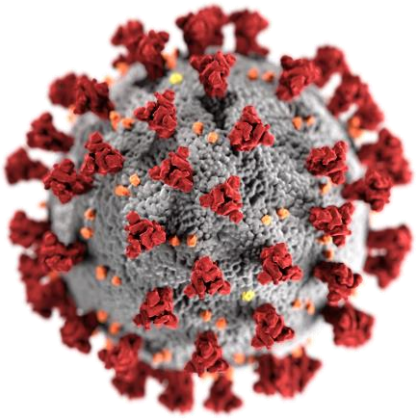
UNIVERSITAT ROVIRA i VIRGILI

CERCA D'UN MODEL DE PREDICCIÓ D'AFINITAT PER INHIBIDORS NO COVALENTS DE LA PROTEASA PRINCIPAL M-PRO DEL SARS-COV-2 A TRAVÉS D'EINES COMPUTACIONALS

Estel Ferré Aguirre

TREBALL FINAL DEL GRAU BIOTECNOLOGIA



Tutor acadèmic: Gerard Pujadas Anguiano
Doctor en Química, Licenciat en Ciències Químiques,
Departament de Bioquímica i Biotecnologia
gerard.pujadas@urv.cat

En cooperació amb: Grup de recerca en Quimioinformàtica i nutrició
Universitat Rovira i Virgili

Supervisors: Santiago Garcia Vallvé,
Doctor en Bioquímica, Llicenciat en Ciències Químiques,
Departament de Bioquímica i Biotecnologia, santi.garcia-vallve@urv.cat
Guillem Macip Sancho, PhD, Grau en Biotecnologia,
Departament de Bioquímica i Biotecnologia,
guillem.macip@urv.cat

Juny 2022

Jo, “Estel Ferré Aguirre” , amb DNI “49651527R”, 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) (http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/) i afirmo que aquet TFG no constitueixen cap de les conductes considerades com a plagi per la URV.

Tarragona, 6 de juny de 2022

Índex:

Dades del centre:

Aquest estudi s'ha realitzat en el grup de recerca en Quimioinformàtica i Nutrició de la Universitat Rovira i Virgili. Es tracta d'un grup format per diversos professors a temps complet i diversos becaris predoctorals del departament de Bioquímica i Biotecnologia. Està situat al Campus Sescelades, al carrer Marcel·lí Domingo num. 1 a Tarragona.

El grup es centra en la recerca i ús d'eines computacionals per així desenvolupar nous mètodes enfocats en la predicció de la bioactivitat de diferents compostos bioactius i donar noves funcions a molècules. Aquests resultats poden servir per al disseny de nous aliments funcionals o nutricèutics o per desenvolupar nous productes cosmètics entre altres. Aquest grup combina una part computacional (*in silico*) i una part experimental (*in vitro in vivo*) que permet buscar nous compostos bioactius i demostrar activitats predites.

Durant els darrers anys han publicat diversos articles en les millors revistes internacionals, alguns dels seus estudis han permès desenvolupar nous ingredients en companyies de disseny de biomolècules. Els seus estudis han tractat sobre la cerca de molècules naturals i el desenvolupament de metodologies computacionals per millorar les prediccions de bioactivitat i per predir nous usos per molècules o fàrmacs ja existents.

Resum:

El Sars-CoV-2 és el virus causant de la COVID-19, una malaltia que durant els últims dos anys ha causat grans danys sanitaris a nivell mundial. Durant aquest temps s'han desenvolupat diverses estratègies i vacunes per tractar amb aquest virus. La cerca d'inhibidors no covalents contra la proteasa principal del Sars-CoV-2 M-pro, una proteasa essencial per al cicle viral del Sars-CoV-2, pot ser una bona estratègia per desenvolupar noves teràpies. L'ús de les eines computacionals cada dia avança i facilita el disseny de fàrmacs. En aquest estudi utilitzem el Kdeep i KdeepTrainer, dos eines computacionals proporcionades pel PlayMolecule, per cercar el millor mètode de predicció d'afinitat proteïna-lligand utilitzant dades experimentals d'afinitat (pIC50,Ki,Kd), i així poder trobar nous inhibidors no covalents de la proteasa principal i poder desenvolupar nous antivirals.

Paraules clau:

Sars-CoV-2, M-pro, Kdeep, KdeepTrainer, predicció d'afinitat, inhibidors no covalents.

Introducció:

El SARS-CoV-2 és un virus causant de la malaltia produïda per un tipus de coronavirus de 2019, aquest patogen va provocar una pandèmia mundial que va començar a finals del 2019 fins a l'actualitat. La seva primer aparició va ser a Wuhan Xina a finals del 2019 i es va anar expandint fins a convertir-se en una crisi sanitària en l'àmbit mundial(1). Té una gran facilitat per propagar-se i ho fa principalment a través dels aerosols. Els seus símptomes més comuns són febre, tos, malestar físic i perduda de gust entre d'altres, en els casos més extrems pot arribar a provocar greus problemes respiratoris, pneumònia i fins i tot la mort(2). La COVID-19 s'ha propagat al voltant de 215 països sent USA , Espanya i UK els països més afectats. Ha causat més de 100.000 morts només a Espanya i més de 6 milions de morts al voltant del món (3)Hi han diverses teories sobre com es va originar, com per exemple a través d'una zoonosi entre ratpenats i humans(4)(5).

Els CoV son un grup de virus genotípicament i fenotípicament diversos, embolcallats i de sentit positiu. Contenen RNA monocatenari i pertanyen a la família Coronaviridae, aquest grup de virus indueix complicacions respiratòries, entèriques, hepàtiques i neurològiques de severitat variable en una àmplia gamma d'espècies animals i humans(6). Els CoV es divideixen en quatre tipus: α-coronavirus (α-CoV), β-coronavirus (β-CoV), γ-coronavirus (γ-CoV) i δ-coronavirus (δ-CoV) (7). L'agent etiològic de la COVID-19 prové del gènere β-coronavirus i és un coronavirus-2 (SARS-CoV-2) (8). La seqüència del genoma dels coronavirus ocupa entre 26 i 32 kb de llargada, és la seqüència més llarga entre els virus d'RNA (9).

El genoma del SARS-CoV-2 codifica més de 20 proteïnes estructurals (com la proteïna Spike encarregada de regular el reconeixement del receptor de la cèl·lula hoste, l'entrada de la cèl·lula hoste i la resposta immune) i no estructurals (com RNA-dependent RNA polimerasa (RdRp)) (10). Conte dos marcs de lectura oberts, ORF1a i ORF1ab, ajuden a traduir les poliproteïnes pp1a i pp1ab, necessàries per a la replicació i transcripció viral. Les poliproteïnes son escindides mitjançant el

processament proteolític per la proteasa viral semblant a la papaïna (PLpro) i la proteasa cisteïna semblant a 3C (M-pro) (8)(11).

S'espera que la campanya massiva de vacunació contra la COVID-19 generi immunitat de ramat. Aquestes vacunes tenen com a objectiu la proteïna Spike del virus SARS-CoV-2. Aquesta proteïna del virus és altament mutable com ja s'ha vist en les noves variants del SARS-CoV-2, comparteix un 76% de seqüència amb la proteïna Spike del SARS-CoV. Tot i que es poden desenvolupar vacunes de reforç per a noves variants, els antivirals de molècules petites contra dianes menys mutables tindran més èxit que una vacuna(12). Actualment no hi ha cap medicament específic disponible per al SARS-CoV-2 i, per tan, s'estan investigant diferents estratègies com la reutilització de fàrmacs existents (9), o l'ús de proteases com la M-pro o PLpro com a objectius per al desenvolupament de fàrmacs (13).

M-pro:

La M-pro és la proteasa principal del SarS-CoV-2. És essencial per al cicle de vida viral, ja que és necessària per a generar la majoria de proteïnes no estructurals del virus(14). Conté una diada de cisteïna-histidina al seu centre catalític i escindeix els seus substrats en llocs que comprenen una glutamina seguida d'un residu amb una petita cadena lateral. La M-pro està altament conservada dins de la família dels coronavirus (15). També és coneguda com a la proteasa 3C-*like* (3CLpro), aquesta consta de 306 aminoàcids de llarg (16). En jugar un important paper en el processament de la poliproteïna i la maduració del virus, és considerat un important objectiu per al disseny de medicaments antivirals per combatre el SARS-CoV-2 (8).

La replicació del Sars-CoV-2 està regulada per un complex format per dues poliproteïnes que són traduïdes de l'RNA viral. Aquestes poliproteïnes escindeixen en un mínim d'11 llocs al voltant de C-terminal i la regió central per l'acció dels residus catalítics, alliberant les proteïnes vitals necessàries per a la replicació viral. Inhibir l'activitat d'aquest enzim bloquejaria la replicació viral (16).

Per el que respecta la seva estructura la M-pro del SARS-CoV-2 compren de tres dominis, el domini I (residus 8-101), domini II (residus 102-184) i domini III (residus (201-303). Els primers dos dominis tenen una estructura en β-barril antiparal·lela. El tercer domini compta amb cinc α-hèlix, formant un conglomerat antiparal·lel connectat al domini mitjançant un bucle llarg en la regió 185-200. La M-pro conte una diada catalítica Cys-His amb el lloc d'unió al substrat situat entre dominis (figura 1) (9).

Estudis previs han demostrat que tots el CoV M-pro comparteixen la mateixa regió d'unió al substrat entre dominis i com a resultat la conservació de l'estructura. La M-pro existeix com un homodímer en solució i es diu que aquesta forma de dímer és molt activa en comparació a la forma monòmer(9). Les principals diferencies entre les seqüències de SARS-CoV-2 i SARS-CoV M-pro resideixen a la superfície de la proteïna, s'espera que els inhibidors en contra SARS-CoV M-pro inhibeixin igual a SARS-CoV-2 M-pro (16) . A més a més com no es coneixen proteases humanes amb una especificitat d'escissió similar, la probabilitat que aquests inhibidors siguin tòxics per als humans és molt baixa, per això la recerca d'inhibidors per a la proteasa M-pro ens interessa tant per al desenvolupament de nous medicaments contra la COVID-19(17) .

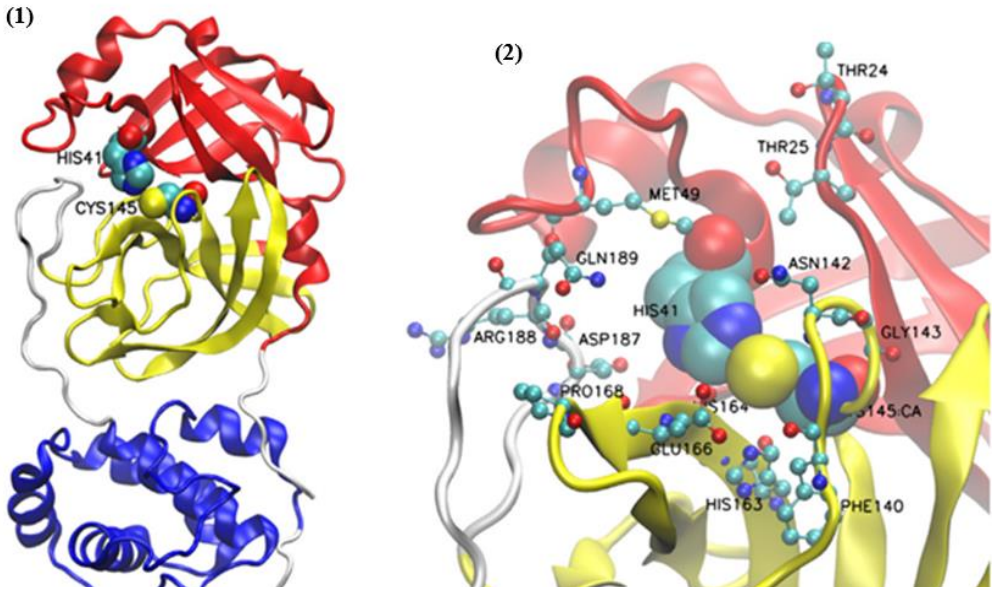


FIgura 1: (1) Representació de l'estructura cristal·lina de la Mpro del SARS-CoV-2, els dominis I, II i II es mostren en Vermell, groc i blau respectivament. La regió de connexió entre II i II és de color blanc i els residus de la diada catalítica (His41 i Cys14) és troben en esferes sòlides. (2) Representen els residus del lloc actiu de la proteasa que estan implicats en les interaccions inhibidores (9).

Actualment, el temps i cost per dissenyar un fàrmac s'ha reduït significativament. L'ús de les eines computacionals, la reutilització de fàrmacs, el cribratge de bases de dades i el disseny de diferents inhibidors s'estan convertint amb una opció més ràpida al descobriment de nous fàrmacs(18). El disseny de fàrmacs assistits per ordinador s'utilitza habitualment per detectar ràpidament inhibidors no covalents i per prevenir la funció biològica d'un enzim específic(4)(19).

Els inhibidors bloquegen l'acció de les proteïnes mitjançant la unió del lligand a la proteïna a través de la seva afinitat química, la mesura de la força dels enllaços químics(20), contra més afinitat química tingui aquest inhibidor amb la proteasa M-pro més forta serà la unió i millor actuarà l'inhibidor.

Predicció d'afinitat, Kdeep i KdeepTrainer:

La predicció de l'afinitat d'unió proteïna-lligand s'ha convertit en un objectiu en la química computacional. Estudis previs han demostrat que pot accelerar el descobriment de noves medicines a través del "virtual screening" i "lead optimization"(21), dos tècniques computacionals que es basen en la identificació de compostos biològics per al desenvolupament de fàrmacs(22).

L'afinitat d'unió mostra la força amb la qual el fàrmac s'uneix al seu objectiu. Generalment, s'expressa en termes de Kd (constant de dissociació), Ki (constant d'inhibició) o IC50 (la meitat de la concentració inhibidora màxima). Per fer aquest estudi utilitzarem dades experimentals sobre l'IC50, es una mesura depèn de la concentració de la diana i del lligand. Els valors baixos d'IC50 signifiquen una alta afinitat d'unió, de la mateixa forma, els valors de Ki baixos signifiquen una gran afinitat d'unió(23), normalment s'expressa com a concentració molar. Les dades d'IC50 aporten una gran quantitat de coneixement sobre la bioactivitat de les proteïnes. El seu ús facilita el desenvolupament de mètodes per a descobrir nous medicaments(24).

Actualment, desenvolupar una teràpia eficaç es una tasca urgent que requereix estimar amb precisió l'energia lliure d'unió de lligands a Sars-CoV-2 M-pro. Tanmateix, s'ha de tenir en compte que la precisió d'un mètode per predir l'afinitat

d'unió depèn de la proteïna, si es trobés un mètode que pugues predir com un fàrmac s'uneix a una proteïna o com de forta és aquesta unió s'acceleraria el procés de descobriment de nous medicaments i disminuirien els costos de desenvolupar nous fàrmacs(4)(25).

Kdeep: Predictor d'afinitat proteïna-lligand:

El Kdeep és un predictor d'afinitat proteïna-lligand. El programa genera prediccions i permet visualitzar l'estructura 3D de la interacció entre la proteïna i el lligand. Aquestes prediccions els fa a través d'un model DCNN (Deep Neural Convuntuional Networks)(26). És una eina que és troba disponible públicament en el PlayMolecule per a que tots els usuaris puguin testar els seus complexes proteïna-lligand (21).

Per utilitzar el Kdeep primer es carrega una estructura de proteïnes en format pdb, el més convenient és que l'estructura contingui hidròdens i càrregues, en cas que no en contingui, el programa protonarà automàticament l'estructura a un pH de 7,4, després es carrega un conjunt de lligands acoblats en format sdf(26). Cal tenir amb compte que aquest programa es molt sensible a la proteïna que utilitza. En aquest estudi utilitzarem proteïnes i lligands d'inhibidors no covalents de la M-pro.

El programa dona l'opció d'afegir un model de predicció d'afinitat ja entrenat, aquest model pot estar ja en la base del Kdeep o pot estar creat pel mateix usuari amb el KdeepTrainer. Els models s'utilitzarien per predir l'afinitat d'unió entre proteïna i lligand(21). Per últim, es selecciona el camp de correlació que es vulgui usar, les opcions d'aquest apartat dependran de la informació del lligand que contingui l'arxiu sdf(26).

Els resultats mostren una taula amb el llistat de tots els lligands utilitzats, inclou alguns descriptors químics dels lligands carregats i les prediccions de totes les mètriques d'afinitat amb les quals es va entrenar inicialment el model Kdeep(26), com per exemple la desviació estàndard o el pkd. El programa et permet visualitzar la unió lligand-proteïna. Els seus resultats és poden descarregar en diferents formats i així utilitzar les dades de les prediccions per a futurs estudis, a més a més,

si es tenen dades experimentals el programa calcula la correlació de les dades per veure la fiabilitat de les prediccions.

KdeepTrainer:

El KdeepTrainer és una aplicació que permet entrenar la teva xarxa d'aprenentatge automàtic per predir les afinitats d'unió proteïna-lligand i així crear un model de predicció d'afinitat d'enllaç a partir de dades pròpies (model DCNN). Aquests models es poden utilitzar posteriorment en l'aplicació Kdeep per avaluar l'afinitat d'unió per a noves conformacions de proteïnes-lligands, l'aplicació és facilitada pel Play Molecule(27). En el KdeepTrainer introduïm les nostres pròpies dades i mitjançant l'aprenentatge automàtic, una rama de la intel·ligència artificial basada en el reconeixement de patrons i algoritmes que permeten deduir quin és el resultat més òptim per a un determinat problema (28), és crearan nous models de predicció d'afinitat a través de les dades que l'usuari proporcioni.

Per utilitzar el KdeepTrainer és necessari reunir totes les dades en un arxiu en format zip i crear el anomenat conjunt d'entrenament. Aquest conjunt ha de presentar una carpeta per proteïna i dins d'aquesta un fitxer en format pdf de l'estructura proteica i un altre del lligand o lligands en format sdf. L'arxiu sdf ha de contenir els valors d'afinitat de la proteïna, els quals es poden trobar en pKd, pKI o pIC50, però ha d'estar correctament indicat en l'arxiu sdf (27).

El KdeepTrainer dona diverses opcions per entrenar el model amb dades personalitzades. Primer, tries el número de repeticions que vols utilitzar per entrenar el model (number of epoch)(27). El número de repeticions és un terme utilitzat per indicar el nombre de passades que tot el conjunt de dades ha completat el cicle complet de l'algoritme del programa, escollir el número de repeticions que és vol per entrenar un model depèn sobretot de la quantitat de dades que es tingui(29). El autors del KdeepTrainer recomanen utilitzar 500 repeticions, però en el cas de tenir molt poques dades és millor utilitzar 50. Després, tenim l'opció d'afegir els pesos inicials i un conjunt d'entrenament addicional que pot servir per reentrenar un nou model de Kdeep (29). Un cop fet tot l'entrenament amb les dades, el KdeepTainer

proporciona 4 gràfics que mostren com de bo és el model creat i si té un bon funcionament (Figura 2).

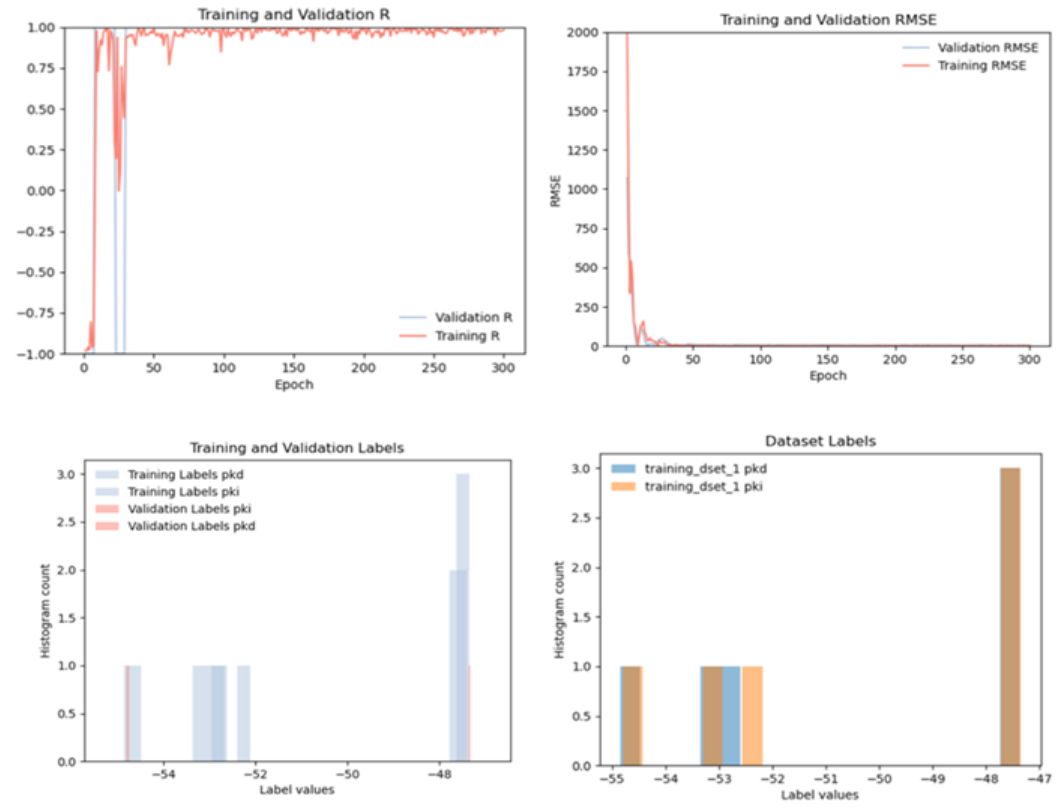


Figura 2: Representació dels resultats del KdeepTrainer amb un dels exemples proporcionats per la web: Els primers dos gràfics mostren com el model ha aprés, interessa que el valor de correlació sigui alt, un valor de correlació menor de 0,7 no seria útil. Els dos últims gràfics mostren la distribució dels valors entrenats i validats, el que interessa és que els valors de la validació siguin semblants que els del training set. L'últim gràfic mostra la quantitat de dades que tenim segons els seus valors

Models de predicció d'afinitat:

Els models DCNN (Deep Neureal Convuntioal Networks) són una xarxa neuronal aplicada per analitzar virtualment diferents tipus de dades. A través d'algoritmes informàtics, funcionen amb xarxes neuronals artificials dissenyades per imitar com pensen i aprenen els humans. Aquestes xarxes permeten als ordinadors observar, aprendre i reaccionar a situacions complexes més ràpid que els humans(30,31). A través d'aquests tipus de models realitzarem totes les prediccions d'afinitat proteïna-lligand.

Tant el Kdeep com el KdeepTrianer donen l'opció d'utilitzar diferents models entrenats fets pel programa amb les seves pròpies dades. Aquests models estan entrenats amb dades del PDBbind de l'any 2016, una base de dades amb l'objectiu de proporcionar una col·lecció completa de dades d'afinitat d'unió mesurades experimentalment per a tots els complexos biomoleculars dipositats al Banc de dades de proteïnes (PDB). Aquesta base de dades proporciona un enllaç essencial entre la informació energètica i estructural d'aquests complexos, el qual és útil per a diversos estudis computacionals i estadístics sobre el reconeixement molecular i descobriment de fàrmacs(32). Apart de les dades del PDBbind, els models d'aquesta aplicació també compten amb dades del ACEbind, una versió curada del conjunt de dades del PDBbind que ha creat internament Acellera, una empresa que es dedica a aportar tecnologies i algoritmes amb l'objectiu de generar nous fàrmacs(33) . És el conjunt de dades que es sol utilitzar per entrenar el predictor Kdeep predeterminat.

Aquestes dades poden estar en format "Model Ensemble" o "Best Model". El "Model Ensemble" consta de 8 models diferents entrenats en un subconjunt diferent de PDBbind o ACEbind, després es fa una mitjana dels resultats dels models donant una predicció mitjana i també informen de la desviació estàndard entre aquests models, això ajuda a mostrar si hi ha un desacord entre els models per així veure si la predicció es correcta o no. En canvi el "Best Model" es el millor dels 8 models del "Model Ensemble". El programa recomana utilitzar els models Best Model per utilitzar el KdeepTrainer i els models de Ensemble per utilitzar el Kdeep.

FEgrow:

El 13 d'abril de 2022 va sortir un estudi realitzat per la universitat de Newcastle. En aquest estudi els autors van desenvolupar una nova eina computacional, el FEgrow, una eina per crear series congènites de lligands i introduir càlculs d'energia lliure, el FEgrow enumera i optimitza les conformacions bioactives dels grups funcionals, amb aquesta eina es poden fer prediccions d'afinitat de diferents estructures.

Per comprovar el funcionament del FEgrow van modelar un conjunt de 13 inhibidors de la proteasa principal del Sars-CoV-2 M-pro(34), i van calcular les bioactivitats dels diferents compostos. El FEgrow es una combinació de diferents eines computacionals: (1) mols2grid, una eina que permet veure estructures en 2D(35) (2) RDkit un software quimioinformàtic amb diferents aplicacions (36); (3)OpenMM per fer simulacions moleculars (37); (4) Gnina un programa de docking molecular que utilitza xarxes neuronals convencionals per optimitzar lligands(38); BioSimSpace/SOMD una eina que s'utilitza per simular processos biològics(39). En aquest treball compararem els resultats obtinguts en el FEgrow amb els nostres.

Hipòtesis i objectius:

S'ha vist la recerca d'inhibidors per a la proteasa principal M-pro del Sars-CoV-2 es interesant per al disseny de noves teràpies per tractar la COVID-19. Els inhibidors bloquegen l'acció de les proteïnes mitjançant l'enllaç proteïna-lligand, com més afinitat d'enllaç, més eficaç serà l'inhibidor. Per trobar aquests inhibidors és importat generar un mètode de predicció d'afinitat fiable. En cas d'aconseguir-se es predir aquests valors de forma precisa i s'acceleraria el procés de descobriment de nous medicaments.

En aquest estudi intentem esbrinar si es possible generar una metodologia per predir eficientment l'afinitat dels inhibidors no covalents de la proteasa principal M-pro del Sars-CoV-2. Per a respondre a aquesta pregunta vam utilitzar diferents eines computacionals, com el Kdeep i el KdeepTrainer. Els objectius que ens permetrien resoldre aquesta hipòtesis son el següents:

- Crear un model de predicció d'afinitat a través del KdeepTrainer mitjançant les dades experimentals de la bioactivitat d'inhibidors no covalents de la M-pro.
- Validar aquest model de predicció d'afinitat mitjançant el Kdeep i cercar la millor metodologia de predicció d'afinitat.

Metodologia:

Entrenament de models amb el KdeepTrianer:

En aquest estudi hem creat diferents models de predicció d'afinitat proteïna-lligand a través del KdeepTrainer. Primer vam preparar el nostre fitxer zip amb totes les dades de 111 inhibidors no-covalents de la proteasa principal del Sars-CoV-2 M-pro. Aquestes estructures han estat obtingudes amb el COVID Moonshoot(40), un projecte amb l'objectiu d'accelerar el desenvolupament de fàrmacs contra la COVID-19.

Primer adjuntem el fitxer preparat en el KdeepTrainer i escollim com volem que siguin els nostres models. Per veure com varien els models segons les repeticions vam crear models de 300, 500 i 1000 repeticions. Després vam triar si volíem entrenar el model amb dades de models addicionals o no, en aquest cas vam crear models entrenats amb dades del PDBbind i l'ACEbind i models sense ser entrenats amb dades addicionals. Vam obtenir un total de 9 models diferents entrenats, analitzar els resultats i escollir el millor model de predicció d'afinitat.

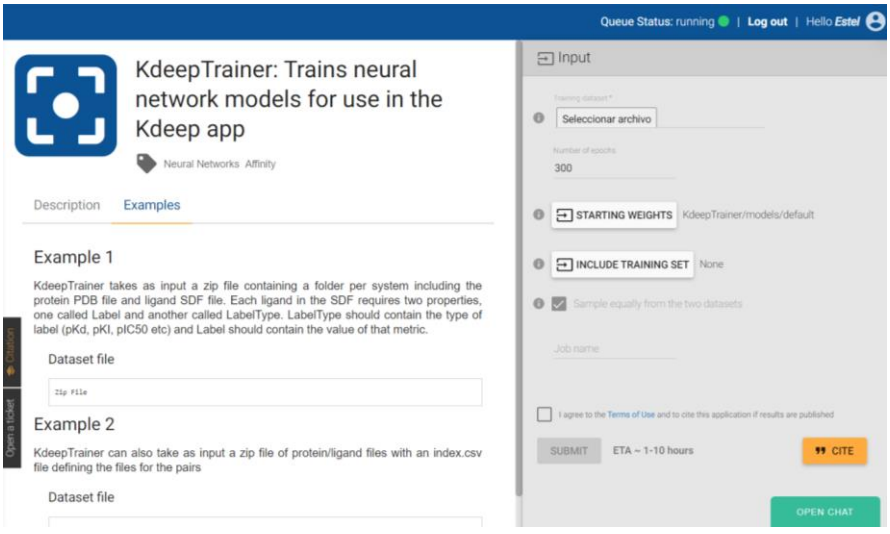


Figura 3: Imatge de la pagina d'inici del Kdeeptrainer.

Predicció d'afinitat proteïna-lligand amb el Kdeep:

Per veure la validesa dels diferents models creats amb el KdeepTrainer i veure si ha funcionat vam utilitzar el Kdeep, amb aquesta eina vam predir l'afinitat de la proteïna-lligand de diverses dades d'inhibidors no covalents. Primer de tot vam fer servir les mateixes dades que s'han utilitzat prèviament amb el KdeepTrainer, vam unir tots els arxius sdf de cada proteïna en un únic document i després vam afegir la proteïna M-pro-0689. Després es fa córrer el programa amb els models que es vulgui veure el seu funcionament, en el nostre cas vam usar tan models que hem fet amb el KdeepTrainer, com els models que proporciona la pàgina del Kdeep.

Vam descarregar els resultats de tots els models i amb aquests vam crear un Excel que recollia les dades dels valors experimentals (exp_affinity) i predits (predicted_pkd_mean) per cada model, amb aquestes dades vam fer un gràfic de dispersió per model i vam calcular els valors de: el coeficient de determinació (r^2), el coeficient de Pearson (R) el coeficient de Spearman (ρ) i l'error quadràtic mitjà (RMSE). Amb aquests càlculs volem veure si hi ha una correlació entre les dades predites i experimentals per veure si el model que s'ha utilitzat és bo o no.

Vam descarregar els resultats de tots els models i amb aquests vam crear un Excel que recollia les dades dels valors experimentals (exp_affinity) i predits (predicted_pkd_mean) per cada model, amb aquestes dades vam fer un gràfic de dispersió per model i vam calcular els valors de: el coeficient de determinació (r^2), el coeficient de Pearson (R) el coeficient de Spearman (ρ) i l'error quadràtic mitjà (RMSE). Amb aquests càlculs volem veure si hi ha una correlació entre les dades predites i experimentals per veure si el model que s'ha utilitzat és bo o no.

Cerca de nous valors de pIC50:

Per poder fer més proves i mirar millor com funcionen els diferents models amb diferents dades, vam fer una recerca bibliogràfica per trobar els valors de pIC50 de 80 inhibidors no covalents de la M-pro extrets del PANDDA. Vam buscar en les entrades del PDB i en els articles citats en aquestes mateixes entrades els valors

de pIC50, dels 80 compostos totals vam trobar els valors de pIC50 de 28 estructures.

Amb aquestes 28 noves dades vam tornar a fer noves prediccions amb el kdeep, unint tots els arxius sdf de les dades dels 28 compostos, en aquest cas es va utilitzar l'estructura 7M90, vam descarregar els resultats i vam tornar a fer nous Excel com els explicats en l'apartat anterior. Després, vam analitzar els resultats per buscar quin és el model que dona millors prediccions.

Cerca de la millor proteïna:

En el programa de predicció d'afinitat Kdeep és necessari adjuntar una proteïna amb format pdb i un model de predicció d'afinitat per poder predir la bioactivitat, depenent de l'estructura proteica el model donarà uns resultats diferents. En els apartats anteriors hem buscat quina és el millor model de predicció d'afinitat, un cop ja sabem aquest model ens interessa saber amb quina proteïna funciona millor.

Vam tornar a utilitzar el Kdeep, però aquest cop vam utilitzar totes les dades utilitzades en els apartats anteriors, tenint un total de 139 compostos, creant un nou arxiu sdf amb tots els lligands. Amb aquests compostos vam fer córrer el Kdeep utilitzant els models escollits segons els resultats anteriors, es repeteix el procés amb totes les estructures proteiques que tenim. Per a cada estructura es crea un Excel com s'ha fet en els apartats anteriors i calculem el percentatge de valors que estan a una distància -1 +1 de la lineal diagonal, d'aquesta forma podrem veure la relació lineal entre les dos variables, vam fer un total de 139 excels per veure quina és la proteïna que dona millors resultats depenent del model.

Finalment, vam fer una taula recollint tots els coeficients de determinació (r^2), els coeficients de correlació d'spearman (ρ) i els percentatges (%) dels 139 excels fets anteriorment. També vam fer un gràfic amb la mitjana de les 139 dades comparant totes les dades experimentals amb les predites, afegint la desviació estàndard.

Resultats i discussió:

Resultats models de predicció d'afinitat KdeepTrainer:

En aquest estudi es vol trobar un mètode que pugui predir l'afinitat entre una proteïna i un lligand, primer vam entrenar aquests models amb el KdeepTrainer, per fer això es van utilitzar diferents dades de 111 estructures proteiques d'inhibidors no covalents de la proteasa principal M-pro del Sars-CoV-2, vam entrenar un total de 9 models amb diferents característiques.

La figura 4. mostra els valors del coeficient de correlació dels diferents models entrenats amb el KdeepTrainer. En primer lloc, es pot observar com el nombre de repeticions en els models no influeix en els resultats dels gràfics, el coeficient de Pearson va augmentant fins a un punt on s'estabilitza encara que el número de repeticions es dupliqui.

En els models entrenats només amb les dades dels inhibidors no covalents de M-pro (figura 1. a, b i c) la "r validation" es troba entre 0,25 i 0,5, aquests valors són bastant baixos mostrant que no hi ha una correlació entre les dades, per a ser un model útil s'ha d'arribar a un valor mínim de R de 0,7 per tant aquests models no serien vàlids per utilitzar amb posteriorment amb el Kdeep. En canvi, els models entrenats amb les dades dels inhibidors i els models entrenats amb les dades addicionals del PDBbind i ACEbind van donar millors resultats mostrant uns valors de "R validation" entre 0,7 i 0,75 indicant una millor correlació de les dades. Si comparem els resultats dels models que han utilitzat el PDBbind i els models que han utilitzat l'ACEbind, podem observar com el model entrenat amb el PDBbind presenten un valor de "R validation" lleugerament més alt que la del ACEbind. En general els resultats no mostren uns valors en el coeficient de correlació molt alts, encara així els millors per utilitzar amb el Kdeep són els entrenats amb les dades del PDBbind.

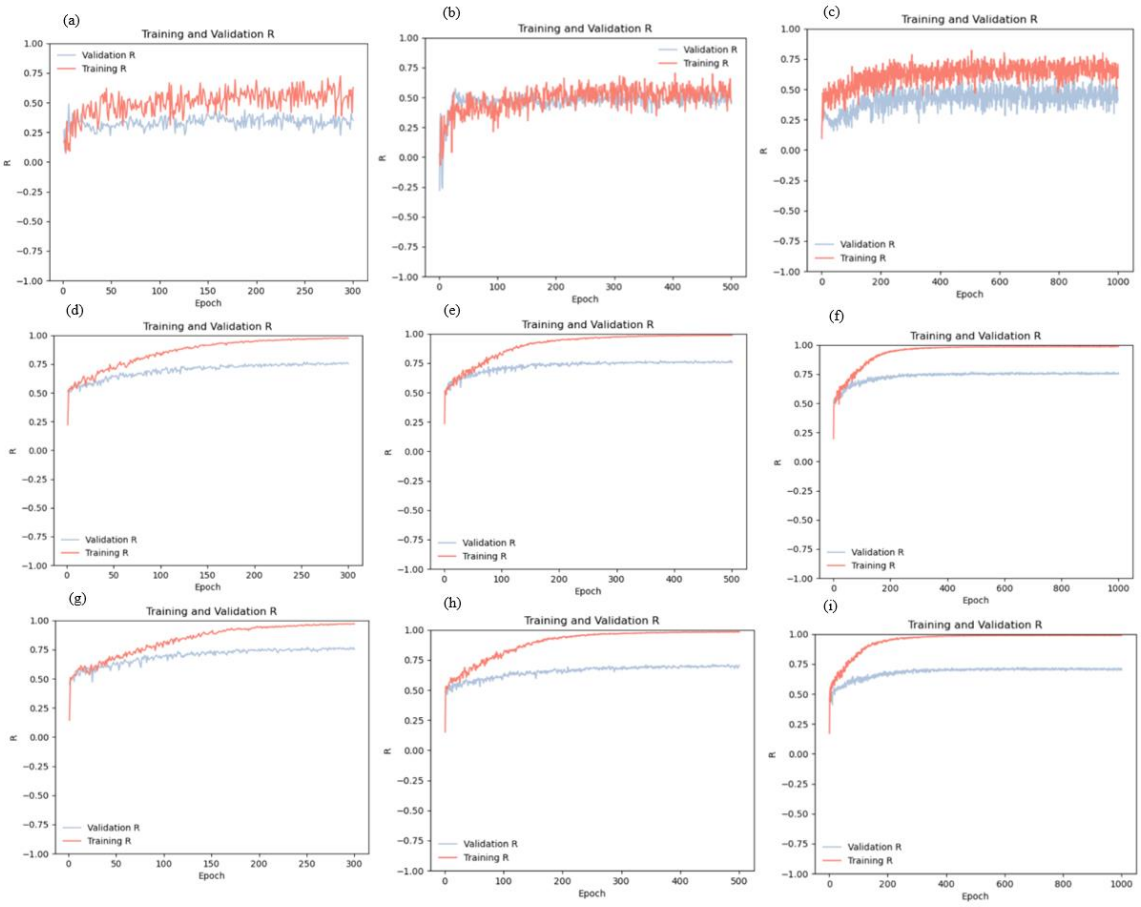


Figura 4: Representació del coeficient de correlació de Pearson (R) amb els diferents models entrenats amb el KdeepTrainer. (a) model amb 300 repeticions (b) model amb 500 repeticions (c) model amb 1000 repeticions (d) model 300 repeticions amb les dades del PDBbind (e) 500 model amb 500 repeticions amb les dades del PDBbind (f) Model amb 1000 repeticions més les dades del PDBbind (g) Model amb 300 repeticions amb les dades del ACEbind (h) Model amb 500 repeticions més dades del ACEbind (i) Model amb 1000 repeticions més les dades del ACEbind.

Millor metodologia de predicció d'afinitat:

Estudi amb 111 estructures proteiques:

Per comprovar la fiabilitat dels models del KdeepTrainer fets amb el PDBbind els vam utilitzar posteriorment amb el Kdeep per veure com funcionaven. Per fer això vam fer diferents experiments amb diferents dades per veure quin es el millor model per predir l'afinitat proteïna-lligand.

En primer lloc, es va utilitzar el conjunt de lligands dels 111 inhibidors no covalents. Amb aquestes dades es van fer prediccions d'afinitat amb els models entrenats del KdeepTrainer (model 300, 500 i 1000 repeticions més PDBbind) i els models predeterminats del Kdeep (PDBbind Ensemble, PDBbind Best Model, ACEbind Ensemble, ACEbind Best Model).

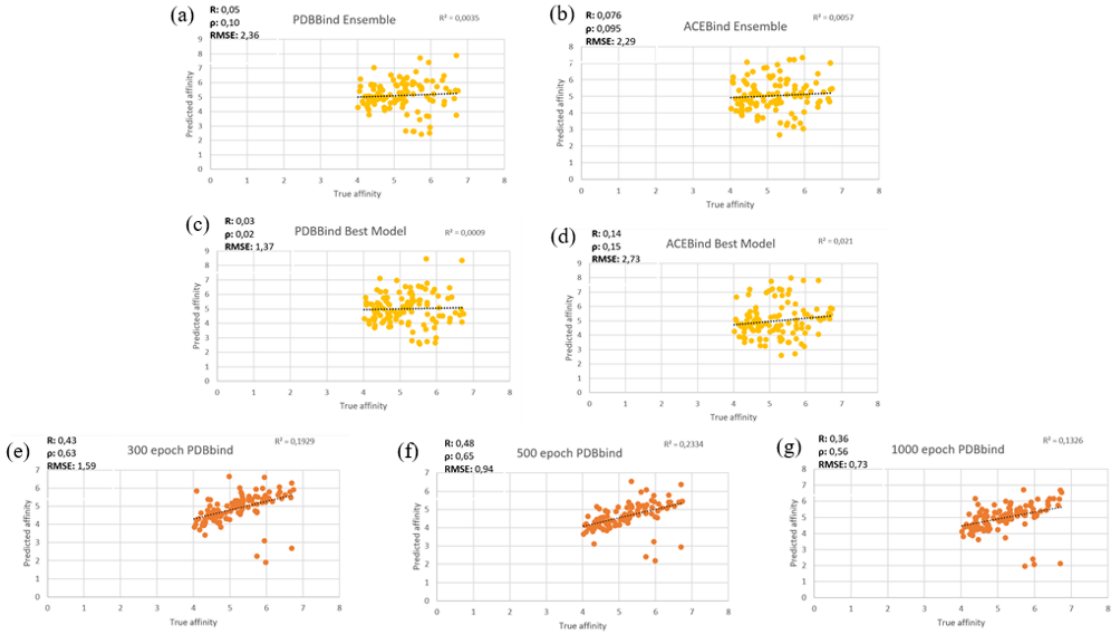


Figura 5: Representació de l'afinitat predita (pkd predit) i experimental (pkd experimental) juntament amb el coeficient de determinació (r^2), el coeficient de correlació de Pearson (r), el coeficient de correlació d'spearman (ρ), i l'error quadràtic mitjà (RMSE) en els diferents models utilitzats amb les dades de 111 inhibidors no covalents de la proteasa principal del Sars-CoV-2 M-pro (a) PDBbind Ensemble (b) ACE Bind Best Model (c) PDBbind Best Model (d) ACEbind Best Model (e) KdeepTrainer 300 repeticions PDBbind (f) KdeepTrainer 500 repeticions PDBbind (g) KdeepTrainer 1000 repeticions PDBbind.

La figura 5. representa gràfics on es compara l'afinitat real amb l'afinitat predita pel model escollit en cada gràfic. En el cas dels models predeterminats del Kdeep (figura 2. a, b, c, d) els gràfics mostren uns Coeficients de Determinació (r^2) 0,0035 (a), 0,0057 (b) i 0,0009 (c), i uns Coeficients de correlació de Spearman (**ρ)** de 0,1, 0,095 i 0,02, l'error quadràtic mitjà es troba entre 2,36 i 1,36. En canvi, els models entrenats amb el KdeepTrainer (figura 2. e, f, g) mostren uns valors de r^2 de 0,19, 0,23 i 0,13 i de **ρ** de 0,63, 0,65 i 0,56 l'error quadràtic mitjà oscil·la entre 0,73 i 1,59. Aquests resultats mostren, en general, molt poca correlació entre els resultats, és a dir, les dades predites pels models no s'ajusten a les dades experimentals que hem donat, tot i això, els models que donen millors resultats són els entrenats amb el KdeepTrainer, encara que aquest resultat més alt es podria donar a causa de que les dades utilitzades per entrenar el model en el KdeepTrainer són les mateixes que les utilitzades amb el Kdeep, per tan els resultats no són fiables. En tots els casos, cap dels models utilitzats presenta uns resultats favorables, presentant una R quadrada que no supera el 0,5 en cap dels casos.

Estudi amb 28 estructures proteiques:

Vam repetir l'experiment amb un set de 28 estructures proteiques noves d'inhibidors no covalents de la M-pro del Sars-CoV-2 per veure si els models donen millors prediccions i és poden utilitzar posteriorment. La figura 6. mostra els resultats de l'estudi, s'observa una millora dels resultats en comparació als gràfics anteriors (figura 5), en aquest cas la R^2 augmenta notablement, en canvi, l'error quadràtic mitjà (RMSE) és manté i inclús augmenta en el cas dels models computats amb el KdeepTrainer.

En aquest cas els models que donen millors resultats són els predeterminats amb el Kdeep, presentant un valor del coeficient de determinació (r^2) màxim de 0,4852 amb el model de ACE Bind Best Model (figura 3. d) i un valor mínim de 0,4753 representant el model fet amb el PDB Bind Best Model (figura 3. b). El coeficient de correlació d'Spearman entre 0,7 i 0,67 mostra un augment significatiu en la correlació dels resultats en comparació amb les dades anteriors i un RMSE de 0,7 aproximadament en tots els casos. Els resultats obtinguts amb els models entrenats

en el KdeepTrainer mostren una r^2 màxima de 0,43 (figura 3. b) i una r^2 mínima de 0,41 (figura 3. a). Per la seva part el coeficient de correlació d'Spearman (ρ) presenta un màxim entre 0,64 i 0,76 i un valor mitja de RMSE de 2,5.

En general podem observar com amb menys dades el programa genera millors prediccions, obtenint un major ajust de les dades. També podem veure com en aquest experiment els models que millor han funcionat han estat els que et proporciona el programa Kdeep, que presenten uns valors de R quadrada i de coeficient de correlació d'Spearman més alts i un error quadràtic mitjà més baix. Els models que més ens interessen per a futurs estudis de predicció serien els que contenen les dades del PDBbind i ACEBind, no els que hem fet prèviament amb el KdeepTrainer.

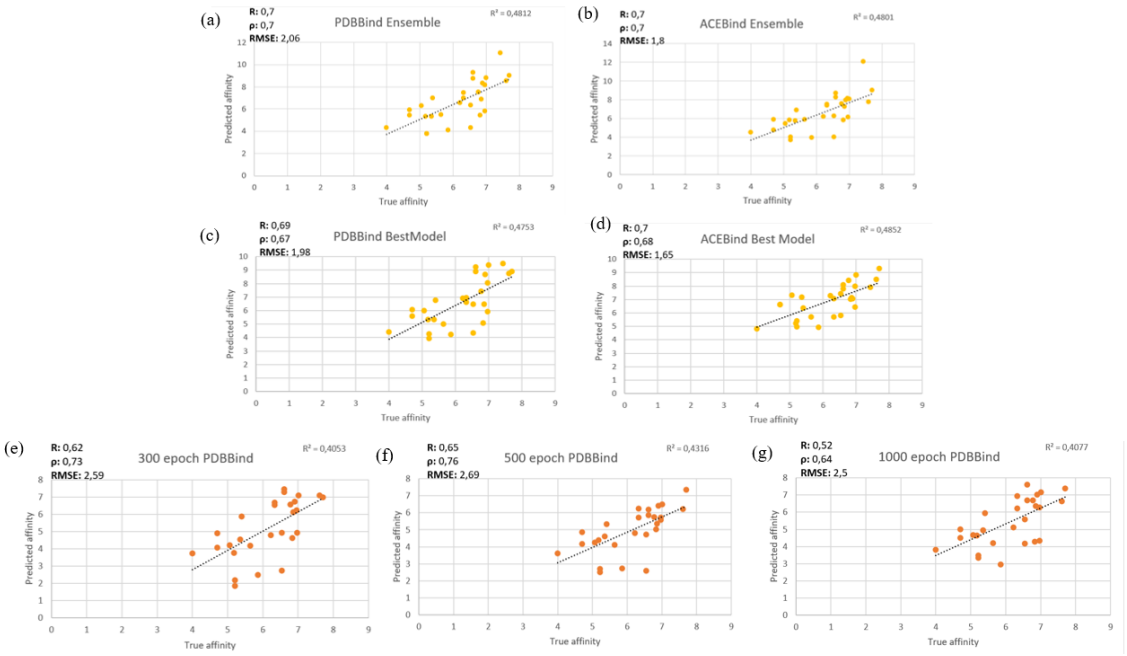


Figura 6 : Representació de l'afinitat predita (pkd predit) i experimental (pkd experimental) juntament amb el coeficient de determinació (r^2), el coeficient de correlació de Pearson (r), el coeficient de correlació d'spearman (ρ), i l'error quadràtic mitjà (RMSE) en els diferents models utilitzats amb les dades de 28 inhibidors no covalents de la proteasa principal del Sars-CoV-2 M-pro (a) PDBbind Ensemble (b) ACEbind Best Model (c) PDBbind Best Model (d) ACEbind Best Model (e) KdeepTrainer 300 repeticions PDB Bind (f) KdeepTrainer 500 repeticions PDBbind (g) KdeepTrainer 1000 repeticions PDBbind.

Estudi amb 139 estructures proteiques:

Els resultats del model solen variar depenent de la proteïna en format pdb que s'utilitza, en els experiments anteriors hem utilitzat la mateixa proteïna en tots els casos perquè volíem veure quins models funcionaven millor amb el Kdeep, ara volem veure quina és la millor proteïna que funciona amb els models que donen millors resultats. Per fer això vam repetir l'experiment amb totes les dades, en aquest cas només utilitzem els models ja entrenats del Kdeep (PDBbind Ensemble, PDBbind Best Model, ACEbind Ensemble i ACEbind Best Model).

La figura 6. mostra una mitjana dels resultats de cada model del Kdeep, juntament amb els càlculs estadístics pertinents. A simple vista observem com els gràfics que han utilitzat les dades del ACEbind presenten una relació més lineal entre les dos variables que els models amb les dades PDBbind. Representant el PDBbind els gràfics a i b mostren una r quadrada de 0,35 aproximadament, en canvi, els gràfics c i d mostren uns valors de r^2 entre 0,37 i 0,39. La desviació estàndard en els dos primers gràfics (a, b) és d'1,87 i 1,8 i dels dos últims d'1,61 i 1,65 (c i d). Respecte al valor dels percentatges podem observar com en els dos primers gràfics en cap dels casos arriba al 10% i en els dos últims, els que utilitzen els models entrenats amb les dades del ACEbind el valor d'aquest percentatge es casi el triple arribant a valors quasi del 30%. En tots dos la correlació entre les dades es baixa, encara que els models amb les dades del ACEbind tenen menys dispersió i un major ajust de les dades, és a dir, les prediccions realitzades amb aquests models s'ajusten més a les experimentals.

En la taula 1. dels annexos és mostra un recull de tots els coeficients de determinació (r^2), d'spearman (ρ) i el percentatge de valors que es troben a una distància de +1 o -1 de la línia diagonal (%). En tots els models els resultats dels valors de r^2 i ρ tenen una variació de dècimes, en canvi, el valor dels percentatges varia més o menys segons la proteïna utilitzada amb el Kdeep. En el cas del PDBbind Ensemble les estructures que millor s'adapten al model són la 7L12 i la 7L14 amb un percentatge del 19,35% en tots dos casos. En canvi, en els altres 3 models la proteïna que millor s'ajusta al model és la 7RME presentant un

percentatge del 34,54% i 36,69% en els models de l'ACEbind Ensemble i ACE Best Model, mostrant com amb aquesta proteïna i aquests models hi ha una major relació lineal entre les dades i els models de predicció són més fiables.

El model que dona unes millors prediccions d'afinitat és l'ACEbind Best Model amb la proteïna 7RME, però els programadors del Kdeep recomanen utilitzar el Model Ensemble per a estudis fets amb el Kdeep, ja que amb el Best Model els resultats poden ser incerts, per tant el millor model de predicció d'afinitat és l'ACEbind Ensemble amb la proteïna 7RME.

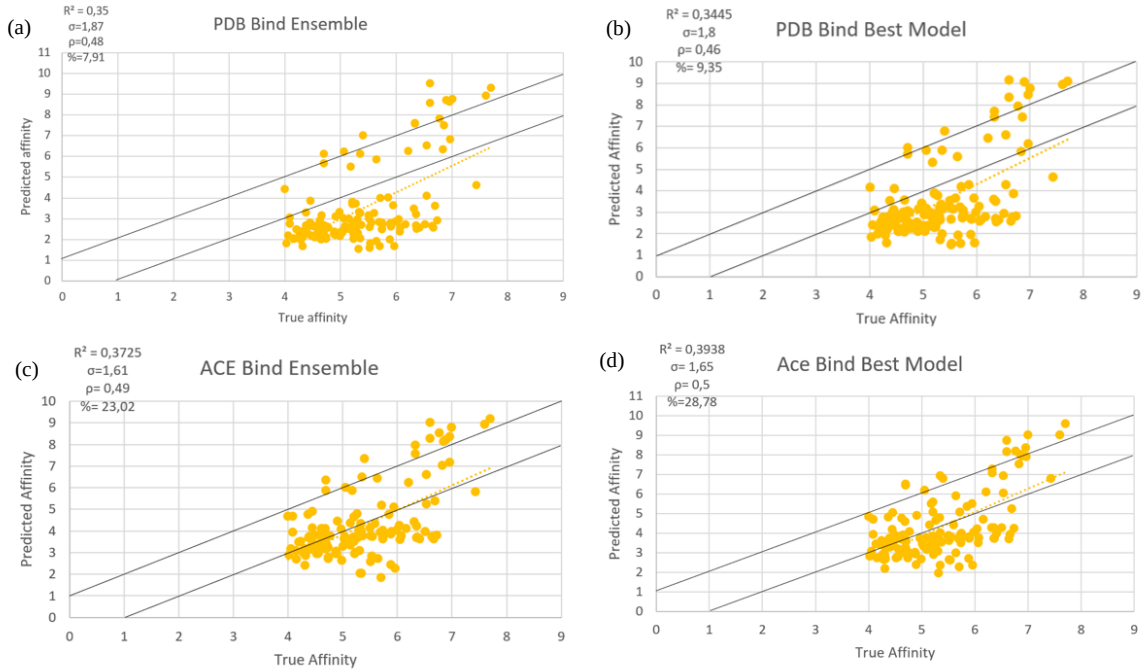


Figura 3: : Representació de la mitjana de l'afinitat predita (pkd predit) i expermental (pkd experimental) juntament amb el coeficient de determinació (r^2), el coeficient de correlació de Pearson (r), el coeficient de correlació d'spearman (ρ), i l'error quadràtic mitjà (RMSE) en els diferents models utltizats amb les dades de 139 inhibidors no covalents de la proteasa principal del Sars-CoV-2 M-pro (a) PDBbind Ensemble (b) ACEbind Best Model (c) PDBbind Best Model (d) ACEbind Best Model (e) KdeepTrainer 300 repeticions PDBbind (f) KdeepTrainer 500 repeticions PDBbind (g) Kdeeptrainer 1000 repeticions PDBbind.

FEgrow:

Finalment, hem comparat els resultats del estudi realitzat recentment per la universitat de Newcastle amb els nostres. La figura 5a esta representada la comparació entre els càlculs de l'energia lliure experimental i la simulació presentant un coeficient de determinació (R^2) de 0,53.(34).

La figura 5 mostra una comparació entre l'estudi realitzat per la universitat de Newcastle i l'estudi realitzat amb el Kdeep. En primer lloc, s'observa com el coeficient de determinació dona millors resultats amb el FEgrow que amb el Kdeep, s'ha de tenir en compte que en l'estudi del FEgrow utilitzen menys dades i més eines bioinformàtiques, s'hauria de fer un treball més extensiu per veure com funciona amb diverses dades i analitzar totes les eines que s'utilitzen.

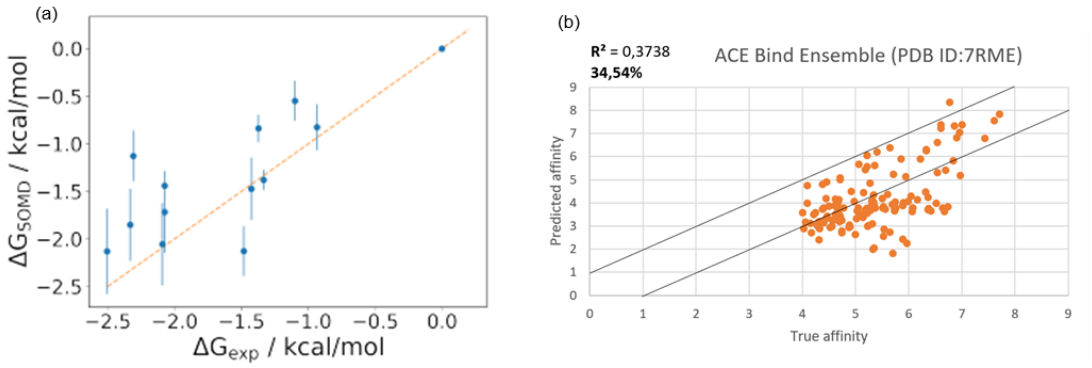


Figura 4: (a) Representació entre l'energia de Gibbs experimental i predita de 13 inhibidors del Sars-CoV-2(34) ((b) Representació de l'afinitat experimental i predita de 139 inhibidors del Sars-CoV-2.

Conclusions:

En aquest estudi s'intenta trobar una eina eficaç per poder avaluar l'afinitat entre diferents compostos vers la Sars-CoV-2 M-pro. Amb aquet objectiu s'han fet diferents experiments amb múltiples eines computacionals per tal de trobar un mètode de predicció d'afinitat i així poder desenvolupar un tractament efectiu contra la COVID-19. Les eines que principalment hem utilitzat son el Kdeep i el KdeepTrainer.

Primer vam desenvolupar diferents models de predicció d'afinitat amb el KdeepTrainer a través de dades experimentals de valors de pIC50. S'ha vist que els models que millors resultats han donat han estat els models que han utilitzat un model de dades addicionals per ser entrenats. Malauradament els valors del coeficient de Pearson eren una mica baixos, però suficient per ser utilitzats en futurs estudis d'optimització.

Amb el Kdeep vam utilitzar aquests models per predir les diferents bioactivitats dels enllaços. Amb aquest experiment vam arribar a les conclusions, que els models que donen millors prediccions d'afinitat són els que et proporciona el servidor donada la gran quantitat d'entrenament que es va utilitzar per a la creació del model. Per poder arribar a valors similars o millor seria necessari una major quantitat de dades d'entrenament. Per el moment es millor utilitzar el Kdeep sense necessitat de crear models amb el KdeepTrainer. En concret les millors prediccions son de l'ACEbind Ensemble utilitzant la proteïna 7RME com a receptor. Tot i així els resultats que dona a l'hora d'avaluar la fiabilitat de les prediccions no són molt bons, el que significa que amb aquesta metodologia obtindrem una predicció d'afinitat aproximada però no exacta. Per futurs estudis podríem utilitzar aquesta metodologia per tenir una idea aproximada de la bioactivitat dels compostos d'interès, tenint en compte les característiques del model.

Finalment, dins els estudis recents que tracten el mateix el de la universitat de Newcastle fan un experiment molt semblant al nostre amb uns resultats

prometedors. No utilitzen moltes dades, però podria ser un bon mètode de predicció d'afinitat. Donant que aquest mètode va sortir quan estàvem acabant l'estudi, no vam poder investigar amb més profunditat la seva metodologia, la qual podríem haver integrat en els nostres experiments.

Bibliografia:

1. Luttens A, Gullberg H, Abdurakhmanov E, Vo DD, Akaberi D, Talibov VO, et al. Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. J Am Chem Soc. 2022 Feb 23;144(7):2905–20.
2. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020 Feb 15;395(10223):497–506.
3. COVID Live - Coronavirus Statistics - Worldometer [Internet]. [cited 2022 Jun 5]. Available from: https://www.worldometers.info/coronavirus/#countries
4. Ngo ST, Tam NM, Pham MQ, Nguyen TH. Benchmark of Popular Free Energy Approaches Revealing the Inhibitors Binding to SARS-CoV-2 Mpro. Journal of Chemical Information and Modeling. 2021 May 24;61(5):2302–12.
5. Zhou P, Yang X lou, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature [Internet]. 2020 Mar 12 [cited 2022 Jun 1];579(7798):270–3. Available from: https://pubmed.ncbi.nlm.nih.gov/32015507/
6. Rai P, Kumar BK, Deekshit VK, Karunasagar I, Karunasagar I. Detection technologies and recent developments in the diagnosis of COVID-19 infection. Appl Microbiol Biotechnol [Internet]. 2021 Jan 1 [cited 2022 Jun 1];105(2):441–55. Available from: https://pubmed.ncbi.nlm.nih.gov/33394144/
7. Kanhed AM, Patel D v., Teli DM, Patel NR, Chhabria MT, Yadav MR. Identification of potential Mpro inhibitors for the treatment of COVID-19 by using systematic virtual screening approach. Molecular Diversity. 2021 Feb 1;25(1):383–401.
8. Ghosh R, Chakraborty A, Biswas A, Chowdhuri S. Evaluation of green tea polyphenols as novel corona virus (SARS CoV-2) main protease (Mpro) inhibitors–an in silico docking and molecular dynamics simulation study. Journal of Biomolecular Structure and Dynamics. 2021;39(12):4362–74.

9. Banerjee R, Perera L, Tillekeratne LMV. Potential SARS-CoV-2 main protease inhibitors. Vol. 26, Drug Discovery Today. Elsevier Ltd; 2021. p. 804–16.
10. Goyal B, Goyal D. Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. ACS Combinatorial Science [Internet]. 2020 Jun 8 [cited 2022 Jun 1];22(6):297–305. Available from: https://pubs.acs.org/doi/full/10.1021/acscombsci.0c00058
11. Douangamath A, Fearon D, Gehrtz P, Krojer T, Lukacik P, Owen CD, et al. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. Nature Communications. 2020 Dec 1;11(1).
12. Szliszka E, Czuba ZP, Domino M, Mazur B, Zydowicz G, Krol W. Ethanolic Extract of Propolis (EEP) Enhances the Apoptosis- Inducing Potential of TRAIL in Cancer Cells. Molecules. 2009 Feb 13;14(2):738–54.
13. Narayanan A, Narwal M, Majowicz SA, Varricchio C, Toner SA, Ballatore C, et al. Identification of SARS-CoV-2 inhibitors targeting Mpro and PLpro using in-cell-protease assay. Communications Biology. 2022 Dec 1;5(1).
14. Citarella A, Scala A, Piperno A, Micale N. SARS-CoV-2 M pro: A Potential Target for Peptidomimetics and Small-Molecule Inhibitors. Biomolecules [Internet]. 2021 Apr 1 [cited 2022 Jun 1];11(4). Available from: https://pubmed.ncbi.nlm.nih.gov/33921886/
15. Rossetti GG, Ossorio MA, Rempel S, Kratzel A, Dionellis VS, Barriot S, et al. Non-covalent SARS-CoV-2 Mpro inhibitors developed from in silico screen hits. Scientific Reports. 2022 Dec 1;12(1).
16. Amin SA, Banerjee S, Ghosh K, Gayen S, Jha T. Protease targeted COVID-19 drug discovery and its challenges: Insight into viral main protease (Mpro) and papain-like protease (PLpro) inhibitors. Bioorganic and Medicinal Chemistry. 2021 Jan 1;29.
17. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved a-ketoamide inhibitors [Internet]. Available from: https://www.science.org

18. Tam NM, Nam PC, Quang DT, Tung NT, Vu V v., Ngo ST. Binding of inhibitors to the monomeric and dimeric SARS-CoV-2 Mpro. RSC Advances. 2021 Jan 13;11(5):2926–34.
19. Qiao J, Li YS, Zeng R, Liu FL, Luo RH, Huang C, et al. SARS-CoV-2 M pro inhibitors with antiviral activity in a transgenic mouse model. Science [Internet]. 2021 Mar 26 [cited 2022 Jun 1];371(6536):1374–8. Available from: https://pubmed.ncbi.nlm.nih.gov/33602867/
20. Limpiti T, Intarapanich A, Assawamakin A, Shaw PJ, Wangkumhang P, Piriyapongsa J, et al. Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and structure. BMC Bioinformatics. 2011 Jun 23;12.
21. Jiménez J, Škalič M, Martínez-Rosell G, de Fabritiis G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. Journal of Chemical Information and Modeling. 2018 Feb 26;58(2):287–96.
22. Good A. Virtual Screening. Comprehensive Medicinal Chemistry II [Internet]. 2007 Jan 1 [cited 2022 May 31];459–94. Available from: https://linkinghub.elsevier.com/retrieve/pii/B008045044X002625
23. Shim J, Hong ZY, Sohn I, Hwang C. Prediction of drug–target binding affinity using similarity-based convolutional neural network. Scientific Reports. 2021 Dec 1;11(1).
24. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P. Comparability of Mixed IC50 Data - A Statistical Analysis. PLoS ONE. 2013 Apr 16;8(4).
25. Nath V, Rohini A, Kumar V. Identification of Mpro inhibitors of SARS-CoV-2 using structure based computational drug repurposing. Biocatalysis and Agricultural Biotechnology. 2021 Oct 1;37:102178.
26. Kdeep: a protein-ligand binding affinity predictor [WEB APP] [Internet]. [cited 2022 May 31]. Available from: https://playmolecule.com/Kdeep/
27. KdeepTrainer: Trains neural network models for use in the Kdeep app [Internet]. [cited 2022 May 31]. Available from: https://playmolecule.com/KdeepTrainer/

28. Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine Learning Methods in Drug Discovery. Vol. 25, Molecules (Basel, Switzerland). NLM (Medline); 2020.
29. PlayMolecule® KDeep and KDeepTrainer: predicting ligand binding affinities | by Nikolai Schapin | PlayMolecule | Medium [Internet]. [cited 2022 Jun 1]. Available from: https://medium.com/playmolecule/playmolecule-kdeep-and-kdeeptrainer-predicting-ligand-binding-affinities-2246ca7b0fe2
30. Hou J, Gao T. Explainable DCNN based chest X-ray image analysis and classification for COVID-19 pneumonia detection. Scientific Reports. 2021 Dec 1;11(1).
31. AlGhamdi M, Abdel-Mottaleb M. DV-DCNN: Dual-view deep convolutional neural network for matching detected masses in mammograms. Comput Methods Programs Biomed [Internet]. 2021 Aug 1 [cited 2022 Jun 1];207. Available from: https://pubmed.ncbi.nlm.nih.gov/34058629/
32. Welcome to PDBbind-CN database [Internet]. [cited 2022 May 31]. Available from: http://www.pdbbind.org.cn/index.php
33. About Us – Acellera [Internet]. [cited 2022 May 31]. Available from: https://www.acellera.com/about/
34. Bieniek MK, Cree B, Pirie R, Horton JT, Tatum NJ, Cole DJ. An Open-Source Molecular Builder and Free Energy Preparation Workflow [Internet]. Available from: https://github.com/cole-
35. mols2grid · PyPI [Internet]. [cited 2022 Jun 1]. Available from: https://pypi.org/project/mols2grid/
36. RDKit [Internet]. [cited 2022 Jun 1]. Available from: https://www.rdkit.org/
37. OpenMM [Internet]. [cited 2022 Jun 1]. Available from: https://openmm.org/
38. GitHub - gnina/gnina: A deep learning framework for molecular docking [Internet]. [cited 2022 Jun 1]. Available from: https://github.com/gnina/gnina
39. BioSimSpace — BioSimSpace 2022.2.1+88 documentation [Internet]. [cited 2022 May 31]. Available from: https://biosimspace.org/
40. PostEra | COVID Moonshot [Internet]. [cited 2022 Jun 2]. Available from: https://postera.ai/moonshot

Autoavaluació:

Primer de tot agrair al grup de recerca de Quimioinformatica i Nutrigenómica per deixar-me fer el meu treball de fi de grau amb ells. Durant aquests últims mesos he pogut veure com funciona un grup de recerca i he vist com és el món de la investigació des de una nova perspectiva.

Al principi no sabia molt bé com fer un treball de recerca o per on començar, però a mesura que anava investigant i buscant informació guanyava més seguretat, encara així, els dubtes de si estarà bé o no sempre han estat presents durant tot el treball, i a base de prova i error he anat descobrint com s'han de fer les coses i com no s'han de fer.

Poder participar dins d'un grup de recerca a estat una gran experiència on he après moltes coses d'un camp científic que no coneixia tant, però sempre he volgut veure més a fons i ara he tingut l'oportunitat, he descobert noves eines i metodologies noves, com la ciència cada cop avança més. Per mi fer aquest treball ha estat un repte, ja que no moltes coses les estava coneixent i hi ha molts conceptes nous que he après durant el treball.

Annexes:

Annex 1:

Taula 1: Recull de tots els resultats extrets del últim experimento on s'han utilitzat les dades de les 139 estructures.

	PDBbind Ensemble			PDBbind Best Model			ACEbind Ensemble			ACEbind Best Model		
PROTEÏNA	**R^2**	**ρ**	**%**	**R^2**	**ρ**	**%**	**R^2**	**ρ**	**%**	**R^2**	**ρ**	**%**
5RF7	0,37	0,48	8,63	0,34	0,45	9,35	0,37	0,49	23,02	0,38	0,49	28,06
6W63	0,37	0,46	12,95	0,37	0,45	12,23	0,39	0,49	24,46	0,37	0,48	10,79
7L0D	0,38	0,49	9,35	0,35	0,45	11,51	0,38	0,49	25,89	0,38	0,5	30,21
7L10	0,37	0,48	10,07	0,35	0,45	12,23	0,38	0,49	25,89	0,38	0,5	35,25
7L11	0,39	0,45	13,67	0,37	0,45	9,35	0,36	0,49	25,18	0,38	0,5	30,21
7L12	0,37	0,5	19,35	0,36	0,45	10,07	0,37	0,49	25,18	0,38	0,5	33,81
7L14	0,37	0,48	19,35	0,36	0,45	11,51	0,37	0,49	24,46	0,4	0,5	28,78
7LMD	0,38	0,49	10,07	0,36	0,45	10,79	0,38	0,49	23,74	0,38	0,5	28,78
7LME	0,39	0,49	12,23	0,36	0,45	11,51	0,38	0,49	25,17	0,38	0,5	18,7
7LMF	0,39	0,49	11,51	0,36	0,45	12,23	0,39	0,49	23,02	0,38	0,5	30,21
7M8M	0,37	0,48	10,79	0,39	0,45	11,51	0,37	0,49	26,61	0,38	0,5	26,61
7M8N	0,38	0,49	12,23	0,36	0,45	13,66	0,38	0,5	26,61	0,39	0,5	30,93
7M8O	0,34	0,48	10,07	0,32	0,45	10,07	0,35	0,5	31,65	0,35	0,5	29,5
7M8P	0,38	0,49	10,07	0,36	0,45	7,91	0,38	0,5	23,74	0,39	0,5	27,34
7M8X	0,37	0,48	10,79	0,36	0,45	12,94	0,38	0,49	25,9	0,4	0,5	30,94
7M8Y	0,37	0,48	10,79	0,34	0,44	11,51	0,37	0,49	30,94	0,37	0,49	31,65
7M8Z	0,38	0,48	11,51	0,37	0,45	12,23	0,37	0,49	25,9	0,38	0,5	33,09
7M90	0,37	0,48	10,79	0,33	0,45	12,23	0,37	0,49	15,83	0,37	0,5	31,65
7M91	0,37	0,48	8,63	0,34	0,45	12,23	0,37	0,49	25,9	0,8	0,51	12,94
7RLS	0,38	0,49	10,07	0,35	0,45	11,51	0,39	0,49	28,78	0,35	0,54	8,63
7RM2	0,33	0,49	13,67	0,31	0,45	13,67	0,37	0,5	28,78	0,37	0,52	35,97
7RMB	0,38	0,48	10,79	0,35	0,45	10,79	0,4	0,49	25,9	0,31	0,52	9,35
7RME	0,34	0,49	17,26	0,34	0,45	15,83	0,37	0,49	34,54	0,36	0,51	36,69
7RMT	0,33	0,49	14,38	0,31	0,46	14,38	0,35	0,49	33,81	0,3	0,51	33,09
7RMZ	0,33	0,49	14,38	0,33	0,47	13,67	0,37	0,5	30,93	0,34	0,52	31,65
7RN4	0,38	0,49	9,35	0,37	0,45	10,07	0,39	0,49	27,54	0,37	0,5	31,65
7RNH	0,38	0,48	9,35	0,36	0,45	12,23	0,4	0,49	28,06	0,36	0,5	34,55
7RNK	0,39	0,49	10,79	0,38	0,45	15,83	0,37	0,48	10,07	0,4	0,5	35,25
M-pro_Nterm-x0029	0,39	0,48	10,07	0,38	0,45	11,51	0,38	0,48	17,98	0,36	0,5	27,3
M-pro_Nterm-x0050	0,37	0,48	10,07	0,35	0,45	8,63	0,38	0,49	25,17	0,35	0,5	28,78

M-pro_Nterm-x0066	0,38	0,48	10,07	0,37	0,45	8,63	0,36	0,49	1,07	0,35	0,49	29,49
M-pro_Nterm-x0077	0,38	0,48	13,67	0,37	0,45	9,35	0,38	0,49	28,06	0,38	0,51	30,21
M-pro_Nterm-x0689	0,32	0,47	6,47	0,32	0,45	5,75	0,67	0,49	23,74	0,31	0,51	10,07
M-pro_Nterm-x0691	0,34	0,47	7,91	0,34	0,45	7,91	0,37	0,49	22,3	0,39	0,51	28,78
M-pro_Nterm-x0755	0,33	0,47	6,47	0,33	0,42	6,47	0,37	0,49	23,02	0,38	0,5	27,33
M-pro-x0770	0,33	0,47	7,91	0,32	0,45	12,95	0,36	0,49	25,9	0,38	0,51	29,5
M-pro-x0830	0,33	0,47	7,19	0,33	0,45	7,91	0,36	0,49	25,89	0,38	0,51	28,05
M-pro-x10236	0,33	0,47	8,63	0,33	0,45	11,51	0,37	0,49	24,46	0,37	0,5	30,21
M-pro-x10322	0,33	0,47	6,47	0,32	0,45	10,07	0,35	0,49	23,74	0,39	0,51	30,21
M-pro-x10338	0,32	0,47	7,91	0,32	0,45	12,23	0,36	0,49	26,62	0,36	0,5	30,21
M-pro-x10371	0,32	0,47	6,47	0,33	0,46	7,19	0,36	0,49	28,06	0,38	0,5	30,21
M-pro-x10387	0,32	0,47	7,91	0,32	0,46	8,63	0,36	0,49	25,18	0,4	0,51	29,5
M-pro-x10417	0,32	0,47	8,63	0,33	0,45	10,79	0,36	0,49	25,18	0,38	0,51	28,06
M-pro-x10422	0,33	0,47	7,91	0,33	0,46	12,23	0,36	0,49	23,74	0,37	0,5	25,18
M-pro-x10423	0,33	0,47	9,35	0,31	0,45	12,23	0,36	0,49	26,62	0,36	0,5	30,93
M-pro-x10466	0,31	0,47	7,19	0,33	0,45	9,35	0,36	0,49	25,18	0,38	0,5	28,78
M-pro-x10535	0,33	0,47	7,19	0,32	0,45	9,35	0,36	0,49	25,18	0,38	0,5	29,5
M-pro-x10565	0,34	0,47	8,63	0,34	0,46	11,51	0,4	0,49	25,18	0,39	0,5	28,78
M-pro-x10638	0,33	0,47	8,63	0,32	0,45	12,23	0,36	0,49	25,18	0,39	0,5	30,94
M-pro-x10679	0,34	0,47	9,35	0,34	0,46	10,79	0,38	0,5	26,62	0,38	0,5	29,49
M-pro-x10789	0,33	0,47	7,91	0,32	0,45	8,63	0,35	0,5	23,74	0,39	0,5	28,05
M-pro-x10820	0,33	0,47	7,19	0,32	0,45	7,91	0,36	0,5	24,46	0,38	0,5	26,62
M-pro-x10870	0,33	0,47	8,63	0,32	0,45	10,07	0,36	0,49	25,18	0,38	0,51	30,21
M-pro-x10871	0,33	0,47	7,19	0,32	0,45	9,35	0,35	0,49	24,46	0,38	0,51	28,78
M-pro-x10876	0,33	0,47	7,19	0,32	0,45	7,19	0,36	0,5	25,9	0,38	0,5	30,94
M-pro-x10942	0,33	0,47	8,63	0,33	0,45	12,23	0,36	0,49	25,18	0,36	0,5	28,78
M-pro-x10959	0,33	0,47	9,35	0,34	0,45	11,51	0,36	0,49	23,74	0,37	0,5	29,5
M-pro-x11011	0,33	0,47	7,19	0,34	0,45	9,35	0,37	0,49	26,62	0,38	0,5	32,37
M-pro-x11271	0,32	0,47	7,19	0,33	0,45	9,35	0,35	0,49	22,3	0,37	0,5	28,78
M-pro-x11276	0,33	0,47	7,91	0,33	0,46	7,91	0,36	0,49	23,02	0,38	0,5	28,05
M-pro-x11294	0,33	0,47	7,91	0,33	0,46	9,35	0,36	0,49	25,18	0,38	0,5	25,9
M-pro-x11313	0,34	0,47	7,91	0,33	0,47	10,07	0,36	0,49	23,74	0,38	0,5	26,62
M-pro-x11317	0,32	0,47	7,19	0,32	0,45	8,63	0,36	0,49	27,34	0,39	0,5	30,94
M-pro-x11318	0,32	0,47	7,91	0,33	0,45	8,63	0,36	0,49	23,02	0,38	0,5	26,62
M-pro-x11366	0,32	0,47	6,47	0,32	0,45	9,35	0,36	0,49	23,02	0,36	0,49	26,62
M-pro-x11368	0,32	0,47	8,63	0,33	0,45	10,79	0,36	0,5	27,34	0,39	0,5	30,94
M-pro-x11454	0,32	0,47	7,91	0,33	0,45	10,79	0,37	0,49	26,62	0,37	0,5	29,5
M-pro-x11458	0,33	0,47	10,07	0,31	0,45	13,67	0,36	0,49	24,46	0,38	0,5	26,61
M-pro-x11488	0,33	0,47	7,19	0,32	0,45	9,35	0,36	0,49	28,78	0,39	0,5	32,37

M-pro-x11498	0,33	0,47	7,91	0,33	0,45	10,07	0,36	0,49	23,74	0,39	0,5	26,62
M-pro-x11499	0,32	0,47	7,91	0,3	0,45	9,35	0,35	0,49	24,46	0,39	0,5	28,78
M-pro-x11501	0,33	0,47	7,91	0,3	0,45	9,35	0,36	0,49	23,02	0,38	0,5	26,62
M-pro-x11507	0,32	0,47	8,63	0,32	0,45	9,36	0,36	0,49	23,74	0,4	0,5	23,74
M-pro-x11508	0,32	0,47	8,63	0,34	0,46	9,35	0,36	0,49	25,9	0,4	0,51	30,22
M-pro-x11530	0,32	0,47	7,91	0,33	0,45	10,78	0,36	0,49	24,46	0,39	0,5	28,06
M-pro-x11541	0,33	0,47	7,19	0,34	0,45	10,07	0,36	0,49	23,74	0,38	0,5	28,06
M-pro-x11542	0,33	0,47	7,19	0,32	0,45	10,79	0,36	0,5	24,46	0,39	0,5	28,78
M-pro-x11543	0,32	0,47	8,63	0,33	0,45	14,39	0,35	0,49	23,74	0,38	0,5	28,78
M-pro-x11548	0,32	0,47	10,07	0,33	0,46	11,51	0,35	0,49	23,74	0,37	0,5	28,05
M-pro-x11562	0,33	0,47	7,91	0,32	0,45	9,35	0,36	0,49	26,62	0,39	0,51	31,65
M-pro-x11564	0,35	0,45	9,35	0,33	0,45	11,51	0,36	0,49	31,65	0,39	0,5	25,9
M-pro-x11609	0,32	0,47	8,63	0,32	0,45	7,91	0,36	0,49	23,74	0,36	0,5	30,21
M-pro-x11612	0,33	0,47	7,91	0,33	0,45	11,79	0,36	0,49	22,3	0,37	0,5	28,06
M-pro-x11616	0,33	0,47	8,63	0,33	0,46	14,39	0,36	0,49	23,02	0,36	0,5	30,21
M-pro-x11641	0,32	0,47	7,91	0,33	0,45	10,07	0,35	0,49	23,02	0,37	0,5	30,21
M-pro-x11642	0,33	0,47	6,47	0,33	0,45	7,91	0,36	0,49	24,46	0,38	0,5	28,78
M-pro-x11723	0,33	0,47	7,91	0,32	0,45	12,95	0,36	0,49	25,9	0,38	0,51	29,5
M-pro-x11742	0,32	0,47	7,91	0,33	0,45	10,07	0,36	0,49	23,74	0,4	0,5	28,78
M-pro-x11743	0,33	0,47	7,19	0,34	0,45	12,94	0,35	0,46	8,63	0,34	0,45	12,64
M-pro-x11757	0,32	0,47	7,19	0,31	0,45	10,07	0,35	0,49	25,18	0,38	0,51	29,5
M-pro-x11764	0,33	0,47	7,19	0,34	0,45	10,07	0,36	0,49	26,62	0,38	0,51	27,34
M-pro-x11789	0,32	0,47	7,91	0,32	0,49	11,51	0,35	0,49	24,46	0,38	0,51	28,06
M-pro-x11790	0,33	0,47	6,47	0,31	0,45	7,19	0,36	0,49	23,74	0,39	0,51	27,34
M-pro-x11797	0,33	0,47	6,47	0,31	0,45	6,47	0,37	0,49	25,18	0,39	0,51	28,06
M-pro-x11798	0,31	0,47	7,91	0,32	0,45	10,79	0,34	0,49	25,18	0,37	0,5	28,78
M-pro-x11801	0,34	0,47	11,51	0,35	0,45	6,47	0,37	0,49	27,34	0,38	0,51	31,65
M-pro-x11810	0,32	0,47	7,91	0,32	0,45	12,23	0,35	0,49	23,74	0,38	0,51	30,21
M-pro-x11812	0,33	0,47	7,91	0,32	0,45	9,35	0,35	0,49	23,74	0,38	0,51	30,94
M-pro-x11813	0,33	0,47	7,91	0,32	0,45	9,35	0,35	0,49	23,74	0,37	0,5	28,78
M-pro-x11831	0,33	0,47	9,35	0,34	0,45	10,07	0,37	0,49	26,62	0,39	0,51	28,78
M-pro-x12000	0,33	0,47	8,63	0,33	0,45	10,07	0,36	0,49	25,18	0,38	0,51	28,78
M-pro-x12073	0,33	0,47	7,19	0,32	0,45	10,07	0,35	0,49	23,02	0,39	0,5	28,34
M-pro-x12143	0,33	0,47	6,47	0,32	0,45	7,19	0,35	0,49	9,35	0,39	0,51	27,34
M-pro-x12171	0,33	0,47	7,19	0,33	0,45	8,63	0,36	0,49	23,74	0,38	0,51	24,46
M-pro-x12177	0,33	0,47	7,91	0,31	0,45	7,19	0,36	0,49	9,35	0,38	0,51	29,46
M-pro-x12202	0,32	0,47	9,35	0,33	0,45	7,19	0,36	0,49	23,74	0,38	0,5	28,78
M-pro-x12207	0,33	0,47	7,19	0,33	0,45	10,79	0,37	0,49	24,46	0,39	0,51	26,62
M-pro-x12300	0,33	0,47	7,19	0,33	0,45	10,07	0,37	0,49	24,46	0,4	0,51	31,65
M-pro-x12321	0,32	0,47	7,91	0,34	0,45	10,79	0,34	0,49	23,74	0,4	0,5	28,78
M-pro-x12419	0,34	0,47	8,63	0,32	0,45	7,91	0,37	0,49	25,17	0,37	0,51	29,5
M-pro-x12423	0,33	0,47	7,19	0,31	0,45	7,91	0,37	0,49	25,17	0,38	0,51	30,94
M-pro-x12582	0,33	0,47	9,35	0,34	0,46	12,95	0,35	0,49	24,46	0,38	0,51	31,65

M-pro-x12587	0,32	0,45	8,63	0,32	0,46	12,95	0,36	0,49	24,46	0,38	0,51	26,62
M-pro-x12659	0,33	0,47	9,35	0,34	0,45	11,51	0,36	0,49	25,18	0,37	0,5	26,62
M-pro-x12661	0,33	0,47	7,19	0,32	0,45	9,35	0,37	0,49	23,74	0,4	0,5	28,06
M-pro-x12674	0,36	0,47	6,47	0,34	0,45	10,79	0,36	0,49	23,02	0,39	0,51	30,94
M-pro-x12679	0,34	0,47	10,79	0,34	0,46	10,79	0,36	0,49	23,74	0,37	0,5	28,05
M-pro-x12686	0,35	0,46	10,07	0,32	0,45	8,63	0,37	0,49	25,9	0,39	0,51	26,61
M-pro-x12692	0,33	0,47	7,91	0,32	0,45	9,35	0,36	0,49	24,46	0,38	0,51	28,06
M-pro-x12695	0,32	0,47	7,19	0,34	0,45	9,35	0,35	0,49	24,46	0,37	0,5	30,93
M-pro-x12696	0,33	0,47	6,47	0,34	0,45	10,79	0,37	0,49	23,74	0,39	0,51	25,18
M-pro-x12698	0,33	0,47	7,91	0,33	0,45	11,51	0,36	0,49	23,74	0,38	0,5	28,78
M-pro-x12699	0,33	0,47	7,19	0,32	0,45	10,07	0,36	0,49	23,74	0,36	0,5	26,62
M-pro-x12710	0,33	0,47	6,47	0,34	0,45	9,35	0,37	0,49	23,74	0,39	0,51	29,5
M-pro-x12715	0,34	0,47	6,47	0,34	0,45	10,79	0,35	0,49	23,74	0,38	0,5	27,34
M-pro-x12716	0,33	0,47	6,47	0,33	0,45	10,79	0,36	0,49	23,74	0,39	0,51	28,06
M-pro-x12731	0,33	0,47	5,75	0,32	0,45	6,47	0,36	0,49	23,74	0,38	0,5	25,9
M-pro-x12740	0,32	0,47	6,47	0,33	0,45	10,79	0,36	0,49	23,74	0,39	0,51	29,5
M-pro-x1336	0,33	0,47	10,07	0,33	0,45	10,07	0,37	0,49	25,71	0,34	0,48	8,63
M-pro-x1386	0,33	0,47	7,19	0,34	0,45	10,79	0,37	0,49	25,17	0,38	0,5	29,5
M-pro-x1418	0,33	0,47	7,19	0,32	0,45	7,19	0,36	0,49	22,3	0,38	0,5	26,61
M-pro-x2563	0,33	0,47	7,19	0,34	0,46	10,07	0,36	0,49	25,9	0,38	0,51	27,33
M-pro-x2572	0,33	0,47	7,91	0,33	0,45	10,79	0,37	0,5	23,74	0,37	0,5	30,21
M-pro-x2646	0,32	0,47	7,91	0,33	0,45	12,94	0,36	0,49	25,17	0,38	0,5	29,49
M-pro-x2649	0,32	0,47	7,91	0,35	0,45	12,23	0,36	0,49	25,28	0,38	0,5	30,21
M-pro-x2908	0,33	0,47	6,47	0,34	0,45	11,51	0,37	0,49	26,61	0,39	0,5	26,61
M-pro-x2910	0,34	0,47	7,91	0,32	0,45	10,07	0,37	0,49	24,46	0,39	0,5	26,61
M-pro-x2912	0,33	0,47	9,35	0,34	0,45	12,23	0,37	0,49	24,46	0,37	0,5	28,78
M-pro-x3303	0,33	0,47	5,75	0,33	0,45	10,07	0,35	0,49	22,3	0,37	0,5	28,05