha hagut un error en la representació. De totes maneres, serveix per observar que en aquest cas els residuals sempre estan per baix, cosa que indica una petita tendència a predir valors superiors de GED.

9.2.6 Model GED FreeSolv GPU

Aquest model serà entrenat amb el dataset obtingut a l'apartat 9.1.3.2.

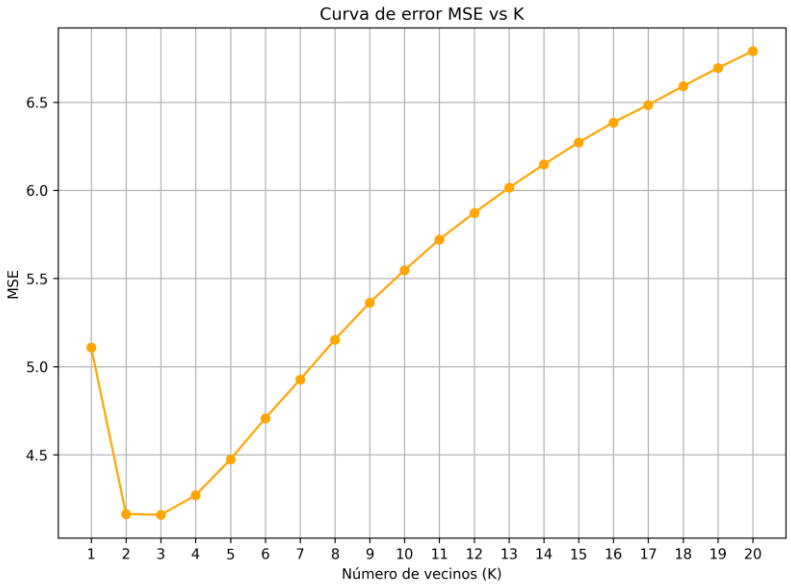


Figura 26 Representació del Mean Squared Error per avaluar els diferents valors del paràmetre K (FreeSolv GPU)

En el cas de la GED del dataset de FreeSolv calculada mitjançant GPU, la K escollida ha estat 3.

Posteriorment, es tornarà a entrenar el model amb aquest valor de K = 3, i aquí és on es realitzaran les prediccions del conjunt de test.

Les mètriques d'error obtingudes per avaluar el model han estat les següents:

- MSE = 4,158
- MAE = 1,401
- R^2 = 0,880

Per aquest cas, s'ha trigat aproximadament 3 hores i 40 minuts per a la validació creuada, 7 minuts per tornar a entrenar el model i realitzar les prediccions, i 2 hores per obtenir els gràfics els resultats.

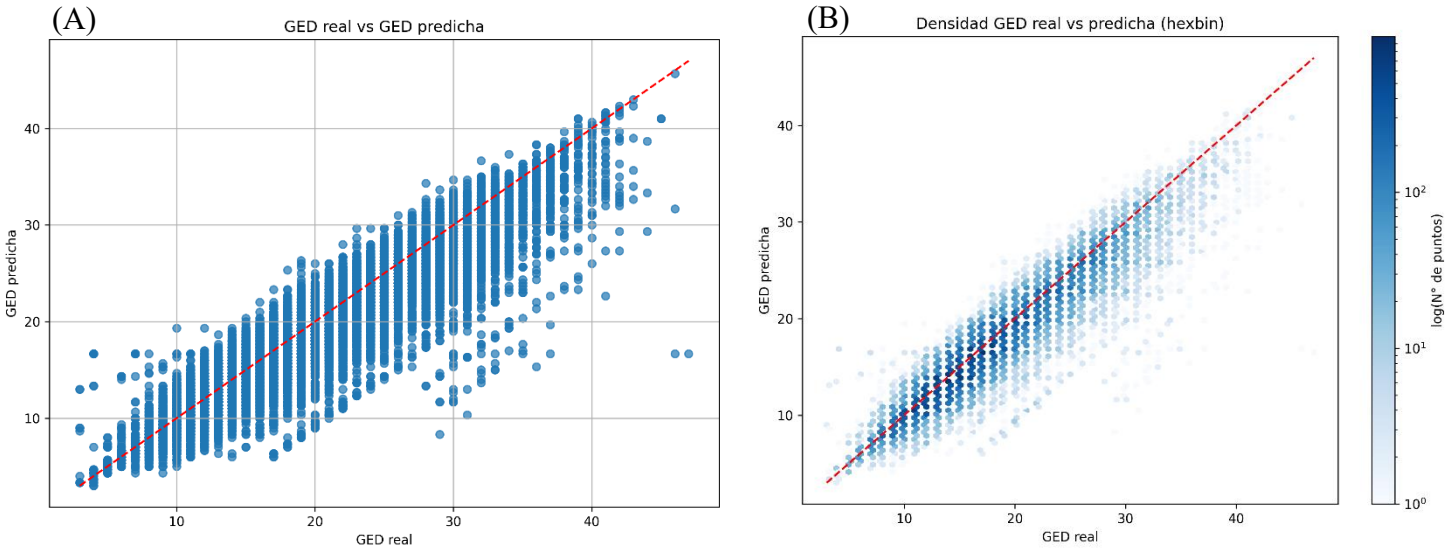


Figura 27. Representació dels valors i la densitat (logaritme del nº de punts) de punts de GED reals (continguts al dataset de test) respecte dels valors de GED predits pel mètode KNN. (FreeSolv GPU)

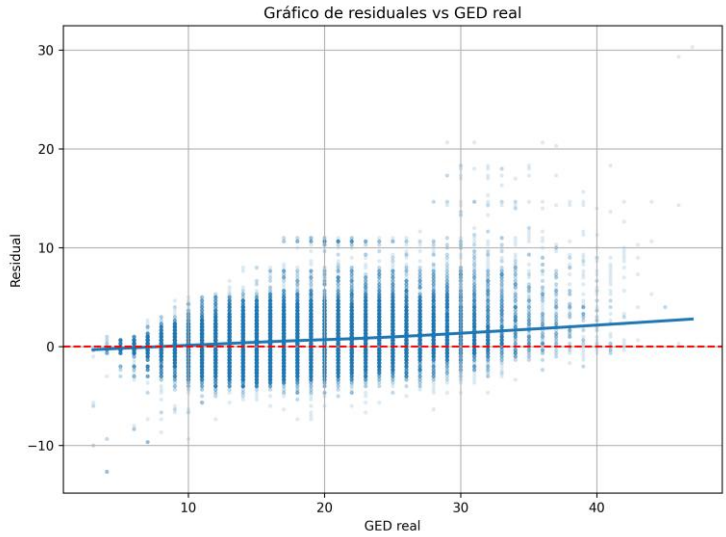


Figura 28. Representació dels valors de GED reals (continguts al dataset de test) respecte dels valors dels residuals $y_i - \hat{y}_i$ (FreeSolv GPU)

Igual que abans s'observa un petit error de representació, però de manera orientativa en aquest cas podem observar que s'ha tendit a predir amb valors de GED inferiors als veritables.

9.2.7 Model GED FreeSolv Reduït

Aquest model serà entrenat amb el dataset obtingut a l'apartat 9.1.5.2.

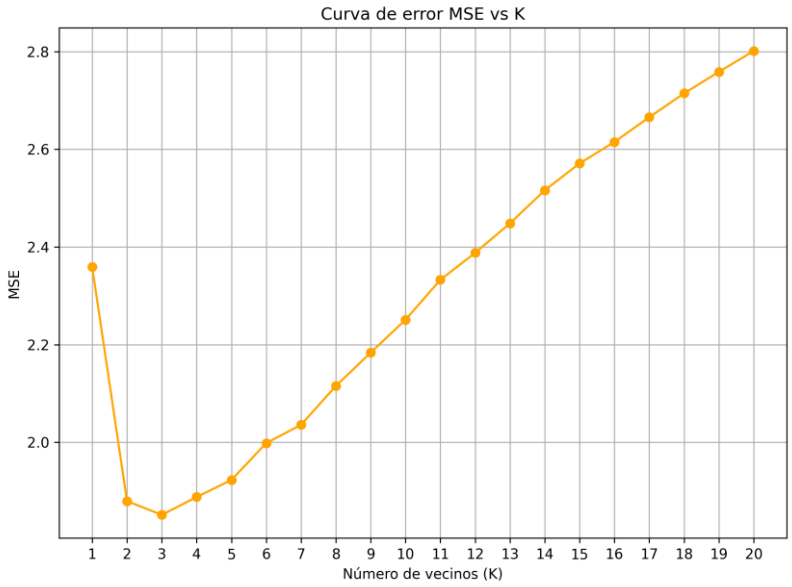


Figura 29 Representació del Mean Squared Error per avaluar els diferents valors del paràmetre K (FreeSolv Reduït)

En el cas de la GED del dataset de FreeSolv Reduït, la K escollida ha estat 3.

Posteriorment, es tornarà a entrenar el model amb aquest valor de K = 3, i aquí és on es realitzaran les prediccions del conjunt de test.

Les mètriques d'error obtingudes per avaluar el model han estat les següents:

- MSE = 1,888
- MAE = 0,7125
- R^2 = 0,959

Per aquest cas, s'ha trigat aproximadament 20 minuts per a la validació creuada, menys d'un minut per tornar a entrenar el model i realitzar les prediccions, i 10 minuts per obtenir els gràfics els resultats.

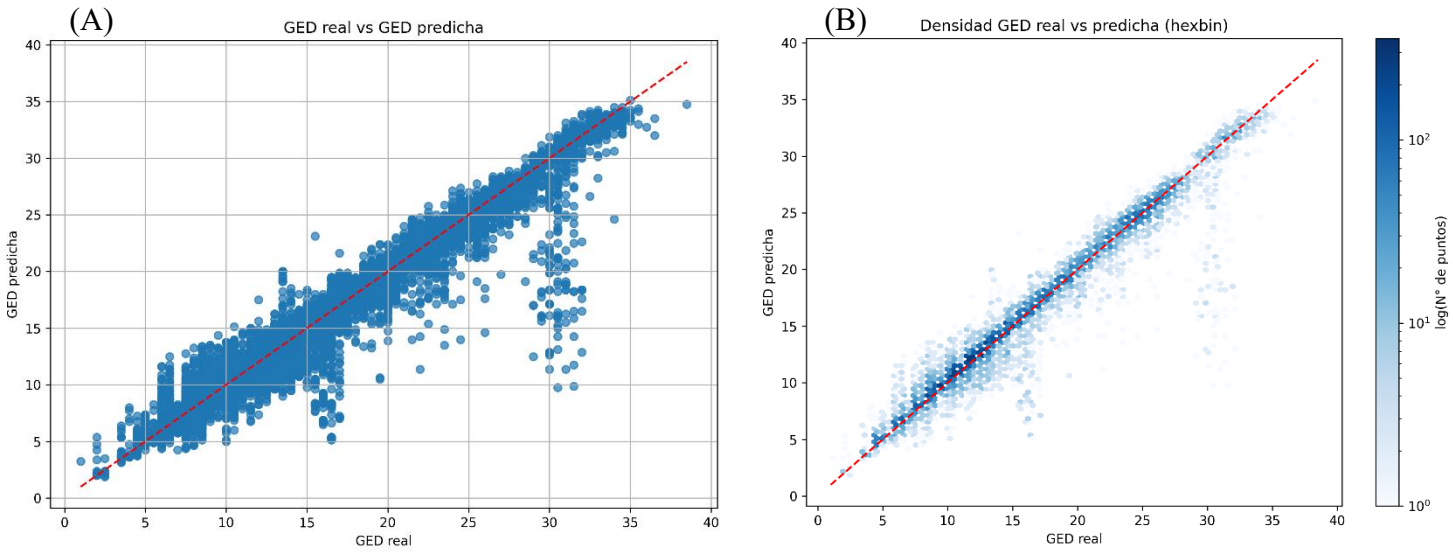


Figura 30. Representació dels valors(A) i la densitat(B)(logaritme del nº de punts) de punts de GED reals (continguts al dataset de test) respecte dels valors de GED predits pel mètode KNN. (FreeSolv Reduït)

En aquest cas, podem observar que els valors de les prediccions son més semblants als valors reals. Hi ha un parell d'excepcions, al voltant del 15 i del 30, on podem observar que es prediu un rang de valors inferior al que veritablement és. Això pot ser causat per algun outlier, que degut a com són les dades no s'ha mirat d'eliminar.

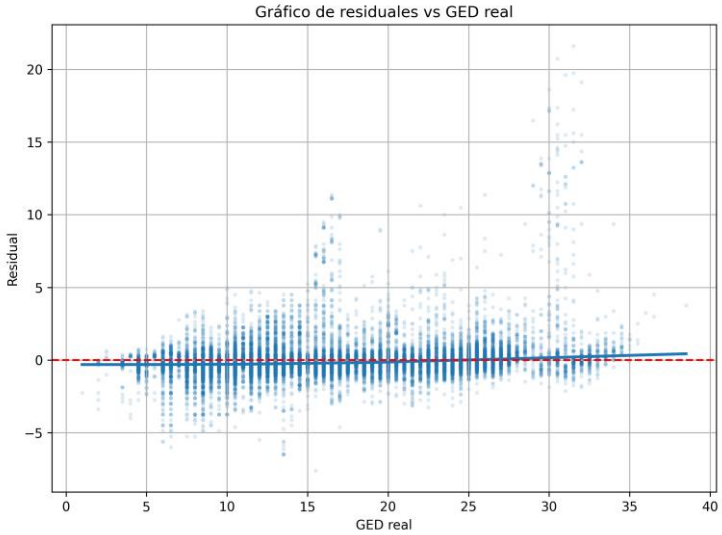


Figura 31. Representació dels valors de GED reals (continguts al dataset de test) respecte dels valors dels residuals $y_i - \hat{y}_i$ (FreeSolv Reduït)

El gràfic de residuals volta al 0 excepte el cas de les excepcions al valor de GED 15 i 30, que es torna a veure que s'està estimant un valor inferior de GED al que correspondria. Això és indicatiu d'una bona predicció.

9.2.8 Model Mordred FreeSolv

Aquest model serà entrenat amb el dataset obtingut a l'apartat 9.1.4.

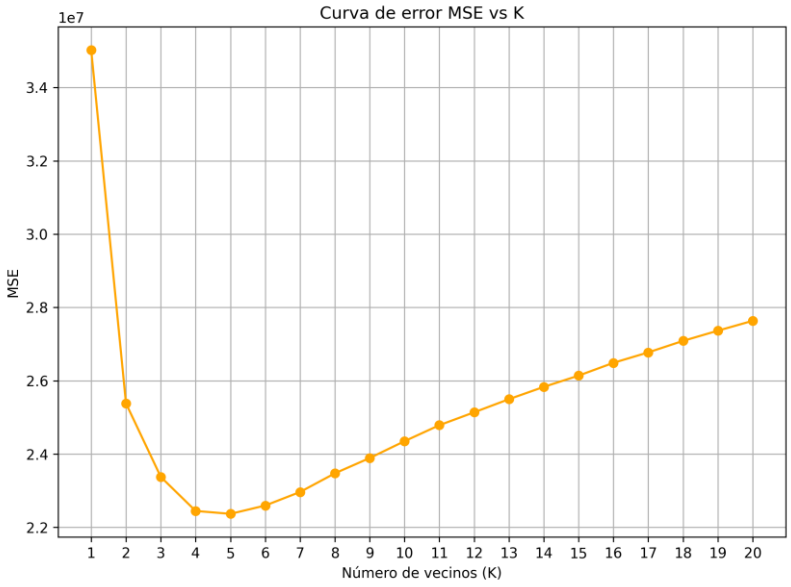


Figura 32 Representació del Mean Squared Error per avaluar els diferents valors del paràmetre K (FreeSolv Mordred)

En el cas de la distància Mordred per a les parelles del dataset FreeSolv, la K escollida ha estat 5.

Posteriorment, es tornarà a entrenar el model amb aquest valor de K = 5, i aquí és on es realitzaran les prediccions del conjunt de test.

Les mètriques d'error obtingudes per avaluar el model han estat les següents:

- MSE = 22.368.884,25
- MAE = 3289,85
- R^2 = 0,846

Per aquest cas, s'ha trigat aproximadament 4 hores per a la validació creuada, 7 minuts per tornar a entrenar el model i realitzar les prediccions, i 2 hores per obtenir els gràfics els resultats.

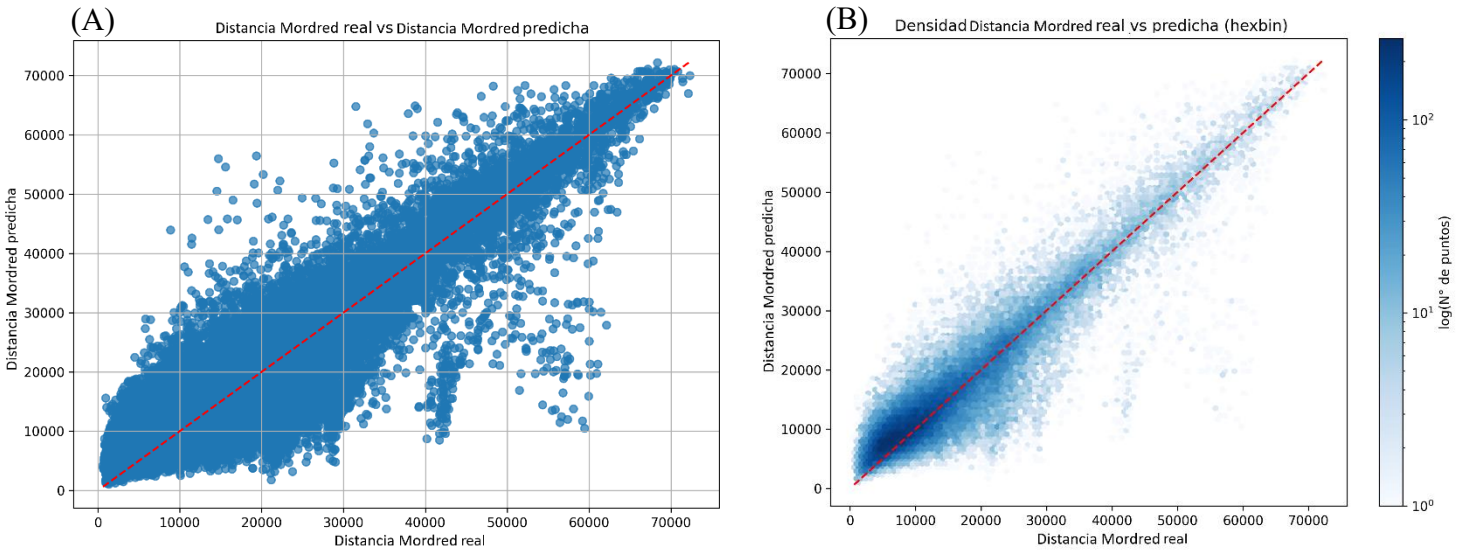


Figura 33. Representació dels valors i la densitat (logaritme del nº de punts) de punts de distància Mordred reals (continguts al dataset de test) respecte dels valors de GED predits pel mètode KNN. (FreeSolv)

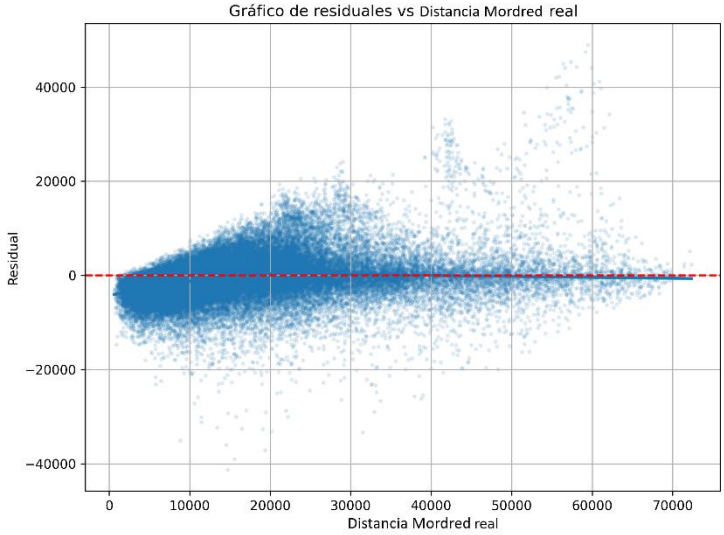


Figura 34. Representació dels valors de distància Mordred reals (continguts al dataset de test) respecte dels valors dels residuals
$y_i - \hat{y}_i$ (FreeSolv)

Es torna a veure una gran densitat de punts degut a la gran densitat de dades, i la distribució dels residuals és prou igual. Es torna a veure una cascada de punts en aquest cas a les distàncies Mordred de 40.000 i de 60.000, la qual cosa pot confirmar que pot ser degut a un outlier.

9.3 Discussions sobre les prediccions

Com es pot observar a les figures que representen el valor real respecte al valor predit (A) o la densitat de punts (B), tots els models estimen de manera correcta predient per aquelles parelles de molècules amb una GED real més gran un valor de GED predita també superior, uns amb més precisió que altres. S'observa que els enfocaments per CPU realitzen prediccions més pessimistes, és a dir, valors de GED superiors als que són veritablement, veient que hi ha més densitat de punts per sobre de la línia vermella discontínua que representaria una predicció perfecta. En el cas dels enfocaments per GPU, la tendència és tot el contrari, predir valors inferiors, ja que veiem més densitat de punts per sobre d'aquesta línia de predicció perfecta.

Això també es pot veure observant la línia de tendència dels gràfics dels residuals, si aquesta línia blava contínua està per sobre de la línia vermella (residuals = 0) vol dir que el valor real menys el valor predit ($y_i - \hat{y}_i$) ha donat més gran que zero, indicant que el valor real és superior al predit. Això és el que s'observa als gràfics de GPU. El cas contrari és el que passa als de CPU.

Dels subdatasets que contenen els millors valors de GED calculats, aquesta distribució és menyspreable, el gràfic de residuals té la línia de tendència molt propera al 0, i la densitat de punts està molt al voltant de la línia de predicció perfecta. Això també es veu reflectit a la precisió del model, observant les mètriques.

Dels resultats d'entrenar els models, en podem extreure les següents taules comparatives amb les mètriques de cadascun dels 8 models:

Taula 10 Conté els valors de les mètriques per avaluar el rendiment de la KNN per predir la GED a la base de dades ESOL.

ESOL - GED	**MSE**	**MAE**	**R^2**
CPU	23,24	3,36	0,867
GPU	10,87	2,16	0,885
Dataset Reduït	3,92	0,945	0,952

Taula 11 Conté els valors de les mètriques per avaluar el rendiment de la KNN per predir la GED a la base de dades FreeSolv.

FreeSolv - GED	**MSE**	**MAE**	**R^2**
CPU	10,03	2,28	0,849
GPU	4,158	1,401	0,880
Dataset Reduït	1,888	0,713	0,959

Taula 12 Conté els valors de les mètriques per avaluar el rendiment de la KNN per predir les distàncies Mordred.

Mordred	MSE	MAE	R^2
ESOL	83.563.847,82	5881,77	0,906
FreeSolv	22.368.884,25	3289,85	0,846

En general, podem observar que la mètrica de MSE no ha estat la millor opció per a les distàncies Mordred, ja que escala molt ràpidament amb números d'error molt grans. Per les distàncies GED, les tres funcions de pèrdua representen molt bé els errors en les prediccions. Aquestes taules es faran servir per extreure conclusions per als diferents models entrenats.

9.4 Correlació GED / Distància Mordred

Una alternativa a la KNN per obtenir la GED de grafs a partir de molècules químiques, ja que el càlcul de la distància Mordred té un cost menyspreable comparat amb el cost de càlcul de la GED, es vol investigar com a objectiu secundari si hi ha alguna correlació entre aquesta distància entre vectors de descriptors moleculars respecte a la GED entre dos grafs. Per simplificar el cost computacional, es realitzarà l'estudi només sobre les dades del dataset extret de la base de dades FreeSolv, de mida menor.

Primerament, s'ha calculat el coeficient de correlació de Pearson (R) per veure si es troba una forta correlació lineal entre aquestes variables. Aquest valor pot prendre valors entre [-1,+1]. Com més proper a 0, més feble serà la correlació, mentre que com més proper als extrems major correlació hi haurà entre aquestes variables. S'espera que, com parlem de distàncies, siguin directament proporcionals, per tant, un valor de R proper a +1.

El valor obtingut ha estat de 0,841. Això indica una correlació prou forta entre les distàncies Mordred i GED. Això vol dir que podem continuar endavant amb l'estudi.

S'ha representat la GED respecte la distància Mordred per tal de calcular la regressió lineal que millor ajusta els punts del dataset. Els resultats obtinguts són els de la figura 35, on podem observar l'equació de la regressió lineal, juntament amb el valor R^2.

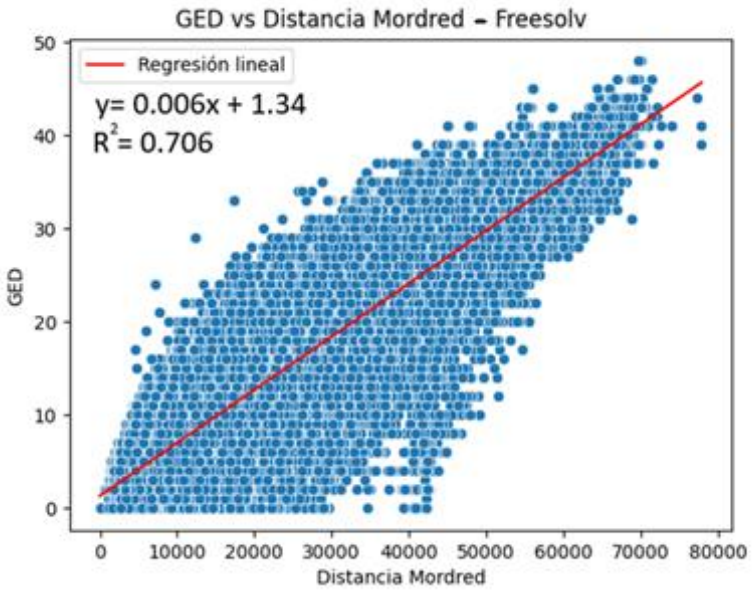


Figura 35. Regressió lineal dels valors de GED respecte la distància Mordred

Com es pot observar, tot i ser una bona correlació, molts valors es dispersen, i com l'escala de la distància Mordred és d'aproximadament 10^5 i la de la GED és de desenes, això pot produir error.

10. Conclusions

Principalment, podem observar que, per als càlculs de la GED, dins d'una mateixa base de dades, l'algorisme Fast Bipartite (GPU) ha calculat els valors de GED per entrenar un model que realitza unes prediccions més robustes a partir de les Morgan Fingerprints, possible indicador de què l'algorisme, a part de ser més ràpid, pot arribar a ser més eficient en referència a la fidelitat del graf molecular respecte a la proposta de la llibreria NetworkX (CPU).

Tret d'això, també es veu molt bé a simple vista que aquelles molècules que tant l'algorisme de CPU com de GPU han calculat valors de GED semblants, pertanyents a la categoria del dataset reduït a les taules 10 i 11 han servit per entrenar un model que ajusta de manera quasi perfecta la distància entre dues molècules. Com les molècules dels dos datasets eren de diferents característiques, la varietat d'empremtes moleculars amb què han estat entrenats els models és prou gran per a usar-los per predir grafs moleculars de tot tipus, sempre i quan es disposi dels codis SMILES. D'aquesta manera, podem concloure que s'ha complert l'objectiu principal d'aquesta tesi.

Respecte a les distàncies dels vectors de descriptors Mordred, també s'ha aconseguit entrenar dos models, un per a cada base de dades, amb resultats prou favorables. En aquest cas, com els valors de distàncies podien arribar a ser molt elevats, potser es podria millorar la robustesa del model realitzant alguna normalització, per exemple extraient el logaritme. Això podria ser objecte de futures investigacions sobre el tema.

Finalment, respecte a la correlació entre la GED i la distància Mordred, tot i haver obtingut uns resultats favorables, el valor de R^2 de 0,706 encara és inferior a les mètriques que demostren la qualitat dels millors predictors. D'aquesta manera, podem concloure que, a partir de la distància Mordred es podria aproximar el valor de la GED dels grafs moleculars de manera ràpida i senzilla aplicant la regressió lineal, però sacrificant precisió. De totes maneres, es podria realitzar una estimació la GED a partir de la distància Mordred a partir de, per exemple, un interval de confiança, motiu de futurs estudis. Actualment, el millor mecanisme per predir la GED és fent servir el KNN entrenat amb els subdatasets on la GED ha estat calculada amb els mateixos resultats per la CPU i la GPU. Per al cas de la regressió, es podria, en futurs estudis, proposar intervals de confiança per a casos on no es necessiti el valor exacte de la GED, o potser millorar la robustesa de la regressió incloent altres variables.

11. Bibliografia

[1] Rubén Megía González (Coordinador del área de formación). “De Gen A Carácter: El Dogma Central de La Biología Molecular.” *Genotipia*, 26 Mar. 2025, https://genotipia.com/dogma-central-bm/

[2] Amador, Samuel Antonio Sánchez. “Biomoléculas: Qué Son, Tipos, Funciones y Características.” *Psicología y Mente*, 3 Feb. 2025, https://psicologiaymente.com/salud/biomoleculas

[3] “7.13c: Homólogos, Ortólogos Y Paralogos.” *LibreTexts Español*, Libretexts, 2 Nov. 2022, https://espanol.libretexts.org/Biologia/Microbiologia/Libro%3A_Microbiolog%C3%ADa_(Sin_l%C3%ADmites)/7%3A_Gen%C3%A9tica_Microbiana/7.13%3A_Bioinform%C3%A1tica/7.13C%3A_Hom%C3%B3logos%2C_Ort%C3%B3logos_y_Paralogos

[4] Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Science 28: 31-36. https://doi.org/10.1021/ci00057a005

[5] “International Union of Pure and Applied Chemistry.” *IUPAC*, 4 Apr. 2025, https://iupac.org/

[6] “2.11: Fuerzas Intermoleculares y Puntos de Ebullición Relativos (PB).” *LibreTexts Español*, Libretexts, 2 Nov. 2022, https://espanol.libretexts.org/Quimica/Qu%C3%ADmica_Org%C3%A1nica/Mapa%3A_Qu%C3%ADmica_Org%C3%A1nica_(Wade)/02%3A_Estructura_y_Propiedades_de_Mol%C3%A9culas_Org%C3%A1nicas/2.11%3A_Fuerzas_intermoleculares_y_puntos_de_ebullici%C3%B3n_relativos_(pb)

[7] NetworkX Developers. “Graph_edit_distance#.” *Graph_edit_distance - NetworkX 3.5 Documentation*, 29 May 2025, https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.similarity.graph_edit_distance.html

[7-1]Abu-Aisheh, Zeina, et al. “An exact graph edit distance algorithm for solving pattern recognition problems.” *Proceedings of the International Conference on Pattern Recognition Applications and Methods*, 2015, https://doi.org/10.5220/0005209202710278.

[8] Serratosa, Francesc. “Fast computation of bipartite graph matching.” *Pattern Recognition Letters*, vol. 45, Aug. 2014, pp. 244–250, https://doi.org/10.1016/j.patrec.2014.04.015.

[9] “P, NP, CoNP, NP Hard and NP Complete: Complexity Classes.” *GeeksforGeeks*, GeeksforGeeks, 22 Feb. 2025, https://www.geeksforgeeks.org/types-of-complexity-classes-p-np-conp-np-hard-and-np-complete/

[10] Communication, ProtoQSAR. “Descriptores Moleculares.” *ProtoQSAR*, 19 Aug. 2024, https://protoqsar.com/publicaciones/descriptores-moleculares/

[11] Moriwaki, Hirotomo, et al. “Mordred: A molecular descriptor calculator.” *Journal of Cheminformatics*, vol. 10, no. 1, 6 Feb. 2018, https://doi.org/10.1186/s13321-018-0258-y.

[12] “Molecular Fingerprints and Similarity Searching” *Molecular Fingerprints and Similarity Searching - Open Babel 3.0.1 Documentation*, 5 May 2020, https://open-babel.readthedocs.io/en/latest/Fingerprints/intro.html

[13] Rogers, David, and Mathew Hahn. “Extended-connectivity fingerprints.” *Journal of Chemical Information and Modeling*, vol. 50, no. 5, 28 Apr. 2010, pp. 742–754, https://doi.org/10.1021/ci100050t

[14] Google Cloud Team. “¿Qué Es El Aprendizaje Automático? Tipos y Usos | Google Cloud.” *Google*, Google, https://cloud.google.com/learn/what-is-machine-learning?hl=es-419

[15] IBM team. “¿Qué Es El Aprendizaje Supervisado?” *IBM*, 20 Mar. 2025, https://www.ibm.com/es-es/topics/supervised-learning

[16] IBM team. “What Is the K-Nearest Neighbors Algorithm?” *IBM*, 17 Apr. 2025, https://www.ibm.com/think/topics/knn

[17] Raafat, Ahmed. “K-Nearest Neighbor (KNN) Explained.” *Machine Learning Archive*, 8 Apr. 2023, https://mlarchive.com/machine-learning/k-nearest-neighbor-knn-explained/

[18] Jain, Deepak. “Cross-Validation Using KNN.” *Towards Data Science*, 21 Jan. 2025, https://towardsdatascience.com/cross-validation-using-knn-6babb6e619c8/

[19] Bergmann, Dave, and Cole Stryker. “What Is Loss Function?” *IBM*, 17 Apr. 2025, https://www.ibm.com/think/topics/loss-function

[20] Alake, Richmond. “Loss Functions in Machine Learning Explained.” *DataCamp*, DataCamp, 4 Dec. 2024, https://www.datacamp.com/tutorial/loss-function-in-machine-learning

[21] IBM team. “¿Qué Es El Sobreajuste?” *IBM*, 23 Jan. 2025 https://www.ibm.com/es-es/think/topics/overfitting

[22] Encord Team. “Mean Square Error (MSE): Machine Learning Glossary: Encord.” *Encord*, https://encord.com/glossary/mean-square-error-mse/

[23] Ahmed, M Waqar. “Understanding Mean Absolute Error (MAE) in Regression: A Practical Guide.” *Medium*, Medium, 24 Aug. 2023, https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide-26e80ebb97df

[24] “Coeficiente de Determinación.” *Wikipedia*, Wikimedia Foundation, 22 Feb. 2024, https://es.wikipedia.org/wiki/Coeficiente_de_determinaci%C3%B3n

[25] “Papers with Code - ESOL (Estimated Solubility) Dataset.” *Dataset | Papers With Code*, https://paperswithcode.com/dataset/esol-scaffold, presented in [25-1]

[25-1] Li, Yuquan, et al. “An adaptive graph learning method for automated molecular interactions and properties predictions.” *Nature Machine Intelligence*, vol. 4, no. 7, 23 June 2022, pp. 645–651, https://doi.org/10.1038/s42256-022-00501-8

[26] Mobley DL, Guthrie JP. FreeSolv: a database of experimental and calculated hydration free energies, with input files. J Comput Aided Mol Des. 2014 Jul;28(7):711-20. DOI: 10.1007/s10822-014-9747-x

[27] Larrosa, Juan MC. *ARS 101: Clique*, 25 Oct 2013,https://ars-uns.blogspot.com/2013/10/ars-101-clique.html

[28] Jimin Khim et. al. “Hungarian Maximum Matching Algorithm: Brilliant Math & Science Wiki.” *Brilliant*, https://brilliant.org/wiki/hungarian-matching/

[29] Prasanna, Chanaka. “Evaluating Model Performance with K-Fold Cross-Validation - a Practical Example.” *Medium*, Medium, 1 Aug. 2024, https://medium.com/@chanakapinfo/evaluating-model-performance-with-k-fold-cross-validation-a-practical-example-485aaeb01dc0

[30] Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

[31] Sayle, Roger. RasMol: program for molecular graphics visualisation http://rasmol.org/

11.1 Llibreries

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. *Array programming with NumPy*. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2.

Okuta Ryosuke, et. al. “CuPy: A NumPy-Compatible Library for NVIDIA GPU Calculations” Proceedings of Workshop on Machine Learning Systems (LearningSys) in

The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS). Preferred Networks 2017 http://learningsys.org/nips17/assets/papers/paper_16.pdf / https://cupy.dev/

The Pandas development team, pandas-dev/pandas: Pandas, 2020, https://doi.org/10.5281/zenodo.3509134 / https://pandas.pydata.org/

Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

Landrum, G. RDKit Documentation. Release 2017.09.1 https://doi.org/10.5281/zenodo.60510 / https://www.rdkit.org/

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/stable/index.html

J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007. https://doi.org/10.5281/zenodo.14940554 / https://matplotlib.org/

Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021.

12. Annexos

12.1 Enllaç al repositori GitHub

https://github.com/XaviVeC/tfgGEI

12.2 Altres

12.2.1 Equacions matemàtiques

Equació 1: Distància euclidiana

$$Distancia\ Euclidiana = \|v_1 - v_2\| = \sqrt{\sum_{i=1}^{n} (v_{1i} - v_{2i})^2}$$

Equació 2: Vector gradient

$$Vector\ Gradient = \nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right)$$

Equació 3: MSE

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}$$

Equació 4: MAE

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \bar{y}|}{n}$$

Equació 5: R^2

$$R^2 = 1 - \frac{\sigma_r^2}{\sigma^2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

12.2.2 Figures

12.2.2.1 Taules que han estat truncades per afavorir la visualització:

	id1	smiles1	id2	smiles2	ged_cpu	ged_gpu	ged_diff	ged_abs_diff	ged_rel_diff
0	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Fenfuram	Cc1occc1C(=O)Nc2ccccc2	41.0	47.0	-6.0	6.0	0.127660
1	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	citral	CC(C)=CCCC(C)=CC(=O)	49.0	43.0	6.0	6.0	0.122449
2	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Picene	c1ccc2c(c1)ccc3c2ccc4c5ccccc5ccc43	34.0	54.0	-20.0	20.0	0.370370
3	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Thiophene	c1ccsc1	58.0	37.0	21.0	21.0	0.362069
4	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	benzothiazole	c2ccc1scnc1c2	51.0	41.0	10.0	10.0	0.196078
5	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	2,2,4,6,6'-PCB	Clc1cc(Cl)c(c(Cl)c1)c2c(Cl)cccc2Cl	41.0	49.0	-8.0	8.0	0.163265
6	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Estradiol	CC12CCC3C(CCc4cc(O)ccc34)C2CCC1O	35.0	52.0	-17.0	17.0	0.326923
7	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Dieldrin	ClC4=C(Cl)C5(Cl)C3C1CC(C2OC12)C3C4(Cl)C5(Cl)Cl	38.0	51.0	-13.0	13.0	0.254902
8	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Rotenone	COc5cc4OCC3Oc2c1CC(Oc1ccc2C(=O)C3c4cc5OC)C(C)=C	18.0	61.0	-43.0	43.0	0.704918
9	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	2-pyrrolidone	O=C1CCCN1	56.0	38.0	18.0	18.0	0.321429

MAE (Mean average error): 10.3341
RMSE (Root Mean Squared Error): 12.6049
R^2: 0.1060

Filtered amb GED<=5: 189581

	id1	smiles1	id2	smiles2	ged_cpu	ged_gpu	ged_diff	ged_abs_diff	ged_rel_diff
10	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	2-Chloronapthalene	Clc1ccc2ccccc2c1	47.0	43.0	4.0	4.0	0.085106
18	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	2-Undecanol	CCCCCCCCCC(C)O	43.0	44.0	-1.0	1.0	0.022727
21	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Phorate	CCOP(=S)(OCC)SCSCC	45.0	45.0	0.0	0.0	0.000000
22	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Phenacetin	CCOc1ccc(NC(=O)C)cc1	44.0	45.0	-1.0	1.0	0.022222
25	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Theophylline	Cn1c(=O)n(C)c2nc[nH]c2c1=O	47.0	45.0	2.0	2.0	0.042553
26	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Butethal	CCCCC1(CC)C(=O)NC(=O)NC1=O	42.0	47.0	-5.0	5.0	0.106383
28	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Methyl octanoate	CCCCCCCC(=O)OC	45.0	43.0	2.0	2.0	0.044444
30	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Terbufos	CCOP(=S)(OCC)SCSC(C)(C)C	43.0	47.0	-4.0	4.0	0.085106
33	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	1-Methylfluorene	Cc1cccc2c1Cc3ccccc32	42.0	46.0	-4.0	4.0	0.086957
35	Amigdalin	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)...	Diphenylamine	N(c1ccccc1)c2ccccc2	43.0	45.0	-2.0	2.0	0.044444

Figura 36. Mostra de les comparatives de les 10 primeres parelles de molècules de la base de dades ESOL amb la seva GED calculada mitjançant CPU, mitjançant GPU i els càlculs de les diferències, diferències absolutes i relatives. A sota les 10 primeres molècules del dataset ESOL Reduït posterior a aplicar el filtre de GED inferior o igual a 5.

	id1	smiles1	id2	smiles2	ged_cpu	ged_gpu	ged_diff	ged_abs_diff	ged_rel_diff
0	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	methanesulfonyl chloride	CS(=O)(=O)Cl	19.0	18.0	1.0	1.0	0.052632
1	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	3-methylbut-1-ene	CC(C)C=C	17.0	18.0	-1.0	1.0	0.055556
2	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	2-ethylpyrazine	CCc1cnccn1	10.0	21.0	-11.0	11.0	0.523810
3	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	heptan-1-ol	CCCCCCCO	11.0	21.0	-10.0	10.0	0.476190
4	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	3,5-dimethylphenol	Cc1cc(cc(c1)O)C	10.0	22.0	-12.0	12.0	0.545455
5	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	2,3-dimethylbutane	CC(C)C(C)C	15.0	19.0	-4.0	4.0	0.210526
6	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	2-methylpentan-2-ol	CCCC(C)(C)O	15.0	20.0	-5.0	5.0	0.250000
7	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	1,2-dimethylcyclohexane	C[C@@H]1CCCC[C@@H]1C	12.0	21.0	-9.0	9.0	0.428571
8	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	butan-2-ol	CC[C@H](C)O	17.0	18.0	-1.0	1.0	0.055556
9	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	dibromomethane	C(Br)Br	21.0	16.0	5.0	5.0	0.238095

MAE (Mean average error): 7.9794
RMSE (Root Mean Squared Error): 9.3774
R^2: -0.3028

Filtered amb GED<=5: 72875

	id1	smiles1	id2	smiles2	ged_cpu	ged_gpu	ged_diff	ged_abs_diff	ged_rel_diff
0	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	methanesulfonyl chloride	CS(=O)(=O)Cl	19.0	18.0	1.0	1.0	0.052632
1	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	3-methylbut-1-ene	CC(C)C=C	17.0	18.0	-1.0	1.0	0.055556
5	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	2,3-dimethylbutane	CC(C)C(C)C	15.0	19.0	-4.0	4.0	0.210526
6	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	2-methylpentan-2-ol	CCCC(C)(C)O	15.0	20.0	-5.0	5.0	0.250000
8	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	butan-2-ol	CC[C@H](C)O	17.0	18.0	-1.0	1.0	0.055556
9	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	dibromomethane	C(Br)Br	21.0	16.0	5.0	5.0	0.238095
18	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	2,2-dimethylpentane	CCCC(C)(C)C	15.0	20.0	-5.0	5.0	0.250000
24	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	pentanenitrile	CCCCC#N	15.0	19.0	-4.0	4.0	0.210526
25	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	2-methylpropan-2-ol	CC(C)(C)O	19.0	18.0	1.0	1.0	0.052632
27	4-methoxy-N,N-dimethyl-benzamide	CN(C)C(=O)c1ccc(cc1)OC	propanal	CCC=O	19.0	17.0	2.0	2.0	0.105263

Figura 37. Mostra de les comparatives de les 10 primeres parelles de molècules de la base de dades FreeSolv amb la seva GED calculada mitjançant CPU, mitjançant GPU i els càlculs de les diferències, diferències absolutes i relatives. A sota les 10 primeres molècules del dataset FreeSolv Reduït posterior a aplicar el filtre de GED inferior o igual a 5.

12.2.2.2 Resultats de les prediccions

```
(base) andromeda@andromeda-System-Product-Name:~/Documentos/TFG_Xavi/tfg-inf$ /home/andromeda/
Documentos/TFG_Xavi/tfg-inf/.venv2/bin/python /home/andromeda/Documentos/TFG_Xavi/tfg-inf/knnG
ED.py
[2025-05-06 16:00:20] Crear generador de Morgan fingerprint (nuevo método)
[2025-05-06 16:00:20] Calcular fingerprints
[2025-05-06 16:02:55] Eliminar filas inválidas
[2025-05-06 16:02:56] Crear vectores de diferencia absoluta
[2025-05-06 16:02:59] Train-test split (70% entrenamiento, 30% test)
[2025-05-06 16:02:59] Búsqueda del mejor valor de K con GridSearchCV (usando MSE)
[2025-05-07 08:39:51] Mejor valor de K encontrado: 4
[2025-05-07 08:39:51] Entrenar modelo con el mejor K
[2025-05-07 09:05:20] Cálculo métricas de evaluación

Resultados del modelo KNN (K=4):
MSE: 23.2436
MAE: 3.3635
R²: 0.8689
[2025-05-07 09:05:26] Graficar MSE vs K
[2025-05-08 03:02:39] Fin ejecución
```

Figura 38. Registre de la consola al executar la KNN per a la GED de l'ESOL amb CPU. Conté els timestamps de quan ha ocorregut cada event, i finalment els resultats dels errors i per obtenir els gràfics.

```
(base) andromeda@andromeda-System-Product-Name:~/Documentos/TFG_Xavi/tfg-inf$ /home/andromeda/
Documentos/TFG_Xavi/tfg-inf/.venv2/bin/python /home/andromeda/Documentos/TFG_Xavi/tfg-inf/knnG
ED.py
[2025-05-07 11:01:21] Crear generador de Morgan fingerprint (nuevo método)
[2025-05-07 11:01:21] Calcular fingerprints
[2025-05-07 11:03:33] Eliminar filas inválidas
[2025-05-07 11:03:33] Crear vectores de diferencia absoluta
[2025-05-07 11:03:36] Train-test split (70% entrenamiento, 30% test)
[2025-05-07 11:03:36] Búsqueda del mejor valor de K con GridSearchCV (usando MSE)
[2025-05-08 15:32:42] Mejor valor de K encontrado: 3
[2025-05-08 15:32:42] Entrenar modelo con el mejor K
[2025-05-08 16:13:09] Cálculo métricas de evaluación

Resultados del modelo KNN (K=3):
MSE: 10.8738
MAE: 2.1636
R²: 0.8848
[2025-05-08 16:13:17] Graficar MSE vs K
[2025-05-09 05:20:51] Fin ejecución
```

Figura 39. Registre de la consola al executar la KNN per a la GED de l'ESOL amb GPU. Conté els timestamps de quan ha ocorregut cada event, i finalment els resultats dels errors i per obtenir els gràfics.

```
[2025-05-06 15:59:15] Crear generador de Morgan fingerprint (nuevo método)
[2025-05-06 15:59:15] Calcular fingerprints
[2025-05-06 15:59:56] Eliminar filas inválidas
[2025-05-06 15:59:57] Crear vectores de diferencia absoluta
[2025-05-06 15:59:57] Train-test split (70% entrenamiento, 30% test)
[2025-05-06 15:59:57] Búsqueda del mejor valor de K con GridSearchCV (usando MSE)
[2025-05-06 18:16:27] Mejor valor de K encontrado: 4
[2025-05-06 18:16:27] Entrenar modelo con el mejor K
[2025-05-06 18:20:10] Cálculo métricas de evaluación

Resultados del modelo KNN (K=4):
MSE: 3.9182
MAE: 0.9451
R²: 0.9528
[2025-05-06 18:20:14] Graficar MSE vs K
[2025-05-06 19:33:32] Fin ejecución
```

Figura 40. Registre de la consola al executar la KNN per a la GED de l'ESOL Reduït. Conté els timestamps de quan ha ocorregut cada event, i finalment els resultats dels errors i per obtenir els gràfics.

```
(.venv2) (base) andromeda@andromeda-System-Product-Name:~/Documentos/TFG_Xavi/tfg-inf$ time /home/andromeda
[2025-06-03 08:56:39] Crear generador de Morgan fingerprint (nuevo método)
[2025-06-03 08:56:39] Calcular fingerprints
[2025-06-03 08:58:14] Eliminar filas inválidas
[2025-06-03 08:58:15] Crear vectores de diferencia absoluta
[2025-06-03 08:58:16] Train-test split (70% entrenamiento, 30% test)
[2025-06-03 08:58:16] Búsqueda del mejor valor de K con GridSearchCV (usando MSE)
[2025-06-03 23:56:40] Mejor valor de K encontrado: 4
[2025-06-03 23:56:40] Entrenar modelo con el mejor K
[2025-06-04 00:22:23] Cálculo métricas de evaluación

Resultados del modelo KNN (K=4):
MSE: 83563847.8174
MAE: 5881.7664
R²: 0.9059
[2025-06-04 00:22:29] Graficar MSE vs K
[2025-06-04 08:58:16] Fin ejecución
```

Figura 41. Registre de la consola al executar la KNN per a la Distància Mordred al dataset ESOL. Conté els timestamps de quan ha ocorregut cada event, i finalment els resultats dels errors i per obtenir els gràfics.

```
[2025-05-06 15:52:14] Crear generador de Morgan fingerprint (nuevo método)
[2025-05-06 15:52:14] Calcular fingerprints
[2025-05-06 15:52:38] Eliminar filas inválidas
[2025-05-06 15:52:38] Crear vectores de diferencia absoluta
[2025-05-06 15:52:38] Train-test split (70% entrenamiento, 30% test)
[2025-05-06 15:52:38] Búsqueda del mejor valor de K con GridSearchCV (usando MSE)
[2025-05-06 18:27:58] Mejor valor de K encontrado: 4
[2025-05-06 18:27:58] Entrenar modelo con el mejor K
[2025-05-06 18:32:20] Cálculo métricas de evaluación

Resultados del modelo KNN (K=4):
MSE: 10.0323
MAE: 2.2810
R²: 0.8493
[2025-05-06 18:32:24] Graficar MSE vs K
[2025-05-06 19:53:47] Fin ejecución
```

Figura 42. Registre de la consola al executar la KNN per a la GED de FreeSolv amb CPU. Conté els timestamps de quan ha ocorregut cada event, i finalment els resultats dels errors i per obtenir els gràfics.

```
(base) andromeda@andromeda-System-Product-Name:~/Documentos/TFG_Xavi/tfg-inf$ /home/andromeda/
Documentos/TFG_Xavi/tfg-inf/.venv2/bin/python /home/andromeda/Documentos/TFG_Xavi/tfg-inf/knnG
ED.py
[2025-05-07 11:02:42] Crear generador de Morgan fingerprint (nuevo método)
[2025-05-07 11:02:42] Calcular fingerprints
[2025-05-07 11:03:11] Eliminar filas inválidas
[2025-05-07 11:03:11] Crear vectores de diferencia absoluta
[2025-05-07 11:03:12] Train-test split (70% entrenamiento, 30% test)
[2025-05-07 11:03:12] Búsqueda del mejor valor de K con GridSearchCV (usando MSE)
[2025-05-07 14:44:44] Mejor valor de K encontrado: 3
[2025-05-07 14:44:44] Entrenar modelo con el mejor K
[2025-05-07 14:51:18] Cálculo métricas de evaluación

Resultados del modelo KNN (K=3):
MSE: 4.1583
MAE: 1.4009
R²: 0.8803
[2025-05-07 14:51:23] Graficar MSE vs K
[2025-05-07 17:08:38] Fin ejecución
```

Figura 43. Registre de la consola al executar la KNN per a la GED de FreeSolv amb GPU. Conté els timestamps de quan ha ocorregut cada event, i finalment els resultats dels errors i per obtenir els gràfics.

```
[2025-05-06 15:57:49] Crear generador de Morgan fingerprint (nuevo método)
[2025-05-06 15:57:49] Calcular fingerprints
[2025-05-06 15:58:00] Eliminar filas inválidas
[2025-05-06 15:58:00] Crear vectores de diferencia absoluta
[2025-05-06 15:58:00] Train-test split (70% entrenamiento, 30% test)
[2025-05-06 15:58:00] Búsqueda del mejor valor de K con GridSearchCV (usando MSE)
[2025-05-06 16:19:14] Mejor valor de K encontrado: 4
[2025-05-06 16:19:14] Entrenar modelo con el mejor K
[2025-05-06 16:19:53] Cálculo métricas de evaluación

Resultados del modelo KNN (K=4):
MSE: 1.8878
MAE: 0.7125
R²: 0.9590
[2025-05-06 16:19:54] Graficar MSE vs K
[2025-05-06 16:32:15] Fin ejecución
```

Figura 44. Registre de la consola al executar la KNN per a la GED de FreeSolv Reduït. Conté els timestamps de quan ha ocorregut cada event, i finalment els resultats dels errors i per obtenir els gràfics.

```
(base) andromeda@andromeda-System-Product-Name:~/Documentos/TFG_Xavi/tfg-inf$ /home/andromeda/
Documentos/TFG_Xavi/tfg-inf/.venv2/bin/python /home/andromeda/Documentos/TFG_Xavi/tfg-inf/knnM
ordred.py
[2025-05-07 11:51:34] Crear generador de Morgan fingerprint (nuevo método)
[2025-05-07 11:51:34] Calcular fingerprints
[2025-05-07 11:52:10] Eliminar filas inválidas
[2025-05-07 11:52:10] Crear vectores de diferencia absoluta
[2025-05-07 11:52:11] Train-test split (70% entrenamiento, 30% test)
[2025-05-07 11:52:11] Búsqueda del mejor valor de K con GridSearchCV (usando MSE)
[2025-05-07 15:54:57] Mejor valor de K encontrado: 5
[2025-05-07 15:54:57] Entrenar modelo con el mejor K
[2025-05-07 16:02:06] Cálculo métricas de evaluación

Resultados del modelo KNN (K=5):
MSE: 22368884.2449
MAE: 3289.8577
R²: 0.8460
[2025-05-07 16:02:10] Graficar MSE vs K
[2025-05-07 18:05:12] Fin ejecución
```

Figura 45. Registre de la consola al executar la KNN per a la Distància Mordred al dataset de FreeSolv. Conté els timestamps de quan ha ocorregut cada event, i finalment els resultats dels errors i per obtenir els gràfics.

12.2.2.3 Consola de la correlació entre distàncies Mordred i GED

```
--- Distancia Mordred vs GED Freesolv ---
Correlación de Pearson: 0.8407 (p=0.0000e+00)
R² del modelo: 0.7067
Ecuación: GED ≈ 0.0006 x distancia_mordred + 1.3399
```

Figura 46. Registre de la consola al calcular el valor de R, R^2 i l'equació de la regressió.