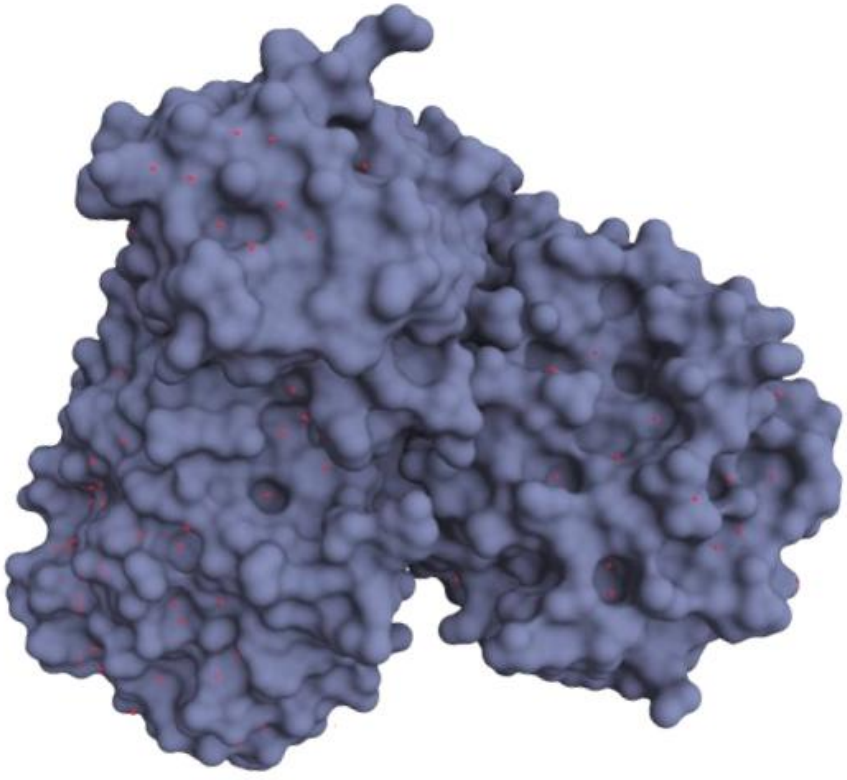
UNIVERSITAT ROVIRA i VIRGILI

Aplicació de xarxes neuronals de grafs a la predicció d'afinitat lligand–proteïna: impacte de l'entorn estructural i les interaccions febles en la Mpro del SARS-CoV-2

Arnau Dastis Fonoll

TREBALL FINAL DE GRAU BIOTECNOLOGIA



Tutor acadèmic: Santiago Garcia-Vallve

En cooperació amb: Cheminformatics & Nutrition

Supervisor/s: Santiago Garcia-Vallve

Data de convocatòria; Juny 2025

Jo, Arnau Dastis Fonoll , amb DNI 47748913-Q, sóc coneixedor de la guia de prevenció del plagi a la URV Prevenció, detecció i tractament del plagi en la docència: guia per a estudiants (aprovada el juliol 2017) i afirmo que aquet TFG no constitueixen cap de les conductes considerades com a plagi per la URV.

Tarragona 05 de 06 de 2025

(signatura)

Índex

1.Dades del Centre

Aquest treball s'ha desenvolupat dins del marc del grup de recerca Cheminformatics & Nutrition, un grup consolidat de la Universitat Rovira i Virgili (URV) inscrit al Departament de Bioquímica i Biotecnologia.

L'activitat científica del grup se centra en el desenvolupament i aplicació de tècniques de quimioinformàtica, bioinformàtica estructural, modelatge molecular i intel·ligència artificial per a l'estudi de la interacció entre compostos bioactius i dianes moleculars. A més, el grup col·labora activament en projectes relacionats amb la nutrició personalitzada, la seguretat alimentària, i el disseny racional de fàrmacs, tant en l'àmbit de la salut com en el de la prevenció de malalties.

2. Resum i paraules clau

2.1. Resum

El SARS-CoV-2, causant de la pandèmia global de la COVID-19, ha evidenciat la necessitat urgent de desenvolupar noves estratègies computacionals per a la identificació de fàrmacs antivirals. Aquest treball explora l'ús de xarxes neuronals de grafs (GNN) per predir l'afinitat entre lligands i la proteïna Mpro del SARS-CoV-2, una diana clau per al desenvolupament de fàrmacs antivirals. A partir de 386 estructures experimentals de complexes prtoteïna-lligand d'inhibidors no covalents de la Mpro, s'han generat i avaluat diferents models que incorporen informació del lligand, del seu entorn proteic i de les interaccions febles. Mitjançant validació creuada i la mètrica de l'error absolut mitjà (MAE), s'ha determinat que l'addició de l'entorn estructural millora la capacitat predictiva dels models, mentre que la incorporació explícita de les interaccions febles presenta resultats variables segons l'arquitectura i la configuració. Els resultats obtinguts subratllen el potencial de les GNN com a eina eficient per al cribratge virtual i la descoberta de fàrmacs, especialment en contextos d'emergència sanitària.

2.2. Paraules clau

SARS-CoV-2, Mpro, xarxes neuronals gràfiques, predicció d'afinitat, bioinformàtica estructural, interaccions febles, GNN, cribratge virtual, intel·ligència artificial, modelatge molecular.

3. Introducció

3.1 Context i motivació

L'any 2020 el SARS-CoV-2 va endinsar el món en una crisi sanitària global. La seva ràpida propagació i l'impacte socioeconòmic que va suposar, va posar en evidència la necessitat de tenir eines que facilitin poder combatre de manera ràpida i eficaç aquest tipus d'emergències. En aquest context neix el projecte de recerca Next- pandemics format per un grup multidisciplinari d'investigadors de la URV que té com a objectiu generar noves eines computacionals que ajudin la comunitat científica a buscar fàrmacs antivirals contra pandèmies actuals i futures.

3.2 El SARS-CoV-2: característiques i genoma

El SARS-CoV-2 és el virus responsable de la malaltia coneguda com a COVID-19, declarada pandèmia per l'Organització Mundial de la Salut (OMS) el març de 2020. Aquest virus pertany a la família Coronaviridae, dins de l'ordre Nidovirales, i al gènere Betacoronavirus. Es tracta d'un virus d'ARN monocatenari de sentit positiu (+ssRNA), això significa que el seu material genètic pot ser traduït directament pels ribosomes de la cèl·lula hoste (Bhat et al., 2021).

El genoma del SARS-CoV-2 té una longitud aproximada de 29.9 kb, cosa que el situa entre els virus d'ARN amb els genomes més llargs coneguts (Bhat et al., 2021). Aquest genoma es troba embolcallat dins d'una càpsida helicoidal i protegit per una membrana lipídica derivada de la cèl·lula hoste, on s'insereixen diverses proteïnes estructurals (Brant et al., 2021).

El genoma conté diversos marcs oberts de lectura (open reading frames, ORF). ORF1a codifica la poliproteïna pp1a i compta amb un *Frameshift Stimulation Element* (FSE) cap al final, aquest element és una seqüència de nucleòtids que estimulen la possibilitat que el ribosoma canviï la pauta de lectura, si això passa continuarà traduint fins a arribar al final de ORF1b generant la poliproteïna pp1ab. Tant pp1a com pp1ab posteriorment són processades per generar les diferents proteïnes no estructurals (nsps) essencials per a la replicació i transcripció de l'ARN viral (Brant et al., 2021; Wu et al., 2022).

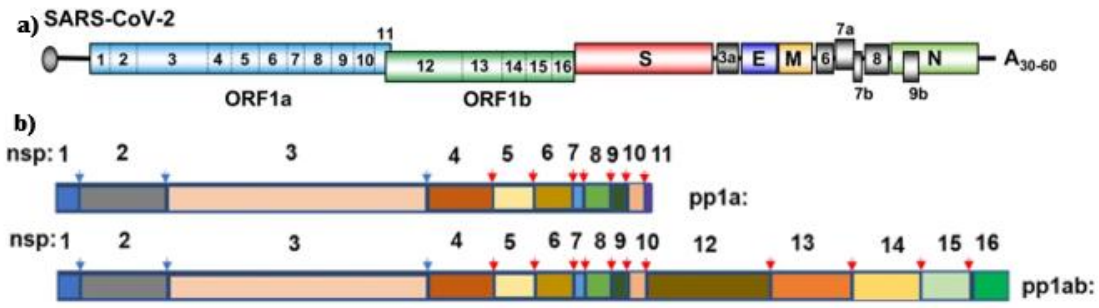


Figura 1. a) Representació esquemàtica del genoma del SARS-CoV-2, amb indicació dels principals marcs oberts de lectura (ORFs) i les regions que codifiquen per a les proteïnes estructurals i no estructurals (adaptat de Brant et al., 2021). b) Representació de les dues poliproteïnes derivades de la traducció dels ORFs (pp1a i pp1ab) i les corresponents proteïnes no estructurals (nsps) que se'n generen mitjançant processos de clivatge (adaptat de Wu et al., 2022).

Entre aquestes proteïnes no estructurals desataquen la Mpro i la PLpro pel seu rol com a proteases, ja que són les encarregades de generar les nsps a partir de pp1a i pp1ab.

La resta del genoma codifica per a les proteïnes estructurals i altres proteïnes accessòries que contribueixen a la capacitat del virus per evadir la resposta immunitària de l'hoste i afavorir la seva propagació (Bhat et al., 2021; Brant et al., 2021)

Estructuralment, el virus està format per quatre proteïnes principals:

- Proteïna S (Spike): responsable de la unió del virus al receptor ACE2 de les cèl·lules humanes, pas inicial per a la seva entrada.
- Proteïna M (Membrana): participa en l'assemblatge i donació de forma a la partícula viral.
- Proteïna E (Envolupant): juga un paper en l'assemblatge, la sortida i la patogenicitat del virus.
- Proteïna N (Nucleocàpside): s'uneix al genoma viral, estabilitzant-lo i contribuint en el procés de replicació i transcripció (Bhat et al., 2021; Brant et al., 2021).

3.3 SARS-Cov2-Mpro estructura funció i inhibidors.

Com explica La SARS-Cov2-Mpro és la principal proteasa del SARS-CoV-2, té un rol clau en la replicació viral, ja que s'encarrega de tallar les poliproteïnes virals en múltiples llocs, generant múltiples unitats funcionals. Aquest fet sumat a què aquesta cisteïna proteasa no tingui homòlegs en l'ésser humà la fa atractiva per a desenvolupar inhibidors que tinguin una alta afinitat per aquesta, ja que evitarien la replicació del virus tenint pocs efectes en el pacient. (Hu et al., 2022; Li et al., 2023; Shawky et al., 2024). La proteasa principal del SARS-CoV-2 és una cisteïna proteasa essencial per la maduració viral, encarregada de tallar les poliproteïnes virals en proteïnes funcionals. Aquesta proteasa té un pes molecular aproximat de 34 kDa per monòmer, i adopta una forma funcional homodimèrica, ja que els monòmers per separat tenen una activitat hidrolítica molt reduïda. (Shawky et al., 2024).

Cada monòmer consta de tres dominis:

- Domini I i II: contenen el lloc catalític.
- Domini III: format per hèlixs, és clau per la dimerització.

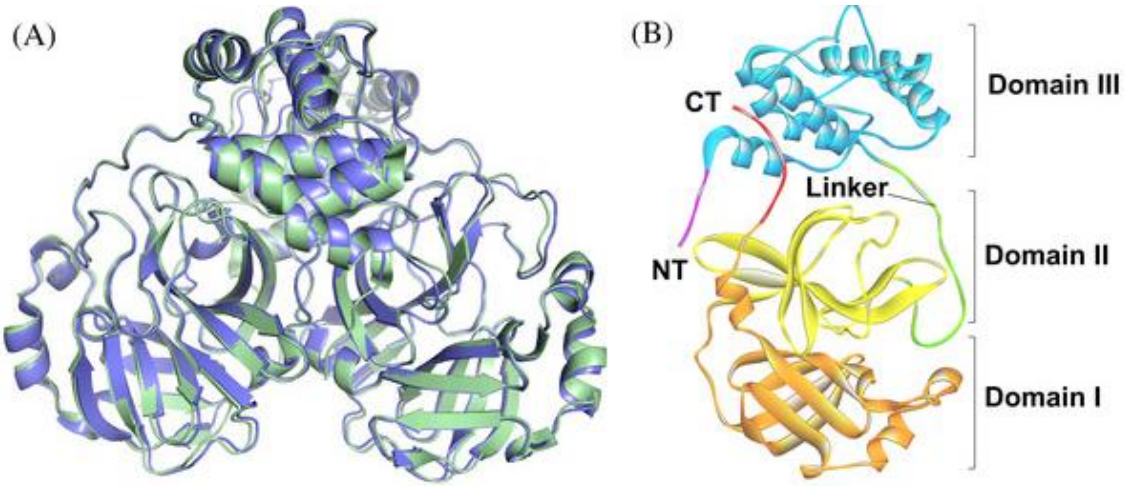


Figura 2. a) Estructura tridimensional de la proteasa principal del SARS-CoV-2 Mpro en la seva forma funcional de dímer. b) Representació esquemàtica d'un monòmer de Mpro amb indicació dels tres dominis estructurals: Domini I ,Domini II, i Domini III (Hu et al., 2022).

El lloc actiu està situat entre els dominis I i II i presenta una diada catalítica formada per His41 i Cys145, a diferència de moltes serines proteases que disposen d'una tríada.

La cavitat catalítica es divideix en cinc subpockets (S1, S2, S3, S4 i S1′), cadascuna amb especificitat per diferents residus del substrat (Fernandes et al., 2021; Hu et al., 2022):

- S1: és el subpocket clau per al reconeixement del substrat, ja que acomoda el residu de glutamina (Gln) en la posició P1. Està format per residus com Phe140, His163 i Glu166, que estableixen interaccions d'hidrogen amb el grup amida del substrat.
- S2: és un subpocket de caràcter hidrofòbic, format per residus com Met49 i His41, que reconeix aminoàcids com Leu o Phe en la posició P2 del substrat.
- S3 i S4: són regions més superficials i flexibles, que poden acomodar cadenes laterals més diverses. Tot i tenir menys impacte en l'activitat catalítica, poden influir en l'estabilització i selectivitat del substrat o inhibidor.
- S1′: es troba immediatament al costat de la diada catalítica. És petit i polar, i acomoda residus petits en la posició P1′ del substrat. Està directament implicat en el posicionament correcte per al clivatge.

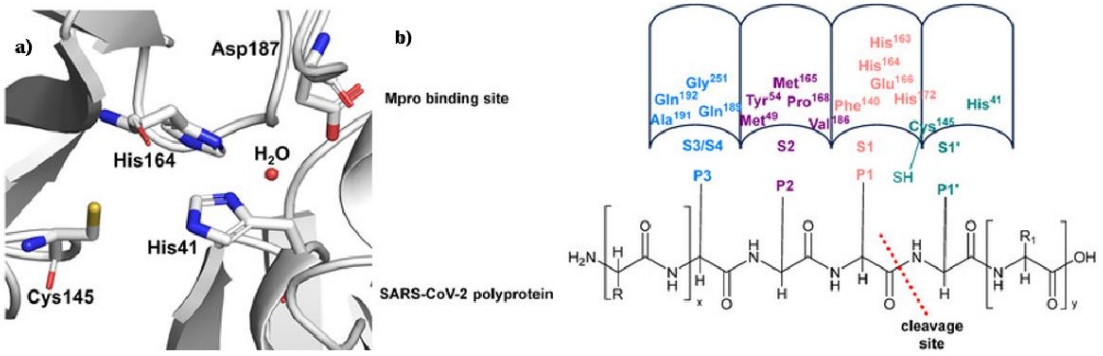


Figura 3. a) Vista detallada del lloc actiu de la proteasa principal del SARS-CoV-2 Mpro, mostrant la diada catalítica formada per His41 i Cys145, juntament amb altres residus clau com His164 i Asp187, implicats en l'activació d'una molècula d'aigua necessària per a la hidròlisi del substrat (adaptat de Hu et al., 2022).
b) Representació esquemàtica dels subpockets del lloc actiu de Mpro (S1, S1′, S2, S3/S4) i de les posicions corresponents del substrat (P1, P1′, P2, P3), destacant els residus implicats en el reconeixement i clivatge específic de la cadena polipeptídica viral (adaptat de Li et al., 2023).

El mecanisme de reacció transcorre en quatre passos seqüencials (Fernandes et al., 2021; Hu et al., 2022; Shawky et al., 2024):

1. Activació del nucleòfil: Inicialment, es forma un parell iònic com a resultat de la transferència del protó del grup tiol de Cys145 al residu His41, que actua com a base general. Això dona lloc a un anió tiolat, altament nucleofílic, capaç d'atacar el grup carbonil del substrat peptídic.
2. Formació de l'intermediari tioèster: El carboni carbonílic de l'enllaç amida del residu Gln (P1) del substrat és atacat pel tiolat. Simultàniament, His41 transfereix

un protó al nitrogen de l'amida, facilitant la ruptura de l'enllaç peptídic i donant lloc a la formació d'un intermediari tioèster covalent entre el substrat i Cys145.

3. Hidròlisi del tioèster: Una molècula d'aigua, activada per His41, realitza un atac nucleofílic sobre el carboni del tioèster. Aquest pas genera un intermediari tetraèdric amb un oxyanion transitori, que és estabilitzat per l'oxyanion hole, format pels residus Gly143, Ser144 i Cys145.
4. Alliberament del producte i regeneració del centre catalític: Finalment, l'enllaç entre l'àtom de sofre (Sγ) de Cys145 i el carboni carbonílic es trenca. His41, ara en forma catiònica, transfereix un protó a Cys145, restaurant la forma original de la diada catalítica i permetent l'alliberament dels productes hidrolitzats (fragments de la poliproteïna viral).

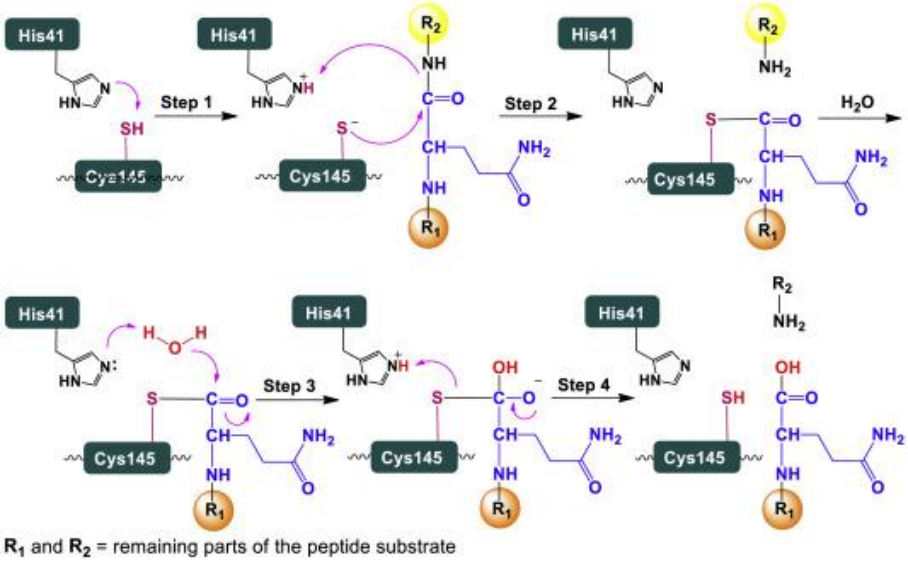


Figura 4 Mecanisme de catàlisi del substrat per part de la proteasa principal del SARS-CoV-2 Mpro. Es mostren les quatre etapes del procés catalític: (1) activació del nucleòfil Cys145 mitjançant la transferència de protó a His41; (2) atac nucleofílic al carboni carbonílic del substrat i formació de l'intermediari tioèster; (3) activació d'una molècula d'aigua per His41 i atac al tioèster; i (4) trencament de l'enllaç tioèster i alliberament dels productes finals, amb la regeneració del centre catalític (adaptat de Shawky et al., 2024).

A causa de l'interès per a inhibir aquesta proteïna, un gran nombre complexos han estat registrats al PDB, els quals es poden classificar segons la cinètica d'enllaç en una de tres categories: inhibidors no covalents, covalents reversibles i covalents irreversibles. Els inhibidors no covalents utilitzen interaccions febles per a unir-se a l'enzim i, per tant, són

capaços de revertir la unió un nombre il·limitat de vegades. En canvi, els inhibidors covalents, en general, segueixen dos passos per a inactivar l'enzim, primerament interactuen de manera reversible amb el binding poket, tot seguit, el grup tiol present en el pocket genera un atac nucleofílic que dona lloc a un enllaç covalent entre el lligand i l'enzim, depenent de la força d'aquest nou lligand l'enllaç serà reversible o irreversible (Zagórska et al., 2024).

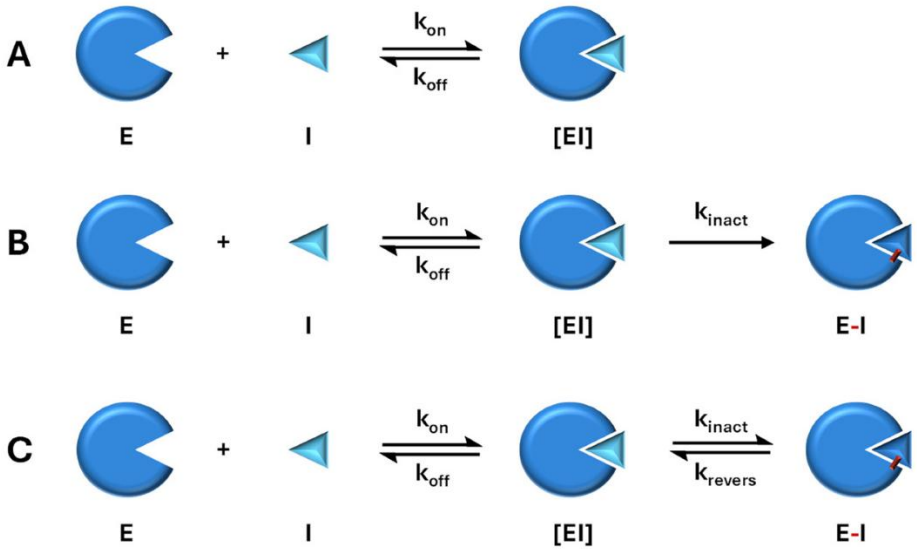


Figura 5. Tipus d'inhibició enzimàtica segons la naturalesa de la interacció entre l'enzim (E) i l'inhibidor (I).
a) Inhibidor no covalent: s'uneix de manera reversible a l'enzim formant el complex EI, regulat pels constants d'associació i dissociació k_{on} $k_{of.}$
b) Inhibidor covalent irreversible: després de la unió reversible, el complex EI pateix una reacció irreversible k_{inact} que dona lloc a un complex E–I permanent.
c) Inhibidor covalent reversible: el complex E–I format pot desfer-se, ja que l'enllaç covalent és susceptible de dissociació mitjançant k_{revers}, mantenint un equilibri amb [EI] (adaptat de Zagórska et al., 2024).

Tot i el gran nombre d'estructures publicades, avui dia només hi ha dos fàrmacs al mercat que actuen inhibint Mpro: nirmatrelvir (en combinació amb ritonavir, comercialitzat com a Paxlovid) i ensitrelvir (Xocova). Aquesta limitació es deu a diversos factors. En primer lloc, molts dels inhibidors candidats presenten problemes de farmacocinètica, com ara baixa biodisponibilitat oral o metabolisme hepàtic ràpid, que en dificulten l'ús clínic. A més, el risc de resistències víriques, com la mutació E166V a Mpro, pot reduir l'eficàcia d'aquests tractaments i comprometre la seva durabilitat. Tot i que Mpro no té cap homòleg directe en humans, i per tant és un objectiu molt selectiu, algunes proteases humanes com les catepsines poden reconèixer seqüències semblants, cosa que pot donar lloc a

interaccions fora diana i toxicitat inespecífica si els compostos no estan prou optimitzats. Finalment, el desenvolupament d'aquests fàrmacs és costós i lent, i sovint es prioritzen molècules amb perfils clínics ja coneguts. Per tot això, malgrat l'interès i el nombre elevat de candidats, només un petit nombre ha assolit l'aprovació reguladora (Li et al., 2023).

3.4 Bioinformàtica estructural i predicció de l'afinitat lligand-proteïna

La bioinformàtica estructural és una branca de la bioinformàtica dedicada a l'anàlisi de les estructures tridimensionals de molècules, especialment proteïnes i àcids nucleics. Aquesta disciplina permet entendre com la forma i les propietats químiques d'una proteïna determinen la seva funció biològica i la seva capacitat d'interaccionar amb altres molècules.

En el context del disseny de fàrmacs, un dels objectius principals és la predicció de l'afinitat entre una proteïna i un lligand, és a dir, quan fortament una molècula pot unir-se a una diana biològica. Aquesta afinitat depèn de múltiples factors estructurals i fisicoquímics, incloent-hi no només la complementarietat geomètrica i electrònica, sinó també un conjunt de forces febles que estabilitzen el complex format.

Les interaccions febles tenen un paper fonamental en la formació i estabilitat dels complexos lligand-proteïna. Tot i que aquestes interaccions tenen una energia individual relativament baixa, el seu efecte acumulatiu pot ser determinant per al reconeixement molecular específic i l'activitat biològica resultant.

Aquestes són les principals interaccions:

- Ponts d'hidrogen: Aquestes interaccions es donen entre un àtom d'hidrogen unit covalentment un àtom electronegatiu (generalment N, O o F) i un altre àtom electronegatiu amb un parell d'electrons lliure.
- Interaccions hidrofòbiques: quan dues molècules apolars es troben en un medi aquos tendeixen a agrupar-se per minimitzar la seva superfície exposada al medi polar. Aquestes interaccions estabilitzen el complex lligand-proteïna en medis aquosos, promovent una unió més estreta.
- Forces de Van der Waals: són forces intermoleculars febles degudes a fluctuacions en la densitat electrònica que generen dipols instantanis. Tot i ser

molt dèbils, la seva contribució pot ser rellevant quan es donen múltiples punts de contacte entre el lligand i la proteïna.

- Ponts salins: són interaccions electroestàtiques entre grups amb càrregues oposades, com ara entre un grup carboxilat ($–COO^-$) i un grup amoni ($–NH_3^+$). Aquestes interaccions poden tenir una força considerable i contribueixen a l'estabilitat i especificitat de la unió lligand-proteïna, especialment en ambients de pH fisiològic.
- Interaccions π-π: es donen entre sistemes aromàtics, com anells benzènics, i impliquen l'apilament dels núvols electrònics dels orbitals π. Aquestes interaccions són freqüents en interaccions entre residus aromàtics de proteïnes (com fenilalanina, tirosina o triptòfan) i grups aromàtics dels lligands, aportant estabilitat i orientació en la unió.

En el cas de la proteasa principal del SASRS-CoV-2 (Mpro), aquestes interaccions són especialment rellevants, ja que el lligand s'uneix a diversos subpockets dins el lloc actiu de l'enzim, cadascun amb característiques químiques específiques. La capacitat de captar i representar aquestes interaccions és, per tant, essencial per a predir de manera fiable la potència d'un inhibidor.

La capacitat de predir l'afinitat d'un lligand i una proteïna té una gran importància en el desenvolupament de fàrmacs, ja que permet prioritzar candidats amb major probabilitat d'èxit, reduint costos i temps. Això és especialment rellevant en situacions d'emergència sanitària, com la pandèmia de la COVID-19, on cal actuar amb rapidesa.

3.5 Xarxes Neuronals Gràfiques (GNN): concepte, aplicacions i estudis previs en predicció d'afinitat molecular

En els darrers anys les xarxes neuronals han explotat en popularitat en l'àmbit de la intel·ligència artificial. Aquestes xarxes fan servir una base de dades amb les característiques d'entrada i un valor associat a aquestes. Les xarxes s'entrenen través d'un procés iteratiu, el model intenta predir el valor associat a partir de les entrades i fa servir la diferència entre la predicció i el valor real per a ajustar paràmetres interns minimitzant l'error en cada iteració.

Per tal d'aplicar xarxes neuronals al context bioquímic, és necessari transformar les molècules d'estudi en una representació numèrica que sigui comprensible per al model.

Aquesta representació és possible gràcies a l'ús de grafs, una estructura matemàtica formada per nodes (que representen els àtoms) i arestes (que representen els enllaços químics). Tant cada node com cada aresta es poden descriure mitjançant un vector de característiques que recull informació rellevant, com ara el tipus d'àtom, l'electronegativitat, la càrrega parcial o el tipus d'enllaç entre d'altres.

Les Xarxes Neuronals Gràfiques (Graph Neural Networks, GNN) són un tipus especialitzat de xarxa neuronal dissenyada per operar sobre aquest tipus d'estructures. Aquestes xarxes estan formades per capes que processen la informació de cada node conjuntament amb la del seu entorn local, actualitzant iterativament els vectors de característiques a partir de les dades dels nodes veïns i de les arestes que els connecten. Així, a mesura que s'apliquen capes consecutives, els nodes acumulen informació més global de l'estructura molecular.

En arribar a l'última capa, s'utilitza una funció d'agregació global (com ara la mitjana o la suma) per combinar tots els vectors de node en una única representació vectorial de tota la molècula. Aquest vector agregat s'introdueix finalment en una xarxa neuronal convencional per generar una predicció de la propietat d'interès.

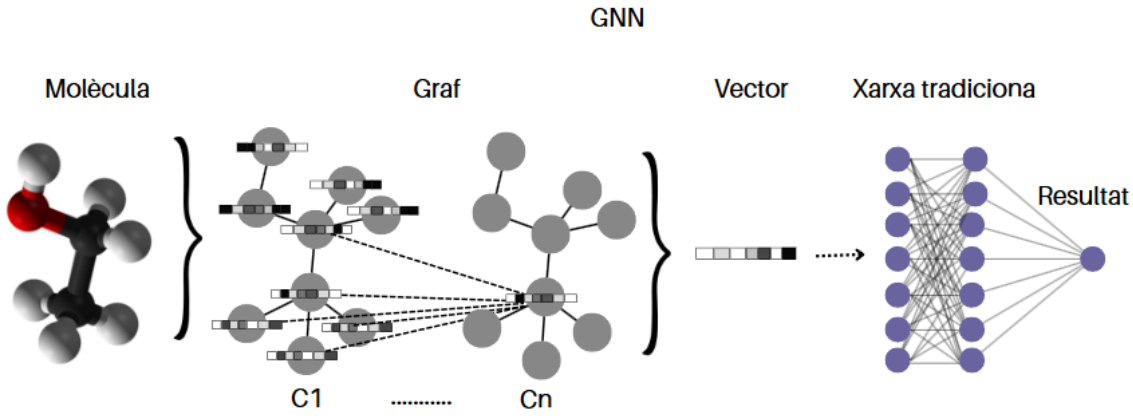


Figura 6. Representació del procés que segueix una xarxa neuronal de tipus GNN per predir un valor a partir d'una molècula:
La molècula s'expressa com un graf on els àtoms són nodes i els enllaços són arestes, cadascun amb el seu vector de característiques. A través de diverses capes de la GNN, cada node actualitza la seva representació vectorial combinant la seva informació amb la dels seus veïns. Al final, es genera un únic vector que encapsula la informació global de la molècula. Aquest vector s'introdueix en una xarxa neuronal tradicional per obtenir la predicció final.

Diversos estudis han demostrat l'eficàcia de les GNN en la predicció de propietats moleculars i, concretament, en la predicció de l'afinitat lligand-proteïna Huang et al.,

2023. Models com PLANET (Zhang et al., 2023) i GraphscoreDTA (Wang et al., 2023) han estat recentment desenvolupats per optimitzar aquesta tasca, demostrant que les representacions basades en grafs permeten capturar millor les interaccions complexes entre molècules i proteïnes que els models tradicionals. Aquests models, però, tenen un enfocament generalista, pretenent predir la interacció entre qualsevol parella proteïna-lligand a diferència dels nostres que s'enfoca únicament en l'afinitat Mpro-lligand. Amb aquest canvi esperem necessitar menys estructures per a entrenar, ja que a diferència de Zhang et al (2023) que comptaven amb més de 20000 estructures nosaltres només en tenim unes 400.

4. Hipòtesi de treball

La incorporació de l'entorn del lligand i les interaccions febles en els models de xarxes neuronals de grafs (GNN) millorarà el rendiment en la predicció de l'afinitat entre els inhibidors i la proteïna SARS-CoV-2 Mpro.

4.1. Objectiu general

Utilitzar les xarxes neuronals de grafs (GNN) per a generar un mètode que inclogui les interaccions febles entre la proteïna i el lligand, capaç de predir l'afinitat de diferents lligands amb la proteïna SARS-CoV-2 Mpro.

Per assolir l'objectiu general, es plantegen les següents tasques:

T1. Recopilació de dades

Recollir i preparar dades de lligands i de la proteïna SARS-CoV-2 Mpro, incloent-hi informació estructural i d'afinitat, així com identificar i codificar les possibles interaccions febles presents.

T2. Desenvolupament del model

Dissenyar i implementar diferents models de GNN per tal de comprovar si la incorporació de l'entorn molecular així com els enllaços febles afecten la predicció.

T3. Avaluació del model

Avaluar el rendiment dels models generats mitjançant MAE i comparar els resultats obtinguts per determinar la contribució de les interaccions febles en la capacitat predictiva del sistema. El MAE (Mean Absolute Error) és una mètrica que mesura l'error mitjà entre les prediccions del model i els valors reals; valors més baixos indiquen millor precisió.

5. Metodologia

La metodologia seguida en aquest treball es pot dividir en tres fases principals: recopilació de dades, generació de models i avaluació dels resultats.

5.1 Recopilació de dades

Al febrer de 2025 es va realitzar una cerca exhaustiva a la base de dades Protein Data Bank (PDB) (Berman, 2000) per identificar totes les estructures cristal·logràfiques que contenen la proteïna principal del SARS-CoV-2 (Mpro). Aquesta cerca va donar com a resultat un total de 1.590 estructures.

Per destriar aquest conjunt inicial, es va utilitzar el programa PDB-CAT (Llop-Peiró et al., 2024), desenvolupat pel grup QiN de la URV, que permet classificar i analitzar complexes proteïna-lligand de forma eficient. Amb aquest programari es van filtrar aquelles estructures que complien simultàniament les següents condicions:

- Presentaven un lligand unit a la proteïna.
- El lligand no estava unit covalentment a la Mpro.
- La seqüència de la Mpro no presentava cap mutació respecte a la seqüència de referència.

A continuació, es va dur a terme una cerca bibliogràfica per obtenir els valors de potència d'inhibició (IC_{50}) dels compostos corresponents. El IC_{50} indica la concentració necessària per inhibir el 50% de l'activitat de la proteïna, i es va convertir a escala logarítmica negativa (pIC_{50}), on valors més alts indiquen inhibidors més potents. Després d'aquest procés de selecció i conversió, es va obtenir un conjunt definitiu de 386 inhibidors no covalents, dels quals es coneix com s'uneixen experimentalment a la Mpro i es disposa de valors quantitatius de la seva activitat biològica.

Per a cadascun d'aquests 386 complexos proteïna-lligand, es van calcular les interaccions intermoleculars mitjançant el programa Arpeggio (Jubb et al., 2017), una eina que permet identificar les interaccions interatòmiques a partir de l'estructura tridimensional. Aquesta informació sobre les interaccions febles ha estat clau per a la generació dels descriptors emprats en la construcció dels models predictius desenvolupats posteriorment.

5.2 Generació de models

Amb l'objectiu d'avaluar diferents aproximacions en la predicció de l'activitat dels lligands, es van generar tres models amb arquitectures diferents: una xarxa atencional GAT (Graph Attention Network), una xarxa convolucional GIN (Graph Isomorphism Network) i una xarxa híbrida que combina elements de les dues anteriors, utilitzant la llibreria de Python MolGraph (Kensert et al., 2022). Cada model va ser entrenat amb 3 conjunts de dades diferents per comprovar si l'addició de l'entorn i les interaccions febles repercutia en la capacitat predictiva:

Entrada 1: Lligand

Aquest primer set es va confeccionar exclusivament amb la informació estructural del lligand, sense considerar l'entorn proteic.

Entrada 2: Lligand + Entorn

El segon set incorpora tant el lligand com l'entorn immediat de la proteïna, considerant els àtoms a distàncies de 4 Å, 5 Å, 6 Å, 7 Å,8 Å,9 Å respecte al lligand. Aquesta informació es va extreure a partir de les estructures del PDB obtingudes a l'apartat 7.1 utilitzant la llibreria ProDy (Bakan et al., 2011). L'objectiu d'aquesta configuració era determinar si la distància de l'entorn proteic influeix en la capacitat predictiva del model.

Entrada 3: Lligand + Entorn + Enllaços febles

Finalment, el tercer set afegeix als elements anteriors la informació sobre interaccions febles com:

- Enllaços d'hidrogen
- Interaccions hidrofòbiques
- Interaccions aromàtiques
- Van der Waals

Interaccions polars i apolars

Aquestes interaccions es van detectar mitjançant el programari Arpeggio i es van incorporar com a característiques químiques personalitzades dels enllaços mitjançant la llibreria RDKit i l'API de MolGraph. Concretament, es van registrar com a *features* amb la classe @bond_features.register, donant accés als models de manera explicita a les interaccions

5.3 Avaluació dels models

Per poder garantir la robustesa i fiabilitat dels resultats, els models van ser avaluats utilitzant mitjançant una estratègia de validació creuada de tipus k-fold (k=10). Aquesta tècnica consisteix a dividir el conjunt de dades en deu particions de mida similar i avaluar el model 10 cops fent servir 9 particions per a entrenar i una per a avaluar. Per a cada iteració la partició d'avaluació és una diferent d'aquesta manera cada dada és avaluada un cop. Això ens permet obtenir una avaluació del rendiment més fiable, ja que evita el sobreajustament i la dependència d'una única partició de les dades.

Com a mètrica principal de rendiment es va utilitzar l'error absolut mitjà (MAE, Mean Absolute Error), una mètrica habitual en la predicció de valors continus com la pIC_{50}, perquè proporciona una mesura intuïtiva i directament interpretable de la diferència mitjana entre les prediccions del model i els valors reals.

Addicionalment, per a cada iteració del k-fold, es van generar gràfics de dispersió (predicció vs. valor real) amb zones d'error marcades (±0.3, ±0.7 i ±1.0 pIC_{50}).

Aquest procediment d'avaluació es va aplicar a cadascuna de les combinacions de models (Lligand, Lligand + Entorn, Lligand + Entorn + Enllaços febles) i arquitectures (GAT, GIN i híbrid), cosa que permet comparar quantitativament l'impacte de cada configuració d'entrada i estructura neuronal sobre la capacitat predictiva final.

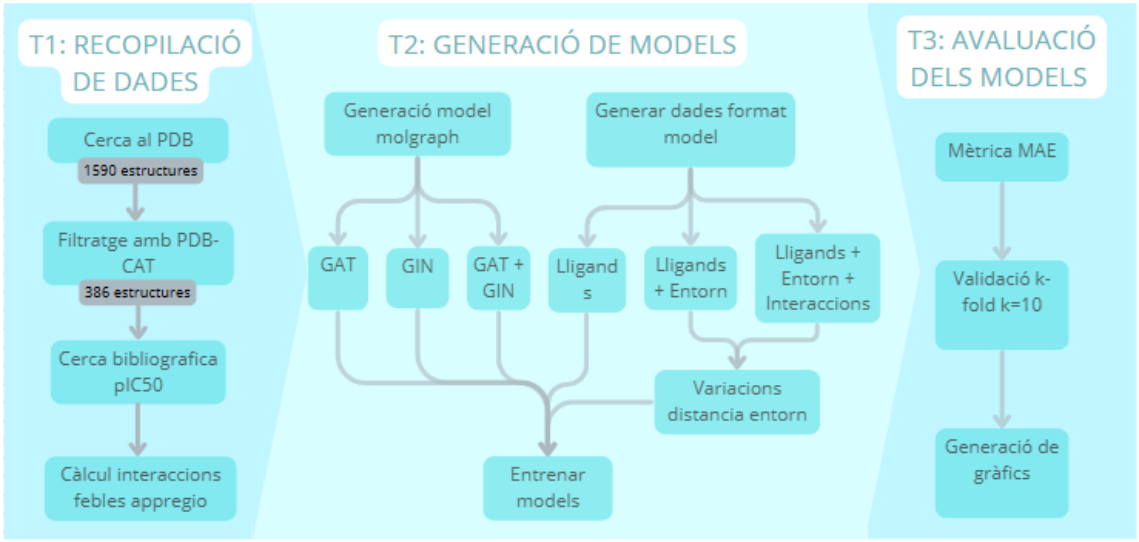


Figura 7. Diagrama de flux dels passos seguits durant la metodologia del projecte. Es divideix en tres etapes principals: T1) Recopilació de dades, incloent la cerca d'estructures al PDB, el filtratge mitjançant PDB-CAT, la recollida de valors de pIC50 i el càlcul d'interaccions febles; T2) Generació de models, on es construeixen diferents arquitectures basades en GAT, GIN i la seva combinació, amb dades de lligands, entorn i interaccions; i T3) Avaluació dels models, utilitzant la mètrica d'error absolut mitjà (MAE) i validació creuada amb k=10, juntament amb la generació de gràfics per analitzar el rendiment.

6.Resultats i discussió i relació amb els objectius

6.1. Resultats

Després d'entrenar amb el conjunt de 386 complexos proteïna-lligand d'inhibidors no covalents de la Mpro cadascuna de les arquitectures provades (GAT (Graph Attention Network), GIN (Graph Isomorphism Network) i la combinació GIN+GAT) amb les diferents configuracions d'entrada (lligand; lligand + entorn, variant les distàncies en Ångstroms; i lligand + entorn + interaccions febles, també variant distàncies), s'ha generat un gràfic de dispersió per a cada combinació.

Aquests gràfics mostren la relació entre els valors predits de pIC_{50} i els valors experimentals. Les diferents iteracions de la validació creuada (*k-fold cross-validation*) es representen amb colors diferenciats, i una línia vermella marca la recta identitat (x = y), que simbolitza la predicció perfecta. La proximitat dels punts a aquesta línia és indicativa de la qualitat predictiva del model: com més a prop s'hi troben, menor és l'error de predicció.

Per facilitar la interpretació visual dels resultats, s'han establert tres franges d'error:

- ± 0.3 pIC_{50}: Corresponent a una variació inferior al doble de concentració (2x)
- ± 0.7 pIC_{50}: Equivalent a una variació de cinc vegades la concentració (5x)
- (± 1 pIC_{50}: Representa errors de fins a deu vegades la concentració (10x)

Aquest sistema de representació visual permet avaluar de manera intuïtiva i comparativa l'eficiència de cada model i arquitectura provada. Addicionalment, el títol de cada gràfic incorpora el valor del MAE (Mean Absolute Error) associat, una mètrica quantitativa que reflecteix l'error mitjà absolut del model i permet comparar objectivament el rendiment entre les diferents arquitectures i configuracions d'entrada.

Observant els gràfics de dispersió ens podem fer una idea de com de fiables són els diferents models:

En primer lloc, els models entrenats únicament amb la informació del lligand (Figura 8) mostren valors de MAE de 0.5201 per a l'arquitectura GIN, 0.5411 per a GIN+GAT i 0.5524 per a GAT. En introduir l'entorn proteic a diferents distàncies, es poden observar millores clares en el rendiment. Per al model GAT (Figura 9), els valors de MAE

comencen a millorar entre 6 Å i 9 Å, amb una reducció progressiva fins a arribar a un mínim de 0.5088 a 8 Å. Això representa una millora de 0.0436 punts respecte al valor obtingut amb lligand sol (0.5524). En el cas de GIN (Figura 10), la millora s'inicia a partir de 6 Å i culmina a 9 Å amb un MAE de 0.4819, que suposa una reducció de 0.0382 punts en comparació amb el model basat només en el lligand (0.5201). Per a la combinació GIN+GAT (Figura 11), els resultats comencen a millorar a partir dels 7 Å, assolint el millor valor a 8 Å amb una MAE de 0.5095, és a dir, 0.0316 punts menys que el valor inicial (0.5411).

Quan s'incorporen també les interaccions febles, les dades mostren variacions addicionals. Amb l'arquitectura GAT (Figura 12), s'obté un MAE mínim de 0.5056 a 9 Å, millorant en 0.0468 punts respecte al model de lligand sol. Per al model GIN amb interaccions (Figura 13), el millor resultat torna a aparèixer a 9 Å amb un MAE de 0.4819, mantenint exactament la mateixa millora respecte al model inicial sense entorn ni interaccions. En el cas de la combinació GIN+GAT (Figura 14), el valor mínim de MAE es troba a 8 Å amb 0.4985, que representa una millora de 0.0426 punts respecte al model original entrenat només amb el lligand.

Per veure la resta de resultats de manera conjunta consulteu la Taula 1 ,on es recullen tots els valors de MAE per a les diferents arquitectures i configuracions.

En quant a la dispersió de les prediccions, a simple vista no es perceben diferències evidents entre els models observant els gràfics de dispersió. No obstant això, aquesta anàlisi es desenvoluparà amb més detall en figures posteriors.

		LLIGAND+ENTORN						LLIGAND+ENTORN+INTERACCIONS					
	LLIGAND	4Å	5Å	6Å	7Å	8Å	9Å	4Å	5Å	6Å	7Å	8Å	9Å
GAT	0.5524	0.5627	0.607	0.5358	0.5388	0.5088	0.5262	0.5803	0.635	0.5104	0.531	0.5189	0.5056
GIN	0.5201	0.5447	0.5898	0.5143	0.5024	0.5012	0.4819	0.5917	0.5789	0.5209	0.5132	0.4922	0.5511
GIN+GAT	0.5411	0.532	0.5353	0.5655	0.5095	0.5117	0.5068	0.5193	0.5602	0.5437	0.5191	0.4985	0.5119

Taula 1. Taula resum dels resultats de MAE obtinguts per les diferents arquitectures (GAT,GIN i GAT+GIN) entrenades amb els conjunts de dades lligand, lligand + entorn i lligand + entorn + interaccions a 4,5,6,7,8 i 9 Å

Lligand

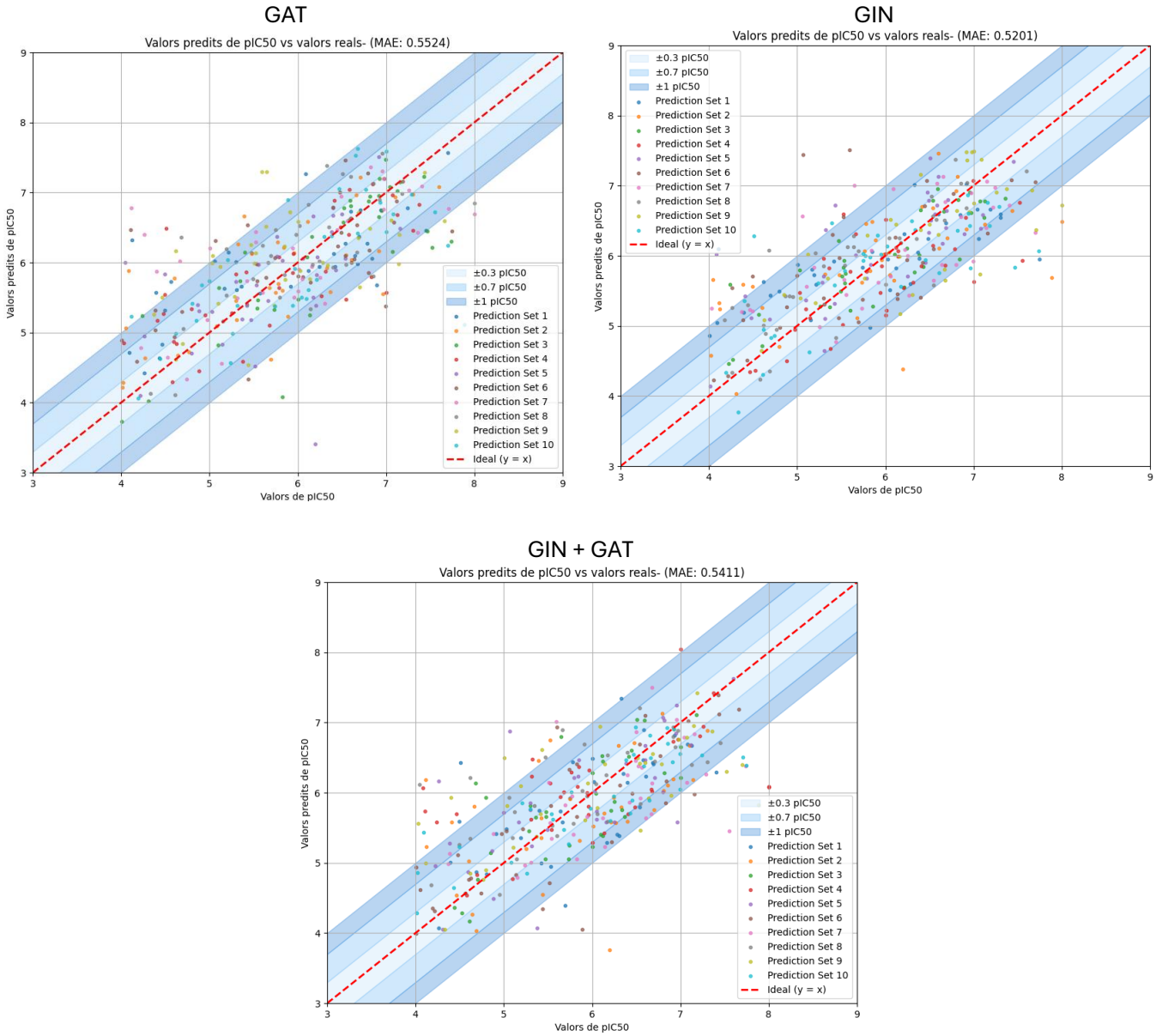


Figura 8. Dispersió dels valors de pIC_{50} predits versus els valors reals per als models GAT, GIN i GIN+GAT (considerant només el lligand). La línia vermella representa la predicció ideal (y = x), mentre que les bandes blaves indiquen els marges d'error: ±0.3, ±0.7 i ±1 unitat de pIC_{50} (aproximadament equivalents a errors de predicció de 0–2x, 2–5x i 5–10x, respectivament). Els valors de MAE (Mean Absolute Error) obtinguts són: 0.5524 per al GAT, 0.5201 per al GIN i 0.5411 per a la combinació GIN+GA

Lligand + entorn

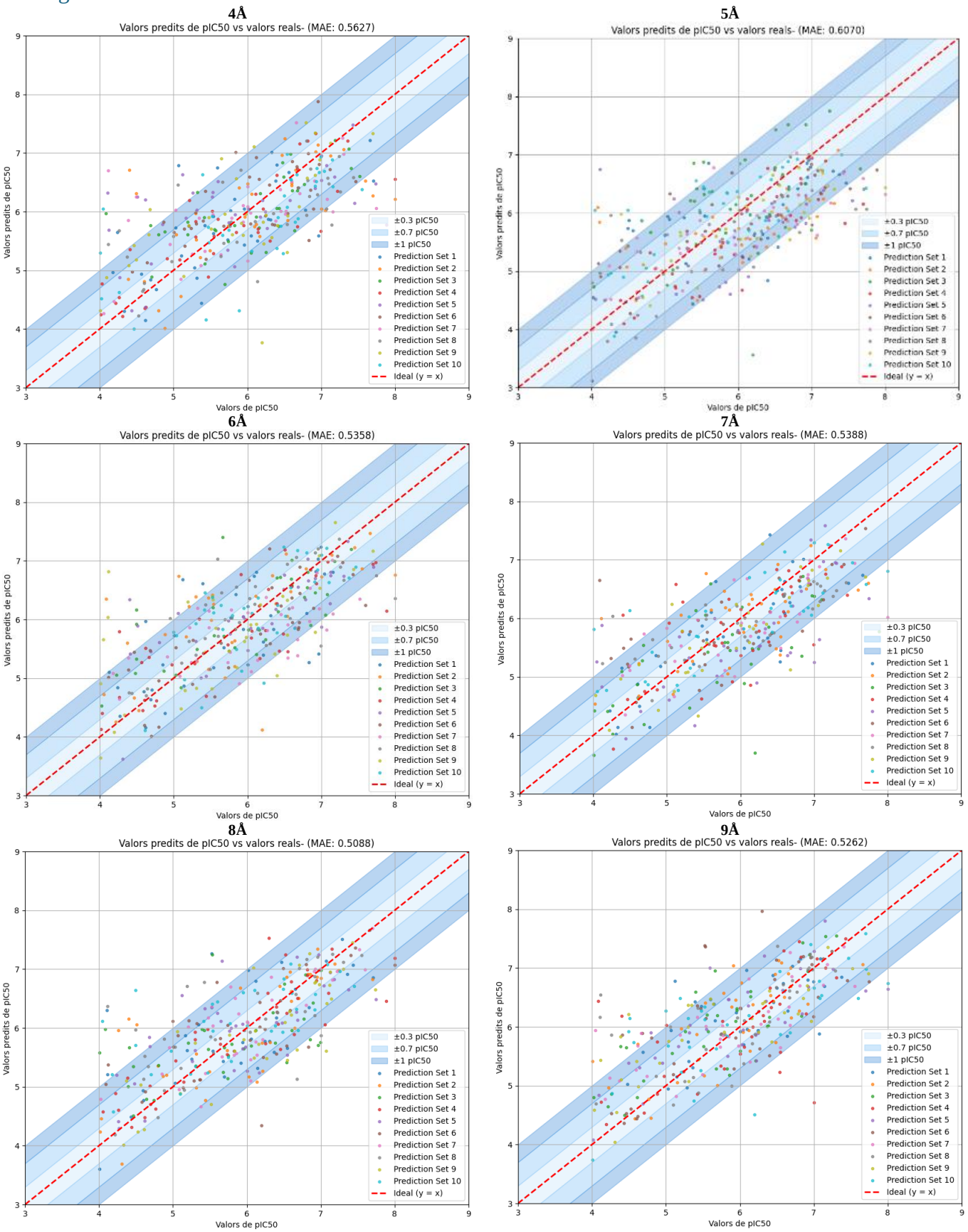


Figura 9 Gràfics de dispersió dels valors predits de pIC50 versus els valors reals per GAT entrenat amb Lligand + Entorn. Cada gràfic representa el resultat obtingut a certa distància . Els valors de Mean Absolute Error (MAE) obtinguts són els següents: 0.5627(4Å), 0.6070(5Å), 0.5358(6Å), 0.5388(7Å), 0.5088(8Å) i 0.5262(9Å).

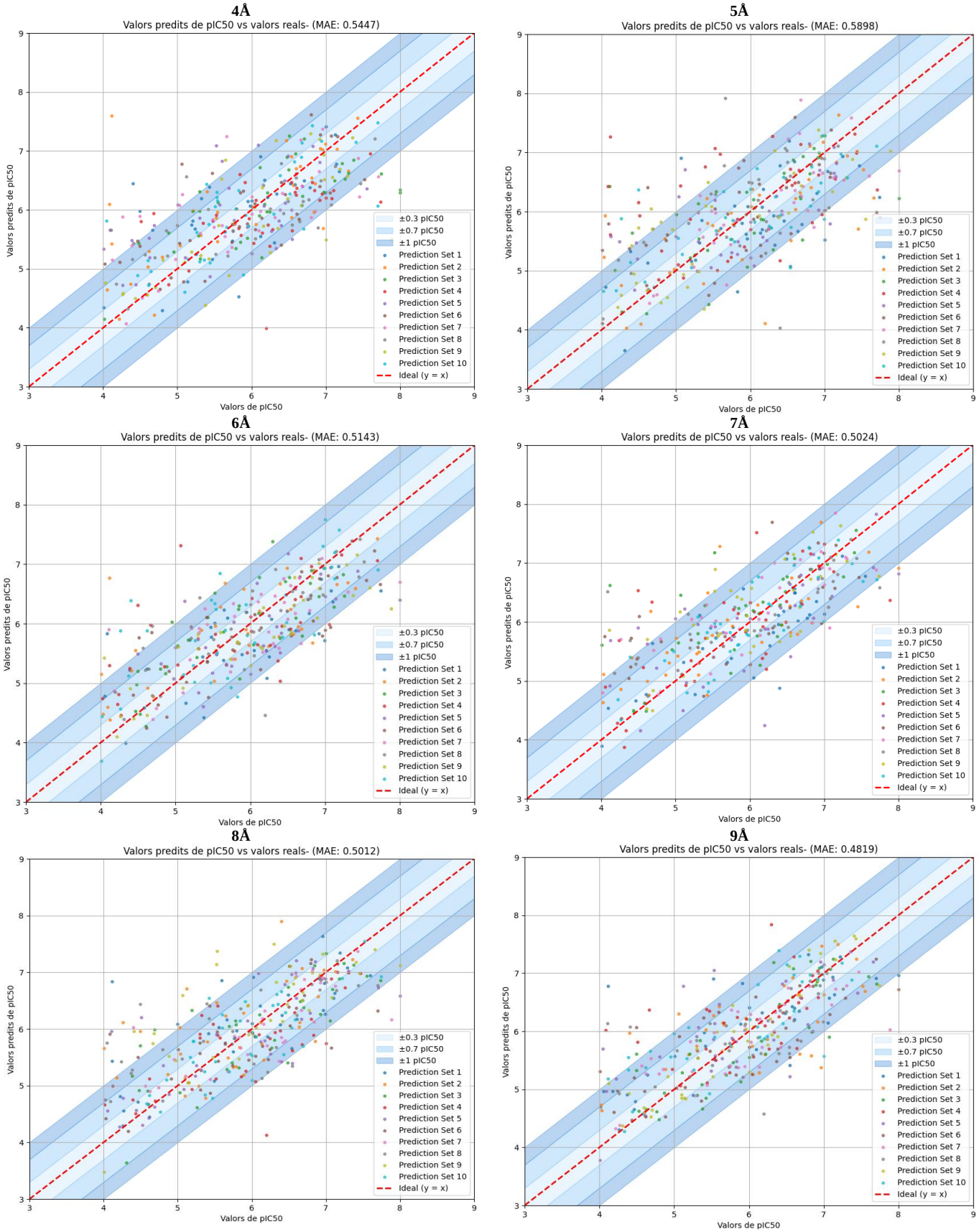


Figura 10. Gràfics de dispersió dels valors predits de pIC50 versus els valors reals per GIN entrenat amb Lligand + Entorn. Cada gràfic representa el resultat obtingut a certa distància . Els valors de Mean Absolute Error (MAE) obtinguts són els següents: 0.5447 (4 Å), 0.5898 (5 Å), 0.5143 (6 Å), 0.5024 (7 Å), 0.5012 (8 Å) i 0.4819 (9 Å)

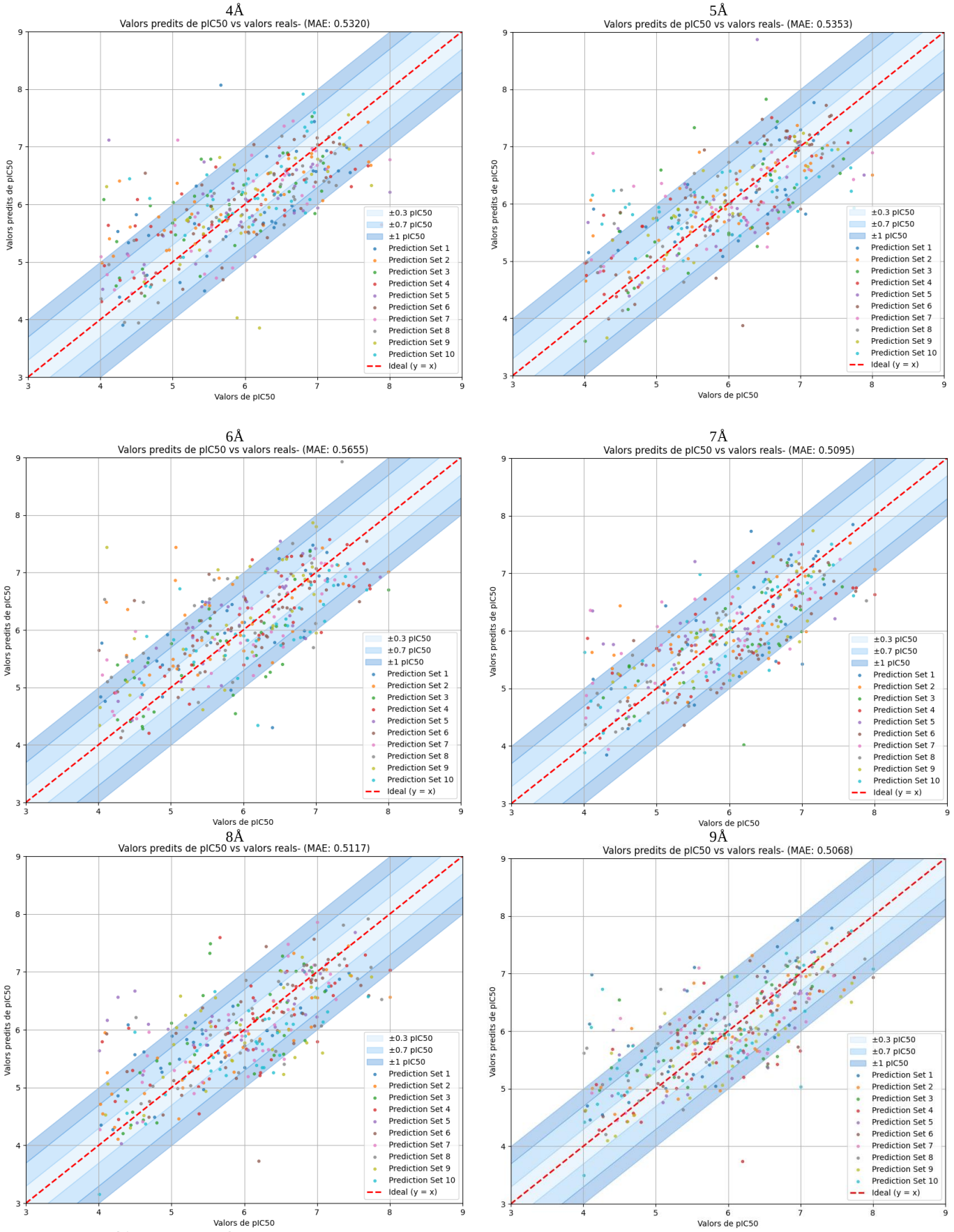


Figura 6. Gràfics de dispersió dels valors predits de pIC50 versus els valors reals per GIN+GAT entrenat amb Lligand + Entorn. Cada gràfic representa el resultat obtingut a certa distància . Els valors de Mean Absolute Error (MAE) obtinguts són els següents: 0.5320(4 Å), 0.5353(5 Å), 0.5655(6Å), 0.5095(7Å), 0.5117(8 Å) i 0.5068(9 Å)

Lligand + entorn + interaccions

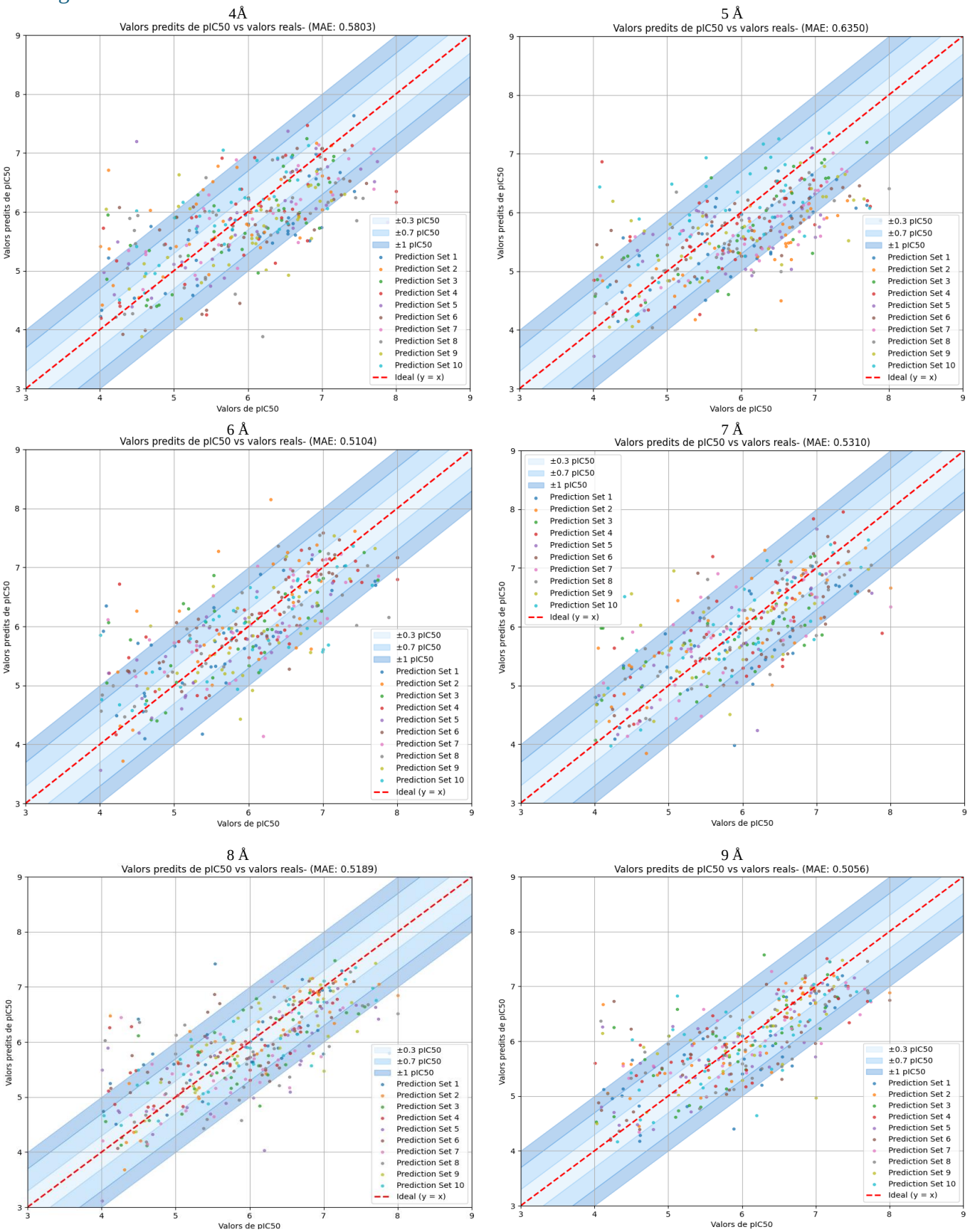


Figura 12. Gràfics de dispersió dels valors predits de pIC50 versus els valors reals per GAT entrenat amb Lligand + Entorn + Interaccions Cada gràfic representa el resultat obtingut a certa distància . Els valors de Mean Absolute Error (MAE) obtinguts són els següents: 0.5803(4 Å), 0.6350(5 Å), 0.5104(6 Å), 0.5310(7 Å), 0.5189(8 Å) i 0.5056 (9 Å)

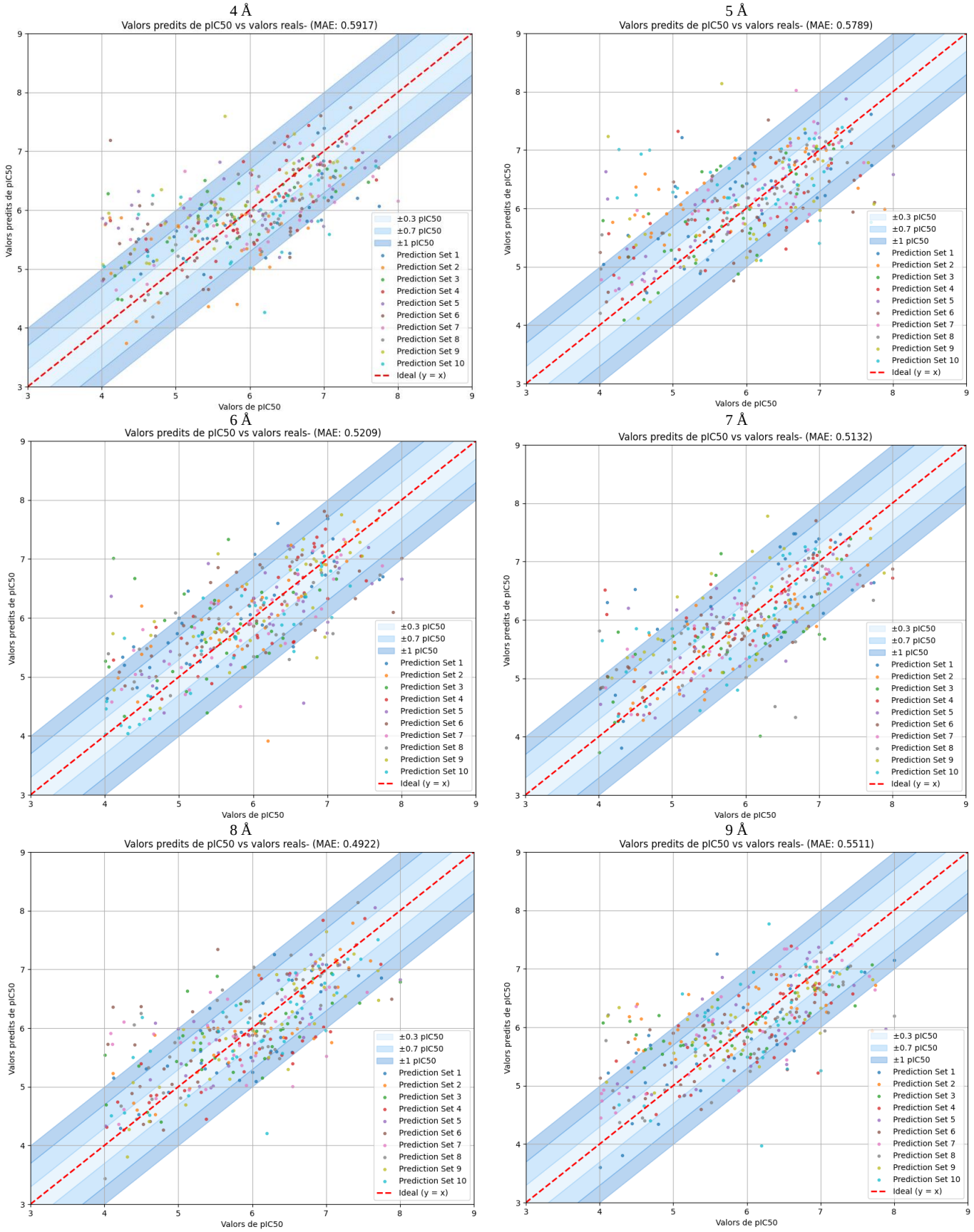


Figura 13. Gràfics de dispersió dels valors predits de pIC50 versus els valors reals per GIN entrenat amb Lligand + Entorn. Cada gràfic representa el resultat obtingut a certa distància . Els valors de Mean Absolute Error (MAE) obtinguts són els següents: 0.5917(4 Å), 0.5789(5 Å), 0.5209(6 Å), 0.5132(7 Å), 0.4922(8 Å) i 0.5511(9 Å)

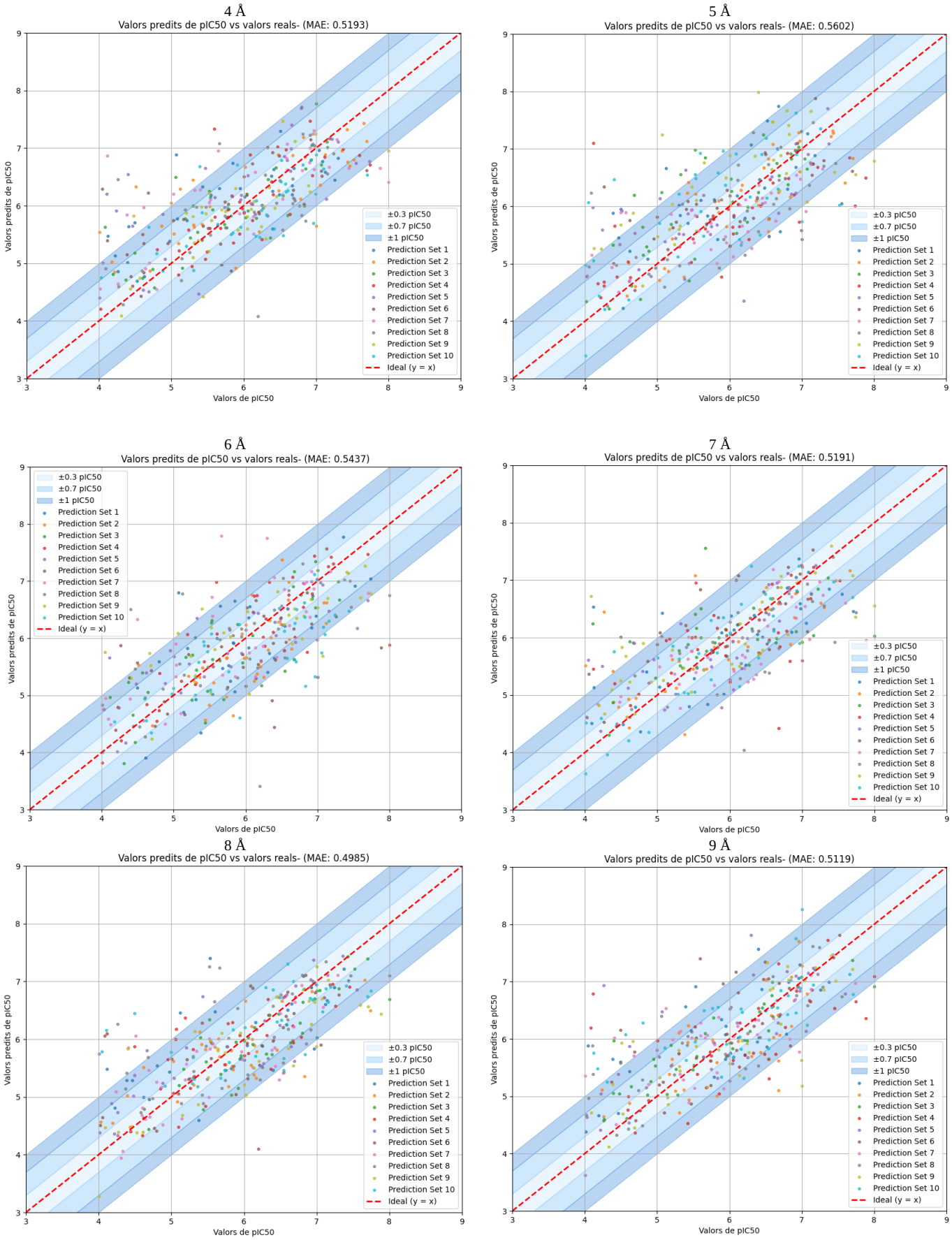


Figura 14. Gràfics de dispersió dels valors predits de pIC50 versus els valors reals per GIN+ GAT entrenat amb Lligand + Entorn Cada gràfic representa el resultat obtingut a certa distància . Els valors de Mean Absolute Error (MAE) obtinguts són els següents: 0.5193(4 Å), 0.5602(5 Å), 0.5437(6 Å), 0.5191(7 Å), 0.4985(8 Å) i 0.5119(9 Å)

6.2. Discussió

Per tal d'analitzar com afecta la distància al lligand de l'entorn a la capacitat predictiva del model hem generat un gràfic (Figura 15) que mostra la variació del MAE per a cada arquitectura (GAT, GIN i GIN+GAT) en funció de la distància de l'entorn proteic considerat (de 4 Å a 9 Å), distingint entre models amb només entorn (E) i models amb entorn més interaccions febles explícites (E+I). És important remarcar que aquest gràfic representa la diferència del MAE respecte al model base amb només el lligand per a cada arquitectura, de manera que no es pot utilitzar per comparar el rendiment entre arquitectures diferents, sinó únicament per avaluar com evoluciona el rendiment de cada model quan se li afegeix més informació.

Els resultats mostren una clara tendència: en general, l'addició d'un entorn proteic millora el rendiment dels models, especialment quan es considera un radi de 7 Å o 8 Å, amb variacions negatives del MAE en gairebé totes les arquitectures. Això indica que un entorn de mida mitjana–àmplia proporciona informació rellevant per a la predicció d'afinitat. En canvi, a distàncies més curtes com 4 Å o 5 Å, l'efecte sovint és neutre o fins i tot lleugerament negatiu, probablement perquè l'entorn és massa limitat per aportar context químic útil o bé perquè introdueix soroll estructural.

Quant a la comparació entre els models E i E+I, s'observa que la incorporació explícita de les interaccions no sempre condueix a una millora sistemàtica. En alguns casos, com GAT E+I a 6 Å i 8 Å, es registra una millora clara, mentre que en altres (per exemple, GIN E+I a 5 Å) l'impacte és neutre o lleugerament desfavorable. Això posa de manifest que l'eficàcia de representar explícitament aquestes interaccions depèn fortament del context i del tipus de xarxa.

Tanmateix, cal tenir en compte que, encara que les interaccions febles no s'hagin afegit explícitament en els models E, moltes d'elles ja estan implícitament representades en l'estructura del graf a través de la geometria local, la distància entre àtoms i les propietats dels nodes i arestes. És a dir, els models E ja poden aprendre patrons relacionats amb aquestes interaccions febles de manera indirecta, fet que podria explicar per què, en alguns casos, la inclusió explícita no genera una millora tan significativa com s'esperaria.

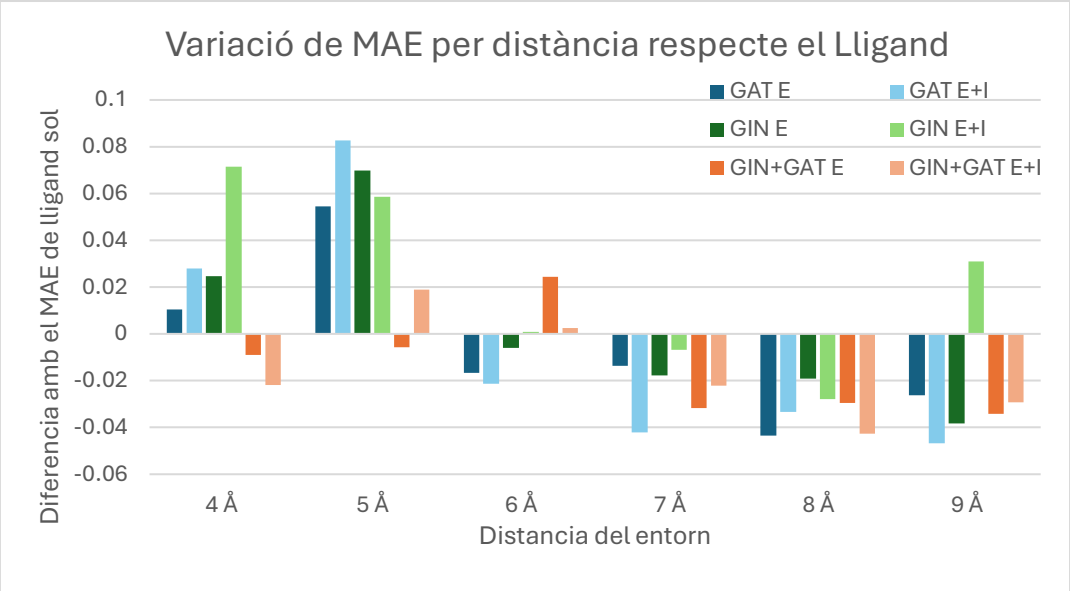


Figura 15. Variació del MAE respecte al model base (només lligand) en funció de la distància de l'entorn proteic considerat (de 4 Å a 9 Å), per a cadascuna de les arquitectures testades (GAT, GIN i GIN+GAT), tant amb entorn (E) com amb entorn més interaccions febles (E+I).

Respecte a quina arquitectura és millor per a la predicció en el *heatmap* de la Figura 16 s'observa que totes les arquitectures milloren el seu rendiment quan s'inclou l'entorn proteic, especialment a distàncies d'entre 7 Å i 9 Å. En aquest context, els models GIN E (0.481 a 9 Å) i GIN+GAT E+I (0.498 a 8 Å) són els que mostren els valors de MAE més baixos, tot i que la diferència respecte a la resta és relativament petita (de l'ordre de 0.02–0.05). Aquestes dades suggereixen que, tot i que GIN i la combinació GIN+GAT són lleugerament més precises, cap arquitectura destaca de manera contundent sobre les altres.

A més, un cop més s'evidencia que afegir interaccions febles (E+I) pot tenir un efecte positiu o neutre depenent de l'arquitectura i la distància, amb casos on la incorporació d'interaccions generen un MAE menor i altres on l'efecte és neutre o inclús negatiu.

En conclusió, totes les arquitectures presenten rendiments comparables, però es pot considerar que GIN E i GIN+GAT E+I amb entorns amplis (8–9 Å) ofereixen un lleuger avantatge en precisió, tot i que la diferència no és prou gran per a declarar un guanyador clar i universal.

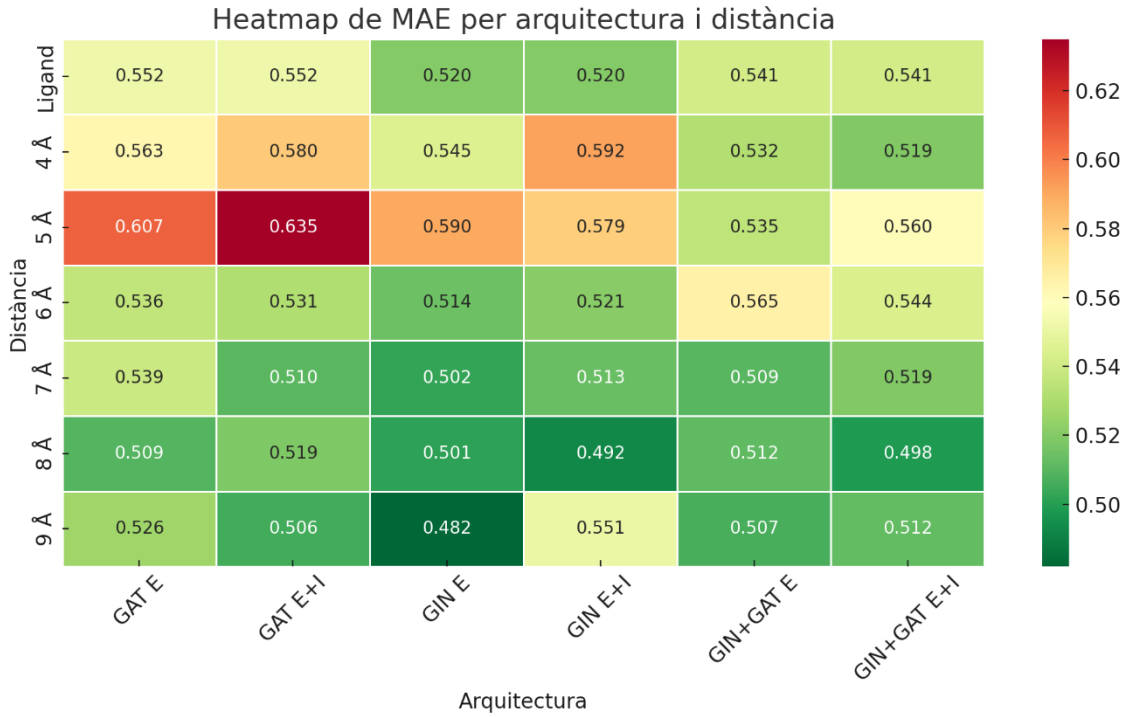


Figura 16. Valors absoluts del MAE obtinguts per cada combinació d'arquitectura i distància d'entorn en la predicció de l'afinitat lligand–Mpro. Els valors més baixos (indicats en verd) representen millor rendiment predictiu, mentre que els valors més alts (vermell intens) indiquen una major desviació respecte als valors reals.

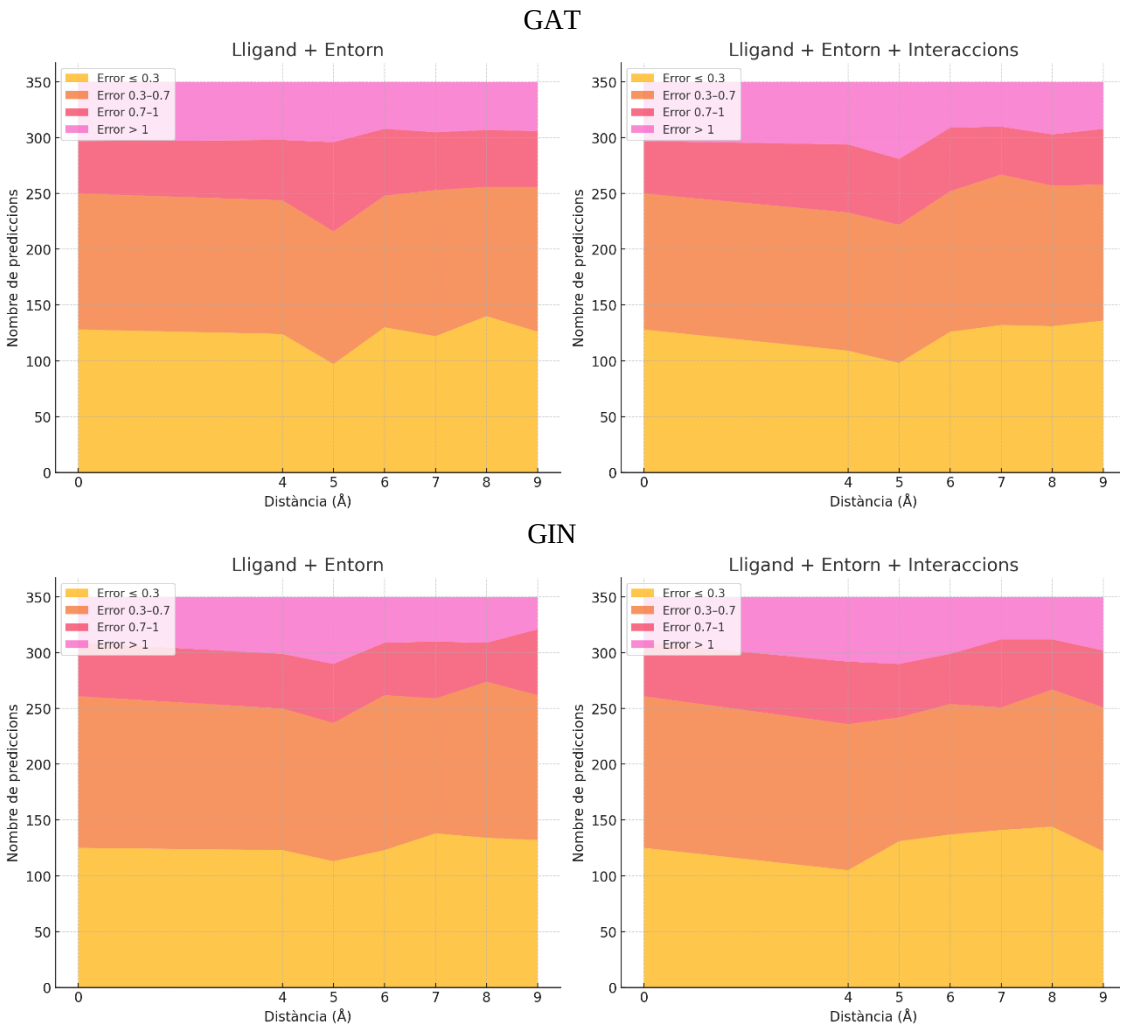
Tot i que la mètrica MAE ens proporciona una visió general del rendiment mitjà dels models, la Figura 17 permet analitzar amb més detall la distribució de les prediccions en funció dels intervals d'error, l'arquitectura emprada i el set de dades d'entrenament. Aquesta anàlisi és especialment útil per detectar patrons específics que poden quedar amagats darrere de les mitjanes generals. En general, es pot observar que, per a totes les arquitectures analitzades (GAT, GIN i GAT+GIN), la configuració "Lligand + Entorn + Interaccions" tendeix a presentar una proporció més elevada de prediccions amb error menor o igual a 0.3, sobretot en distàncies compreses entre 6 i 8 Å. Això suggereix que la incorporació d'interaccions febles pot millorar la precisió del model. No obstant això, cal destacar que aquestes diferències, tot i ser consistents, no sempre són marcadament significatives. En moltes distàncies, la diferència entre les arquitectures amb i sense interaccions és moderada, la qual cosa indica que els models bàsics ja tenen una capacitat predictiva rellevant.

Pel que fa al comportament específic per arquitectura, en el cas del model GAT, l'arquitectura "Lligand + Entorn" mostra una estabilitat inferior, amb una reducció notable en la proporció de prediccions bones (error ≤ 0.3) al voltant dels 5 Å que, tot i ser en els dos gràfics, és menys pronunciada en el segon. Tanmateix, amb la incorporació de

les interaccions febles, s'observa una millora clara en el tram de distàncies mitjanes (6–8 Å), amb un augment del nombre de prediccions de baix error.

El model GIN segueix una tendència similar, tot i que amb un descens més progressiu en les prediccions bones per a "Lligand + Entorn" també presenta una davallada en la zona de 5 Å. En aquest cas, la inclusió d'interaccions permet mantenir una proporció més alta i estable de prediccions ≤ 0.3 al llarg de les diferents distàncies evitant el descens dels 5 Å.

Finalment, la configuració combinada GAT+GIN mostra el millor comportament global: la versió amb interaccions manté una proporció elevada de prediccions bones gairebé constant entre 6 i 9 Å, i redueix lleugerament els errors més grans (>1), indicant una arquitectura més robusta. Així i tot, tal com s'ha indicat anteriorment, aquestes millores no sempre són dràsticament superiors.



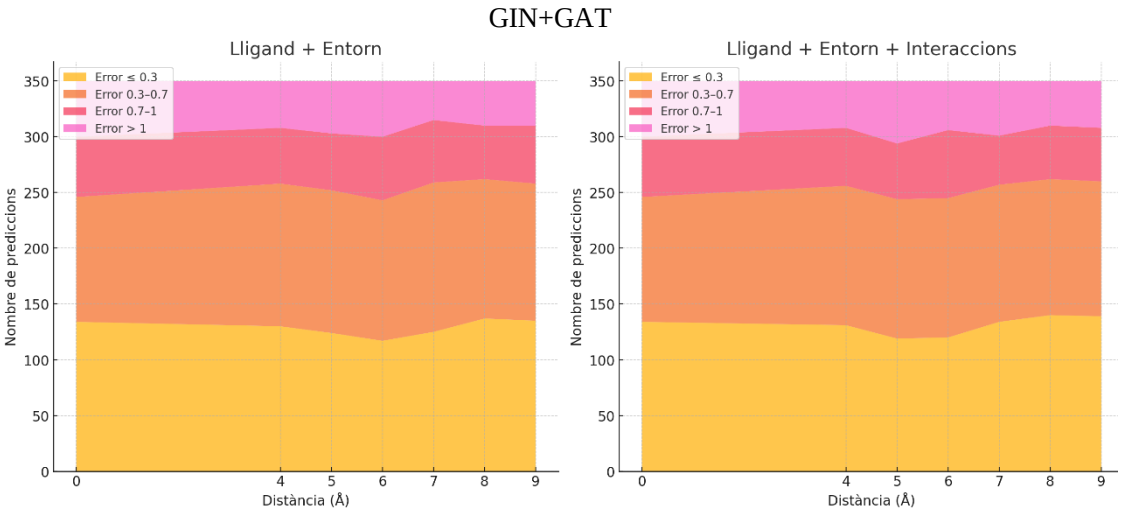


Figura 17. Comparació del nombre de prediccions agrupades per rangs d'error segons la distància de l'entorn proteic considerat (de 3 Å a 9 Å), per als diferents models (GAT, GIN, GIN+GAT) amb configuració "Lligand + Entorn" (esquerra) i "Lligand + Entorn + Interaccions" (dreta). Els errors es classifiquen en quatre intervals: ≤ 0.3, entre 0.3–0.7, entre 0.7–1 i > 1 unitats de pIC50 que representen un error en la predicció de la concentració d'entre 0-2x, 2x-5x i 5x-10x respectivament.

7. Conclusions

L'anàlisi del rendiment dels models GNN aplicats a la predicció de l'afinitat entre lligands i la proteïna Mpro del SARS-CoV-2 ha permès extreure diverses conclusions rellevants:

En primer lloc, s'ha confirmat que la incorporació de l'entorn proteic al voltant del lligand millora de forma generalitzada la capacitat predictiva del model, especialment quan es consideren distàncies mitjanes i àmplies (entre 6 Å i 9 Å). Aquesta millora es manifesta tant en la reducció del MAE com en l'augment de la proporció de prediccions amb error baix (≤ 0.3), posant de manifest la importància del context estructural en la predicció d'afinitat.

Pel que fa a la incorporació explícita de les interaccions febles, els resultats mostren que el seu impacte és més variable i depèn fortament del tipus d'arquitectura i de la distància considerada. En alguns casos concrets, es detecten millores clares respecte als models sense interaccions. Tanmateix, en altres configuracions l'efecte pot ser neutre o fins i tot lleugerament negatiu. Això suggereix que la representació explícita d'aquestes interaccions no garanteix per si sola un millor rendiment.

És rellevant destacar que moltes d'aquestes interaccions febles poden estar ja representades de manera implícita en els models que només compten amb l'entorn, a

través de la geometria local dels grafs i les propietats dels nodes i arestes. Aquesta representació implícita pot explicar per què, en alguns casos, afegir informació explícita no comporta un avantatge clar. Així doncs, la capacitat dels models GNN d'aprendre patrons estructurals complexos sense necessitat d'anotació explícita pot ser suficient per capturar bona part del comportament d'interacció.

Finalment, en relació amb les arquitectures comparades, els models GIN i GIN+GAT amb entorns amplis (8–9 Å), especialment en la configuració E+I, han mostrat un lleuger avantatge en termes de MAE absolut. Tot i això, les diferències entre arquitectures són modestes (de l'ordre de 0.02–0.05), i no permeten identificar un model clarament superior de manera universal. Aquesta observació reforça la idea que el rendiment no depèn exclusivament de l'arquitectura escollida, sinó de la sinergia entre aquesta, la mida de l'entorn i el tractament de les interaccions febles.

Tot i que el rendiment dels models obtinguts és raonablement bo, cal tenir en compte que el conjunt d'entrenament utilitzat estava format per aproximadament 360 estructures de complexos lligand–enzim. Aquest volum de dades, tot i ser insòlitament alt pel que fa a estructures disponibles per a una mateixa proteïna dins del PDB, pot no ser suficient per permetre als models una comprensió profunda i generalitzable dels patrons d'interacció. Aquesta limitació en la quantitat de dades disponibles pot explicar part de la variabilitat observada.

En conjunt, es conclou que la inclusió de l'entorn proteic ajuda als models a generar per a una predicció més precisa de l'afinitat, mentre que la utilitat d'afegir interaccions febles de manera explícita ha de ser valorada cas per cas. Aquestes troballes proporcionen una base sòlida per a futures optimitzacions de models GNN en predicció bioquímica estructural.

8. Bibliografia

Bakan, A., Meireles, L. M., & Bahar, I. (2011). ProDy: Protein Dynamics Inferred from Theory and Experiments. Bioinformatics, 27(11), 1575–1577. https://doi.org/10.1093/bioinformatics/btr168

Berman, H. M. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Bhat, E. A., Khan, J., Sajjad, N., Ali, A., Aldakheel, F. M., Mateen, A., Alqahtani, M. S., & Syed, R. (2021). SARS-CoV-2: Insight in genome structure, pathogenesis and viral receptor binding analysis – An updated review. International Immunopharmacology, 95, 107493. https://doi.org/10.1016/j.intimp.2021.107493

Brant, A. C., Tian, W., Majerciak, V., Yang, W., & Zheng, Z. (2021). SARS-CoV-2: from its discovery to genome structure, transcription, and replication. Cell & Bioscience, 11(1). https://doi.org/10.1186/s13578-021-00643-z

Fernandes, H. S., Sousa, S. F., & Cerqueira, N. M. F. S. A. (2021). New insights into the catalytic mechanism of the SARS-CoV-2 main protease: an ONIOM QM/MM approach. Molecular Diversity, 26(3), 1373–1381. https://doi.org/10.1007/s11030-021-10259-7

Hu, Q., Xiong, Y., Zhu, G., Zhang, Y., Zhang, Y., Huang, P., & Ge, G. (2022). The SARS-CoV-2 main protease (Mpro): Structure, function, and emerging therapies for COVID-19. MedComm, 3(3). https://doi.org/10.1002/mco2.151

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., & Blundell, T. L. (2016). Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. Journal of Molecular Biology, 429(3), 365–371. https://doi.org/10.1016/j.jmb.2016.12.004

Kensert, A., Desmet, G., & Cabooter, D. (2022, August 21). MolGraph: a Python package for the implementation of molecular graphs and graph neural networks with TensorFlow and Keras. arXiv.org. https://arxiv.org/abs/2208.09944

Li, G., Hilgenfeld, R., Whitley, R., & De Clercq, E. (2023). Therapeutic strategies for COVID-19: progress and lessons learned. Nature Reviews Drug Discovery, 22(6), 449–475. https://doi.org/10.1038/s41573-023-00672-y

Llop-Peiró, A., Pujadas, G., Gimeno, A., & Garcia-Vallvé, S. (2024). PDB-CAT: A User-Friendly Tool to Classify and Analyze PDB Protein-Ligand Complexes. ChemRxiv. https://doi.org/10.26434/chemrxiv-2024-54073

Shawky, A. M., Almalki, F. A., Alzahrani, H. A., Abdalla, A. N., Youssif, B. G., Ibrahim, N. A., Gamal, M., El-Sherief, H. A., Abdel-Fattah, M. M., Hefny, A. A.,

Abdelazeem, A. H., & Gouda, A. M. (2024). Covalent small-molecule inhibitors of SARS-CoV-2 Mpro: Insights into their design, classification, biological activity, and binding interactions. European Journal of Medicinal Chemistry, 277, 116704. https://doi.org/10.1016/j.ejmech.2024.116704

Wang, K., Zhou, R., Tang, J., & Li, M. (2023). GraphscoreDTA: optimized graph neural network for protein–ligand binding affinity prediction. Bioinformatics, 39(6). https://doi.org/10.1093/bioinformatics/btad340

Wu, C., Yin, W., Jiang, Y., & Xu, H. E. (2022). Structure genomics of SARS-CoV-2 and its Omicron variant: drug design templates for COVID-19. Acta Pharmacologica Sinica, 43(12), 3021–3033. https://doi.org/10.1038/s41401-021-00851-w

Zagórska, A., Czopek, A., Fryc, M., & Jończyk, J. (2024). Inhibitors of SARS-CoV-2 Main Protease (Mpro) as Anti-Coronavirus Agents. Biomolecules, 14(7), 797. https://doi.org/10.3390/biom14070797

Zhang, X., Gao, H., Wang, H., Chen, Z., Zhang, Z., Chen, X., Li, Y., Qi, Y., & Wang, R. (2023). PLANET: A Multi-objective Graph Neural Network Model for Protein–Ligand Binding Affinity Prediction. Journal of Chemical Information and Modeling, 64(7), 2205–2220. https://doi.org/10.1021/acs.jcim.3c00253

9. Autoavaluació

La realització d'aquest Treball de Fi de Grau m'ha permès aplicar de manera pràctica els coneixements adquirits durant el doble grau en Informàtica i Biotecnologia, en un context clarament interdisciplinari. Tot i haver treballat amb tècniques d'intel·ligència artificial al llarg de la carrera d'Informàtica, aquest projecte ha estat la primera vegada que he pogut aplicar xarxes neuronals de grafs (GNN), una àrea que m'interessava especialment i en la qual tenia moltes ganes d'aprofundir.

L'ús de grafs per representar molècules i estudiar la seva afinitat amb proteïnes m'ha ajudat a veure de manera clara com es poden integrar les eines pròpies de la IA amb problemes complexos en biotecnologia. En aquest sentit, considero que aquest TFG ha estat una experiència molt enriquidora i alineada amb la naturalesa del meu perfil doble, consolidant el meu interès per la recerca en bioinformàtica i IA aplicada a la biomedicina.

Vull agrair especialment al professor Santi per la seva direcció, la seva disponibilitat constant al llarg del projecte i la manera com ha compartit la seva experiència i

coneixement en bioinformàtica. El seu tutoratge ha estat clau tant per al desenvolupament tècnic del treball com per a arribar a les conclusions d'aquest.