Universitat de Girona

Validació i optimització de ferramentes bioinformàtiques per la predicció d'estructura de proteïnes i interaccions lligand-receptor

Aplicació a dianes terapèutiques del càncer

Estudiant: Ayman Benyahya Mechouat

Correu electrònic: ayman_1971@live.com

Grau en Biotecnologia

Tutor: Ferran Feixas Gerones

Correu electrònic: ferran.feixas@udg.edu

Data de dipòsit de la memòria a través de la plataforma de TFG: 04/07/2023

ÍNDEX

RESUM

Actualment, s'estan desenvolupant ferramentes bioinformàtiques enfocades a la predicció d'estructures proteiques i la simulació de les interaccions presents entre aquestes proteïnes i diferents molècules. Entre elles, destaca el programa AlphaFold, un innovador sistema d'intel·ligència artificial, que s'utiliza per predir estructures de proteïnes. Tot i això, l'aplicació d'aquests protocols computacionals es troben en procés de validació. En concret, encara no s'ha trobat un protocol general que permeti aplicar aquestes ferramentes de manera rutinària en el procés de disseny de fàrmacs o la medicina personalitzada.

L'objectiu d'aquest treball és, primerament, determinar la validesa del mètode AlphaFold de predicció d'estructura de proteïnes que són dianes terapèutiques contra el càncer. En primer lloc s'ha determinat que les estructures predites per AlphaFold són altament similars a les experimentals quan es parla en termes d'estructura global, donant a entendre el seu potencial i la projecció que té aquest sistema.

No obstant això, una anàlisi més detallada d'aquestes estructures generades amb AlphaFold ha revelat que existeixen variacions respecte a estructures experimentals en certs dominis que, generalment, es troben desordenats, però que són crucials per al funcionament fisiològic de la proteïna. De fet, juguen un paper crucial en la regulació i transmissió de vies de senyalització i, sobretot, afecten a les quinases per la seva naturalesa en si.

Seguidament, s'ha estudiat la interacció d'aquestes proteïnes predites amb diferents molècules amb les quals tenen certa afinitat utilitzant programes de docking molecular. S'ha observat que les estructures AlphaFold en general no ofereixen una millor predicció de la interacció receptor-lligand que les estructures experimentals.

Finalment, s'ha demostrat que les estructures predites per AlphaFold relaxades mitjançant una dinàmica molecular no millora el resultat de la predicció inicial.

Aquest estudi ha mostrat que les eines bioinformàtiques es troben en un desenvolupament que podria resultar molt útil per a la recerca. No obstant això, cal seguir treballant en la millora d'aquestes eines per poder analitzar i comprendre millor les interaccions complexes que es produeixen a nivell fisiològic. Aquest avenç impulsaria el descobriment i desenvolupament de noves teràpies dirigides a proteïnes responsables de l'aparició de malalties com el càncer.

RESUMEN

Actualmente, se están desarrollando herramientas bioinformáticas enfocadas en la predicción de estructuras proteicas y la simulación de las interacciones presentes entre estas proteínas y diferentes moléculas. Entre ellas destaca el programa AlphaFold, un innovador sistema de inteligencia artificial utilizado para predecir estructuras de proteínas. Sin embargo, la aplicación de estos protocolos computacionales se encuentra en proceso de validación. Específicamente, aún no se ha encontrado un protocolo general que permita aplicar estas herramientas de manera rutinaria en el proceso de diseño de fármacos o medicina personalizada.

El objetivo de este trabajo es, en primer lugar, determinar la validez del método AlphaFold en la predicción de estructuras de proteínas que son dianas terapéuticas contra el cáncer. En primer lugar, se ha determinado que las estructuras predichas por AlphaFold son altamente similares a las experimentales en términos de estructura global, lo que indica su potencial y proyección.

Sin embargo, un análisis más detallado de estas estructuras generadas con AlphaFold ha revelado que existen variaciones respecto a las estructuras experimentales en ciertos dominios que, generalmente, están desordenados pero que son cruciales para el funcionamiento fisiológico de la proteína. De hecho, desempeñan un papel crucial en la regulación y transmisión de vías de señalización y, especialmente, afectan a las quinasas debido a su propia naturaleza.

A continuación, se ha estudiado la interacción de estas proteínas predichas con diferentes moléculas con las que tienen cierta afinidad utilizando programas de docking molecular. Se ha observado que, en general, las estructuras de AlphaFold no ofrecen una mejor predicción de la interacción receptor-ligando que las estructuras experimentales.

Finalmente, se ha demostrado que las estructuras predichas por AlphaFold, relajadas mediante una dinámica molecular, no mejoran el resultado de la predicción inicial.

Este estudio ha mostrado que las herramientas bioinformáticas están en desarrollo y podrían resultar muy útiles para la investigación. Sin embargo, es necesario seguir trabajando en la mejora de estas herramientas para poder analizar y comprender mejor las interacciones complejas que ocurren a nivel fisiológico. Este avance impulsaría el descubrimiento y desarrollo de nuevas terapias dirigidas a proteínas responsables de la aparición de enfermedades como el cáncer.

ABSTRACT

Currently, bioinformatic tools focused on predicting protein structures and simulating the interactions between these proteins and different molecules are being developed. Among them, the AlphaFold program stands out as an innovative artificial intelligence system used to predict protein structures. However, the application of these computational protocols is still undergoing validation. Specifically, a general protocol that allows the routine use of these tools in drug design or personalized medicine has not yet been found.

The objective of this work is primarily to determine the validity of the AlphaFold method for predicting the structures of proteins that are therapeutic targets against cancer. Firstly, it has been determined that the structures predicted by AlphaFold are highly similar to the experimental ones in terms of global structure, indicating its potential and the projection that this system has.

However, a more detailed analysis of these structures generated with AlphaFold has revealed that there are variations compared to experimental structures in certain domains that are generally disordered but crucial for the physiological functioning of the protein. In fact, they play a crucial role in the regulation and transmission of signaling pathways and, especially, they affect kinases due to their inherent nature.

Subsequently, the interaction of these predicted proteins with different molecules, with which they have a certain affinity, has been studied using molecular docking programs. It has been observed that, overall, the AlphaFold structures do not provide a better prediction of the receptor-ligand interaction than the experimental structures.

Finally, it has been demonstrated that the structures predicted by AlphaFold, relaxed through molecular dynamics, do not improve the outcome of the initial prediction.

This study has shown that bioinformatic tools are in development and could be very useful for research. However, further work is needed to improve these tools in order to better analyze and understand the complex interactions that occur at the physiological level. This advancement would drive the discovery and development of new therapies targeting proteins responsible for diseases such as cancer.

REFLEXIÓ ÈTICA

La tecnologia està en constant evolució i es pot justificar amb la seva influència en els aspectes de les nostres vides. Cada vegada més, es presència com s'integra de manera significativa en la salut humana i en la ciència, brindant avenços i possibilitats que abans semblaven complicades o fins i tot inimaginables.

Un camp que destaca en aquesta convergència, entre la tecnologia i la salut, és la bioinformàtica, una disciplina que combina la biologia i la informàtica per comprendre i analitzar dades biològiques complexes. En aquest treball, concretament, s'avalua el potencial i la fiabilitat d'una de les eines bioinformàtiques més revolucionaries d'aquests temps, AlphaFold.

Fent servir aquesta nova eina, programa d'intel·ligència artificial desenvolupada per DeepMind (Google), es poden obtenir prediccions d' estructures tridimensionals de proteïnes a partir de la introducció de les seves seqüències corresponents. La determinació de la capacitat d'aquest programa permetria confiar en una revolució del sistema actual d'obtenció estructural convertint-la en una molt bona opció en diferents camps de la ciència i la salut[1]. En concret es podria aplicar en el camp de la medicina personalitzada.

Tanmateix, a mesura que avancem en aquest camp, és imprescindible reflexionar sobre les qüestions ètiques que sorgeixen en relació amb l' ús de programes bioinformàtics, ja que es manipulen dades genètiques des del punt de vista computacional.

Un dels temes centrals o més preocupants que s'han de considerar és el de la privacitat i la confidencialitat. En utilitzar programes bioinformàtics, normalment es requereix d'accés a dades personals i genètiques dels individus, com podria ser en el cas del registre a un programa en línia per poder-lo fer servir. Per tant, resulta fonamental garantir la protecció de la privacitat i la confidencialitat d' aquestes dades, evitant el seu mal ús o divulgació no autoritzada. La pirateria informàtica certament és capaç de superar el Firewall protector d'un servidor bioinformàtic el qual pot contenir dades personals importants. Un exemple serien els codis genètics utilitzats per a la predicció d'estructures i, respecte a aquest tema sorgeixen diverses hipòtesis i ideologies. Personalment, la informació genètica de la població supera de manera destacable, per exemple, la informació d'una targeta de crèdit. Aquesta comparativa es fa per fer entendre que la informació genètica és la base de la nostra identitat i que en cas de modificar-se, pot modificar a l'individu en si[2].

Utilitzar dades genètiques per a anàlisis implica establir mesures de seguretat sòlides i obtenir el consentiment informat de les persones involucrades.

La qüestió de la propietat intel·lectual i l'accés obert també mereix una reflexió ètica. La bioinformàtica depèn de l' intercanvi de dades i eines entre investigadors, però la qüestió de l'accés

i la propietat dels programes bioinformàtics pot plantejar dilemes ètics. Fomentar la col·laboració i l'accés obert als recursos bioinformàtics és important, alhora que es protegeixen els drets de propietat intel·lectual i es reconeixen els esforços dels creadors (s'eviten plagis).

La responsabilitat i la transparència són aspectes ètics fonamentals en l'ús de programes . Aquests programes poden tenir un impacte significatiu en la presa de decisions clíniques i de recerca, per la qual cosa és essencial que els desenvolupadors i usuaris siguin responsables i transparents quant als mètodes utilitzats, els resultats obtinguts i els possibles límits o biaixos dels programes. L'explicació clara i la documentació adequada són cabdals per garantir la integritat i la confiança en els resultats bioinformàtics[3].

En conclusió, la bioinformàtica s'ha convertit en una disciplina revolucionària i en ràpid creixement en els últims anys. La seva capacitat per processar i analitzar grans quantitats de dades biològiques ha impulsat avenços significatius en la genòmica, la medicina personalitzada i la recerca biomèdica. Tanmateix, a mesura que la bioinformàtica continua transformant la ciència i la salut humana, és essencial reconèixer la importància de l'ètica i la seva aplicació en aquest camp.

L' ètica exerceix un paper fonamental en la bioinformàtica, ja que les decisions preses a partir dels resultats obtinguts poden tenir implicacions significatives per als individus i la societat en general. El coneixement i la comprensió dels principis ètics, com la privacitat, l' equitat i la responsabilitat, són indispensables per garantir un enfocament ètic en l' aplicació de la bioinformàtica[4].

REFLEXIÓ PERSPECTIVA DE GÈNERE

Com ja s'ha comentat, la bioinformàtica és considerada una àrea interdisciplinària que ha generat gran interès aquests últims anys. De fet, és considerada com la combinació de coneixements i habilitats de les àrees de les famoses STEM (Science, Technology, Engineering and Mathematics)[4].

Segons un informe de l'European Center for the development of vocational training (CEDEPOF, Eurofound) de 2018, el 2025 el 85% dels llocs de treball estaran relacionats directament o indirectament amb els comentats STEM[5]. Aquest fet és degut a que la gran majoria dels treballs, en un termini mig-llarg, tindran un element comú i és la implementació dels coneixements de les diferents àrees STEM, sent base per tots ells. Per exemple, en la medicina ja s'està incorporant la tecnologia convertint-se en un aliat que serà imprescindible. Però, actualment sols el 7% de les alumnes cursen una carrera tecnològica sent una dada molt a tenir en comte. Un dels possibles motius va relacionat amb la perspectiva de gènere en aquestes àrees.

Històricament, les dones han estat subre-presentades en enginyeries i tecnologies, a més de la informàtica, i la bioinformàtica no és una excepció. En Espanya, del total de persones que es troben cursant una titulació STEM, sols un 28% són dones sent més inferior en el cas de les enginyeries (20%). Són proporcions a priori que podrien anar millorant al llarg del temps però no és el cas ja que, en Espanya, el percentatge de dones que han decidit cursar estudis d'informàtica, ha disminuït d'un 30 % a un 12% en els darrers anys. Però no sols hi és present aquest problema en Espanya ja que en Europa, sols un 30% de les persones que treballen en el sector TIC són dones.

S'han realitzat estudis que poguessin explicar aquesta situació i s'han establert uns factors que podrien influir com podrien ser els estereotips i els rols de gènere associat a les dones[6]. Per exemple, "falses" creences com podria ser que les dones són menys competents en matemàtiques i/o tecnologies, podria generar una alteració en la percepció d'habilitats i capacitats de les dones en aquestes àrees.

A més, les expectatives culturals i socials poden generar barreres per a la participació de les dones en àrees STEM. La falta de models a seguir femenins en aquestes àrees poden influir en l'elecció de carreres i en les oportunitats de desenvolupament professional.

Els ambients de treball en aquestes àrees sovint tenen una cultura dominada per homes, la qual cosa pot generar barreres per a la participació i l'avenç de les dones. Aquestes barreres poden incloure manca de suport, manca d' oportunitats de mentoria, així com discriminació i biaixos en les decisions de promoció. Finalment, les expectatives tradicionals de gènere relacionades amb la

cura de la família i els rols domèstics són un important fet a tenir en comte ja que podrien afectar a les decisions de les dones.

De forma general, aquestes possibles causes poden variar segons les circumstàncies, els contextos i les regions d'estudi o anàlisi però tenen conseqüències comunes (la desigualtat d'oportunitats, una bretxa salarial i segregació ocupacional, violència de gènere o afectació a la salut mental). A més, relacionat amb el treball professional, aquests problemes podrien causar una manca de diversitat ideològica impedint la percepció d'altres perspectives.

Tot i així, en el camp de la bioinformàtica destaquen professionals femenines que han fet desenvolupar aquesta disciplina de manera formidable, com és el cas de Shoshana Wodak i Margaret Oakley Dayhoff, totes dues aportant un coneixement i una aplicació essencials en la bioinformàtica. Però, tot i així, situació actual promou la necessitat immediata de poder solucionar aquests problemes de discriminació de gènere.

En conclusió, la perspectiva de gènere en la bioinformàtica és un aspecte a tenir en comte per poder eliminar les possibles desigualtats actuals. Al treballar de manera conjunta i cooperativa es pot avançar a un camp més divers i eficaç gràcies a la valoració dels talents i professionals, independentment del seu gènere, amb la finalitat de beneficiar a la ciència, a la salut humana i la societat.

REFLEXIÓ SOSTENIBILITAT

Un fet molt important d'aquests últims anys és la pujada de preu de quasi totes les fonts d'energia utilitzades de manera quotidiana. En el cas dels oficis en el que es requereixen maquinàries de grans capacitats i grans fonts d'energia, aquest augment es quasi inevitable que afecti a nivell econòmic i de planificació per poder estalviar el màxim possible.

A més, aquest increment ha generat una preocupació i una repercussió en la necessitat d'investigar altres alternatives sostenibles i eficients per poder assolir un benefici econòmic i una prevenció de contaminació a nivell ambiental. Actualment es troba en desenvolupament diferents tipus de fonts d'energia renovables i netes però cal esperar un temps per poder-les introduir en la vida de les persones.

En el àmbit científic, l'experimentació és essencial per poder assolir els objectius marcats segons el projecte que es vulgui desenvolupar, però, no es pot negar que aquesta pràctica, també té la seva part negativa. Basant-se en aquest projecte, la determinació d'una estructura proteica, es pot obtenir mitjançant diferents tècniques, en especial, la cristal·lització per raig X.

Independentment de la manera i de la tècnica utilitzada, totes tenen certes característiques comunes. En primera instància, aquesta aplicació experimental pot requerir de grans quantitats de reactius i materials per poder dur a terme la cristal·lització, a més, de l'energia requerida pels instruments que s'utilitzen. En conseqüència, es podria generar un impacte negatiu en termes de consum de recursos inicials i en problemes de generació de residus.

De fet, en la formació de cristalls d'una proteïna, es soles fer servir diferents solvents i agents precipitants, com podrien ser, l'etanol, sulfat d'amoni, polietilenglicol, d'entre altres (alguns d'aquests reactius poden aparèixer registrats dintre del PDB de la estructura corresponent). A més, alguns enzims, poden necessitar ions metàl·lics que actuen com a cofactors o estabilitzadors d'estructures proteiques. En resum, es podria dir que aquesta pràctica experimental destaca per l'ús de diferents compostos químics i reactius que podrien generar riscos per la salut i per al medi ambient si no es tracten correctament ni es practiquen les mesures de seguretat necessàries.

Per contra, l'aplicació de la bioinformàtica en aquesta pràctica resol parts dels problemes esmentats. Tot i fer servir instruments electrònics per poder determinar les estructures moleculars, el consum energètic és molt menor, generant un estalvi destacable. El fet de no fer servir reactius i, per tant, no generar residus que podrien ser tòxics, aporta una aplicació de bioremediació i s'evita possibles contaminacions ambientals. Finalment, es pot comentar que no es fa servir cap experiment amb radiació evitant possibles problemes amb aquest raig X que es fa servir en un laboratori.

Per tant, tot i que s'implementen mesures de seguretat durant la manipulació experimental i la eliminació de residus per tal de minimitzar els risc i reduir possibles impactes ambientals, l'aplicació de la computació en l'àrea de la investigació i salut humana, mostra una major capacitat de reduir i evitar aquests problemes, considerant-se una tecnologia neta.

INTRODUCCIÓ

- BIOINFORMÀTICA I SITUACIÓ ACTUAL

Evolucionant de manera paral·lela amb l'avanç revolucionari de la computació d'aquests darrers anys, la biologia computacional s'està convertint en un camp científic, en auge i amb un increment de coneixement al respecte. La biologia, en aquesta era digital, requereix de computació i col·laboració amb altres camps, destacant la química. Actualment, un projecte d'investigació pot incloure múltiples sistemes models, l'ús de diverses tecnologies d'assaig i, el més important, la recopilació de diferents i nombroses tipus de dades. Amb aquestes característiques, es requereixen estratègies computacionals complexes que, de manera combinada, fan que aquest disseny i la seva execució sigui molt complicada per a un/a científic/a individual[8].

Aquesta biologia computacional es troba íntimament relacionada amb la bioinformàtica, sent la segona, un camp interdisciplinari que involucra i engloba diverses ciències i enginyeries, com la biologia molecular, genètica, informàtica, d'entre altres[9].

Inicialment, l'aparició de la bioinformàtica va ser donada amb la finalitat de poder resoldre diferents incògnites i problemes, sobretot genètics, com per exemple la manera d'emmagatzemar i organitzar seqüències de DNA, recerca d'introns i exons d'una seqüència genòmica, estudi de l'estructura d'una proteïna i moltes més pràctiques que, prèviament a la seva aparició, eren conceptes que no es podien fer servir en aquesta àrea[10].

Aquests problemes van començar uns anys posteriors a la postulació del model de doble hèlix (Watson i Crick) al 1953[11] ja que, a partir d'aquest descobriment, es comencen a proposar solucions i ferramentes innovadores per fer possible aquest anàlisi i resolució de la pròpia estructura de DNA, de la informació genètica codificant de proteïnes, propietats d'aquestes, factors associats a la regulació gènica i la evolució de rutes metabòliques[12,13].

Per tant, la idea que es té respecte a que l'aplicació de la bioinformàtica és recent no és del tot certa. És més, cap al 1960, degut a l'increment de dades relacionades amb la química proteica, va provocar que nombrosos professionals científics col·laboraren amb la finalitat de poder combinar diferents disciplines científiques amb la computació per poder superar aquest obstacle que en aquell moment hi era present. Durant aquesta etapa, a més, d'haver-hi uns avenços molt importants en la determinació d'estructures proteiques per mitjà de la cristal·lografia[14], es va publicar la primera seqüència completa d'una proteïna, la insulina, per part de Frederick Sanger i altres professionals de la Universitat de Cambridge[15,16].

A partir d'aquesta troballa, va ser un fet determinant en el camp de la bioinformàtica ja que va evidenciar la necessitat de poder interpretar la informació present i continguda en les diferents seqüències (DNA, RNA i proteïnes). A més, va fomentar el desenvolupament de mètodes més eficients i òptims per l'obtenció de seqüències proteiques, com per exemple, el mètode de degradació d'Edman[17].

Un problema que es va observar és que la seqüenciació de proteïnes grans es realitzava amb una fragmentació prèvia en pèptids de menor mida provocant un estudi individual de cada resultat de la divisió, generant errors o dificultats en el assemblatge de la seqüència completa de la proteïna. Aquest problema es va poder solucionar gracies al desenvolupament d'un dels primers software bioinformàtics de la època, COMPROTEIN, un programa informàtic capaç de determinar la seqüència primària d'una proteïna fent servir dades obtingudes a partir de seqüenciacions mitjançant pèptids d'Edman[18].

Per tant, es podria considerar que l'anàlisi de proteïnes va ser el punt de partida d'aquest nou camp ja que, tot i haver-hi un desenvolupament en temes relacions amb el DNA, es van necessitar més de 10 anys des de la resolució de l'estructura de doble hèlix del DNA[19] i més de 20 anys per a l'aparició dels primers mètodes disponibles de seqüenciació de DNA[20] fent que l'aplicació de la bioinformàtica en aquest tipus d'anàlisi s'endarrerís. Aquesta diferència de temps d'un anàlisi respecte a l'altre be donat per un major coneixement de la estructura i el comportament bioquímic de la proteïna respecte dels àcids nucleics.

Durant aquest desenvolupament de la bioinformàtica cal destacar a la doctora Margaret Oakley Dayhoff (1925-1983), una fisicoquímica nord-americana que va aplicar mètodes computacional en el camp de la bioquímica. La doctora, junt amb el físic Robert S. Ledley, van desenvolupar el programa computacional esmentat anteriorment (COMPROTEIN). El treball realitzat per la doctora es va recopilar en llibre esmentat "Atles de seqüència i estructura de proteïna" on hi eren presents totes les seqüències proteiques que en aquell moment van ser determinades[21].

Entre el 1980 i 1990, el treball de la doctora Dayhoff va ser el origen de les bases de dades primaries com GenBank i BLAST[22]. A partir d'aquest punt, hi ha una revolució en aquest camp destacant el modelatge automatitzat d'estructura de proteïnes en base a homologia amb la creació del servidor SWISS-MODEL, donant com a resultat un creixement exponencial d'aquesta disciplina[23].

Aquestes ferramentes computacionals que han anat creant-se i desenvolupant-se, han estat orientades, no sòls al tractament de dades sinó també, a la caracterització de gens, determinació de propietats proteiques, anàlisi filogenètic i la realització de simulacions amb la finalitat de poder estudiar les interaccions de biomolècules presents en una cèl·lula viva de manera computacional[24].

Algunes ferramentes a destacar són BLAST i CLUSTER OMEGA utilitzades en la identificació de gens i anàlisi de seqüències sent la primera una ferramenta de recerca de seqüències de DNA i/o proteïnes[24] i la segona, un programa capaç de poder realitzar múltiples alineaments de seqüències[25]. Relacionats amb anàlisis filogenètics, MEGA és un dels programes amb més renom, ja que s'utilitza per construir arbres filogenètics, d'entre altres, per estudiar possibles relacions evolutives[26]. Per tal de poder estudiar les biomolècules d'interès, es requereixen bases de dades on hi és present tota la informació sobre aquestes molècules biològiques ja sigui DNA, RNA, proteïnes, etc. Les més utilitzades i amb major aplicació són GenBank (recurs de seqüencies nucleotídiques[27]), UniProt (com a base de seqüències proteiques[28] on hi ha una sub secció, SWISS PROT, que conté les anotacions d'aquestes seqüències)[29], PDB (Protein Data Bank, que es diferencia d'altres bases per contenir informació sobre estructures determinades experimentalment)[30] i ENSEMBL (relacionat amb anotacions de genomes eucariotes)[31].

- APLICACIONS DE LA BIOINFORMÀTICA

Per tant, l'aplicació de la bioinformàtica, es podria resumir en 5 àrees. El modelatge molecular que correspon a la predicció estructural 3D i de funcionalitat (sobretot de proteïnes), la interacció molecular on s'estudia la relació que hi ha entre diferents biomolècules, l'anàlisi filogenètic que es centra en la construcció de la història evolutiva i identificació de regions conservades, l'anàlisi de seqüències tant de DNA com de proteïnes, la simulació de dinàmica molecular per poder posar en pràctica la informació obtinguda a través de les interaccions moleculars i visualitzar-les i, finalment, el disseny de fàrmacs.

Aquestes aplicacions, de manera tradicional, soles ser lentes i costoses generant un mercat que pressiona de manera constant la recerca i troballa de processos més automatitzats, generant una innovació en un temps relativament curt amb el mínim de risc possible.

Un exemple molt clar és la determinació d'una estructura proteica i les seves interaccions. Es determina a partir de cristal·lografia de raig X, normalment, i la espectroscòpia de ressonància magnètica nuclear. Aquestes estructures cristal·lines, mostren una estructura i unes interaccions estàtiques, sent un resultat no del tot correcte ja que la relació i combinació entre molècules és més complexa i, per tant, no es podria definir i representar amb una sola estructura sense moviment. Aquest moviment es defineix com els diferents canvis conformacionals que presenta una proteïna al llarg del temps que provoca, per exemple, una variació en l'afinitat entre una proteïna i el seu lligand.

Un altre exemple amb més repercussió és el disseny i descobriments de fàrmacs. La bioinformàtica ha permès una major facilitat per aquesta aplicació pel fet de ser més ràpid l'anàlisi de molècules a nivell computacional que de manera experimental. Degut al gran impacte en aquest sector, han sorgit sub àrees dintre de la farmacologia, com el disseny de fàrmacs assistits per computadores basat en eines com el *docking* molecular [32].

L'adaptabilitat es tan gran que, fins i tot, aquestes ferramentes són capaces de predir propietats de nous fàrmacs basats en d'altres ja desenvolupats, a més de les propietats ADMET (absorció, distribució, metabolisme, excreció i toxicitat)[33].

Es pot apreciar que la bioinformàtica actua en diferents camps i no sòls els que estan relacionats en la salut humana. De fet, hi ha aplicacions d'aquesta disciplina dedicada a objectius i àrees vegetals, identificant gens clau per explotar de manera eficient les plantes com un recurs biològic per assolir una millor qualitat, una optimització de costos econòmics i una reducció en problemes ambientals. Sobretot, l'interès principal, és la generació de vegetals resistents a patògens i a estrès abiòtic[34].

Finalment, aquests últims anys la bioinformàtica ha sofert una revolució davant de l'auge de d'intel·ligència artificial. En concret en el camp de predicció d'estructura de proteïnes s'han desenvolupament de ferramentes com el programa AlphaFold que han mostrat una major precisió en la predicció d'estructures proteiques que eines bases en modelatge d'homologia fet que obra les portes cap a l'ús d'aquesta metodologia a diversos camps com la medicina personalitzada. Tot i això, aquesta ferramenta es troba encara en procés de validació. Valorar el seu ús per predir estructures de proteïnes involucrades en el càncer serà un dels objectius d'aquest treball.

- MOTIVACIÓ PERSONAL, EL CÀNCER.

El càncer és un problema actual i molt important que afecta a la salut pública, sent la segona causa de mort a nivell mundial. El càncer es podria definir com un conjunt de malalties degudes a una descontrolada i anòmala divisió cel·lular. Aquesta malaltia pot aparèixer per factor externs, com podrien ser l'estil de vida o la alimentació, i/o factors interns com mutacions genètiques. Degut a que l'origen i la causa poden ser diferents, aquests són claus per poder seleccionar i desenvolupar una teràpia personalitzada[35]. Gràcies a aquesta nova aplicació personalitzada, a més de la educació i l'avenç en les tècniques de detecció, tot i haver-hi un augment d'incidències, la mortalitat degut al càncer ha anat disminuint[36].

S'ha demostrat que una de les causes principals del desenvolupament d'un fenotip de càncer agressiu i resistent és la sobre expressió de receptors i factors de creixement. Els receptors de factors de creixement epidèrmic (EGFR) són una família de tirosina quinasa (RTK) que, amb una expressió i un funcionament anòmal, s'expressen en diferents tipus de càncer com el càncer de mama, de pulmó, d'esòfag, colorectal, d'entre altres[37]. Aquesta sobre expressió i activació afavoreix les diferents rutes metabòliques implicades en a proliferació, angiogènesi, migració i adhesió cel·lular[38].

La integració de la bioinformàtica en la oncologia permetria una millora en l'anàlisi gràcies a una integració de dades i proporció d'informació variada d'aquesta malaltia. Per aquest motiu s'ha decidit realitzar un anàlisi de diferents proteïnes involucrades en la via de senyalització d'EGFR.

- VIA DE SENYALITZACIÓ (EGFR)

Quan EGFR s'uneix al domini extracel·lular d'EGFR, es produeix una dimerització del receptor generant la activació de l'activitat tirosina-quinasa intrínseca d'aquest receptor. Aquesta activació provoca una fosforilació dels residus tirosina en el domini citoplasmàtic d'EGFR. La proteïna adaptadora GrB2 es pot unir a les tirosines fosforil·lades en EGFR gràcies al seu domini SH2. Aquesta unió provoca el reclutament i l'activació de PI3K el qual fosforil·la PIP2 a PIP3 en la membrana cel·lular. Aquesta fosforilació permet l'unió d'AKT a la membrana i, ja fosforil·lat i activat per PDK1, es desencadenarien una sèrie de cascades de senyalització que promouen la supervivència i proliferació cel·lular.

Una altra via és la regulació de l'activitat de les Cdks, enzims claus per al control del cicle cel·lular. Gràcies a l'acció d'AKT, els inhibidors presents en aquests enzims s'alliberen permetent formar complexes amb ciclines específiques.

K-Ras és una proteïna de la família RAS que actua com un interruptor molecular en aquesta via. Posteriorment a l'activació de PI3K i AKT, aquesta via es pot bifurcar per a activar la via de RAS-MAPK, que de manera fisiològica K-Ras es troba inactiva amb la unió de GTP, però al unir-se GTP, es promou la seva activació provocant l'inici de la cascada de les MAPK. Aquesta via, al igual que les altres, pot influir en la diferenciació, migració i la supervivència cel·lular, a més de la regulació d'expressió de gens implicats en el creixement cel·lular (Elk-1).

En el cas d'haver-hi algun problema o dany amb el DNA, actua p53, conegut com el guardià del genoma, ja que deté el cicle cel·lular i promou la reparació d'aquest DNA i, en cas de no ser possible, l'apoptosi per poder evitar la proliferació de cèl·lules afectades.

Actualment, es disposa d'estructures cristal·logràfiques d'aquestes proteïnes obtingudes a partir de raigs X de manera de que aquestes proteïnes involucrades en la via de senyalització EGFR poder servir com a bones estructures de referència per a la validació de ferramentes bioinformàtiques de predicció d'estructures i disseny de fàrmacs en el camp de l'oncologia[37,38].

OBJECTIVES

In recent years, a series of bioinformatics tools have been developed that use artificial intelligence for the prediction of protein structures. However, the application of these protocols are in the process of validation. In particular, a general protocol that allows these tools to be routinely applied in the drug design process or personalized medicine has not yet been found.

The aim of this work is, firstly, to determine the validity of the AlphaFold method of predicting protein structures at a global and local level, useful for delimiting areas and positions of drug binding against a series of proteins considered targets against cancer. For validation, structures obtained from X-rays will be used as a reference.

Following this prediction, we want to determine the reliability of the interaction simulations between a protein and its ligand, using two molecular docking bioinformatics tools and the analysis of the results of this simulation.

Finally, due to the fact that proteins are dynamic entities, we want to validate whether the prediction of protein structures from AlphaFold improves if a subsequent refinement of the structure obtained from molecular dynamics is carried out.

In summary, the purpose of this project is the validation of this computational protocol for the study of proteins and drug design focused on a field, in this case oncology, applying knowledge, in particular, on molecular biology and computational and bioinformatics.

MATERIALS I MÈTODES

En aquest treball s'han realitzat una sèrie d'experiments de validació per poder assolir l'objectiu marcat d'establir un protocol computacional per a l'estudi de proteïnes considerades dianes contra el càncer que es podrien resumir en tres fases. La primera fase consta de la predicció d'una sèrie d'estructures de proteïnes que es troben involucrades en diferents vies de senyalització utilitzant el programa AlphaFold. La segona fase pretén validar el potencial d'aquestes estructures predites a partir d'AlphaFold per a predir la correcte unió d'un lligand. D'aquesta manera, la finalitat d'aquesta part és la simulació de les possibles interaccions presents entre les proteïnes seleccionades i els seus lligands i validar si la disposició del lligand concorda amb l'obtinguda experimentalment. Finalment, la última part, implica utilitzar simulacions de dinàmica molecular a partir de les estructures d'AlphaFold per tal de poder determinar si la nova estructura obtinguda millora les estructures predites en fases anteriors.

Tot seguit es detallen els passos seguits i els mètodes utilitzats per portar a terme les tres fases del treball.

- SELECCIÓ DE PROTEÏNES

Les estructures proteiques de referència que s'han fet servir per a aquest projecte s'han obtingut a partir del servidor web anomenat *Protein Data Bank*. En aquest cas, s'ha decidit escollir la via de senyalització d'EGFR per motius estadístics i de manipulació ja que interessava fer servir proteïnes amb una anotació acceptable i que es tingués informació al respecte. Aquesta via de senyalització, al igual que moltes altres, presenta com diferents subvies en les que hi ha involucrades diferents proteïnes que es consideren dianes terapèutiques pel càncer (veure Introducció). Per aquest motiu, les biomolècules seleccionades formen part de diferents subvies per poder obtenir uns resultats variats i que permetin una ampliació visual de l'aplicació que es vol fer servir.

Les proteïnes i els fitxers PDB corresponents amb les que s'ha treballat són: EGFR (1IVO)[39], PI3Kα (5XGJ)[40], AKT1 (3O96)[41], Cdk2 (1B38)[42], Cdk4 (3G33)[43], Cdk6 (1JOW)[44], p53 (3TS8)[45], K-RAS (4OBE)[46] i K-RAS-G12C (8AZX)[47].

- PREDICCIÓ D'ESTRUCTURES PROTEIQUES - ALPHAFOLD

Per tal de poder fer la predicció de les proteïnes seleccionades s'ha fet servir un sistema, AlphaFold[48], un programa desenvolupat per DeepMind que permet realitzar prediccions de les estructures proteiques a partir de la seva seqüència. Les seqüències de les proteïnes s'han obtingut en format FASTA a partir del PDB. Per a la predicció d'estructures s'ha fet servir la versió oberta del programa AlphaFold que s'anomena Alphafold2Colab, amb una predicció suposadament més precises i una precisió quasi experimental[48,49,50].

La idea de validar el potencial i la coherència dels resultats que es generaran a partir del programa AlphaFold, es basa en la predicció utilizant els paràmetres per defecte que proposa el programa, com podrien ser el mètode de múltiple alineament de seqüències o paràmetres de resolució i manipulació d'imatges.

En la predicció d'algunes proteïnes, s'ha optat per utilitzar la funció multimèrica, una opció que permet integrar el lligand, obligatòriament proteic, junt amb el receptor. Amb aquest plantejament s'ha pretès simular el comportament de la proteïna en el seu estat d'oligomerització fisiològic.

Els models estructurals generats a partir d'AlphaFold s'han comparat amb les estructures cristal·logràfiques del PDB mesurant el *Root-mean-square deviation (RMSD)* entre les dues estructures. Aquesta mesura s'ha calculat a partir de les diferències entre les posicions atòmiques dels carbonis alfa de la proteïna obtinguda experimentalment (PDB) i el model computacional (AlphaFold). Les unitats de l'RMSD són ens *àngstroms* i com més gran és aquest valor indica més desviació respecte les estructures. En aquest projecte es considera que valors d'RMSD inferiors a 2 Å indiquen una predicció acurada a nivell global de l'estructura, valors entre 2 Å i 5 Å una predicció mitjanament acurada i valors superiors a 5 Å una predicció poc acurada. Tot i això, es pot obtenir una predicció acurada a nivell global però pot haver-hi diferències a nivell local (orientació zones desordenades o cadenes laterals) que poden tenir importància a l'hora de determinar la interacció amb altres molècules.

- DOCKING MOLECULAR

Un cop predita l'estructura de la proteïna a partir d'AlphaFold, s'ha validat quina capacitat tenen l'estructura del PDB i l'estructura de l'AlphaFold per predir correctament la interacció amb els seus respectius lligands. Per fer-ho s'han realitzat càlculs de docking molecular amb dos programes (HADDOCK[51] i Cluspro[52], veure més avall) per predir la orientació del lligand respecte a la proteïna d'interès. La finalitat d'aquesta part és doncs la simulació de les possibles interaccions presents entre les proteïnes seleccionades i els seus lligands (proteics i molècules petites com inhibidors). A partir d'aquestes interaccions es pot observar la disposició espaial que adopta el lligand de la proteïna d'interès. Al igual que la fase anterior, s'han realitzat els càlculs de docking molecular amb els paràmetres per defecte, a excepció d'un, la delimitació a priori de la zona activa del receptor.

- PREPARACIÓ DEL MATERIAL - PRE-DOCKING

Prèviament a realitzar els càlculs de docking molecular, s'ha hagut de realitzar un tractament dels arxius en format PDB, ja que aquest programa únicament és capaç de reconèixer aquest tipus de format o en format mmCIF. Aquest tractament consta de la preparació de dos arxius en format PDB, un primer en el que es troba el receptor i un altre on hi és el seu lligand. Tots dos arxius s'han fet passar per ChimeraX[53] per tal de afegir hidrògens, càrregues als àtoms que conformen les estructures d'entre altres opcions.

De forma paral·lela, s'ha realitzat un estudi i un anàlisi de les proteïnes i els seus lligands. Amb aquest estudi es va voler determinar, primerament, la zona activa del receptor i, posteriorment, els residus que la conformen. La zona activa correspon a la zona a on es pot unir el lligand. D'aquesta manera es dirigeix el docking a una zona concreta de la proteïna. Tot aquest procés, s'ha fet utilitzant el programa de visualització PyMOL[54].

A partir d'un PDB, en el que hi és present un receptor i el seu lligand, es separen en dos objectes dintre del mateix arxiu PDB. Fent servir la comanda de PyMOL *show sticks, byres all within 5 of NOM DE L'OBJECTE DEL LLIGAND*, es mostren tots els residus, tant del receptor com del lligand, que poden interactuar entre ells delimitant una distància màxima de 5 Å.

A més, s'ha fet servir la comanda *select active, byres all within 5 of NOM DE L'OBJECTE DEL LLIGAND* per poder seleccionar automàticament els residus de la zona activa del receptor. Aquesta acció ha fet estalviar temps ja que, en cas de no haver fet servir aquesta comanda, s'hauria haver seleccionat tots els residus manualment generant un possible error experimental en la determinació de la zona activa.

Un cop delimitada la zona activa, aquests residus es plasmen i es seleccionen en el programa HADDOCK ja que, també, s'ha volgut comparar les diferents disposicions que pot adquirir el lligand segons si es delimita o no la zona activa determinant si el programa és capaç de simular aquestes interaccions de forma correcta i amb bones aproximacions respecte a l'obtenció estructural experimental.

En cas de les proteïnes que presenten lligands proteïcs s'ha optat per utilitzar també el programa de docking molecular ClustPro, i comparar els resultats obtinguts amb aquests dos programes.

Un cop realitzat el càlcul de docking molecular, tots dos programes generen diferents arxius en format PDB on hi són presents els diferents resultats dels dockings, a més, de les energies d'interacció que hi ha segons la posició que adopta el lligand a la zona activa. Aquests PDBs obtinguts a partir del docking molecular realitzat a partir de l'estructura experimental i l'estructura provinent d'AlphaFold s'han comparat amb la orientació del lligand corresponent a les estructures PDB experimentals utilitzant una mesura d'RMSD.

- DINÀMICA MOLECULAR

Fisiològicament, les proteïnes es troben en constant moviment, generant una sèrie de estructures a mesura que passa el temps. L'objectiu de les simulacions de dinàmica molecular és simular aquest moviment molecular per tal de poder generar una sèrie d'estructures, diferents entre si, d'una mateixa proteïna. En aquest projecte, s'ha realitzar una simulació de dinàmica molecular a partir de l'estructura predita amb AlphaFold i s'ha generat una estructura mitjana a partir de les diferents estructures generades amb el programa ChimeraX. Un cop obtinguda aquesta estructura mitjana s'ha comparat l'RMSD amb l'estructura PDB experimental.

Finalment, fent servir els programes de docking esmentats anteriorment, s'ha volgut determinar si aquesta nova estructura genera prediccions amb millor resultat respecte a les anteriors.

Per realitzar les simulacions de dinàmica molecular s'ha utilitzat el següent protocol:

Preparació del sistema per a les simulacions de dinámica molecular. Un cop obtinguts els models estructurals de les proteïnes a partir d'AlphaFold es procedeix a la preparació del sistema per a la realització de simulacions de dinàmica molecular. Els sistemes s'han simulat utilitzant el programa AMBER20 amb el camp de forces ff14SB. Les simulacions s'han realitzat a les GPUs de l'Institut de Química Computacional i Catàlisi. S'escull el camp de forces ff14SB, ja que permet descriure correctament l'estructura secundària i presenta alta transferibilitat entre sistemes amb diferents característiques estructurals (hèlix alfa, làmina beta i loop). Per tal de preparar el sistema s'empra TLEAP (un mòdul del programari AMBER20) que, partint dels fitxers en format PDB obtinguts en l'apartat anterior, utilitza la informació continguda al camp de forces ff14SB per generar els dos fitxers necessaris per iniciar la simulació amb el programa AMBER20. Un fitxer conté la topologia, és a dir, els tipus d'àtoms i els paràmetres que s'utilitzen per calcular l'energia del sistema. L'altre conté les coordenades del sistema. A més, el TLEAP addiciona ions de sodi (Na+) amb els paràmetres estàndard d'AMBER fins a la neutralització i solvata el sistema en una caixa cúbica d'aigua de 10 Å amb el camp de forces TIP3P per tal de simular l'entorn característic de la proteïna en medi aquós.

Protocol de simulació. Per a començar pròpiament la simulació de dinàmica molecular (MD) a partir dels models estructurals generats amb AlphaFold cal equilibrar prèviament el sistema a les condicions de simulació. El procediment d'equilibració, consta de tres fases:

• Minimització. Procés que es duu a terme en dos passos: i) en el primer s'optimitza la posició del solvent (aigua) i dels ions amb l'estructura proteica fixada, ii) en el segon, s'optimitza la posició dels àtoms del sistema complet per tal d'obtenir una conformació inicial més estable.

• "Heating". Procés que consisteix en augmentar progressivament la temperatura de 0K a 300K, donant energia cinètica al sistema minimitzat per tal d'iniciar el moviment dels àtoms que composen el sistema dinàmic. Aquest procés es realitza en condicions de volum constant.

• Equilibració. Procés que es realitza a pressió constant permetent que el volum canviï de manera que la densitat del solvent (aigua) es relaxi i s'adapti a les condicions del sistema (temperatura 300 K).

Dinàmica molecular convencional. Una vegada el sistema està equilibrat es procedeix a realitzar pròpiament la simulació de dinàmica molecular. Aquesta fase és equivalent a l'equilibració realitzada anteriorment, però es realitza a volum constant i per un període de temps de simulació major. En aquest pas, es pretén relaxar l'estructura obtinguda a partir de l'AlphaFold fins a arribar a una conformació més estable de la proteïna. S'envien simulacions de dinàmica molecular convencional de 50 ns per les diferents proteïnes. Un cop acabades aquestes simulacions es clusteritzen per extreure l'estructura mitjana.

RESULTATS I DISCUSSIÓ

- PREDICCIÓ ALPHAFOLD

En primer lloc, s'ha realitzat una predicció de l'estructura de les proteïnes seleccionades utilizant el programa AlphaFold: EGFR, PI3K, AKT, Cdk2, Cdk4, Cdk6, p53, K-RAS i K-RAS-G12C. Un cop realitzada la predicció, el programa AlphaFold genera una sèrie de fitxers que contenen diferent informació. El primer i més important són cinc diferents estructures predites d'una mateixa proteïna, anomenades com a rank1, rank2, rank3, rank4 i rank5 (de més a menys acurat segons els criteris AlphaFold). Aquestes es troben en dos formats diferents segons l'ús que es vulgui fer servir per a aquesta predicció.

En aquet projecte tot arxiu que s'ha fet servir es troba en format PDB. A més, es generen diferents imatges (en format PNG) que poden ser útils per tal de comprovar la qualitat del model. Una d'aquestes correspon al nombre de seqüències que s'han fet servir per realitzar la predicció de la proteïna a partir d'un alineament múltiple. A més, es disposa d'un plot que representa el possible error d'alineament d'aquesta estructura predita.

Finalment, la imatge que s'ha fet servir és la prova IDDT, un test de diferencia de distancia local, que correspon a una puntuació sense superposició que avalua les diferents distancies locals de tots els àtoms en un model. Per tant, aquest test és considerat un marcador per avaluar la qualitat del model segons la posició, fins i tot, en dominis que presenten un comportament dinàmic. Es considera que si aquests valors són superiors a 80 la predicció és fiable.

De manera complementària i per tal de validar la qualitat del model, s'ha determinat el RMSD, una mesura que s'ha fet servir per avaluar i analitzar la superposició i l'alineament de les estructures d'AlphaFold respecte a les determinades experimentalment. Aquesta mesura és una desviació i mesura la qualitat d'aquesta comparació i s'expressa en unitats de longitud, en aquest cas en àngstroms (Å).

Prèviament a l'anàlisi, s'ha fet un alineament de les cinc estructures generades a partir d'AlphaFold (rank1-rank5) per determinar si la variació d'aquestes prediccions és lo suficientment gran com per ser necessari un estudi de cada una de les estructures. Amb aquest anàlisi, s'ha observat que per totes les proteïnes analitzades no hi ha quasi variació entre aquestes a excepció dels extrems, zones que no influencien de manera important en el resultat del projecte.

Per aquest motiu s'ha decidit seleccionar l'estructura denominada com a rank 1 ja que és considerat la estructura amb millor predicció pel programa AlphaFold.

Tot seguit es comenten els resultats de les prediccions per a les diferents proteïnes seleccionades.

EGFR

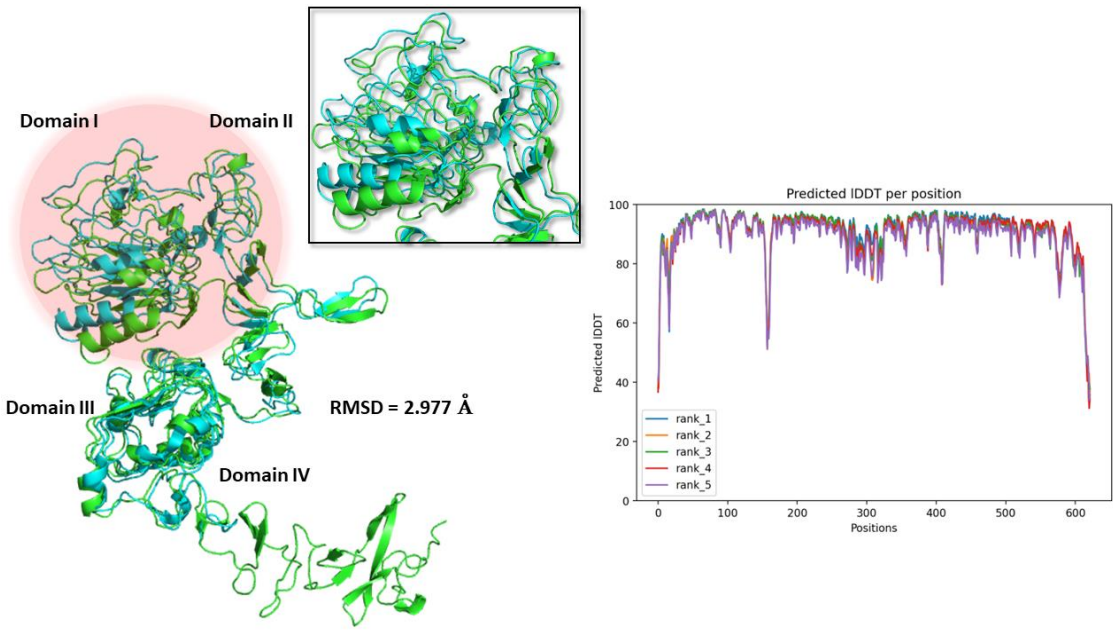


Figura 1. *Comparativa de l'estructura global de la proteïna EGFR predita (verd) respecte a l'estructura determinada experimentalment.(blau). A més, es representa la gràfica amb la puntuació IDDT, per cada posició, de la predicció.*

A partir de l'estructura d'EGFR predita per AlphaFold, s'ha realitzat un alineament entre aquesta estructura i la determinada de manera experimental (PDB). Primerament es pot observar el valor de la Desviació Quadràtica Mitjana, RMSD, que en aquest cas, correspon a un valor de 2.977 Å. Aquest valor es troba entre 2 i 5 Å donant a entendre que hi ha una similitud pel que fa als elements d'estructura secundària però que hi ha pot haver desplaçaments en la disposició espacial de les dues estructures.

Com es pot observar a la Figura 1, es poden apreciar diferències significatives al visualitzar la comparació entre les dues estructures (PDB (blau) i AlphaFold (verd)). Per exemple, a la zona C-terminal, s'observa una prolongació de la estructura AlphaFold que no s'aprecia en la estructura experimental. Aquest fet és degut a que experimental no s'ha pogut determinar l'estructura d'aquesta regió a partir de l'aspartat 513 fet que indica que l'estructura d'aquesta regió pot ser altament flexible i desordenada. Tot i això, el model AlphaFold genera una estructura que presenta un elevat nombre de làmines beta. En general, el valor de qualitat IDDT és superior a 70 per aquesta regió.

Fixant-se en la zona central d'EGFR, s'observa que hi ha certes diferències ja que no hi ha un solapament complet. De fet, hi ha una part (delimitada amb l'esfera vermella), on hi ha una diferència més destacable. Aquesta zona (Domini I i Domini II), correspon a un conjunt de residus que construeixen estructures desordenades anomenades *loop*. Aquests *loops* són molt interessants ja que són les zones amb major moviment i, per tant, una zona complicada de predir. Tot i haver-hi un desplaçament del Domini I, es pot comprovar que els *loops* tenen una conformació similar en les dues estructures.

Tot i així, s'ha vist que l'estructura experimental d'EGFR (1IVO), s'ha obtingut interactuant amb el seu lligand proteic, EGF. Aquesta unió receptor-lligand podria generar un canvi conformacional del receptor sent la causa d'aquest desplaçament dels dominis observat al PDB 1IVO respecte l'estructura AlphaFold. Per tant, l'estructura AlphaFold no és capaç de capturar l'estat conformacional de l'EGFR amb el lligand unit, de manera que possiblement s'ha obtingut l'estructura d'EFGR sense presència de lligand. Tot i això, actualment no hi ha cap estructura PDB d'EFGR que no tingui lligand, per tant, no es pot validar aquesta predicció d'AlphaFold. Per aquest motiu, s'ha realitzat una predicció de la proteïna EGFR amb el seu lligand a través de la opció multimèrica d'Alphafold, per demostrar si la predicció del receptor amb el seu lligand genera una estructura amb millor relació a la obtinguda experimentalment.

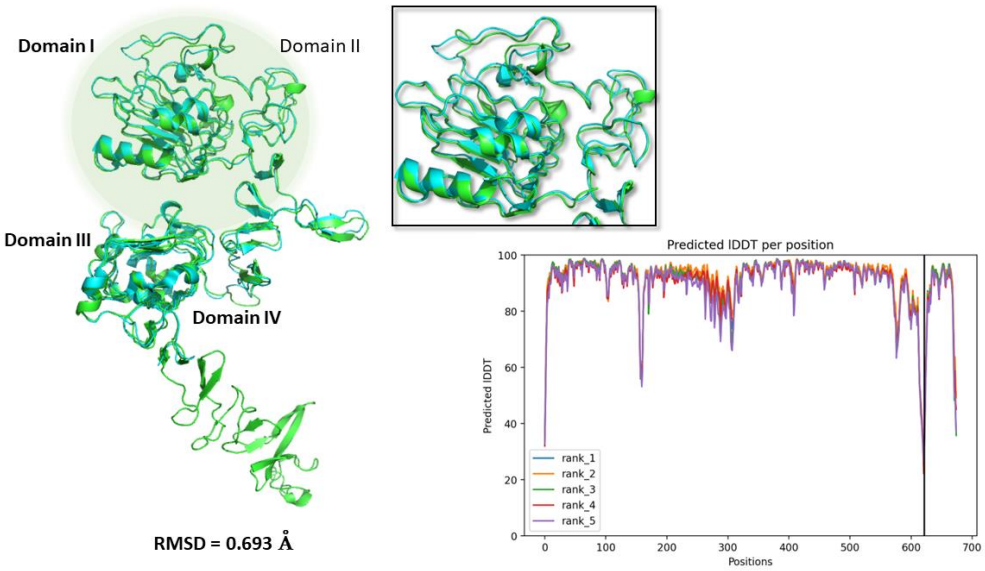
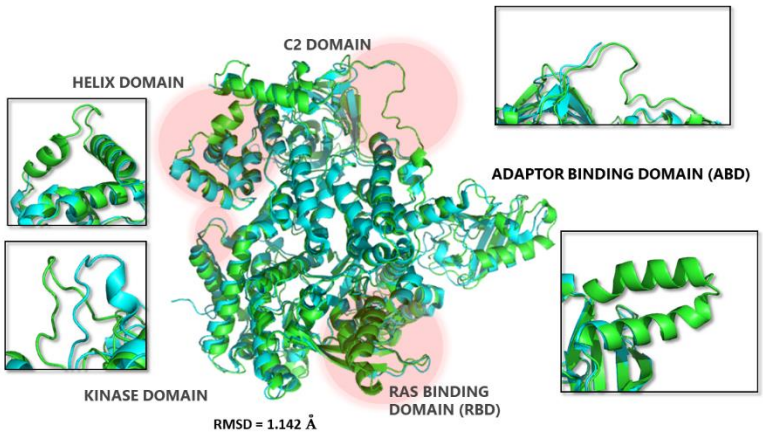


Figura 2. *Comparativa de l'estructura global d'EGFR predita amb la funció multimètica (verd)) respecte a l'estructura determinada experimentalment.(blau). A més, es representa la gràfica amb la puntuació IDDT, per cada posició, de la predicció*

Fent servir l'opció multimèrica, s'ha pogut observar una millora notable tant en l'alineament de les dues estructures com en el valor de RMSD, donant a entendre que una predicció de la proteïna amb el seu lligand genera una estructura molt similar a la cristal·logràfica. Aquests resultats es mostren a la Figura 2 i confirmen que a través d'AlphaFold es pot obtenir una estructura acurada per la proteïna EGFR en presència de lligand.

PI3Kα



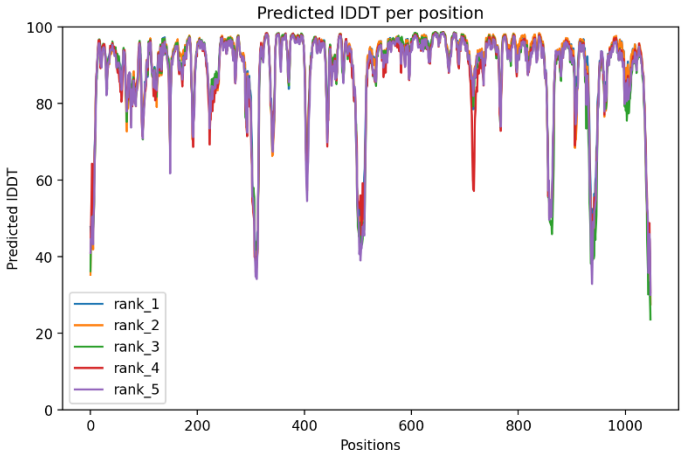


Figura 3. *Comparativa de l'estructura global de la subunitat catalítica de la proteína PI3Kα predita (verd)) respecte a l'estructura determinada experimentalment.(blau). A més, es representa la gràfica amb la puntuació IDDT, per cada posició, de la predicció*

A la Figura 3, es mostra el resultat de la proteïna PI3Kα, una de les quatre isoformes que consten la família PI3K, on cada isoforma està formada per dues subunitats: una subunitat catalítica (p110) i una subunitat reguladora (p85). S'ha vist que la subunitat p85 pot exhibit una certa plasticitat que podria anar relacionada amb una regulació de les propietats funcionals de la proteïna. La subunitat p110 es considera essencial per a PI3K perquè és la responsable de la fosforilació d'altres molècules generant una cadena de fosforilació que produeix diferents funcions cel·lulars.

S'ha volgut realitzar la predicció de les dues subunitats però, malauradament, no ha sigut possible ja que el servidor generava un codi d'error, segurament per la llargada de la seqüència i per tant, degut a la mida global de la biomolècula. Per aquest motiu s'ha seleccionat la subunitat amb major importància i la que pot desencadenar, amb una mutació, l'aparició de càncer.

Comparat la predicció de la subunitat p110α amb l'estructura cristal·logràfica, es pot observar que hi ha una gran similitud entre aquestes dues, no sols pel valor del RMSD (1.142 Å), considerat un valor que reflexa una bona aproximació, sinó que, visualment, es poden apreciar on hi ha un solapament quasi complet. Tot i així, hi ha zones que no es poden comparar perquè, de manera experimental, no s'ha pogut determinar ni resoldre certs aminoàcids que probablement corresponen a una regió flexible.

A més, una zona de molt d'interés en les quinases, és la presència d'un *loop*, anomenat *loop* d'activació. Aquesta zona, segons si hi és un lligand o no, canvia la seva conformació. Amb PI3Kα, s'observa una diferència significativa entre l'estructura PDB i l'obtinguda a partir d'AlphaFold pel que fa a aquest *loop* ja que la predicció s'ha fet sense cap lligand mentre que la estructura cristal·logràfica conté un inhibidor en aquesta zona activa de la proteïna, sent la possible causa d'aquest canvi conformacional de PI3Kα. Això significa, que tot i que AlphaFold pot predir correctament l'estructura global pot haver-hi certs canvis a nivell local que poden ser difícils d'identificar si es desconeix el lligand.

Pel que fa a la confiança de la predicció basada en el valor d'IDDT, s'ha de comentar que hi ha diferents posicions en que no es supera el 50 % donant a entendre que en aquestes zones es podria donar un plegament de la proteïna que no fos totalment correcte. Aquestes posicions, en aquest cas, coincideixen amb les zones que no s'han resolt experimentalment.

AKT1/PKD

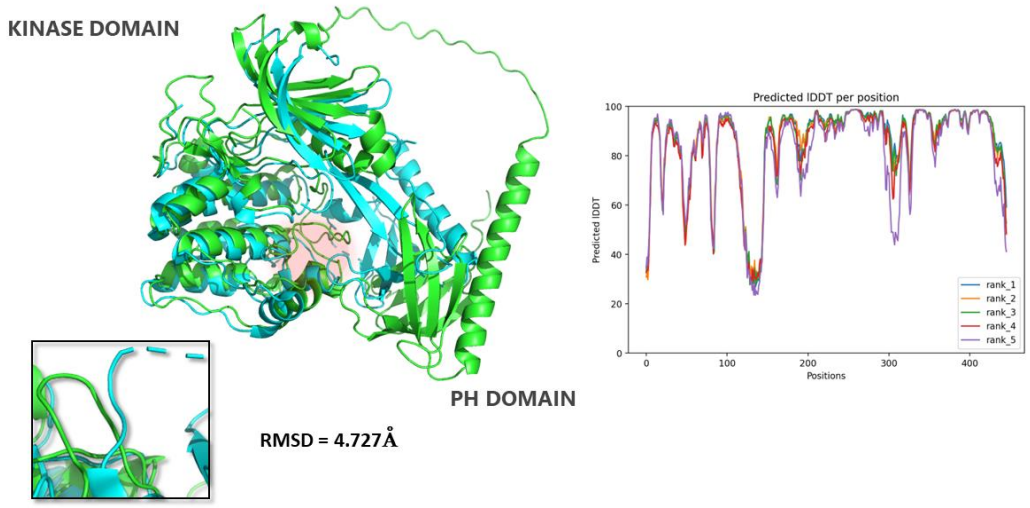


Figura 4. *Comparativa de l'estructura global de la proteína AKT1 predita (verd)) respecte a l'estructura determinada experimentalment.(blau). A més, es representa la gràfica amb la puntuació IDDT, per cada posició, de la predicció*

La comparativa de predicció de la proteïna AKT1 respecte a l'experimental no ha resultat ser la que millor resultats ha donat, tal i com es reflexa a la Figura 4, ja que hi ha diverses zones que s'observen molt distants entre si, la més destacada és la hèlix present al domini PH, que es troba unida a un *loop* format per més de 20 residus. Aquests residus no s'ha resolt de manera experimental i aquest fet pot implicar el gran canvi d'orientació del domini PH.

Una part molt important en aquest tipus de proteïnes és el *loop* d'activació, comentat prèviament. S'ha vist que els residus que formen aquest *loop* no es troben resolts a l'estructura experimental, per aquest motiu surten uns residus de manera discontinua sense acabar de formar aquesta zona. A més, no hi ha una aproximació notable d'aquest *loop* al comparar el model computacional amb la part resolta del *loop* d'activació a l'estructura cristal.logràfica.

On més desordre s'observa és en el domini d'homologia Pleckstrin (PH), un domini que consta d'uns 120 residus i que es troba involucrat en senyalitzacions internes i amb funció estructural. De fet aquest domini és molt important en les quinases ja que s'uneixen a productes fosfolípids, normalment presents a la membrana. En canvi, el domini quinasa s'observa amb major homologia tot i haver-hi petites variacions en les conformacions proteiques d'alguns elements estructurals.

En aquest cas, el valor de RMSD, aporta un valor superior a les altres proteïnes (pràcticament 5 Å) però no acaba descartant que la distància d'alineament entre aquestes dues estructures sigui prou propera. Un fet a destacar d'aquesta estructura, determinada per AlphaFold, és que hi han moltes zones que no superen el 50 % de confiança, donant a entendre que hi ha un gran possibilitat de que en les posicions marcades, no es corresponguin a conformacions fiables de la proteïna.

p53

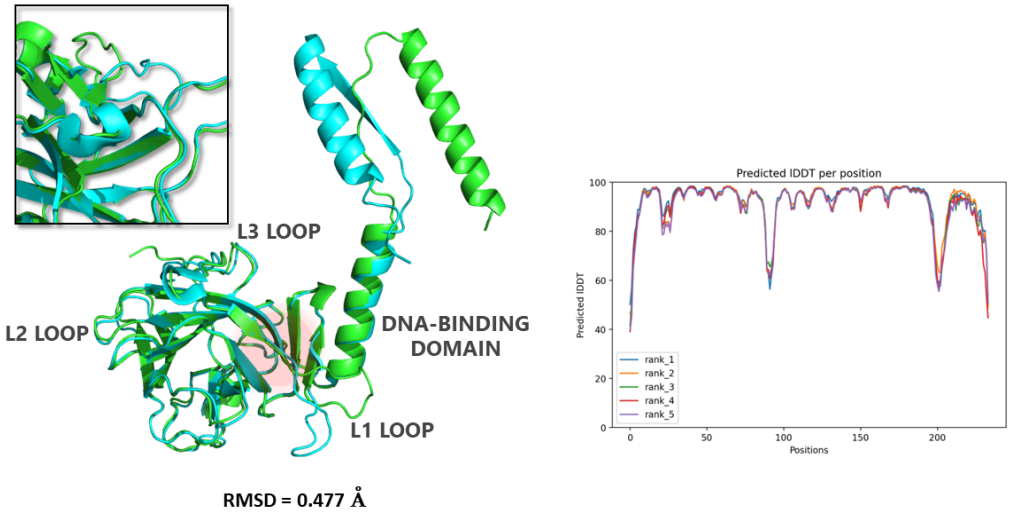


Figura 5. *Comparativa de l'estructura global de la proteïna p53 predita (verd)) respecte a l'estructura determinada experimentalment.(blau). A més, es representa la gràfica amb la puntuació IDDT, per cada posició, de la predicció*

Seguidament, s'analitza la predicció per la proteïna p53, representada a la Figura 5. L'estructura predica consta del domini d'unió al DNA, ja que aquest domini és el responsable de la interacció directa (a partir de dos monòmers formant un dímer) a la cadena doble del DNA. Aquesta proteïna reconeix dues zones d'unió definides com a RRRCWWGYYY (R = A, G; W = A, T; Y = C, T) i separades per 0-13 parells de bases.

Pel que fa a la comparativa, s'observa que la zona central de la proteïna es conserva de manera molt bona sent una aproximació a destacar i que es va perdent al llarg de la C- terminal, veient-se una distribució no coordina de la hèlix α final.

S'ha senyalat un zona on hi ha certa variabilitat entre les estructures, que correspon a una zona propera al L1 LOOP. Aquest *loop* és el responsable de la unió amb el solc major del DNA. Hi ha dos *loops* més (L2 i L3) que interactuen en el solc menor del DNA. Com que p53 s'ha determinat experimentalment unit a la CDKN1A(p21), aquesta unió, podria generar aquesta petita variació que es presenta. De fet, de manera fisiològica i sense interacció, el domini central de p53 humà, posseeix una baixa estabilitat termodinàmica intrínseca provocant desplegaments ràpids de la estructura. Aquesta baixa plasticitat es troba relacionada amb la disposició requerida per poder facilitar la unió amb la zona d'interacció. Per tant, la predicció d'AlphaFold, podria estar generant una estructura amb una certa flexibilitat lo suficientment gran com per fer variar part de l'estructura d'aquesta proteïna. Tot i així, s'observa que el plegament del L1 és diferent respecte a l'experimental, però, L2 i L3 si s'aproximen, tot i estar tots 3 *loops* interactuant amb el DNA.

No s'ha pogut fer la predicció de p53 amb el seu lligand ja que, al ser DNA, aquesta seqüència es traduiria automàticament en una proteïna, sent una comparativa poc informativa.

K-RAS

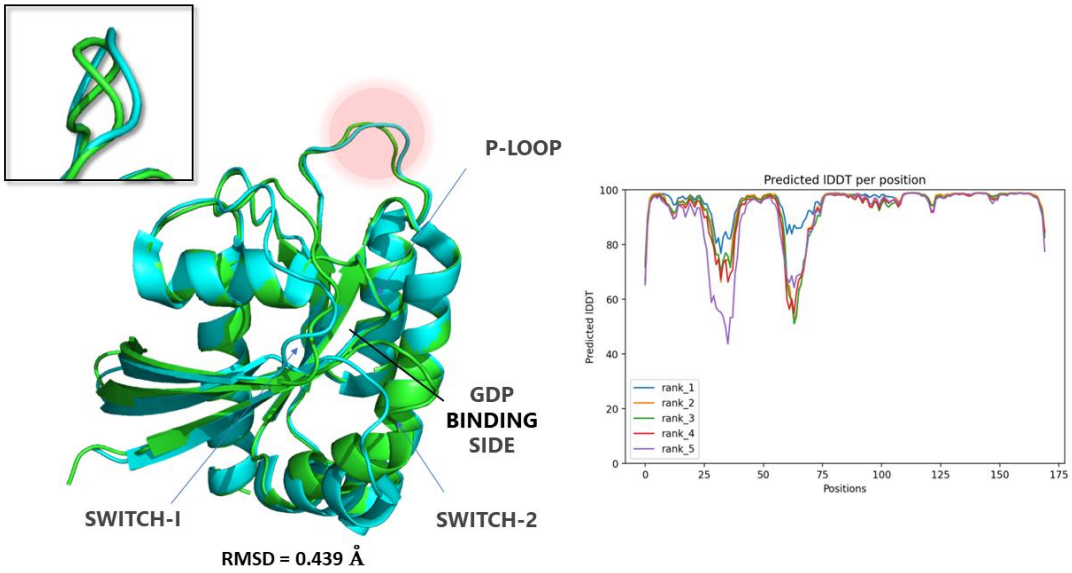


Figura 6. *Comparativa de l'estructura global de la proteína K-RAS predita (verd)) respecte a l'estructura determinada experimentalment.(blau). A més, es representa la gràfica amb la puntuació IDDT, per cada posició, de la predicció*

Amb la proteïna K-RAS, representada a la Figura 6, s'observa una comparativa molt diferent de la resta ja que la major part l'estructura AlphaFold coincideix amb el PDB. El valor de RMSD determinat, és inferior a 0.5 Å i de manera visual es corrobora aquesta similitud. Tot i així, hi ha zones en les que hi pot haver una petita variació, sobretot en les zones més desordenades. Aquestes zones corresponen als diferents *loop* que presenta aquesta proteïna. Segons la posició que presenti, es considera que K-RAS es troba de manera activa o inactiva sent una diana terapèutica per diferents fàrmacs. Teòricament, els *loops* presents a l'estructura predita deurien ser iguals d'aquestes mateixes zones de l'estructura experimental. Aquest fet s'explica perquè la predicció s'ha realitzat sense lligand, per tant, s'hauria de considerar com a receptor inactiu, al igual que l'experimentalment, ja que s'ha obtingut K-RAS unit a GDP. K-RAS per poder estar activa, ha d'estar unit a la seva zona activa GTP que, per a accedir a aquests residus, hi ha una modificació conformacional dels *loops* propers a la zona d'unió. Estudiant els diferents *loops*, s'observa que hi ha tres *loops* que no coincideixen els quals són el P-*loop*, el *loop* SWITCH-1 i SWITCH-2.

Per tant, les zones més importants de la proteïna K-RAS no s'ha pogut predir de la manera més informativa ja que no coincideix amb els loops, considerats en un estat inactiu, de la proteïna determinada experimentalment. Curiosament, aquestes zones concorden amb la poca confiança de la predicció realitzada ja que el SWITCH-I es delimita d'entre els residus 30-40 i SWITCH-2 entre els residus en la posició 55-70. El *loop* P es troba definit pels residus en posició 5-15. Per tant, aquest gràfic és molt important ja que si la confiança de cada posició no és molt elevada, aquesta posició resolta podria no ser tan correcta com s'esperaria i aquest és un cas a destacar.

QUINASA DEPENENT DE CICLINA (CDK2, CDK4 O CDK6)

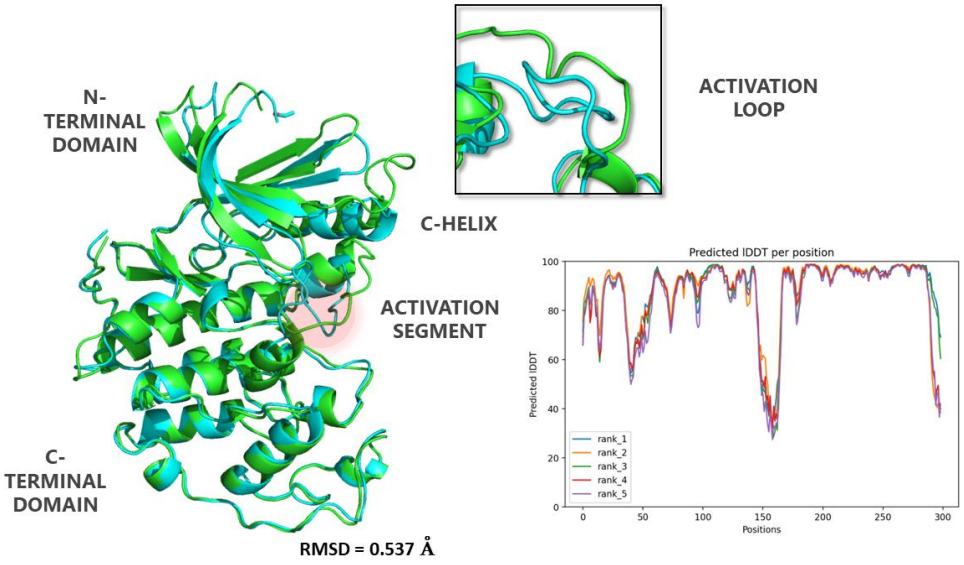


Figura 7. *Comparativa de l'estructura global de quinasa Cdk 2 predita (verd)) respecte a l'estructura determinada experimentalment.(blau). A més, es representa la gràfica amb la puntuació IDDT, per cada posició, de la predicció*

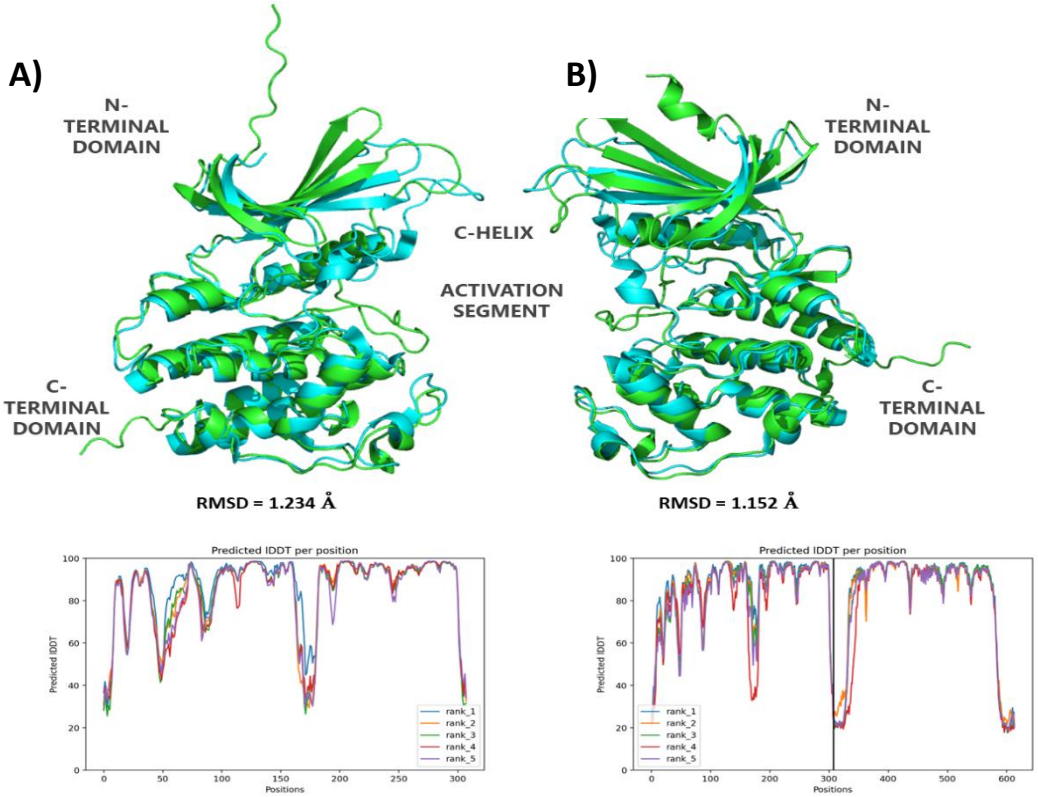


Figura 8. *Comparativa de l'estructura global de la quinasa Cdk 4 predita (verd)), sense l'opció multímer, respecte a l'estructura determinada experimentalment.(blau). A més, es representa la gràfica amb la puntuació IDDT, per cada posició, de la predicció. B) Comparativa de l'estructura global de la quinasa Cdk 4 predita (verd)), fent servir l'opció multímer, respecte a l'estructura determinada experimentalment.(blau). A més, es representa la gràfica amb la puntuació IDDT, per cada posició, de la predicció.*

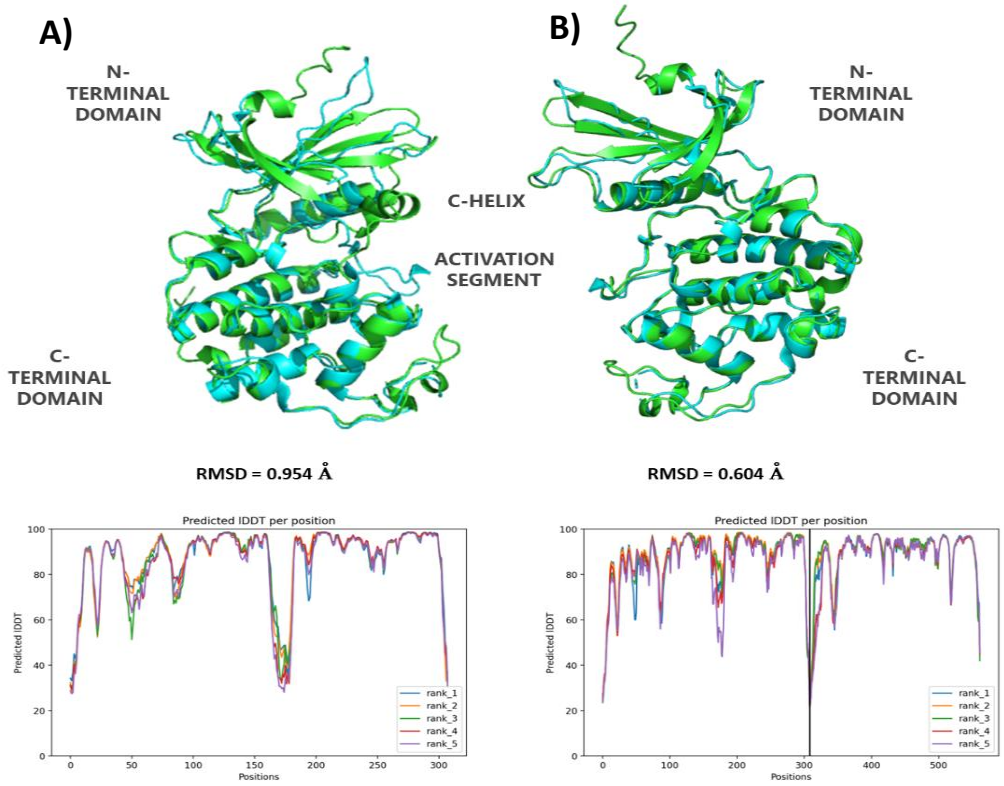


Figura 9. A)*Comparativa de l'estructura global de la quinasa Cdk 6 predita (verd)), sense l'opció multímer, respecte a l'estructura determinada experimentalment.(blau). A més, es representa la gràfica amb la puntuació IDDT, per cada posició, de la predicció. B) Comparativa de l'estructura global de la quinasa Cdk 6 predita (verd)), fent servir l'opció multímer, respecte a l'estructura determinada experimentalment.(blau). A més, es representa la gràfica amb la puntuació IDDT, per cada posició, de la predicció.*

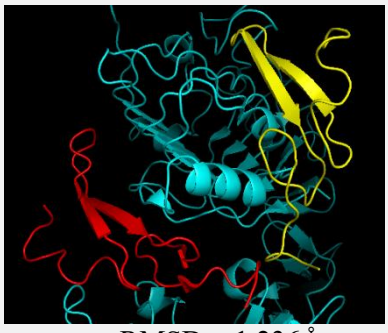
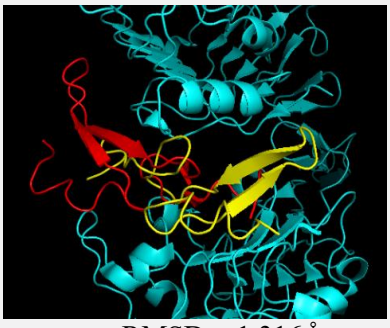
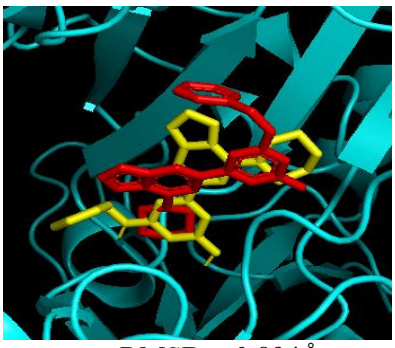
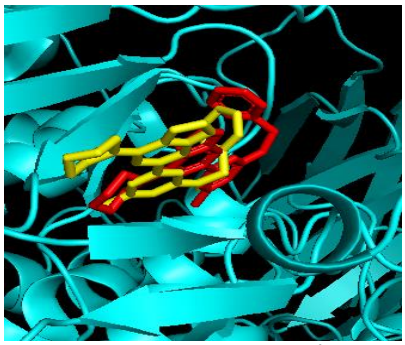
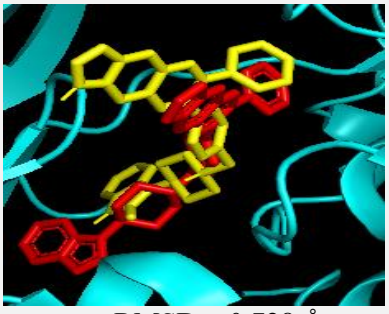
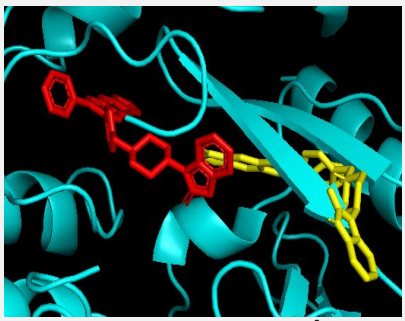
Les Cdks presenten una estructura molt similar entre si, de fet, varien en uns certs residus que els fa canviar l'afinitat per un lligand o un altre.

En el cas de Cdk2, mostrat a la Figura 7, el lligand utilitzat, és un lligand no proteic per tant, no s'ha fet servir la opció multimèrica. Les dues estructures s'assemblen molt, fet que es reflexa en el RMSD (0.537 Å), però, al igual que altres estructures, les zones més desordenades presenten una variació que cal estudiar. En totes tres quinaes, s'observa que la disposició del loop d'activació (loop T), no concorda amb l'estructura experimental. Aquest problema ja s'ha vist i s0ha comentat que podria ser degut a que AlphaFold podria predir la estructura inactiva. En les Cdk 4 y 6, presents en la Figura 8 i Figura 9 respectivament, s'ha fet servir el lligand proteic d'aquestes per determinar si aquest loop canvia de forma y es podria aproximar al loop experimental. Malauradament no s'ha aconseguit aquesta proximitat tot i haver millorat el RMSD en totes tres quinases. Per tant, es podria dir que AlphaFold no és capaç de predir una estructura en una forma activa o almenys diferenciar una proteïna que actuï de manera activa respecte a la mateixa proteïna però que es trobi inactiva.

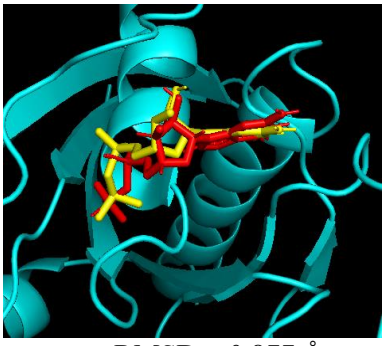
- DOCKING (PROTEÏNA-LLIGAND)

Seguidament es mostren representats els resultats resultats obtinguts al realitzar el docking proteïna-lligand *(Taula 1)*.

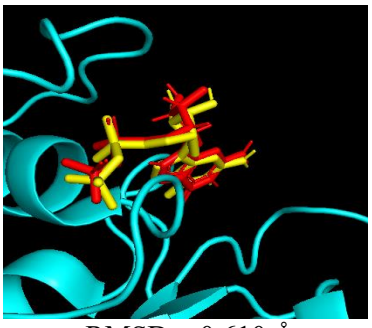
Taua 1. *Representació dels resultats obtinguts realitzant un càlcul docking, amb HADDOCK, entre les diferents proteïnes i els seus respectius lligands. Es mostren el receptor (blau), la posició del lligand determinada per la simulació d'interacció (groc) i la posició determinada experimentalment (vermell). La proteïna utilitzada com a receptor és l'aïllada a partir de la determinació cristal·logràfica (PDB) o la predita (AlphaFold).*

PROTEÏNA	RECEPTOR PDB	RECEPTOR ALPHAFOLD
EGFR **Lligand: EGF**	 RMSD = 1.236Å	 RMSD = 1.316Å
PI3Kα **Lligand:** **Inhibidor 84X**	 RMSD = 0.894Å	 RMSD = 0.710 Å
AKT1 **Lligand:** **Inhibidor IQO**	 RMSD = 0.728 Å	 RMSD = 0.608 Å

K-RAS
Lligand: GDP
(no-Mg)

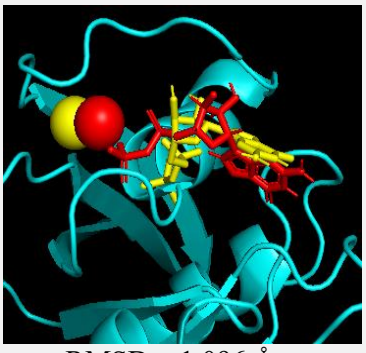


RMSD = 0.877 Å

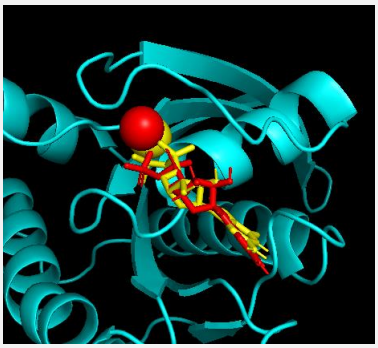


RMSD = 0.610 Å

K-RAS
Lligand: GDP
(Mg+lligand)

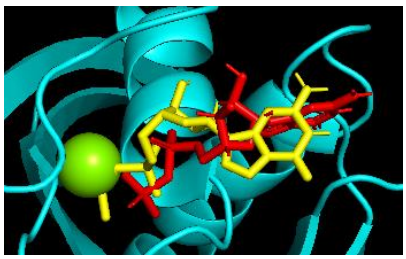


RMSD = 1.096 Å

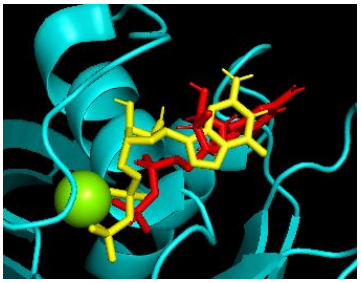


RMSD = 0.872 Å

K-RAS
Lligand: GDP
(mg+receptor)

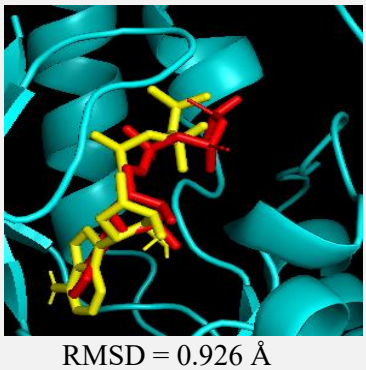


RMSD =1.423 Å

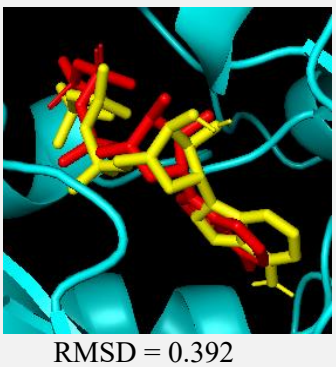


RMSD = 1.054 Å

CDK2
Lligand: ATP

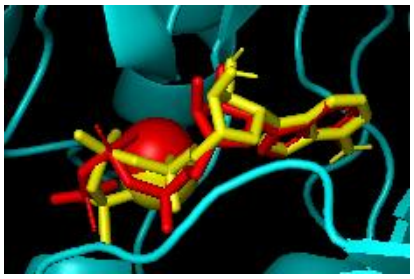


RMSD = 0.926 Å

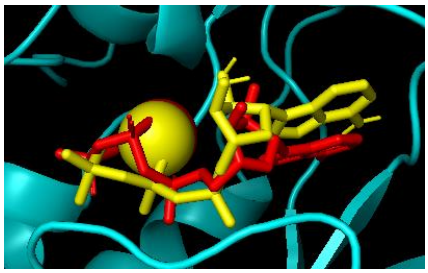


RMSD = 0.392

CK2
Lligand: ATP
(Mg+lligand)

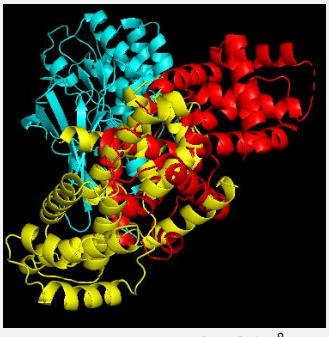


RMSD = 0.678 Å

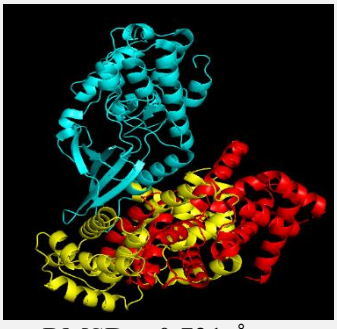


RMSD = 1.427 Å

CDK4
Lligand:
Ciclina

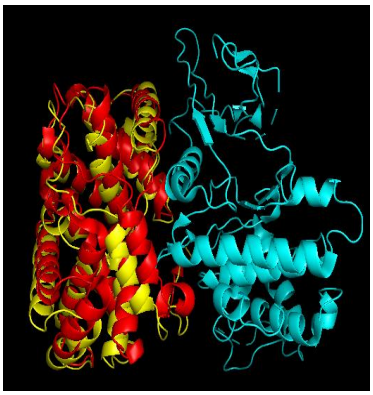


RMSD = 0.731 Å

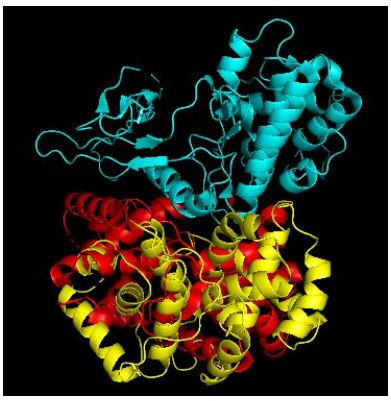


RMSD = 0.731 Å

CDK6
Lligand:
Ciclina



RMSD = 0.948 Å



RMSD = 0.779 Å

Un cop obtingudes les estructures de les proteïnes, s'ha realitzat una validació del poder predictiu d'aquestes estructures obtingudes a partir d'AlphaFold per descriure correctament interaccions proteïna lligand. Per aconseguir-ho, s'han realitzat càlculs de docking molecular amb el programa HADDOCK utilitzant els lligands co-cristal.litzats juntament amb les proteïnes seleccionades. Per tal de validar el càlcul de docking s'ha utilizat l'orientació del lligand al PDB de referència. Com que la predicció errònia de l'orientació del lligand pot venir donada pel propi càlcul de docking també s'ha realitzat aquest càlcul utilitzant l'estructura del PDB com a receptor en el procés de docking. A partir dels resultats obtinguts, tot seguit es comenten de manera general els aspectes més rellevants de les interaccions proteïna lligand.

En primer lloc, amb la proteïna EGFR, el càlcul de docking de verificació amb l'estructura del PDB no ha sigut l'esperada ja que l'anàlisi estructural ha mostrat que el programa HADDOCK,

no ha disposat el lligand de la mateixa manera en la que s'ha determinat experimentalment. Utilitzant l'estructura d'AlphaFold, s'ha vist una millora pel que fa al posicionament d'EGF però no lo suficient ja que no hi ha quasi solapament entre els dos lligands. En canvi amb el lligand d'AKT1 succeeix el contrari, l'estructura de referència del PDB proporciona una millor predicció del càlcul de docking que no la d'AlphaFold. Amb la proteïna PI3Kα, tant en la verificació com en la simulació amb la proteïna predita, s'ha vist una similitud en la ubicació de tots dos lligands, tot i això, l'estructura d'AlphaFold ofereix una millor predicció.

Les proteïnes K-RAS i Cdk2 destaquen per la presència del ió magnesi al centre actiu de les seves estructures. S'espera que la presència d'aquest cofactor tingui un efecte en el càlcul de docking a l'hora de predir la interacció proteïna-lligand ja que, de manera fisiològica, aquest magnesi pot influir en l'activitat d'aquestes proteïnes. S'han realitzat diferents proves tinguent en compte la presència de magnesi juntament amb el receptor, juntament amb el lligand i sense presència de magnesi. Efectivament, s'ha vist que la presència de magnesi en el càlcul de docking genera un canvi en la disposició del lligand, sent la millor predicció, la simulació en la que s'ha fet servir el lligand amb magnesi. Tot i així, els càlculs realitzats en presència de magnesi no superen el solapament quasi perfecte de totes dues estructures sense aquest cofactor. En general les estructures d'AlphaFold donen bons resultats a l'hora de predir l'interacció entre proteïna i lligand.

En el cas de les CDK 4 i CDK6, s'ha optat per fer servir uns lligands proteics, les ciclines. Sorprenentment, a diferència d'EGFR, s'ha vist un solapament de certs residus que formen aquestes ciclines però amb la CDK4, curiosament, aquests lligands es disposen en direcció contraria, una respecte l'altra. El càlcul de docking amb la proteïna CDK6 és la que millor resultat ha mostrat pel que fa a lligands proteics ja que la disposició de les ciclines, utilitzant l'estructura de referència del PDB, es disposen de manera molt similar i amb una lleugera diferenciació pel que fa al receptor predit. En aquest cas, les estructures d'AlphaFold donen pitjors resultats.

Pel que fa al RMSD i el HADDOCK SCORE, en aquesta pràctica, no han aportat una informació que pugui ajudar a la lectura del anàlisi realitzat. En el cas del RMSD, s'ha vist que en lligands amb una disposició molt similar a l'experimental, posseïen un RMSD excessivament elevat, per exemple amb la estructura predita de CDK2 i el lligand amb magnesi. La predicció del lloc d'unió ha sigut molt bona però presenta una desviació de 1.427 Å, superior a la que presenta EFGR. Això fa pensar que el valor d'RMSD obtingut no és correcte. Actualment s'estan explorant altres vies per realitzar aquest càlcul d'RMSD per obtenir resultats més acurats.

Referent al HADDOCK SCORE, una mesura d'afinitat entre proteïna i lligand. S'ha vist que per a interaccions que, de manera fisiològica es coneixen que són favorables, posseïen un HADDOCK SCORE excessivament elevat, fins i tot, assolint un valor positiu. Aquest fet, s'ha vist relaxat en algunes de les proteïnes estudiades. Amb el HADDOCK SCORE, s'ha demostrat, en treballs anteriors, que no és un paràmetre precís per mostrar l'afinitat d'un lligand. Una de les possibles causes podria ser la consideració d'excessives energies electrostàtiques en que algunes d'aquestes no tingui una influència directa entre el receptor i el lligand.

A partir d'aquests resultats, es pot dir que el programa HADDOCK genera una predicció de la disposició de lligands relativament bona però no perfecta i que es pot establir un patró clar aplicable a totes les proteïnes. A més, tenir en compte certs cofactors durant el càlcul de docking no ha millorat aquesta col·locació del lligand i que les millors prediccions han sigut fent servir el receptor i el lligand sense cap element extra. Els càlculs de docking han millorat

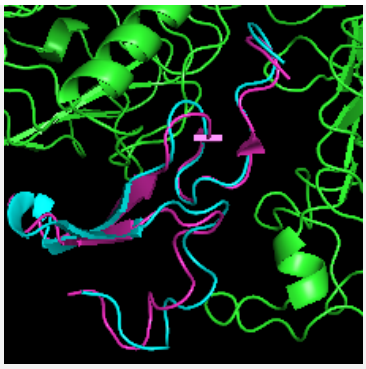
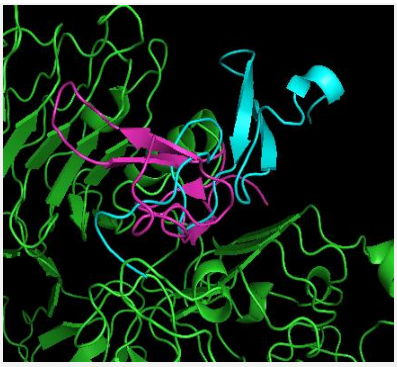
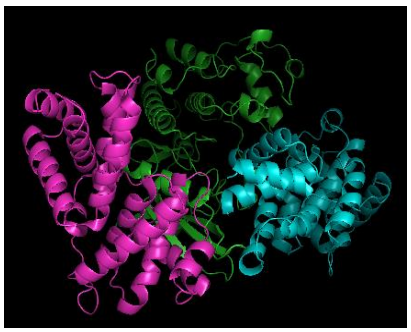
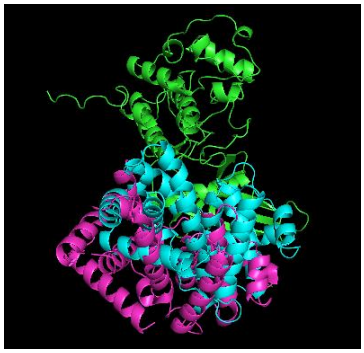
considerablement, quan s'ha realitzat una selecció de la zona activa del receptor. Finalment, cal destacar que on millor s'ha vist aquesta predicció és en lligands químics petits.

Pel que fa a la validesa de les estructures generades a partir d'AlphaFold s'ha vist que en molts casos s'han obtingut resultats similars als de les estructures de referència. Fet que fa pensar que poden ser unes bones estructures per analitzar interaccions proteïna-lligand.

COMPARATIVA HADDOCK - CLUSPRO

Seguidament, s'han realitzat càlculs de docking utilitzant un programa diferent, el ClusPro. D'aquesta manera es vol comprovar si els resultats són dependents del programa de docking utilitzat. El ClusPro únicament permet realitzar acoblaments proteïna-proteïna. Per aquest motiu només s'han analitzat la proteïna EGFR i CDK4.

Taula 2. *Representació dels resultats obtinguts realitzant un càlcul docking, amb ClusPro, entre les diferents proteïnes i els seus respectius lligands. Es mostren el receptor (verd), la posició del lligand determinada per la simulació d'interacció (blau) i la posició determinada experimentalment (rosa). La proteïna utilitzada com a receptor és l'aïllada a partir de la determinació cristal·logràfica (PDB) o la predita (AlphaFold)*

	RECEPTOR PDB	**RECEPTOR ALPHAFOLD**
EGFR Lligand: EGF	 RMSD =1.319 Å	 RMSD = 1.557 Å
CDK 4 Lligand: Ciclina	 RMSD = 1.242 Å	 RMSD = 1.450 Å

A partir dels resultats obtinguts (taula 2), s'ha vist que la verificació amb l'estructura de referència del PDB de EGFR el solapament és quasi perfecta, amb una aproximació molt bona respecte a la disposició experimental. En el cas de l'estructura predita amb AlphaFold d'aquesta proteïna, s'ha vist una variació pronunciada però, tot i així, un millor posicionament respecte als resultats del HADDOCK.

Amb CDK4 no s'ha obtingut el mateix resultat. La verificació d'aquesta proteïna amb l'estructura de referència no ha sigut la esperada, de fet, no hi ha cap coincidència o solapament en algun dels residus que presenta el lligand que, a comparació del HADDOCK, si hi era mínimament present. Amb l'estructura AlphaFold, els resultats han millorat generant un interacció similar a l'experimental.

Per tant, a partir d'aquests resultats, inicialment, semblava que ClusPro era capaç de determinar d'una manera més precisa l'acoblament entre dues proteïnes, sent una el receptor i l'altra el lligand. No obstant, amb la Cdk4, no ha hagut una predicció tan precisa com la obtinguda amb EGFR. Tot i així, aquests dos resultats, són perfectament comparables amb els resultats obtinguts amb HADDOCK.

A diferència del HADDOCK, aquest programa no permet seleccionar la zona activa de la proteïna fet que dona valor al ClusPro per ser capaç de col·locar d'una manera aproximada el lligand amb el seu receptor. Un fet a destacar és la mesura d'afinitat d'aquestes interaccions que presenten aquestes proteïnes amb els seus lligands proteics. Les energies de totes les interaccions són molt negatives donant a entendre que aquestes interaccions són favorables i que, per tant, permet mostrar l'afinitat d'un lligand pel seu receptor que, en el cas del HADDOCK, no era possible.

- DINÀMICA MOLECULAR

Taula 3. *Resultats de la comparació entre les estructures proteiques predites per Alphafold i determinades després de realitzar una dinàmica molecular respecte a l'estructura de la proteïna determinada experimentalment.*

PROTEÏNA	RMSD ALPHAFOLD - PDB (Å)	RMSD DINÀMICA MOLECULAR- PDB (Å)
EGFR	2.977	4.866
PI3Kα	1.142	1.756
AKT1	4.727	6.711
p53	0.477	0.891
K-Ras	0.439	0.818
Cdk2	0.537	1.033
Cdk4	1.234	1.945
Cdk6	0.954	1.231

Finalment, s'ha comprovat si realitzant una simulació de 50 ns a partir de l'estructura AlphaFold es millorava la predicció de l'estructura de la proteïna respecte a la referència del PDB. A partir dels resultats obtinguts (*taula 3)* fent servir una dinàmica molecular, s'observa que aquesta nova estructura determinada, no millora l'estructura prèvia obtinguda per Alphafold en cap dels casos. De fet els valors de RMSD són, en algunes proteïnes, molt superiors que els de la predicció, per tant, l'estructura obtinguda per AlphaFold genera estructures de proteïnes que s'aproximen, de manera precisa, a les estructures determinades experimentalment.

En tot cas el refinament d'estructures d'AlphaFold amb dinàmica molecular requereix una exploració més extensa.

- APLICACIÓ D'AQUESTA VALIDACIÓ

La verificació i validesa d'aquestes ferramentes són essencials per poder generar una confiança en els resultats obtinguts. Per acabar aquest treball es vol mostrar aquesta importància plasmant-ho en una pràctica d'investigació experimental.

K-RAS G12C és una mutació específica en el gen KRAS que s'ha identificat en diversos tipus de càncer, incloent-hi el càncer de pulmó no microcític, el càncer colorrectal i altres tipus de tumors sòlids. Aquesta mutació en particular afecta la proteïna KRAS, que és un regulador clau de la proliferació cel·lular i la supervivència.

La mutació K-RAS G12C implica un canvi en la seqüència genètica que produeix una forma alterada i activada de la proteïna KRAS. La presència d'aquesta mutació fa que les cèl·lules canceroses siguin més propenses a proliferar i sobreviure, cosa que les torna més resistents a les teràpies convencionals.

Durant molts anys, el gen KRAS ha estat considerat "ineludible" en el desenvolupament de tractaments dirigits contra el càncer a causa de la seva complexa estructura i falta de llocs adequats per al disseny de fàrmacs. No ha sigut fins al novembre del 2020 quan l'Administració d'Aliments i Medicaments dels Estats Units (FDA) va aprovar el primer fàrmac dirigit a K-RAS G12C, anomenat sotorasib (comercialitzat com a Lumakras)[55].

Pel que fa al temps que ha portat trobar cures per a aquesta mutació, ha estat un procés prolongat i complex. El gen KRAS va ser identificat inicialment com a oncogèn en la dècada de 1980, i des d'aleshores ha estat objecte d'intensa investigació. Durant dècades, els científics han estat treballant per comprendre millor la funció de KRAS i desenvolupar estratègies terapèutiques efectives per atacar les cèl·lules canceroses que depenen d'aquesta mutació.

A més del sotorasib, s'estan desenvolupant fàrmacs molt interessants. El més recent, BI-2865[47], un potent inhibidor de diferents mutacions de K-RAS (G12C, G12D o G12V), amb un substituent de propilenglicol i un connector de pirimidina.

La idea d'aquesta secció final del treball és la predicció computacional de l'estructura de la proteïna KRAS amb la mutació G12C i realitzar un acoblament entre aquesta proteïna i aquest últim fàrmac esmentat per veure si es pot predir correctament el lloc d'unió i orientació. Primerament, es realitza un càlcul de docking molecular sense delimitar la zona activa de la proteïna ja que un dels problemes que s'ha vist és la determinació de zones d'unió accessibles a fàrmacs. Una de les zones on s'hauria de col·locar el fàrmac és la mateixa que la que s'ha determinat experimentalment. A partir d'aquest punt, es vol saber la variació que presenta la disposició del fàrmac, de manera computacional, respecte la manera experimental.

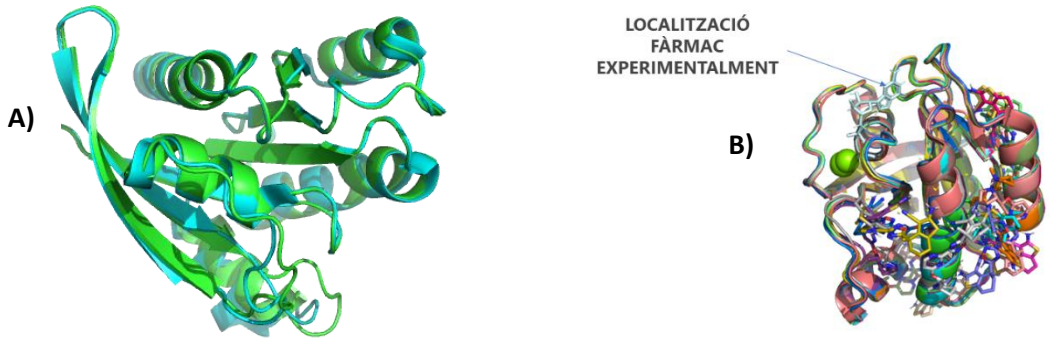


Figura 10. *A) Alineament entre la predicció de l'estructura de la proteïna mutada G12C respecte i aquesta mateixa proteïna determinada experimentalment. B) Resultat de les possibles unions que pot establir el fàrmac BI-2865 segons l'algoritme del programa HADDOCK.*

A partir d'aquesta estructura, s'ha observat una prou bona predicció d'aquesta proteïna mutada respecte a l'experimental, amb petites variacions, sobretot als loops. Com que el fàrmac és un compost químic, s'ha fet servir el HADDOCK per poder realitzar el docking.

A partir del resultats obtingut, s'ha vist que cap disposició d'aquests resultats s'aproxima a la zona d'unió localitzada experimentalment. Aquest resultat demostra la obligació de seleccionar la zona d'unió del fàrmac d'interès. Per aquest motiu, s'ha realitzat un segon docking marcant la zona activa de la proteïna K-RAS mutat.

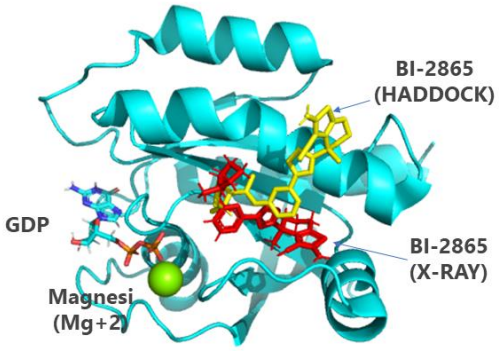


Figura 11.Respresentació de la localització del fàrmac BI-2865 amb un càlcul de docking (HADDOCK) restringint com a lloc d'unió, la zona activa de la proteïna mutada K-RAS G12C.

A partir d'aquest segon resultat, s'ha obtingut una millor simulació però amb una variació en l'orientació del fàrmac.

Per tant, amb aquests resultats, es mostra que es pot realitzar un estudi de la estructura d'una proteïna sabent la seva seqüència, però, per al disseny de fàrmacs, es requereixen coneixements sobre les possibles zones d'unió. Per aquest motiu, es requereix d'una millora en el mètode d'acoblament entre un receptor i el seu lligand, fent èmfasi en els diferents paràmetres i algoritmes utilitzats per desenvolupar aquestes ferramentes bioinformàtiques. En el d'assolir una predicció correcta de la proteïna d'interès i una simulació d'interaccions entre el receptor i el lligand ideals, l'estudi i el disseny de fàrmacs canviaria totalment al que s'ha estat fent fins ara. A partir del descobriment de noves zones d'unió o fins i tot de control al·lostèric, permetria una nova via per poder desenvolupar fàrmacs de manera computacional.

CONCLUSIONS

The field of bioinformatics has made exponential progress in recent years. Thanks to its application, it has been possible to verify and confirm hypotheses in a computational way, offering an alternative to the traditional experimental approach. This allows researchers to decide in advance whether to perform costly experiments or not, based on detailed information provided by bioinformatics. Currently, there is a great interest in this scientific discipline, justified by the increasing development of tools applied to various areas mentioned earlier. Therefore, it is considered necessary to validate these applications and sophisticated systems in order to improve them and generate more accurate and reliable results.

Due to the growing importance of bioinformatics, there is even consideration of introducing it into secondary education. Through research-based activities, students would be able to understand and solve real-world problems using these updated tools, while adapting to the level of biology they present. However, it is essential to validate these tools before using them in different fields such as personalized medicine or education[56].

In this study, the AlphaFold bioinformatics tool has been validated for predicting the structures of cancer therapeutic target proteins and for predicting protein-ligand interactions through molecular docking calculations. The conclusions drawn from this work can be summarized as follows:

- PDB structures of a set of proteins involved in cancer-related signaling pathways have been selected.

- Globally, protein structure predictions from AlphaFold have shown similarities to experimentally determined structures in most cases, making it a valuable tool for predicting 3D structures. However, significant differences have been observed in local regions of the proteins that are relevant for determining their active state. Therefore, AlphaFold tends to obtain the structure of the inactive state of the protein.

- The prediction of protein structure with its protein ligand using AlphaFold shows slight improvement compared to when the protein ligand is not considered. This suggests that AlphaFold can access information about the receptor-ligand binding.

- Regarding the validation of AlphaFold structures for predicting protein-ligand interactions through molecular docking, it can be said that AlphaFold structures show a predictive power similar to those from the PDB. On the other hand, docking programs like HADDOCK and ClusPro simulate protein-ligand interactions relatively accurately depending on the system but with a notable margin for improvement. HADDOCK is considered more versatile as it can handle non-protein molecules in the docking process.

- The addition of a cofactor influences the results of HADDOCK, and these results may vary depending on whether the cofactor is bound to the receptor or the ligand.

- Applying molecular dynamics to the proteins predicted by AlphaFold does not improve the structure or provide a better approximation to experimental protein structures.

Overall, it has been observed that a general protocol cannot be established for using these tools in real cases. Results vary significantly depending on the protein system and the programs used. While these tools are revolutionizing the field of research, improvements are still needed for individual application. Currently, experimental information and bases are still required to effectively use these bioinformatics tools.

BIBLIOGRAFIA

1. Ruff, K. M., & Pappu, R. V. (2021). AlphaFold and Implications for Intrinsically Disordered Proteins. Journal of molecular biology, 433(20), 167208. https://doi.org/10.1016/j.jmb.2021.167208

2. Hongladarom, S. (2006). Ethics of bioinformatics: A convergence between bioethics and computer ethics. Asian biotechnology and development review, 9(1), 37-44.

3. Goodman, K. W., & Cava, A. (2008). Bioethics, business ethics, and science: bioinformatics and the future of healthcare. Cambridge quarterly of healthcare ethics : CQ : the international journal of healthcare ethics committees, 17(4), 361–372. https://doi.org/10.1017/S096318010808050X

4. Bottasso, O., Mendicino, D., Perez, A. R., & Moretti, E. (2021). Bioinformática y bioética. El desafío de complementarlas. Medicina (Buenos Aires), 81(6), 1091-1092.

5. del Val, C., Ruiz, E., Alcalá, R., Fernández, A., Cano, C., Fajardo, W., & Alcala-Fdez, J. Can Bioinformatics close the gender gap in STEM skills?: Reflections from the I Bioinformatics UGR Workshop.

6. CEDEPOF, 2018. European Center for the development of vocational training (CEDEPOF, Eurofound). "Skills forecast: trends and challenges to 2030".

7. Rueda, S., Forte, A., Botella, C., & López-Iñesta, E. (2019). Situación actual de las STEM ('science, technology, engineering and mathematics') y oportunidades de investigación en ómicas, bioinformática, inteligencia artificial y salud desde la perspectiva de género.

8. Way, G. P., Greene, C. S., Carninci, P., Carvalho, B. S., de Hoon, M., Finley, S. D., Gosline, S. J. C., Lê Cao, K. A., Lee, J. S. H., Marchionni, L., Robine, N., Sindi, S. S., Theis, F. J., Yang, J. Y. H., Carpenter, A. E., & Fertig, E. J. (2021). A field guide to cultivating computational biology. *PLoS biology*, *19*(10), e3001419. https://doi.org/10.1371/journal.pbio.3001419

9. Can, T. (2014). Introduction to bioinformatics. *miRNomics: MicroRNA biology and computational analysis*, 51-71.

10. Franco, M. L., Cediel, J. F., & Payán, C. (2008). Breve historia de la bioinformática. *Colombia Médica*, *39*(1), 117-120.

11. Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737-738. https://doi.org/10.1038/171737a0

12. Britten, R. J., & Davidson, E. H. (1969). Gene regulation for higher cells: a theory. *Science (New York, N.Y.)*, *165*(3891), 349–357. https://doi.org/10.1126/science.165.3891.349

13. Horowitz N. H. (1945). On the Evolution of Biochemical Syntheses. Proceedings of the National Academy of Sciences of the United States of America, 31(6), 153–157. https://doi.org/10.1073/pnas.31.6.153

14. Jaskolski, M., Dauter, Z., & Wlodawer, A. (2014). A brief history of macromolecular crystallography, illustrated by a family tree and its N obel fruits. *The FEBS journal*, *281*(18), 3985-4009.

15. Sanger, F., & Thompson, E. O. P. (1953). The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal*, *53*(3), 353.

16. Sanger, F., & Thompson, E. O. P. (1953). The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal*, *53*(3), 366.

17. Edman, P. (1949). A method for the determination of the amino acid sequence in peptides. Arch. Biochem., 22, 475-476.

18. Dayhoff, M. O., & Ledley, R. S. (1962, December). Comprotein: a computer program to aid primary protein structure determination. In *Proceedings of the December 4-6, 1962, fall joint computer conference* (pp. 262-274).

19. Nirenberg, M., & Leder, P. (1964). RNA Codewords and Protein Synthesis: The Effect of Trinucleotides upon the Binding of sRNA to Ribosomes. *Science*, *145*(3639), 1399-1407.

20. Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. Proceedings of the National Academy of Sciences, 74(2), 560-564.

21. Dayhoff, M. O., & Eck, R. V. (Eds.). (1972). Atlas of protein sequence and structure. National Biomedical Research Foundation.

22. Barnetche, J. M. (2007). La bioinformática como herramienta para la investigación en salud humana. *Salud Pública de México*, *49*, 64-66.

23. Peitsch M. C. (1996). ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. Biochemical Society transactions, 24(1), 274–279. https://doi.org/10.1042/bst0240274

24. Mehmood, M. A., Sehar, U., & Ahmad, N. (2014). Use of bioinformatics tools in different spheres of life sciences. *Journal of Data Mining in Genomics & Proteomics*, *5*(2), 1.

24. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

25. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7: 539.

26. Kumar S, Tamura K, Nei M (1994) MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Comput Appl Biosci 10: 189-191.

27. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, et al. (2012) GenBank. Nucleic Acids Res 40: D48-53.

28. UniProt Consortium (2008). The universal protein resource (UniProt). Nucleic acids research, 36(Database issue), D190–D195. https://doi.org/10.1093/nar/gkm895

29. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research, 31(1), 365–370. https://doi.org/10.1093/nar/gkg095

30. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. Nucleic acids research, 28(1), 235–242. https://doi.org/10.1093/nar/28.1.235

31. Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A. K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., … Searle, S. M. (2012). Ensembl 2012. Nucleic acids research, 40(Database issue), D84–D90. https://doi.org/10.1093/nar/gkr991

32. Cordeiro, M. N., & Speck-Planche, A. (2012). Computer-aided drug design, synthesis and evaluation of new anti-cancer drugs. *Current topics in medicinal chemistry*, *12*(24), 2703–2704. https://doi.org/10.2174/1568026611212240001

33. Boruah, L., Das, A., Nainwal, L. M., Agarwal, N., & Shankar, B. (2013). In-Silico Drug Design: A revolutionary approach to change the concept of current Drug Discovery Process. Indian Journal of Pharmaceutical and Biological Research, 1(02), 60.

34. Vassilev, D., Leunissen, J., Atanassov, A., Nenov, A., & Dimov, G. (2005). Application of bioinformatics in plant breeding. *Biotechnology & Biotechnological Equipment*, *19*(sup3), 139-152.

35. Nagai, H., & Kim, Y. H. (2017). Cancer prevention from the perspective of global cancer burden patterns. *Journal of thoracic disease*, *9*(3), 448.

36. Wang, H., Naghavi, M., Allen, C., Barber, R. M., Bhutta, Z. A., Carter, A., ... & Bell, M. L. (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. The lancet, 388(10053), 1459-1544.

37. Seshacharyulu, P., Ponnusamy, M. P., Haridas, D., Jain, M., Ganti, A. K., & Batra, S. K. (2012). Targeting the EGFR signaling pathway in cancer therapy. Expert opinion on therapeutic targets, 16(1), 15–31. https://doi.org/10.1517/14728222.2011.648617

38. Keller, S., & Schmidt, M. H. H. (2017). EGFR and EGFRvIII Promote Angiogenesis and Cell Invasion in Glioblastoma: Combination Therapies for an Effective Treatment. International journal of molecular sciences, 18(6), 1295. https://doi.org/10.3390/ijms18061295.

39. Ogiso, H., Ishitani, R., Nureki, O., Fukai, S., Yamanaka, M., Kim, J. H., Saito, K., Sakamoto, A., Inoue, M., Shirouzu, M., & Yokoyama, S. (2002). Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. Cell, 110(6), 775–787. https://doi.org/10.1016/s0092-8674(02)00963-7

40. Yang, X., Zhang, X., Huang, M., Song, K., Li, X., Huang, M., Meng, L., & Zhang, J. (2017). New Insights into PI3K Inhibitor Design using X-ray Structures of PI3Kα Complexed with a Potent Lead Compound. Scientific reports, 7(1), 14572. https://doi.org/10.1038/s41598-017-15260-5

41. Wu, W. I., Voegtli, W. C., Sturgis, H. L., Dizon, F. P., Vigers, G. P., & Brandhuber, B. J. (2010). Crystal structure of human AKT1 with an allosteric inhibitor reveals a new mode of kinase inhibition. PloS one, 5(9), e12913. https://doi.org/10.1371/journal.pone.0012913

42. Brown, N. R., Noble, M. E., Lawrie, A. M., Morris, M. C., Tunnah, P., Divita, G., Johnson, L. N., & Endicott, J. A. (1999). Effects of phosphorylation of threonine 160 on cyclin-dependent kinase 2 structure and activity. The Journal of biological chemistry, 274(13), 8746–8756. https://doi.org/10.1074/jbc.274.13.8746

43. Takaki, T., Echalier, A., Brown, N. R., Hunt, T., Endicott, J. A., & Noble, M. E. (2009). The structure of CDK4/cyclin D3 has implications for models of CDK activation. Proceedings of the National Academy of Sciences of the United States of America, 106(11), 4171–4176. https://doi.org/10.1073/pnas.0809674106

44. Schulze-Gahmen, U., & Kim, S. H. (2002). Structural basis for CDK6 activation by a virus-encoded cyclin. Nature structural biology, 9(3), 177–181. https://doi.org/10.1038/nsb756

45. Emamzadah, S., Tropia, L., & Halazonetis, T. D. (2011). Crystal structure of a multidomain human p53 tetramer bound to the natural CDKN1A (p21) p53-response element. Molecular cancer research : MCR, 9(11), 1493–1499. https://doi.org/10.1158/1541-7786.MCR-11-0351

46. Hunter, J. C., Gurbani, D., Ficarro, S. B., Carrasco, M. A., Lim, S. M., Choi, H. G., Xie, T., Marto, J. A., Chen, Z., Gray, N. S., & Westover, K. D. (2014). In situ selectivity profiling and crystal structure of SML-8-73-1, an active site inhibitor of oncogenic K-Ras G12C. Proceedings of the National Academy of Sciences of the United States of America, 111(24), 8895–8900. https://doi.org/10.1073/pnas.1404639111

47. Kim, D., Herdeis, L., Rudolph, D., Zhao, Y., Böttcher, J., Vides, A., Ayala-Santos, C. I., Pourfarjam, Y., Cuevas-Navarro, A., Xue, J. Y., Mantoulidis, A., Bröker, J., Wunberg, T., Schaaf, O., Popow, J., Wolkerstorfer, B., Kropatsch, K. G., Qu, R., de Stanchina, E., Sang, B., … Lito, P. (2023). Pan-KRAS inhibitor disables oncogenic signalling and tumour growth. Nature, 10.1038/s41586-023-06123-3. Advance online publication. https://doi.org/10.1038/s41586-023-06123-3

48. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

49. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., … Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic acids research, 50(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

50. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nature methods, 19(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

51. Honorato, R. V., Koukos, P. I., Jiménez-García, B., Tsaregorodtsev, A., Verlato, M., Giachetti, A., Rosato, A., & Bonvin, A. M. J. J. (2021). Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. Frontiers in molecular biosciences, 8, 729513. https://doi.org/10.3389/fmolb.2021.729513

52. Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D., & Vajda, S. (2017). The ClusPro web server for protein-protein docking. Nature protocols, 12(2), 255–278. https://doi.org/10.1038/nprot.2016.169

53. Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., & Ferrin, T. E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein science : a publication of the Protein Society, 30(1), 70–82. https://doi.org/10.1002/pro.3943

54. Rosignoli, S., & Paiardini, A. (2022). Boosting the Full Potential of PyMOL with Structural Biology Plugins. Biomolecules, 12(12), 1764. https://doi.org/10.3390/biom12121764

55. AMG 510 First to Inhibit "Undruggable" KRAS. (2019). Cancer discovery, 9(8), 988–989. https://doi.org/10.1158/2159-8290.CD-NB2019-073

56. Form, D., & Lewitter, F. (2011). Ten simple rules for teaching bioinformatics at the high school level. PLoS computational biology, 7(10), e1002243. https://doi.org/10.1371/journal.pcbi.1002243