



Gáspári Zoltán

■ PPKE ITK

Fehérjeszerkezetek a számítógépből: a 2024-es kémiai Nobel-díj

A 2024-es fizikai és kémiai Nobel-díjak között igen szoros kapcsolat látszik: a fizikai díjakat a ma használt mély neurális hálózatokon alapuló gépi tanulási rendszerek elvi alapjainak lefektetéséért ítelték oda, a kémiai díjak pedig részben egy ebbe a kategóriába tartozó rendszer adott problémára történő sikeres alkalmazásáért adták. Az indoklás szerint *David Baker* a fehérjetervezésben elért, *Demis Hassabis* és *John M. Jumper* pedig a szerkezetpredikció területén felmutatott eredményeivel érdemelte ki a díjat – utóbbi két kutató a mesterséges intelligencián (MI) alapuló AlphaFold két vezető fejlesztője. Fontos ugyanakkor már most leszögezni, hogy egyrészt David Baker tevékenysége a fehérjeszerkezet-predikció területén is korszakalkotónak számít, valamint hogy az AlphaFold és a hozzá hasonló MI-alapú rendszerek jelenleg a fehérjetervezési eljárások szerves részét képezik.

De ne szaladjunk ennyire előre, lássuk először az alapproblémát – már csak azért is, mert 2020-ban az AlphaFold (második, AlphaFold2 verzióját) úgy harangozták be, hogy „megoldotta a biológia egyik 50 éves kihívását”. A fehérjék háromdimenziós szerkezetének pusztán az aminosavsorrendből való becslése valóban a biokémia egyik legrégebbi feladványa. Anfinson kísérletei óta általánosan elfogadott, hogy a fehérjék aminosav-szekvenciája meghatározza a térszerkezetet, sőt, a fehérjemolekulák térszerkezetét stabilizáló kölcsönhatások természetét is ismerjük, mégsem sikerült igen hosszú ideig olyan elméleti – a gyakorlat-

ban számítógépes – eljárást találni, amely képes ezt az összefüggést nagy megbízhatósággal reprodukálni. A probléma fizikokémiai megfogalmazása látszólag egyszerű optimalizációs kérdés: a szekvencia által meghatározott kémiai csoportok között kialakulni képes kölcsönhatási mintázatok közül keressük a legalacsonyabb energiájút, illetve az azt megvalósító konformációt. A kudarc oka többretegű: a fehérjemolekulák óriási konformációs változatossága miatt a megvizsgálandó, fizikailag lehetséges téralkatok száma csillagászati lehet, ráadásul az egyes állapotokban megvalósuló kölcsönhatások száma is igen nagy, ami miatt a gyakorlatban csak közelítő energiaszámítások alkalmazhatóak – így pedig nagyon nehéz a feltételezett globális minimum megtalálása.

Szerencsére a technikai fejlődésnek köszönhetően egyre több fehérjeszerkezetet sikerült kísérleti úton meghatározni, így lehetőség nyílt empirikus összefüggések megfigyelésére. Ma már minden biokémikus tudja, hogy a hasonló aminosav-sorrendű fehérjék háromdimenziós szerkezete is hasonló, sőt, a szerkezet akkor is lehet szinte ugyanolyan, ha a két fehérje evolúciósan rokon, de az aminosavsorrendjeik az évmilliók során nagyon eltérők lettek a mutációk felhalmozódása révén. Az evolúciós rokonság felismerése tehát segíthet a szerkezetbecslésben. Rokonságtól függetlenül is igaz, hogy adott aminosavak, illetve adott aminosav-összetételű rövidebb szakaszok sok esetben rendelkeznek

2024 kémiai Nobel-díjasai (balról jobbra): David Baker, Demis Hassabis és John M. Jumper





valamilyen szerkezeti preferenciával, például egyesek az alfa-helikális, mások a nyújtott béta-redőzött konformációt részesítik előnyben.

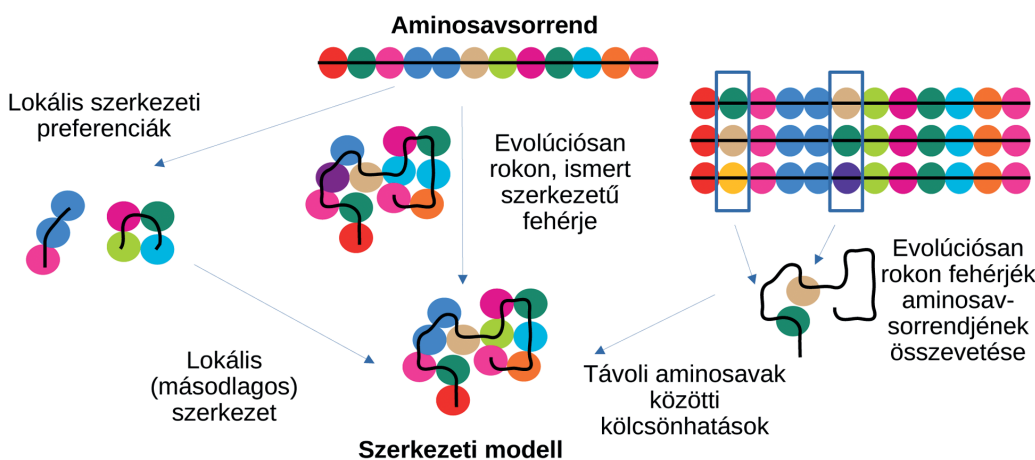
Ezek a megfigyelések már alapot adtak kezdeti predikciós eljárások kidolgozására, melyek értékelésére 1994 óta két évente versenyt is rendeznek az ezzel foglalkozó kutatócsoportok számára CASP (Critical Assessment of protein Structure Prediction) néven. Idén májusban volt szerencsém meghallgatni Andrei Lupas egyik előadását, amelyben elmesélte, hogy az első ilyen megmérettetésen az eredmény annyira csalódást keltő volt, hogy a részt vevő csoportok nagyjából fele egy életre hátat fordított a területnek. Szerencsére újabb csoportok jöttek, a fejlődés lassan, de biztosan elindult, és Lupas professzor szintén ott volt a CASP eseményen 2020 végén, amikor az AlphaFold2 ámulatba ejtette a világot. AlphaFold néven nem ekkor jelent meg módszer először a versenyen, hanem két évvel korábban, 2018-ban. Az első AlphaFold egyébként igen jól teljesített a CASP mezőnyében, különösen annak fényében, hogy a versenyeket akkor már szinte hagyományosan a régóta a területen dolgozó kutatók, köztük David Baker és Yang Zhang kutatócsoportjai dominálták. 2020-ban az alapoktól teljesen átdolgozott, hivatalosan AlphaFold2 nevet viselő módszer szinte letarolta a CASP-ot, alig néhány, a versenyen „szereplő” fehérje esetében maradt alul.

Mire az AlphaFold2 (a továbbiakban AF2) idáig eljutott, a szerkezetbecslésre használt MI-alapú eljárások már évek óta léteztek, és fontos megemlíteni, hogy ezek egy része egy e helyütt korábban nem említett megfontolást is használt: a koevolúciós mintázatot. Egy fehérjeszerkezetben két aminosav által kialakított kölcsönhatás sok esetben meggyengül vagy akár meg is szűnik, ha a kettő közül az egyik aminosavat megváltoztatjuk. De abban az esetben nem feltétlenül, ha a párja is mutálódik! Evolúciós skálán ezért létezik a korrelált mutációk – vagy egymással együtt változó, koevolváló aminosavak – mintázata, ahol a szekvencia két adott pozíciójában történő aminosavcsere között statisztikailag kimutatható kapcsolttság áll fent. Amennyiben megfelelő módon azonosítani tudjuk ezeket a pozíciókat, akkor olyan aminosavakról kapunk listát, amelyek jó eséllyel fizikai közelségben vannak a natív szerkezetben, az ilyen aminosavpárok listája pedig nagyon is jól felhasználható a szerkezetbecslésben.

2020-ra tehát már fejlett, mélytanulással tréningezett rendszerek álltak rendelkezésre, melyek egy adott fehérjeszekvenciát, annak egyszerű fizikokémiai tulajdonságait, valamint a vizsgálttal evolúciósan rokon szekvenciák jellemzőinek megfelelő reprezentációját bemenetként megkapva képesek voltak vagy lokális szerkezeti jellemzők, vagy távoli aminosav-aminosav kontaktusok becslésére (1. ábra). Az első AlphaFold 2018-ban egyébként ezek-

hez igen hasonló alrendszerekkel dolgozott, tehát külön-külön előállított, lokális és globális szerkezeti jellemzőket tartalmazó adatokat generált. Ezeket aztán egy molekuladinamikai számítási lépés „fordította le” valódi 3D atomi koordinátákkal leírt szerkezeti reprezentációra. Megjelentek ugyanakkor az úgynevezett end-to-end predikcióra képes rendszerek is, amelyek a szekvenciát mint bemenetet minimális előfeldolgozási lépések után már neurális hálózattal analizálták, és a hálózat közvetlenül állította elő a kimeneti 3D koordinátákat is.

Az AF2 mindezeket a megfontolásokat tartalmazza, de teljesen újszerű megközelítésben: end-to-end predikciós rendszer, tehát a szerkezet előállítását is a neurális hálózat végzi. Két fő modulból áll, ezek neve EvoFormer és Structure. Az EvoFormer rész felelős a kezdeti reprezentáció kialakításáért, ezt azonban a lokális szerkezet és a kontaktusok tekintetében nem egymástól függetlenül, hanem egymásra épülő iteratív lépések során optimalizálja – és ez kulcsfontosságú! Amikor ugyanis két vagy több szekvenciát összevetünk, illesztünk, akkor nagyon fontos, hogy az egyes egyedi szerkezetekben egymással ekvivalens pozíciókban lévő aminosavak egymás alá kerüljenek. Közeli rokon szekvenciák esetében, ahol csak néhány eltérés van, ez könnyen teljesíthető, de minél több helyen különböznek egymástól, annál nehezebb a feladat – tekintve, hogy magukat a szerkezeteket nem ismerjük, hiszen éppen most akarjuk azokat megbecsülni! Ráadásul a betoldások és deléciók miatt a szekvenciák hossza sem ugyanolyan, néha „ugranunk” kell az egyikben (szakszóval rést [gap] illesztünk az egyik szekvenciába). Sok, egymás alá illesztett rokon szekvencia esetében az egyes szegmensekben található aminosavak kémiai jellege a lokális, másodlagos szerkezetre jellemző információt hordoz, az egyes pozíciók változásainak összefüggéseiben pedig 3D kontaktusok lehetnek kódolva. Az EvoFormer modul ennek a reprezentációnak az optimalizálását végzi el olyan módon, hogy a kétféle szempont, a lokális és a globális szerkezetre vonatkozó információtartalom egyetlen konzisztens mintázatot mutasson. Ennek a lépésnek a jelentőségét a módszer szerzői a 2020-as CASP egyik legnehezebb esetén, a SARS-CoV-2 vírus orf8 nevű fehérjéjén mutatták be. Míg átlagos esetben 48 iteráció (blokk) elegendő volt megfelelő minőségű reprezentáció eléréséhez, az orf8 esetében ennek a négyszeresére, 192 iterációra volt szükség. Ez a gyakorlatban ennyi hálózati blokk összekötését jelenti. Az első AlphaFold esetében egyrészt a lokális és globális információk nem voltak egymással ilyen pontossággal szinkronba hozva, másrészt az a módszer ezeket egy külső programok által előállított reprezentációból, ún. szekvenciaprofilból állította elő. Az AF2 ezeket a lépéseket koncepcióban egyben kezeli, sok energiát fordít a jó megoldás megtalálására, és alapve-



1. ábra. A szerkezetpredikció során használt alapelvek egyszerűsített vázlata. A különböző színű gyöngyök különböző aminosavakat jelölnek. A gyakorlatban a lokális szerkezeti preferenciák elemzésekor is támaszkodunk evolúciósan rokon fehérjék szekvenciáira is



tően ennek köszönhető a sikere. Ez természetesen nagyon komoly számítási teljesítményt is igényel, a módszer részleteinek ismeretése előtt a szakértők értetlenül álltak az előtt, amikor Demis Hassabis elmondta, hogy akár 40(!) GPU is dolgozhatott egy adott fehérjén – ez csak molekuladinamikai számítások esetében tartották volna reálisnak, azt pedig már ekkor lehetett tudni, hogy az AF2 esetében ez nem lehet a magyarázat.

A Structure modul az EvoFormer kimeneteként kapott reprezentációból elkészíti a teljes atomi modellt. Kiindulásként ún. „aminosavgáz” (residue gas) használva, ahol kezdetben minden aminosav egy pontban található. Ha ez még magában nem lenne elég bizarr, a modul a szerkezetépítés során egészen az utolsó lépésekig nem erőlteti az amidkötések sík téralkatának kialakítását. Ez a döntés igencsak elgondolkodtató, különösen annak a történetnek a fényében, amelyet az alfa-hélixek szerkezetére vonatkozó korai elképzelések kapcsán Hargittai István is tárgyal: itt a fizikus Bragg-páros egyes visszaemlékezések szerint azért nem találta a megoldást, mert – a kémikus Linus Paulinggal ellentétben – úgy hitték, hogy a C–N kötés körül is van rotáció. Az AF2 azonban működik úgy is, hogy ezt a feltételt nem követeli meg a szerkezeti modell építésének minden fázisában, és ennek oka, hogy a modell minden más aspektusa így könnyebben felépíthető nagy pontossággal – amely pontosság amúgy is lehetővé teszi, hogy a folyamat végén ezt az optimalizálási lépést biztonsággal meg tudják tenni.

Az AF2 további új tulajdonsága, hogy megtanították neki saját modelljei jószágának becslését is. A tanítási folyamat során, amikor ismert a várt kimenet, kiszámítható a ténylegesen kapott kimenetnek attól való eltérése is – és ezt sikerült megtanítani a hálózaton. Így az AF2 kimenetének része a szerkezeti modell mellett egy lokális jósági becslés minden aminosavra.

Az AF2 egyes esetekben olyan pontosságot ért el, ami még magukat a fejlesztőket is megdöbbenette. Előfordult, hogy egy Zn^{2+} -kötő régió esetében az iont koordináló hisztidin aminosavak térállását szinte tökéletesen becsülte meg, miközben maga a cink-ion nem volt része a szerkezetnek, hiszen az AF2 nem képes nem fehérjetermészetű atomokat kezelni, és nem tanították külön ilyen régiók modellezésére sem. A módszernek ez a magas szintű tel-

jesítőképesége vezetett azokhoz a kijelentésekhez, hogy megoldották a fehérjefeltétkeredés 50 éves problémáját – ami azért nyilvánvalóan túlzás. Az AF2 kiváló eredményeket ad egyedi globuláris domének szerkezetbecslése során, tudnunk kell azonban, hogy nem csodaszer, nem képes minden speciális eset kezelésére – sokszor egyszerűen azért, mert ez nem is volt cél a megalkotása során. Bár sok esetben a lokális szerkezetre jó becslést ad, természeténél fogva nem képes a multimer fibrilláris szerkezetek, például a coiled coil struktúrák felismerésére és modellezésére, emiatt számos, ilyen szerkezetet tartalmazó fehérje esetében hosszú, „semmibe lógó” alfa-helikális régiók találhatók a modellekben (2. ábra). Ugyanígy nehézségei vannak a transzmembrán fehérjék modellezésekor: egyes részek, amelyek esetében ez nem reális, „belóghatnak” a membrán síkjába. Tudnunk kell azt is, hogy az AF2 a modellépítéshez felhasznál ismert szerkezetekből származó fragmenseket, és ezek pontos összetétele befolyásolja a végső modellt. A szerkezetmodellezők kezében az AlphaFold tehát ugyanúgy eszköz, mint minden más szerkezetbecslő módszer, aminek ismerni kell a korlátait.

Mindezzel együtt a tudományos közösség lelkesedése olyan nagy volt, hogy rengeteg fehérje teljes (!) szerkezetét megbecsülték az AlphaFold2-vel, és ezeket különböző adatbázisokban elérhetővé tették. Az AlphaFoldDB több mint 200 millió (!) modellt tartalmaz, köztük az összes ismert emberi fehérjét. Természetesen ezekre is igazak a fent leírt korlátok: a teljes hosszúságú fehérjék esetében a modellek megbízhatósága akár jelentősen más lehet egyes régiókra vonatkozóan.

Az AlphaFold2 logikus továbbfejlesztése volt a több fehérjéből felépülő komplexek modellezése, melyhez először létrehozták az AlphaFold-Multimer módszert – ez az eredeti AF2 viszonylag kisebb mértékű módosítása –, majd 2024 májusában megjelent az AlphaFold3 leírása: az új verzió kifejezetten fehérje–ligandum komplexek, nukleinsavak, valamint kémiaiailag módosított aminosavakat és nukleotidokat tartalmazó szerkezetek modellezésére alkalmas – de ez már másik történet.

Beszéljünk most inkább a Nobel-díj másik feléről! David Baker hosszú utat járt be, és ennek fontos állomása a Rosetta nevű eljárás, mely az évek során igen nagy változásokon ment át. Alap-

2. ábra. Példa egy jó modellre, ami félrevezető lehet. A Homer1 fehérje AlphaFoldDB-ben elérhető modellje (AF-Q86YM7-F1-v4, fent) és az irodalom, valamint az elérhető kísérletes szerkezetek segítségével a kutatócsoportunk által épített előzetes modell (lent).

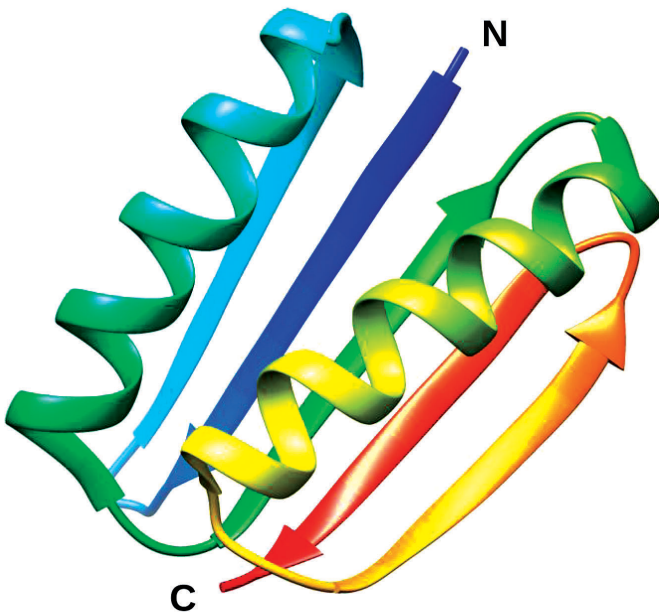
Az AlphaFold2 önmagában nem képes felismerni, hogy a szerkezet multimer, erre nem is tanították, tehát az egymás köré csavarodó alfa-helikális coiled coil szerkezetet nem tudja visszaadni, ahogyan a tetramer szerkezetet sem. Ugyanakkor a helikális szerkezeti preferenciát megfelelően felismeri és modellezi. Az ábra a UCSF Chimera programmal készült





vetően fehérjeszerkezetek predikciójára alkalmas fragmensalapú módszer, de már viszonylag korán felismerték, hogy használható a fehérjetervezésben is. Egy picit itt álljunk meg, mert ez az összefüggés némi magyarázatra szorul. Valójában a szerkezetbecslés és a szerkezettervezés egymás fordított (inverz) problémái: előbbinél egy adott szekvencia által felvehető 3D szerkezetet keresünk, míg utóbbi esetében egy kívánt 3D szerkezethez szeretnénk megtalálni az azt megvalósítani képes szekvenciát. Mindkét feladat megoldásában segít, ha képesek vagyunk olyan kisebb szakaszokat, fragmenseket azonosítani, amelyek adott térszerkezetet preferálnak. Ezekből a fragmensekből a teljes szerkezet felépíthető, összerakható, természetesen a teljes szerkezetben megvalósuló, szekvenciában távoli aminosavak közötti kölcsönhatások figyelembevételével. A gyakorlatban a kapcsolat még ennél is szorosabb: számos esetben a kívánt szerkezethez első körben tervezett szekvenciákat szerkezetbecslő eljárásokon átfuttatva, azok eredményei alapján választják ki azokat, amelyeket a következő körökben újraterveznek – és ezt a folyamatot iteratív módon ismétlik.

A fehérjetervezés egyik mérföldkövének a 2003-ban tervezett, 95 aminosavas „top7” nevű fehérjét tekintik, természetesen ez is a Baker-labor eredménye (3. ábra). Ennek tervezése során egy olyan térszerkezeti sémát kerestek, mely az ismert természetes fehérjékben nem fordul elő. Ezután – a Rosetta programot és egyéb eljárásokat is felhasználva – elvégezték a molekulagerinc és az oldallánc-konformációk geometriai optimalizását egymással összefüggésben, a szekvenciát is változtatva. Ez a gyakorlatban nagyon sok Monte Carlo optimalizációs lépést jelentett, és az első lépésekhez használt kiindulási szekvencia aminosavainak csupán 31%-a maradt változatlan. A puding próbája természetesen az evés: előállították a végső szekvenciát és meghatározták a térszerkezetét, ami kiváló egyezést mutatott a tervezettel.



3. ábra. A tervezett top7 fehérje kísérletileg meghatározott szerkezete (PDB kód: 1QYS). A színskála az N-terminálisról a C-terminális felé változik, a terminálisokat N és C betűk jelölik. Az ábra a UCSF Chimera programmal készült

Az azóta eltelt évek alatt David Baker csoportja számos jelentős előrelépést tett, a kétféle, megfelelően tervezett alegységből felépülő, víruskapszidszerű nanoklaszterektől az allosztérikus át-

menetekre képes fehérjekomplexek létrehozásáig. A terület kétségkívül a felfedező szerkezeti biológiai kutatásokból az alkalmazott bionanotechnológia egyik leginkább látványosan fejlődő ágává nőtte ki magát, de természetesen itt sem beszélhetünk arról, hogy minden egyes aspektus megoldottnak tekinthető. Lásuk, mit gondol minderről maga David Baker, aki egy 2019-es közleményben összefoglalta a legfontosabb addigi tanulságokat!

A fehérjeszerkezet-tervezésben szerzett tapasztalatok szerint a termodinamikai stabilitás elérése a kulcs: kellő stabilitású végállapot mellett nem jellemző, hogy a feltekeredés kinetikailag akadályokba ütközzön. Ráadásul a tervezett szerkezetek tipikusan lényegesen stabilabbak, mint a természetes fehérjék. A tervezés máig egyik legnehezebb aspektusa a megfelelő, feszültségektől mentes gerinckonformáció létrehozása, ami rendszerint a megfelelő helyekre beillesztett flexibilis szakaszok segítségével oldható meg. Érdekes módon a „nagy ugrások” tervezési szempontból egyszerűbbek, mint a kis lépések: teljesen új szerkezetet könnyebb létrehozni, mint célzottan módosítani egy meglévőt. Végül, de nem utolsósorban, a partnerkötés tervezésekor a sikeres megoldások változatossága jóval kisebb jelenleg, mint ami a természetes fehérjékben megfigyelhető – a monomer állapotban stabil, specifikus kötésre képes szerkezeti modulok tervezése egyelőre nem kellően megoldott.

Mindezek a megfigyelések elgondolkodtatóak mind a természetes, funkcionális fehérjék szerveződéséről való tudásunk teljessége, mind a fehérjét stabilizáló kölcsönhatások természetének és azok rendszerének pontos megértése kapcsán. Itt kanyarodhatunk vissza arra az aspektusra, hogy úgy a fehérjeszerkezet-predikció, mint a fehérjetervezés jelentős mértékben támaszkodik az ismert szerkezetekben megfigyelt mintázatokra, azokkal való adott jellegű hasonlóságokra, és messze nincs szó arról, hogy pusztán fizikai és kémiai alapelvek mentén képesek lennénk ezeket a kérdéseket megoldani. A hasonlóságok innovatív felhasználása és fizikai elvekkel való ügyes kombinálása azonban ma már lehetővé teszi ezen problémák jelentős részének gyakorlati megoldását.

Több mint tíz éve, 2013-ban Martin Karplus, Arieh Warshel és Michael Levitt kapták a kémiai Nobel-díjat a fehérjeszerkezetek modellezése terén kifejtett munkásságukért. Ez már egyértelműen jelezte, hogy a fehérjeszerkezet-kutatásba „megérkezett” a számítógépes kémia. A 2024-es kémiai Nobel-díjjal ez a terület mintegy felnőttkorba ért, és mind a természetes, mind a tervezett fehérjeszerkezetek világa minden eddiginél jobban kinyílt a kutatók előtt, új irányokat mutatva az orvosi és egyéb biotechnológiai fejlesztések és alkalmazások számára.

IRODALOM

- Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* (2024) 630, 493–500.
- Baker, D. What has de novo protein design taught us about protein folding and biophysics? *Prot Sci* (2019) 28, 678–683.
- Dobson L. et al. TmAlphaFold database: membrane localization and evaluation of AlphaFold2 predicted alpha-helical transmembrane protein structures. *Nucleic Acids Res* (2023) 51, D517–D522.
- Hargittai, I. Linus Pauling's quest for the structure of proteins. *Struct Chem* (2010) 21, 1–7.
- Hegedűs T. et al. Ins and outs of AlphaFold2 transmembrane protein structure predictions. *Cell Mol Life Sci* (2022) 79, 73.
- Jumper, M. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* (2021) 596, 583–589.
- Kuhlman, B. et al. Design of a novel globular fold with atomic-level accuracy. *Science* (2003) 302, 1364–1368.
- <https://deepmind.google/discover/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology/>
- <https://docs.rosettacommons.org/docs/latest/meta/Rosetta-Timeline>
- <https://moalquarishi.wordpress.com/2020/12/08/alphafold2-casp14-it-feels-like-ones-child-has-left-home/>