

Celebrem el Premi Nobel de Química 2024. La predicció computacional: una revolució real. Mètodes per al disseny i la predicció estructural de proteïnes *Celebrating the 2024 Nobel Prize in Chemistry. Computational prediction: An authentic revolution. Methods for the structural prediction and design of proteins*

Sílvia Osuna

Universitat de Girona. Institut de Química Computacional i Catàlisi (IQCC) i Institució Catalana de Recerca i Estudis Avançats (ICREA)

Resum: Fa aproximadament seixanta-cinc anys la cristal·lografia de raigs X va revelar les primeres estructures terciàries de les proteïnes i evidenciava així la relació entre el plegament i la funció d'aquestes estructures. Des del Premi Nobel d'Anfinsen, el 1972, el gran repte de la biologia estructural ha estat predir l'estructura a partir de la seqüència i, a la inversa, dissenyar seqüències que adoptin una forma determinada. El Premi Nobel de Química del 2024 reconeix David Baker, pels seus mètodes de disseny de proteïnes, i Demis Hassabis i John Jumper, per AlphaFold2 (AF2), una eina capaç de predir l'estructura de les proteïnes amb una precisió elevada, en alguns casos propera a la resolució experimental. Malgrat que AF2 va ser dissenyat per a predir majoritàriament una sola estructura, s'ha vist que, realitzant canvis en els paràmetres inicials, pot proporcionar múltiples estructures, és a dir, diferents conformacions per a una mateixa seqüència d'aminoàcids. El desenvolupament d'aquestes eines basades en la intel·ligència artificial ha promogut el sorgiment i l'aplicació de noves estratègies per al disseny computacional de proteïnes. Tot i els avenços, un dels grans reptes actuals és la predicció computacional de proteïnes amb una elevada activitat catalítica determinada.

Paraules clau: Disseny computacional de les proteïnes, AlphaFold2, intel·ligència artificial, superfície conformacional.

Abstract: *About 65 years ago, X-ray crystallography revealed the first protein tertiary structures, showing the link between the folding and function of such structures. Since Anfinsen's 1972 Nobel Prize, a major challenge of structural biology has been to predict a structure from its sequence — and, conversely, to design sequences that adopt a chosen shape. The 2024 Nobel Prize in Chemistry honored David Baker for his methods of computational protein design and Demis Hassabis and John Jumper for AlphaFold2, a neural network able to predict the folded structure of proteins with high accuracy, in some cases approaching experimental resolution. Although AF2 was built mainly to predict a single structure, researchers have found that by tweaking its initial parameters AF2 can output multiple structures — that is, several conformations for one same amino-acid sequence. These AI-based tools have boosted the application of artificial intelligence-based strategies for computational protein design. Despite the advances, though, one of today's major challenges is the accurate computational prediction of proteins with a specific high catalytic activity.*

Keywords: *Computational protein design, AlphaFold2, artificial intelligence, conformational landscape.*

Introducció

Fa uns seixanta-cinc anys es van determinar per primera vegada, mitjançant cristal·lografia de raigs X, les primeres estructures tridimensionals (3D) de proteïnes [1]. Aquestes primeres estructures van posar de manifest que les cadenes polipeptídiques es pleguen sobre si mateixes mitjançant patrons complexos i ben definits, cosa

que és clau per a la seva funció. Aquest descobriment també va evidenciar que la predicció de l'estructura terciària (3D) d'una proteïna permet predir-ne la funció i les propietats bioquímiques. L'any 1972, Christian Anfinsen va rebre el Premi Nobel de Química pel seu notable descobriment que les estructures 3D de les proteïnes són codificades per la seva seqüència d'aminoàcids a la cadena peptídica [2]. Aquest descobriment va donar lloc al gran repte de la biologia estructural dels darrers cinquanta anys: «predir l'estructura 3D d'una proteïna a partir de la seva seqüència d'aminoàcids».

Correspondència: Sílvia Osuna
Universitat de Girona. Institut de Química Computacional i Catàlisi (IQCC)
i Departament de Química
C. de Maria Aurèlia Capmany, 69. 17003 Girona
Tel.: +34 664 284 535. A/e: silvia.osuna@udg.edu

El problema de la predicció d'estructures terciàries també es pot formular de manera inversa, és a dir, hom pot preguntar-se quines seqüències d'aminoàcids produirien un patró de ple-

gament específic amb una funció determinada. Aquest és el principal objectiu del disseny de proteïnes, un camp en què, donada una estructura objectiu desitjada, s'identifiquen mitjançant mètodes computacionals les seqüències que generarien aquesta estructura amb la funció requerida.

El Premi Nobel de Química del 2024 reconeix els grans avenços assolits per a resoldre aquests dos grans reptes: la predicció d'estructures tridimensionals a partir de la seqüència —reconeixement atorgat a Demis Hassabis i a John Jumper, tots dos de l'empresa Google DeepMind— i la predicció de seqüències a partir de l'estructura —reconeixement concedit al professor David Baker, director de l'Institut de Disseny de Proteïnes de la Universitat de Washington, a Seattle, i investigador de l'Institut Mèdic Howard Hughes (figura 1).

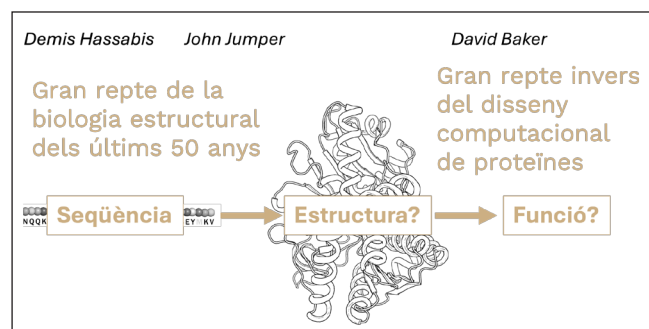


FIGURA 1. El Premi Nobel de Química del 2024 reconeix els avenços realitzats en el repte de predir l'estructura plegada de les proteïnes a partir de la seva seqüència d'aminoàcids i, inversament, en el de predir quina seqüència adoptarà una certa estructura per dur a terme una funció determinada. El premi reconeix la feina del professor David Baker, de la Universitat de Washington a Seattle i de l'Institut Mèdic Howard Hughes, i de Demis Hassabis i John Jumper, de l'empresa Google DeepMind, pel desenvolupament de mètodes computacionals per al disseny i la predicció de l'estructura terciària de les proteïnes. Elaboració pròpia.

Disseny computacional de proteïnes

El 1988, Lynne Regan i William F. DeGrado van construir una proteïna formada per hèlixs α connectades mitjançant bucles que va servir de base per a molts dissenys de proteïnes [3]. I el 1997, Bassil I. Dahiyat i Stephen L. Mayo van publicar el primer disseny computacional d'una proteïna petita [4]. L'objectiu principal era trobar una nova seqüència capaç d'adoptar una determinada estructura, anomenada *dits de zinc* (*zinc-finger*), però sense la presència d'ions metàl·lics. El procés de generació de noves proteïnes amb seqüències no relacionades amb les naturals s'anomena *disseny de proteïnes 'de*

novo'. El gran avanç en aquest camp va arribar el 2003, quan Baker va dissenyar Top7, una nova proteïna de noranta-tres aminoàcids [5]. La clau de l'èxit va ser el programari anomenat *Rosetta*, basat en la combinació de petits fragments de proteïnes no relacionades amb seqüències locals similars de la base de dades d'estructures 3D —el Protein Data Bank (PDB)— i l'optimització conjunta de la seqüència i l'estructura respecte a l'estructura objectiu [6]. El programa Rosetta va ser concebut des del principi tant per a predir l'estructura 3D de proteïnes com per a dissenyar seqüències noves. El 2008, David Baker, Ken Houk i els seus col·laboradors van publicar els primers dissenys d'enzims *de novo* i, tot i que les activitats assolides eren modestes, estudis posteriors van demostrar que aquests dissenys computacionals es podien millorar d'una manera remarcable mitjançant la tècnica experimental de l'evolució dirigida (tècnica per la qual la professora Frances Arnold va rebre el Premi Nobel de Química del 2018) [7]. Al llarg dels anys, el programa Rosetta s'ha anat refinant, amb la incorporació de tècniques d'intel·ligència artificial com a millora més remarcable [8–10]. Gràcies a les eines computacionals desenvolupades, el grup de Baker ha dissenyat proteïnes amb afinitat per a lligands específics [11], nanomaterials autoassemblats que generen estructures altament complexes [12], sensors i interruptors proteics [13, 14] i, fins i tot, proteïnes per a neutralitzar les toxines mortals de determinats verins de serp [15].

Predicció de l'estructura plegada de les proteïnes

Una iniciativa clau en la predicció de l'estructura 3D de les proteïnes és el repte CASP (*critical assessment of protein structure prediction*), organitzat per primera vegada el 1994. Aquest desafiament bianual permet posar a prova els diversos mètodes de predicció d'estructures 3D de les proteïnes, ja que les prediccions es comparen directament amb noves estructures experimentals determinades per cristal·lografia de raigs X o per RMN (ressonància magnètica nuclear). Durant anys el progrés va ser lent, però amb els avenços en la seqüenciació de DNA a l'inici dels noranta, el nombre de seqüències proteiques disponibles també va créixer exponencialment. Aquest fet va permetre alinear i comparar un nombre més gran de seqüències per a una família determinada (alineament múltiple de seqüències, MSA). La mètrica emprada en el CASP per a

avaluar la qualitat dels models generats és el GDT (*global distance test*), que indica el percentatge més alt d'àtoms de carboni α que es troben dins un cert llindar de distància respecte de l'estructura experimental, un cop superposades iterativament ambdues estructures. Durant anys, el GDT es mantingué per sota del 40%, fins al CASP12 (2016). El 2018, el GDT va augmentar fins al 70% gràcies a l'aplicació de mètodes d'aprenentatge profund (o *deep learning*) amb xarxes neuronals convolucionals. En particular, l'empresa DeepMind, fundada per Demis Hassabis, es va presentar amb un programa basat en una xarxa neuronal convolucional que van anomenar *AlphaFold* (AF1) [16]. El 2020, l'equip de DeepMind liderat per Jumper i Hassabis va demostrar l'elevada precisió de la seva nova xarxa neuronal completament redissenyada amb arquitectura *transformer*, *AlphaFold2* (AF2) [17]. En el CASP14, AF2 va aconseguir determinar l'estructura 3D de les proteïnes amb valors de GDT del 90%, considerats comparables a la precisió obtinguda en tècniques experimentals.

L'enginyosa arquitectura d'AF2 de Jumper, Hassabis i els seus col·laboradors pot considerar-se el primer gran avanç científic real de la intel·ligència artificial en biologia. El fet que el codi font d'AF2 fos de domini públic va contribuir decisivament al seu impacte, ja que es va poder provar i validar d'una manera extensiva. I Baker va adoptar ràpidament una arquitectura d'aprenentatge profund semblant en el programa RoseTTAFold (RF) [18].

AF2 s'ha reconegut com una de les fites més importants en la predicció de l'estructura de proteïnes, cosa que ha impulsat l'aplicació i el desenvolupament de múltiples protocols i mètodes d'aprenentatge profund basats en AF2 i en RF per al disseny de proteïnes. I mètodes recents com RFdiffusion [8], ProteinMPNN [10] i aLucinació [9] desenvolupats per Baker han ampliat el camp del disseny de proteïnes per a aplicacions biomèdiques (per exemple, el disseny de vacunes i inhibidors basats en proteïnes), així com de la biologia sintètica [15, 19-21].

Aplicació d'AF2 per a la determinació de múltiples conformacions

Les proteïnes poden adoptar estructures (conformacions) diferents, les quals es poden capturar mitjançant cristal·lografia

de raigs X, raigs X a temperatura ambient i de resolució temporal, criomicroscòpia electrònica (cryo-EM) i per RMN. A més, es poden aplicar diferents tècniques biofísiques i computacionals que proporcionen informació complementària [22]. Aquestes múltiples conformacions d'una mateixa proteïna dipositades al PDB van tenir un paper important tant en l'entrenament d'AF2 com en la seva aplicació per a avaluar l'heterogeneïtat conformacional dels sistemes biològics.

El protocol estàndard d'AF2 requereix la seqüència primària de l'enzim, un MSA generat amb informació de proteïnes evolutivament relacionades i les coordenades 3D d'un petit nombre d'estructures homòlogues anomenades *plantilles* (*templates*). Tot i que AF2 va ser concebut per a predir estructures estàtiques individuals, alguns treballs recents han demostrat que, reduint la profunditat dels MSA d'entrada emprats en l'algoritme d'AF2, es poden generar models precisos en múltiples conformacions diferents [23-25]. Inspirats en aquestes publicacions prèvies que demostren la capacitat d'AF2 de mostrejar estats conformacionals addicionals, vam desenvolupar una aproximació d'AF2 basada en la modificació de les plantilles inicials per tal d'avaluar la varietat conformacional i com aquesta es veu alterada per mutacions a la subunitat β de l'enzim triptòfan sintasa [26-28]. Vam provar l'efecte de reduir el nombre de seqüències proporcionades en l'MSA, però a més vam analitzar com es modifiquen les prediccions d'AF2 quan es faciliten diferents plantilles que mostren múltiples estats conformacionals [27]. Aquestes plantilles provenien tant de conformacions obtingudes mitjançant cristal·lografia de raigs X com de conformacions extretes de simulacions de dinàmica molecular (MD). Amb aquests paràmetres, AF2 va revelar diferències importants en els conjunts d'estructures (conformacions) entre els sistemes analitzats. També vam explorar com executant múltiples simulacions d'MD curtes a partir del conjunt d'estructures d'AF2 i reconstruint les superfícies conformacionals es podien aconseguir estimacions similars a les obtingudes amb simulacions computacionalment més costoses.

Aquesta conclusió és altament engrescadora, ja que demostra el potencial d'AF2, especialment acoblat a simulacions d'MD curtes, per a avaluar de manera ràpida i precisa les diferències conformacionals entre sistemes. Els enfocaments d'aquesta tipologia poden ser de gran utilitat per al desenvolupament de protocols de disseny de proteïnes centrats en els canvis conformacionals clau per a la funció [29, 30]. En aquesta línia, en

el grup hem desenvolupat protocols computacionals basats en MD i eines centrades en l'anàlisi del conjunt d'estructures i moviments correlacionats per a aplicar-los en el camp del disseny racional d'enzims [30, 31]. Mitjançant la nostra eina anomenada SPM (*shortest path map*), hem pogut augmentar l'eficiència catalítica de diferents classes d'enzims i aconseguir activitats fins a 1 400 vegades superiors a la de l'enzim de partida [31–33].

Reptes actuals i perspectives de futur

Tot i els avenços, un dels grans reptes actuals en el disseny computacional de proteïnes és la predicció d'estructures catalíticament funcionals. El disseny racional d'enzims altament eficients mitjançant eines purament computacionals és encara un repte a assolir. La catàlisi enzimàtica és un procés multidimensional complex que requereix la seqüència i l'estructura òptimes per a permetre la unió del(s) substrat(s), a més de catalitzar els passos químics, alliberar el(s) producte(s) i optimitzar les múltiples conformacions necessàries per a desenvolupar la seva funció. Aquesta elevada complexitat fa que la tasca de dissenyar enzims, sobretot orientats a reaccions o substrats no naturals amb altes eficiències, sigui molt complicada.

L'aplicació d'estratègies d'intel·ligència artificial per al disseny computacional d'enzims en qualsevol reacció objectiu i substrat no natural tot just acaba de començar. Durant molts anys, la manca de precisió a l'hora d'incorporar els residus desitjats del centre actiu en estructures proteïques s'ha considerat una de les principals limitacions del procés global. Tanmateix, aquest punt sembla quedar resolt amb el model de difusió RFDifusion basat en RoseTTAFold desenvolupat recentment pel laboratori de Baker [8]. Tot i això, una de les principals limitacions és predir quines estructures són potencialment més adequades per a optimitzar el conjunt conformacional necessari per a la catàlisi que es vol. Tenint en compte els grans avenços que s'han vist aquests darrers temps, especialment en el camp de la predicció d'estructures i en el disseny de proteïnes, la combinació de mètodes d'intel·ligència artificial amb aproximacions basades en la física tindrà un paper clau els pròxims anys per a trobar solucions òptimes per al disseny racional i rutinari d'enzims altament eficients i estables en reaccions i substrats no naturals.

Agraïments

Sílvia Osuna agraeix a la Societat Catalana de Química la seva invitació a parlar del Premi Nobel de Química de l'any 2024.

Referències

- [1] KENDREW, J. C. [et al.]. «A three-dimensional model of the myoglobin molecule obtained by X-ray analysis». *Nature* [en línia], 181 (4610) (1958), p. 662–666. <<https://doi.org/10.1038/181662a0>>.
- [2] ANFINSEN, C. B. «Principles that govern the folding of protein chains». *Science* [en línia], 181 (4096) (1973), p. 223–230. <<https://doi.org/10.1126/science.181.4096.223>>.
- [3] REGAN, L.; DEGRADO, W. F. «Characterization of a helical protein designed from first principles». *Science* [en línia], 241 (4868) (1988), p. 976–978. <<https://doi.org/10.1126/science.3043666>>.
- [4] DAHIYAT, B. I.; MAYO, S. L. «De novo protein design: Fully automated sequence selection». *Science* [en línia], 278 (5335) (1997), p. 82–87. <<https://doi.org/10.1126/science.278.5335.82>>.
- [5] KUHLMAN, B. [et al.]. «Design of a novel globular protein fold with atomic-level accuracy». *Science* [en línia], 302 (5649) (2003), p. 1364–1368. <<https://doi.org/10.1126/science.1089427>>.
- [6] SIMONS, K. T. [et al.]. «Ab initio protein structure prediction of CASP III targets using ROSETTA». *Proteins: Structure, Function, and Bioinformatics* [en línia], 37 (S3) (1999), p. 171–176. <[https://doi.org/10.1002/\(SICI\)1097-0134\(1999\)37:3+<171::AID-PROT21>3.0.CO;2-Z](https://doi.org/10.1002/(SICI)1097-0134(1999)37:3+<171::AID-PROT21>3.0.CO;2-Z)>.
- [7] JIANG, L. [et al.]. «De novo computational design of retroaldol enzymes». *Science* [en línia], 319 (5868) (2008), p. 1387–1391. <<https://doi.org/10.1126/science.1152692>>.
- [8] LISANZA, S. L. [et al.]. «Multistate and functional protein design using RoseTTAFold sequence space diffusion». *Nature Biotechnology* [en línia] (2024). <<https://doi.org/10.1038/s41587-024-02395-w>>.
- [9] ANISHCHENKO, I. [et al.]. «De novo protein design by deep network hallucination». *Nature* [en línia], 600 (7889) (2021), p. 547–552. <<https://doi.org/10.1038/s41586-021-04184-w>>.
- [10] DAUPARAS, J. [et al.]. «Robust deep learning-based protein sequence design using ProteinMPNN». *Science* [en línia], 378 (6615) (2022), p. 49–56. <<https://doi.org/10.1126/science.add2187>>.

- [11] TINBERG, C. E. [et al.]. «Computational design of ligand-binding proteins with high affinity and selectivity». *Nature* [en línia], 501 (7466) (2013), p. 212-216. <<https://doi.org/10.1038/nature12443>>.
- [12] BALE, J. B. [et al.]. «Accurate design of megadalton-scale two-component icosahedral protein complexes». *Science* [en línia], 353 (6297) (2016), p. 389-394. <<https://doi.org/10.1126/science.aaf8818>>.
- [13] BICK, M. J. [et al.]. «Computational design of environmental sensors for the potent opioid fentanyl». *eLife* [en línia], 6 (2017), p. e28909. <<https://doi.org/10.7554/eLife.28909>>.
- [14] LANGAN, R. A. [et al.]. «De novo design of bioactive protein switches». *Nature* [en línia], 572 (7768) (2019), p. 205-210. <<https://doi.org/10.1038/s41586-019-1432-8>>.
- [15] VÁZQUEZ TORRES, S. [et al.]. «De novo designed proteins neutralize lethal snake venom toxins». *Nature* [en línia], 639 (8053) (2025), p. 225-231. <<https://doi.org/10.1038/s41586-024-08393-x>>.
- [16] SENIOR, A. W. [et al.]. «Improved protein structure prediction using potentials from deep learning». *Nature* [en línia], 577 (7792) (2020), p. 706-710. <<https://doi.org/10.1038/s41586-019-1923-7>>.
- [17] JUMPER, J. [et al.]. «Highly accurate protein structure prediction with AlphaFold». *Nature* [en línia], 596 (7873) (2021), p. 583-589. <<https://doi.org/10.1038/s41586-021-03819-2>>.
- [18] BAEK, M. [et al.]. «Accurate prediction of protein structures and interactions using a three-track neural network». *Science* [en línia], 373 (6557) (2021), p. 871-876. <<https://doi.org/10.1126/science.abj8754>>.
- [19] HUANG, B. [et al.]. «Designed endocytosis-inducing proteins degrade targets and amplify signals». *Nature* [en línia] (2024), p. 796-804. <<https://doi.org/10.1038/s41586-024-07948-2>>.
- [20] BERGER, S. [et al.]. «Preclinical proof of principle for orally delivered Th17 antagonist miniproteins». *Cell* [en línia], 187 (16) (2024), p. 4305-4317. <<https://doi.org/10.1016/j.cell.2024.05.052>>.
- [21] EDMAN, N. I. [et al.]. «Modulation of FGF pathway signaling and vascular differentiation using designed oligomeric assemblies». *Cell* [en línia], 187 (14) (2024), p. 3726-3740. <<https://doi.org/10.1016/j.cell.2024.05.025>>.
- [22] BALDWIN, A. J.; KAY, L. E. «NMR spectroscopy brings invisible protein states into focus». *Nature Chemical Biology* [en línia], 5 (11) (2009), p. 808-814. <<https://doi.org/10.1038/nchembio.238>>.
- [23] ALAMO, D. del [et al.]. «Sampling alternative conformational states of transporters and receptors with AlphaFold2». *eLife* [en línia], 11 (2022), p. e75751. <<https://doi.org/10.7554/eLife.75751>>.
- [24] RONEY, J. P.; OVCHINNIKOV, S. «State-of-the-art estimation of protein model accuracy using AlphaFold». *Physical Review Letters* [en línia], 129 (23) (2022), p. 238101. <<https://doi.org/10.1103/PhysRevLett.129.238101>>.
- [25] STEIN, R. A.; MCHAOUB, H. S. «SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with AlphaFold2». *PLOS Computational Biology* [en línia], 18 (8) (2022), p. e1010483. <<https://doi.org/10.1371/journal.pcbi.1010483>>.
- [26] MARIA-SOLANO, M. A.; IGLESIAS-FERNÁNDEZ, J.; OSUNA, S. «Deciphering the allosterically driven conformational ensemble in tryptophan synthase evolution». *Journal of the American Chemical Society* [en línia], 141 (33) (2019), p. 13049-13056. <<https://doi.org/10.1021/jacs.9b03646>>.
- [27] CASADEVALL, G. [et al.]. «Estimating conformational heterogeneity of tryptophan synthase with a template-based AlphaFold2 approach». *Protein Sci.* [en línia], 31 (10) (2022), p. e4426. <<https://doi.org/10.1002/pro.4426>>.
- [28] MARIA-SOLANO, M. A. [et al.]. «In silico identification and experimental validation of distal activity-enhancing mutations in tryptophan synthase». *ACS Catalysis* [en línia], 11 (21) (2021), p. 13733-13743. <<https://doi.org/10.1021/acscatal.1c03950>>.
- [29] MARIA-SOLANO, M. A. [et al.]. «Role of conformational dynamics in the evolution of novel enzyme function». *Chemical Communications* [en línia], 54 (50) (2018), p. 6622-6634. <<https://doi.org/10.1039/C8CC02426J>>.
- [30] OSUNA, S. «The challenge of predicting distal active site mutations in computational enzyme design». *WIREs Computational Molecular Science* [en línia] (2021), p. e1502. <<https://doi.org/10.1002/wcms.1502>>.
- [31] CASADEVALL, G. [et al.]. «The shortest path method (SPM) webserver for computational enzyme design». *Protein Engineering, Design and Selection* [en línia], 37 (2024), p. gzae005. <<https://doi.org/10.1093/protein/gzae005>>.
- [32] CASADEVALL, G. [et al.]. «Designing efficient enzymes: Eight predicted mutations convert a hydroxynitrile lyase into an efficient esterase». *bioRxiv* [en línia] (2023). <<https://doi.org/10.1101/2023.08.23.554512>>.
- [33] DURAN, C. [et al.]. «Altering active-site loop dynamics enhances standalone activity of the tryptophan synthase alpha subunit». *ACS Catalysis* [en línia], 14 (22) (2024), p. 16986-16995. <<https://doi.org/10.1021/acscatal.4c04587>>.



S. Osuna

Sílvia Osuna és doctora en química, professora de la Institució Catalana de Recerca i Estudis Avançats (ICREA) i catedràtica a temps parcial a la Universitat de Girona. La seva recerca està centrada en el desenvolupament de noves eines i de nous protocols computacionals per al disseny racional d'enzims eficients. És autora de més de cent deu publicacions científiques i ha estat reconeguda amb diferents premis de recerca al talent jove d'institucions europees, catalanes i espanyoles.