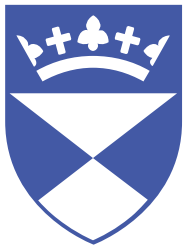

University of Dundee

Computational methods for the characterisation and evaluation of protein-ligand binding sites

Javier Sánchez Utgés

Thesis submitted for the degree of Doctor of Philosophy

Primary Supervisor: Prof Geoffrey J. Barton

Secondary Supervisor: Prof Ulrich Zachariae

Division of Computational Biology

School of Life Sciences

University of Dundee

Dundee, Scotland, UK

February, 2025

Declaration

University of Dundee

School of Life Sciences

I confirm that Javier Sánchez Utgés has carried out the research under my supervision and that he has satisfied all the terms and conditions of the relevant ordinance and regulations of the University of Dundee to qualify to submit this Thesis, entitled *'Computational methods for the characterisation and evaluation of protein-ligand binding sites'*, in application for the degree of Doctor of Philosophy.

Date: 25 February 2025

Research Supervisor: _____

Prof Geoffrey J. Barton

Statement

University of Dundee

School of Life Sciences

I hereby declare that this Thesis is based on results obtained from research which I have personally carried out in the School of Life Sciences at the University of Dundee. I declare that the work described in this Thesis is my own; that I am the author of this Thesis; that it has not previously been put forward in submission for any other degree or qualification; and that I have consulted references herein.

Date: 25 February 2025

Candidate:

Mr Javier Sánchez Utgés

Acknowledgements

This PhD has been quite the journey. Four and a half years that have flown by, but equally feel like such a long time, in which so much happened in both my personal and professional life. I have so many people to thank for their support, without whom I would not be where I am today. I shall start from the beginning.

I first got interested in life sciences during my childhood, at around six years old, whilst living in a small village, called Almedinilla, in the province of Córdoba, Andalucía, Spain. A biologist, whose name I cannot remember, used to take me, school friends and parents on beautiful day trips to explore the local nature. He would tell us all about the plants, lichens, birds, mammals and minerals we could find on these excursions. I was fascinated by the biodiversity of the area and the differences between distinct organisms, their behaviour and adaptation. After moving to Terrassa, a city near Barcelona, Pepita Penalba, my biology school teacher at *Cultura Pràctica* school, would take some students, line us up and ask us questions about what we had been learning about. Students would move along the line depending on the answer to her questions. Accurate answers would move you towards the top of the line, indicating a good notion of the studied subject. I believe this fuelled my competitive spirit and encouraged me to learn and study more, so I could be amongst the top students of the class. Later, in secondary school, my teachers Dr Joel Pascual and Carme Hernández taught me more about physics, chemistry, biology and geology, further increasing my interest in these areas. I found Mendelian genetics particularly interesting and fun, which would lead me later on to study a BSc degree in Genetics at the *Universitat Autònoma de Barcelona* (UAB). During high school at *Institut Montserrat Roig*, my class mate Dr Marc Botifoll got awarded a grant for the *Crazy*

about Biomedicine workshop organised by IRBB. Thanks to this, together with another class mate, Cristina Ortiga, we carried out a project to evaluate potential HIV drugs using computational methods under the supervision of Dr Michela Candotti. This was my first exposure to bioinformatics. During the three years I studied at UAB, I was fortunate to enjoy the lectures of many great readers. Here are some of the ones that really had an impact on my decision to keep on studying after the degree: Profs Antonio Barbadilla, Vicente Martínez, Hafid Laayouni, Alfredo Ruíz, Isaac Salazar, Jesús Piedrafita and Dr Raquel Egea. They taught me about the basis of genetics, animal physiology, biostatistics, population, developmental and quantitative genetics, as well as coding and bioinformatics. It was on the third year of my degree that I realised coding was "my thing" and wet lab was over for me. On that summer, I joined Prof Francesc Calafell's group at PRBB to undertake a 3-month project revolving around population and forensic genetics and R programming. Straight after that, I went on an Erasmus student exchange to Dundee, on the fourth and last year of my degree. During the exchange, I took lectures on Molecular Structure & Interactions with Prof Bill Hunter and Applied Bioinformatics with Dr David Martin, further confirming that bioinformatics was my passion and starting to develop an interest for the structural side of it. The next year, I started an MSc in Bioinformatics for Health Sciences at *Universitat Pompeu Fabra*, in Barcelona. Dr Javier García's lectures on Python programming and Prof Baldo Oliva's on Structural Bioinformatics captivated me, and so I applied for a 1-year internship, part of the MSc, here in the Barton Group, which took place from September 2019 – July 2020. During that time, I secured an EASTBIO DTP studentship to carry out my PhD under Prof Geoff Barton's supervision. I would like to thank all these people that contributed to my education before the beginning of my doctoral training programme, which I have carried out from October 2020 to March 2025.

I would like to thank Prof Geoff Barton, who, I like to believe, saw something in me back in 2019, when I came for an interview without even knowing that STAMP came from the Barton Group, *hahaha*. Thank you so much for the opportunity of joining the Barton Group for that internship project and later on for this doctoral programme. Thanks for the confidence, trust, understanding, patience and flexibility that you have had due

to the various circumstances that have arisen during the length of this project. Thanks for everything that you have taught me in terms of scientific writing, communication, data analysis, project management and other general knowledge such as English phrases, technological applications, broadband network set-up, weather station, how to refurbish windows or wooden floors, as well as your music talent, and countless chats and anecdotes about how science used to be done in the *dark* days when one had to draw graphs by hand, use carbon paper copies and typewriters and paper manuscripts were sent via mail – yes, mail, not *e-mail*. I can say without hesitation that you are the best supervisor I could have had for this PhD. Thanks, Geoff.

I would like to extend my gratitude to another person that has had a massive influence during my PhD: Dr Stuart MacGowan. Stuart first supervised me during my internship, which applied his idea of combining evolutionary divergence with genetic variation to the ankyrin repeat family. Both during the internship and throughout this PhD, Stuart has been a great mentor and shared advice on best practices in data analysis, coding and reproducibility. I also admire his never-ending enthusiasm about science, which is truly contagious, his vision, brilliant research ideas and his readiness to help whenever I have needed it. Thanks, Stuart. Thanks also to the other members, past and present, of the Barton Group, DAG and Jalview team: Mateusz Warowny, Renia Correya, Drs Ben Soares, Carey Metheringham, James Abbot, Jim Procter, Khadija Jabeen, Marek Gierlinski, Matt Parker, Maxim Tsenkov, Michele Tinti, Pete Thorpe and others. It has been a pleasure working with you and having many enjoyable Group Talks, Journal Club sessions and celebration lunches. Thanks also to my secondary supervisor Prof Ulrich Zachariae for his support and for proof-reading a Chapter of this Thesis, along with Drs Ben Soares, Radoslav Krivák, Stuart MacGowan, and Prof Geoff Barton.

To my two Thesis committees: Profs Daan van Aalten, Satpal Virdee, Vicky Cowling and Dr Jorunn Bos for their guidance and feedback during the length of my doctorate. My examiners Profs Alessio Ciuli and David Hoksza, who kindly agreed to read and evaluate this Thesis. Prof Rastko Sknepnek for being convenor of my *viva voce* defence and all other principal investigators in the Computational Biology division: Profs Andrei Pisli-

akov, Ulrich Zachariae, and Drs Gabriele Schweikert, Hajk Drost and Maxim Igaev for making science at such an excellent level, thus elevating this Division, School and University to new heights, making it an ideal destination for cutting-edge research. Thanks also to the IT service of the university for their support of the HPC infrastructure, the work presented in this Thesis was carried out on.

To my past CB PhD colleagues Drs Callum Ives, Dom Gurvik, Marcus Bage, Maxim Tsenkov and Neil Thomson, and present ones: Alp Tegin, Euan MacKay, Peter Ezzat, Rosie Gallagher, Stefan Manolache, Tanmayee Narendra and Yijia Qiang for making the day-to-day work and office routine so comfortable and enjoyable. Special thanks to: Rosie and Carey for their most valued feedback, cakes and dinner parties; Maxim for being such a great supervisor, colleague and friend and having progressed and grown together; Peter for many late evenings in the office and pushing through together. To the wider PhD cohort in the school, PiCLS, my EASTBIO cohort and supervised students. To the postdocs and other staff in the division and school, and the amazing administrative and secretaries team, past and present: Sara Salvaterra, Kirsty Forbes, Jenna Lyons, Paige Nell and Ulla Gingule for their impeccable and efficient work. To the head of postgraduate studies, Prof Carol MacKintosh, who has always been eager and ready to help with a smile on her face. To the EASTBIO DTP administrator, Dr Maria Filippakopoulou, for her kindness and excellent management of the DTP. And, of course to the SLS of the University of Dundee, EASTBIO, BBSRC and UKRI for funding this scholarship.

My deepest gratitude goes also to the amazing musicians and composers Go Shiina, Hans Zimmer, Hiroyuki Sawano, Kohta Yamamoto, Ramin Djawadi, Sofiane Pamart and Yuki Kajiura. Your original sound track and classical music has accompanied me through thousands of hours of exciting work on my research during the last six years. It has filled me with the excitement, strength, determination, sadness, hope and other emotions that you so strongly profess with your beautiful art. Thank you for your magic.

I would too like to thank my dear school friends Ana, Anabella, Cristina, Eric, Lorena and Paulino, my neighbour and friend Rigo and my Dundee friends Alex, Ethan, Karo, Katie, Matt, Maxim, Niamh, Nikita and Sam. Your friendship and emotional support have

been an indispensable pillar during the last six years. You have been the best friends one could wish for and have pulled me up in the darkest of times. I could not be more grateful. To my UPF friends Aina, Altaïr, Luisa and Drs Alexander Gmeiner, Carla Castignani and Pau Badia. You guys are great friends and scientists. We have come a long way since our afternoons at the ruins of UB and our *great* lectures about web design and algorithmics. In a couple of months we will hopefully all be doctors. I am really looking forward to celebrating it together.

To my UAB friends Xavi, Dr Nerea Moreno and my dear *Piñas*: Drs Ferran Garcia, Guillermo Palou, Núria Serna and Sergio Marco. It has been such an honour to share the last decade in academia with you. From the first day of undergrad in Bellaterra in 2014 to the last PhD defence in Dundee in 2025. We have grown and learned so much, and we have done so together. I could not have chosen better companions for this journey. I am so proud and admire every single one of you. I cannot wait to see what the future holds for us. Thanks for being in my life. I love you, guys.

To my Dad, Alfonso, may he rest in peace, my Mother, Alba, and my brothers Héctor and Carlos. Thank you for having raised me as you have, imprinting the values of humility, respect, generosity and perseverance in me, and for always being there for me. To the rest of my family: grandparents, aunts, uncles and cousins for your support, love and the everlasting memories we make when we are together. Specially, to my uncles Alfonso and Paco, my aunts María Elena and Merchi and cousins Alfonso, Blanca, Carmen, Elena and Luis for being the best hosts whilst I worked remotely from your homes. You are always in my heart.

Thanks to Darshan, Hina and Dhyan for welcoming me into your beautiful family and culture, and also for worrying about me and taking care of me. Last, but definitely not least, I would like to thank my amazing partner, Prarthna. I am so thankful to have found you. Thanks for these three years of love, support, advice, patience, faith, encouragement, joy and pure happiness and bliss that you have brought into my life. I can't wait to live whatever comes next *together*. I love you with all my heart.

Agradecimientos

Este doctorado ha sido todo un viaje. Cuatro años y medio que han pasado volando, pero que al mismo tiempo, han parecido tan largos y en los que han ocurrido tantas cosas tanto en mi vida personal como profesional. Tengo muchísimas personas a las que agradecer su apoyo, sin las cuales no estaría donde estoy hoy. Empezaré por el principio.

Mi interés por las ciencias de la vida comenzó en mi infancia, alrededor de los seis años, cuando vivía en un pequeño pueblo llamado Almedinilla, en la provincia de Córdoba, Andalucía, España. Un biólogo, cuyo nombre no recuerdo, solía llevarme a mí, a compañeros de escuela y a nuestros padres a excursiones para explorar la naturaleza local. Nos hablaba sobre las plantas, líquenes, aves, mamíferos y minerales que podíamos encontrar en estas salidas. Me fascinaba la biodiversidad de la zona y las diferencias entre los distintos organismos, su comportamiento y adaptación. Después de mudarme a Terrassa, una ciudad cerca de Barcelona, Pepita Penalba, mi maestra de biología en la escuela *Cultura Pràctica*, solía formar una fila con algunos alumnos y hacernos preguntas sobre lo que habíamos aprendido. Dependiendo de la respuesta, avanzábamos o nos manteníamos en la fila. Las respuestas correctas te acercaban a la parte delantera, indicando un buen dominio del tema. Creo que esto alimentó mi espíritu competitivo y me animó a aprender y estudiar más para estar entre los mejores estudiantes de la clase. Más tarde, en secundaria, mis profesores el Dr. Joel Pascual y Carme Hernández me enseñaron más sobre física, química, biología y geología, aumentando así mi interés por estas áreas. Me parecían especialmente interesantes las leyes de la genética mendeliana, lo que me llevaría posteriormente a estudiar un grado en Genética en la Universidad Autónoma de Barcelona (UAB). Durante el bachillerato en el Instituto Montserrat Roig, mi compañero de clase el Dr. Marc Botifoll

obtuvo una beca para el taller *Crazy about Biomedicine* organizado por el IRBB. Gracias a esto, junto con otra compañera, Cristina Ortiga, llevamos a cabo un proyecto para evaluar posibles fármacos contra el VIH utilizando métodos computacionales, bajo la supervisión de la Dra. Michela Candotti. Este fue mi primer contacto con la bioinformática. Durante los tres años que estudié en la UAB, tuve la suerte de asistir a las clases de grandes docentes. Entre quienes influyeron en mi decisión de seguir estudiando después del grado, quiero destacar a los Profes. Antonio Barbadilla, Vicente Martínez, Hafid Laayouni, Alfredo Ruíz, Isaac Salazar, Jesús Piedrafita y la Dra. Raquel Egea. Me enseñaron sobre las bases de la genética, fisiología animal, bioestadística, genética de poblaciones, del desarrollo y cuantitativa, además de programación y bioinformática. Fue en el tercer año de mi carrera cuando me di cuenta de que programar era "lo mío" y que los experimentos no eran para mí. Ese verano, me uní al grupo del Prof. Francesc Calafell en el PRBB para un proyecto de tres meses sobre genética de poblaciones y forense, trabajando con R. Justo después, realicé un intercambio de estudiantes Erasmus a Dundee durante el cuarto y último año de la carrera. Allí, cursé Estructura Molecular e Interacciones con el Prof. Bill Hunter y Bioinformática Aplicada con el Dr. David Martin, lo que reafirmó mi pasión por la bioinformática y despertó mi interés por su lado estructural. Al año siguiente, comencé un máster en Bioinformática para las Ciencias de la Salud en la Universidad Pompeu Fabra (UPF), en Barcelona. Las clases del Dr. Javier García sobre programación en Python y del Prof. Baldo Oliva sobre Bioinformática Estructural me cautivaron, por lo que solicité una estancia de un año en el *Barton Group* como parte del máster, que realicé de septiembre de 2019 a julio de 2020. Durante ese tiempo, obtuve una beca EASTBIO DTP para llevar a cabo mi doctorado bajo la supervisión del Prof. Geoff Barton. Me gustaría agradecer a todas las personas que contribuyeron a mi formación antes de comenzar este programa de doctorado, que he llevado a cabo de octubre de 2020 a marzo de 2025.

Quiero agradecer especialmente al Prof. Geoff Barton, quien, me gusta pensar, vio algo en mí en 2019, cuando vine a la entrevista sin saber siquiera que STAMP provenía del su grupo, *jajaja*. Muchas gracias por la oportunidad de unirme al *Barton Group* para aquella estancia y, posteriormente, para este doctorado. Gracias por la confianza, la paciencia

y la flexibilidad ante las diversas circunstancias que han surgido durante el transcurso de este proyecto. Gracias por todo lo que me has enseñado en cuanto a redacción científica, comunicación, análisis de datos y otros conocimientos generales, como expresiones en inglés, aplicaciones tecnológicas, cómo configurar una red doméstica, una estación meteorológica, restaurar ventanas o suelos de madera, además de tu talento musical y las incontables charlas y anécdotas sobre cómo se hacía ciencia en los *días oscuros*, cuando había que dibujar los gráficos a mano, usar papel carbón y máquinas de escribir, y los manuscritos se enviaban en papel por correo postal – sí, postal, no *electrónico*. Sin duda alguna, has sido el mejor director de tesis que podría haber tenido. Gracias, Geoff.

También quiero expresar mi gratitud a otra persona que ha tenido una gran influencia durante mi doctorado: el Dr. Stuart MacGowan. Stuart fue mi supervisor en la estancia, en la que aplicamos su idea de combinar la divergencia evolutiva con la variación genética en la familia de repeticiones de anquirinas. Tanto durante esa estancia como a lo largo de mi doctorado, Stuart ha sido un gran mentor y ha compartido consejos sobre buenas prácticas en análisis de datos, programación y reproducibilidad. Admiro su inagotable entusiasmo por la ciencia, su visión, sus brillantes ideas de investigación y su disponibilidad para ayudar siempre que lo he necesitado. Gracias, Stuart. Gracias también a los demás miembros, pasados y presentes, del *Barton Group*, DAG y el equipo de Jalview: Mateusz Warowny, Renia Correya, los Dres. Ben Soares, Carey Metheringham, James Abbot, Jim Procter, Khadija Jabeen, Marek Gierlinski, Matt Parker, Maxim Tsenkov, Michele Tinti, Pete Thorpe y otros. Ha sido un placer trabajar con todos vosotros y disfrutar de numerosas charlas, sesiones de *Journal Club* y comidas de celebración. Gracias también a mi supervisor secundario, el Prof. Ulrich Zachariae, por su apoyo y por revisar uno de los capítulos de esta Tesis, junto a los Dres. Ben Soares, Radoslav Krivák, Stuart MacGowan y el Prof. Geoff Barton.

A mis dos comités de tesis: los Profes. Daan van Aalten, Satpal Virdee, Vicky Cowling y la Dra. Jorunn Bos por su orientación y comentarios durante todo mi doctorado. A mis examinadores, los Profes. Alessio Ciuli y David Hoksza, que amablemente accedieron a leer y evaluar esta Tesis. Al Prof. Rastko Sknepnek por ser el organizador de mi defensa

viva voce y a todos los investigadores principales de la división de Biología Computacional: los Profes. Andrei Pisliakov, Ulrich Zachariae y los Dres. Gabriele Schweikert, Hajk Drost y Maxim Igaev por llevar la ciencia a un nivel tan excelente, elevando así esta División, el Instituto y la Universidad a nuevas alturas, convirtiéndolas en un destino ideal para la investigación de vanguardia. Gracias también al servicio de informática de la universidad por su apoyo a la infraestructura en la que se llevó a cabo este trabajo.

A mis antiguos colegas de doctorado en la división, los Dres. Callum Ives, Dom Gurvik, Marcus Bage, Maxim Tsenkov y Neil Thomson, y a mis compañeros actuales: Alp Tegin, Euan MacKay, Peter Ezzat, Rosie Gallagher, Stefan Manolache, Tanmayee Narendra y Yijia Qiang, por hacer que el día a día y la rutina en la oficina sean tan cómodos y agradables. Un agradecimiento especial para Rosie y Carey por sus valiosos comentarios, pasteles y cenas; Maxim por ser un gran supervisor, colega y amigo, y por haber progresado y crecido juntos; Peter por tantas tardes en la oficina y por seguir adelante juntos. A los compañeros del instittuto, PiCLS, mi cohorte de EASTBIO y los estudiantes a los que he supervisado. A los posdoctorados y al resto del personal de la división e instituto, y al maravilloso equipo administrativo y de secretaría, tanto actual como pasado: Sara Salvaterra, Kirsty Forbes, Jenna Lyons, Paige Nell y Ulla Gingule, por su impecable y eficiente trabajo. A la jefa de estudios de posgrado, la Profa. Carol MacKintosh, siempre dispuesta a ayudar con una sonrisa. A la administradora de EASTBIO, la Dra. Maria Filippakopoulou, por su amabilidad y excelente gestión del programa de doctorado. Y, por supuesto, a la facultad de Biociencias de la Universidad de Dundee, a EASTBIO, BBSRC y UKRI por financiar esta beca.

Mi agradecimiento más profundo también a los increíbles músicos y compositores Go Shiina, Hans Zimmer, Hiroyuki Sawano, Kohta Yamamoto, Ramin Djawadi, Sofiane Pamart y Yuki Kajiura. Vuestra música, tanto bandas sonoras originales como clásica, me ha acompañado durante miles de horas de emocionante trabajo de investigación en los últimos seis años. Me ha llenado de emoción, fuerza, determinación, tristeza, esperanza y otras sensaciones que transmitís con vuestro hermoso arte. Gracias por vuestra magia.

También quiero agradecer a mis queridos amigos de la escuela Ana, Anabella, Cristina, Eric, Lorena y Paulino, a mi vecina y amiga Rigo y a mis amigos de Dundee Alex, Ethan, Karo, Katie, Matt, Maxim, Niamh, Nikita y Sam. Vuestra amistad y apoyo emocional han sido pilares imprescindibles durante estos últimos seis años. Habéis sido los mejores amigos que uno podría desear y me habéis levantado en los momentos más oscuros. No podría estar más agradecido. A mis amigos de la UPF: Aina, Altaïr, Luisa y los Dres. Alexander Gmeiner, Carla Castignani y Pau Badia. Sois grandes amigos y científicos. Hemos recorrido un largo camino desde aquellas tardes en las ruinas de la UB y nuestras *maravillosas* clases de diseño web y algorítmica. En un par de meses, todos seremos doctores. Tengo muchas ganas de celebrarlo juntos.

A mis amigos de la UAB: Xavi, la Dra. Nerea Moreno y mis queridas *Piñas*: los Dres. Ferran Garcia, Guillermo Palou, Núria Serna y Sergio Marco. Ha sido un honor compartir la última década en la academia con vosotros. Desde el primer día del grado en Bellaterra en 2014 hasta la última defensa de doctorado en Dundee en 2025. Hemos crecido y aprendido tanto, y lo hemos hecho juntos. No podría haber elegido mejores compañeros de viaje. Estoy tan orgulloso y os admiro a cada uno de vosotros. Estoy ansioso por ver qué nos depara el futuro. Gracias por estar en mi vida. Os quiero.

A mi padre, Alfonso, que en paz descanse, a mi madre, Alba, y a mis hermanos Héctor y Carlos. Gracias por haberme educado como lo habéis hecho, inculcándome valores de humildad, respeto, generosidad y constancia, y por estar siempre ahí para mí. Al resto de mi familia: abuelos, tíos y primos, por vuestro cariño y por los recuerdos imborrables que creamos juntos. Especialmente, a mis Titos Alfonso y Paco, mis Titas María Elena y Merchi, y mis primos Alfonso, Blanca, Carmen, Elena y Luis, por haber sido los mejores anfitriones mientras teletrabajaba desde vuestras casas. Siempre estáis en mi corazón.

Gracias a Darshan, Hina y Dhyan por acogerme en vuestra preciosa familia y cultura, y también por preocuparos por mí y cuidarme. Por último, quiero dar gracias a mi maravillosa pareja, Prarthna. Estoy tan agradecido de haberte encontrado. Gracias por estos tres años de amor, apoyo, consejo, paciencia, fe, ánimo, alegría y pura felicidad que has traído a mi vida. Deseo vivir lo que venga *juntos*. Te amo con todo mi corazón.

Agraïments

Aquest doctorat ha estat tot un viatge. Quatre anys i mig que han passat volant, però que, alhora, han semblat tan llargs i durant els quals han succeït tantes coses tant en la meva vida personal com professional. Tinc moltes persones a qui agrair el seu suport, sense les quals no estaria on sóc avui. Començaré pel principi.

El meu interès per les ciències de la vida va començar durant la infantesa, al voltant dels sis anys, quan vivia en un petit poble anomenat Almedinilla, a la província de Còrdova, Andalusia, Espanya. Un biòleg, del qual no recordo el nom, solia portar-me a mi, a companys d'escola i als nostres pares d'excursió per explorar la natura local. Ens parlava sobre les plantes, els liquens, ocells, els mamífers i els minerals que podíem trobar en aquelles sortides. Em fascinava la biodiversitat de la zona i les diferències entre els diversos organismes, el seu comportament i adaptació. Després de mudar-me a Terrassa, una ciutat a prop de Barcelona, la Pepita Penalba, mestra de biologia a l'escola Cultura Pràctica, acostumava a formar una fila amb alguns alumnes i a fer-nos preguntes sobre el que havíem après. Segons la resposta, avançàvem o ens manteníem a la fila. Les respostes correctes et feien avançar cap a la part davantera, indicant un bon domini del tema. Crec que això va fomentar el meu esperit competitiu i em va animar a aprendre i estudiar més per estar entre els millors estudiants de la classe. Més endavant, a secundària, els meus professors el Dr. Joel Pascual i la Carme Hernández em van ensenyar més sobre física, química, biologia i geologia, incrementant encara més el meu interès per aquestes àrees. Em resultaven especialment interessants les lleis de la genètica mendeliana, cosa que em portaria posteriorment a estudiar un grau en Genètica a la Universitat Autònoma de Barcelona (UAB). Durant el batxillerat a l'Institut Montserrat Roig, el meu company

de classe el Dr. Marc Botifoll va obtenir una beca per al taller *Crazy about Biomedicine* organitzat per l'IRBB. Gràcies a això, juntament amb una altra companya, la Cristina Ortiga, vam dur a terme un projecte per avaluar possibles fàrmacs contra el VIH fent servir mètodes computacionals, sota la supervisió de la Dra. Michela Candotti. Aquesta va ser la meva primera aproximació a la bioinformàtica. Durant els tres anys que vaig estudiar a la UAB, vaig tenir la sort d'assistir a les classes de grans docents. Entre els que van influir de manera decisiva en la meva tria de continuar estudiant després del grau, vull destacar els Profs. Antonio Barbadilla, Vicente Martínez, Hafid Laayouni, Alfredo Ruíz, Isaac Salazar, Jesús Piedrafita i la Dra. Raquel Egea. Em van ensenyar les bases de la genètica, fisiologia animal, bioestadística, genètica de poblacions, del desenvolupament i quantitativa, a més de programació i bioinformàtica. Va ser al tercer any de la carrera quan em vaig adonar que programar era "el meu camí" i que els experiments al laboratori no eren per a mi. Aquell estiu, em vaig unir al grup del Prof. Francesc Calafell al PRBB per fer un projecte de tres mesos sobre genètica de poblacions i forense, treballant amb R. Tot seguit, vaig fer un intercanvi estudintil d'Erasmus a Dundee durant el quart i darrer any de la carrera. Allà, vaig cursar Estructura Molecular i Interaccions amb el Prof. Bill Hunter i Bioinformàtica Aplicada amb el Dr. David Martin, cosa que va reafirmar la meva passió per la bioinformàtica i va fer créixer el meu interès pel seu vessant estructural. L'any següent, vaig començar un màster en Bioinformàtica per a les Ciències de la Salut a la Universitat Pompeu Fabra (UPF), a Barcelona. Les classes del Dr. Javier García sobre programació en Python i del Prof. Baldo Oliva sobre Bioinformàtica Estructural em van captivar, i per això vaig sol·licitar una estada d'un any al *Barton Group* com a part del màster, que vaig dur a terme de setembre de 2019 a juliol de 2020. Durant aquell període, vaig obtenir una beca EASTBIO DTP per realitzar el meu doctorat sota la supervisió del Prof. Geoff Barton. M'agradaria agrair a totes les persones que van contribuir a la meva formació abans de començar aquest programa de doctorat, que he portat a terme d'octubre de 2020 a març de 2025.

Vull agrair especialment al Prof. Geoff Barton, qui, m'agrada pensar, va veure quelcom en mi al 2019, quan vaig anar a l'entrevista sense ni tan sols saber que STAMP prove-

nia del seu grup, *jajaja*. Moltes gràcies per l'oportunitat d'unir-me al *Barton Group* per aquella estada i, posteriorment, per aquest doctorat. Gràcies per la confiança, la paciència i la flexibilitat davant les diverses circumstàncies que han sorgit durant el transcurs d'aquest projecte. Gràcies per tot el que m'has ensenyat pel que fa a redacció científica, comunicació, anàlisi de dades i altres coneixements generals, com expressions en anglès, aplicacions tecnològiques, com configurar una xarxa domèstica, una estació meteorològica, restaurar finestres o terres de fusta, a més del teu talent musical i les incomptables xerrades i anècdotes sobre com es feia ciència en els *dies foscos*, quan s'havien de dibuixar els gràfics a mà, utilitzar paper carbó i màquines d'escriure, i els manuscrits s'enviaven en paper per correu postal – sí, postal, no *electrònic*. Sens dubte, has estat el millor director de tesi que podria haver tingut. Gràcies, Geoff.

També vull expressar la meva gratitud a una altra persona que ha tingut una gran influència durant el meu doctorat: el Dr. Stuart MacGowan. Durant la meva estada, en Stuart em va supervisar en la tasca d'aplicar la seva idea de combinar la divergència evolutiva amb la variació genètica en la família de repeticions d'ankirines. Tant durant aquella estada com al llarg del doctorat, en Stuart ha estat un gran mentor i ha compartit consells sobre bones pràctiques en l'anàlisi de dades, programació i reproductibilitat. Admiro el seu entusiasme inesgotable per la ciència, la seva visió, les seves idees de recerca brillants i la seva disponibilitat per ajudar sempre que ho he necessitat. Gràcies, Stuart. Gràcies també als altres membres, passats i presents, del *Barton Group*, DAG i l'equip de Jalview: en Mateusz Warowny, la Renia Correya, els Drs. Ben Soares, Carey Metheringham, James Abbot, Jim Procter, Khadija Jabeen, Marek Gierlinski, Matt Parker, Maxim Tsenkov, Michele Tinti, Pete Thorpe i d'altres. Ha estat un plaer treballar amb tots vosaltres i gaudir de nombroses xerrades, sessions de *Journal Club* i dinars de celebració. Gràcies també al meu supervisor secundari, el Prof. Ulrich Zachariae, pel seu suport i per revisar un dels capítols d'aquesta Tesi, juntament amb els Drs. Ben Soares, Radoslav Krivák, Stuart MacGowan i el Prof. Geoff Barton.

Als meus dos comitès de tesi: els Profs. Daan van Aalten, Satpal Virdee, Vicky Cowling i la Dra. Jorunn Bos per la seva orientació i comentaris durant tot el meu doctorat. Als

meus examinadors, els Profs. Alessio Ciuli i David Hoksza, que van acceptar amablement llegir i avaluar aquesta Tesi. Al Prof. Rastko Sknepnek per ser l'organitzador de la meva defensa *viva voce* i a tots els investigadors principals de la divisió de Biologia Computacional: els Profs. Andrei Pisliakov, Ulrich Zachariae i els Drs. Gabriele Schweikert, Hajk Drost i Maxim Igaev per portar la ciència a un nivell excel·lent, elevant així aquesta Divisió, l'Institut i la Universitat a noves altures, convertint-les en una destinació ideal per a la recerca capdavantera. Gràcies també al servei d'informàtica de la universitat pel seu suport a la infraestructura computacional sobre la qual s'ha dut a terme aquest treball.

Als meus antics companys de doctorat en la divisió, els Drs. Callum Ives, Dom Gurvik, Marcus Bage, Maxim Tsenkov i Neil Thomson, i als actuals: l'Alp Tegin, l'Euan MacKay, en Peter Ezzat, la Rosie Gallagher, l'Stefan Manolache, la Tanmayee Narendra i la Yijia Qiang, per fer que el dia a dia i la rutina a l'oficina siguin tan còmodes i agradables. Un agraïment especial per a la Rosie i la Carey pels seus valuosos comentaris, pastissos i sopars; en Maxim per ser un gran supervisor, company i amic, i per haver progressat i crescut junts; en Peter per tantes tardes a l'oficina i seguir endavant plegats. Als companys de l'institut, PiCLS, la meva cohort d'EASTBIO i els estudiants que he supervisat. Als postdoctorands i a la resta de personal de la divisió i l'institut i al meravellós equip administratiu i de secretaria, tant actual com passat: la Sara Salvaterra, la Kirsty Forbes, la Jenna Lyons, la Paige Nell i l'Ulla Gingule, per la seva feina impecable i eficient. A la cap d'estudis de postgrau, la Prof. Carol MacKintosh, sempre disposada a ajudar amb un somriure. A l'administradora d'EASTBIO, la Dra. Maria Filippakopoulou, per la seva amabilitat i excel·lent gestió del programa. I, per descomptat, a la facultat de Biociències de la Universitat de Dundee, a EASTBIO, BBSRC i UKRI per finançar aquesta beca.

El meu agraïment més profund també als increïbles músics i compositors Go Shiina, Hans Zimmer, Hiroyuki Sawano, Kohta Yamamoto, Ramin Djawadi, Sofiane Pamart i Yuki Kajiura. La vostra música, tant bandes sonores originals com clàssica, m'ha acompanyat durant milers d'hores d'apassionant feina d'investigació en els darrers sis anys. M'ha omplert d'emoció, força, determinació, tristesa, esperança i altres sensacions que transmeteu amb el vostre art meravellós. Gràcies per la vostra màgia.

També vull agrair als meus estimats amics de l'escola: l'Ana, l'Anabella, la Cristina, l'Eric, la Lorena i en Paulino, a la meva veïna i amiga Rigo i als meus amics de Dundee: l'Alex, l'Ethan, la Karo, la Katie, en Matt, en Maxim, la Niamh, en Nikita i en Sam. La vostra amistat i suport emocional han estat un pilar indispensable durant aquests últims sis anys. Heu estat els millors amics que algú podria desitjar i m'heu ajudat en els moments més foscos. No podria estar més agraït. Als meus amics de la UPF: l'Aina, l'Altaïr, la Luisa i els Drs. Alexander Gmeiner, Carla Castignani i Pau Badia. Sou grans amics i científics. Hem recorregut un llarg camí des d'aquelles tardes a les runes de la UB i les nostres *meravelloses* classes de disseny web i algorítmica. En un parell de mesos, tots serem doctors. Tinc moltes ganes de celebrar-ho plegats.

Als meus amics de la UAB: en Xavi, la Dra. Nerea Moreno i els meus estimats *Piñas*: els Drs. Ferran Garcia, Guillermo Palou, Núria Serna i Sergio Marco. Ha estat un honor compartir l'última dècada en l'àmbit acadèmic amb vosaltres. Des del primer dia de grau a Bellaterra l'any 2014 fins a l'última defensa de doctorat a Dundee l'any 2025. Hem crescut molt i après molt i ho hem fet junts. No podria haver escollit millors companys de viatge. Estic molt orgullós de vosaltres i us admiro a cadascun. Tinc moltes ganes de veure què ens depara el futur. Gràcies per ser a la meva vida. Us estimo.

Al meu pare, l'Alfonso, que en pau descansi, a la meva mare, l'Alba, i als meus germans, l'Héctor i el Carlos. Gràcies per haver-me educat tal com ho heu fet, inculcant-me valors d'humilitat, respecte, generositat i constància, i per ser sempre al meu costat. A la resta de la meva família: avis, tietes, oncles i cosins, pel vostre suport, estima i pels records inesborrables que creem quan estem junts. Especialment, als meus oncles Alfonso i Paco, a les meves tietes María Elena i Merchi, i als meus cosins l'Alfonso, la Blanca, la Carmen, l'Elena i en Luis per haver estat els millors amfitrions mentre treballava en remot des de casa vostra. Sempre sou al meu cor.

Gràcies a en Darshan, la Hina i la Dhyan per acollir-me a la vostra bonica família i cultura, i també per preocupar-vos per mi i tenir cura de mi. Finalment, però no menys important, vull agrair a la meva meravellosa parella, la Prarthna. Estic molt agraït d'haver-te trobat. Gràcies per aquests tres anys d'amor, suport, consell, paciència, fe, ànim, alegria

i felicitat pura que has portat a la meva vida. Desitjo amb totes les meves forces viure el que vingui *junts*. T'estimo amb tot el meu cor.

Dedication

Para mi Papá. Te echo mucho de menos siempre. Espero que sigas estando orgulloso de mí allá donde estés. Te quiero.

Para mis queridos Tita Carmeli y Tito Paco. Os quiero y os extraño.

Per a la meva molt estimada Padrina Pepita. Et trobo a faltar, però els teus ensenyaments seran amb mi per sempre. T'estimo.

For my Dad. I always miss you so much. I hope you are still proud of me, wherever you may be. I love you.

For my beloved Aunt Carmeli and Uncle Paco. I miss you and love you.

For my much loved Grandma Pepita. I miss you, but your teachings will forever be with me. I love you.

Quote

"I will keep moving forward" – Eren Yeager

Publications

Associated publications

Utgés, J.S., MacGowan, S.A., Ives, C.M., Barton, G.J. Classification of likely functional class for ligand binding sites identified from fragment screening. *Commun. Biol.* **7**, 320 (2024). https://doi.org/10.1038/s42003-024-05970-8.

Utgés, J.S. and Barton, G.J. Comparative evaluation of methods for the prediction of protein-ligand binding sites. *J. Cheminform.* **16**, 126 (2024). https://doi.org/10.1186/s13321-024-00923-z.

Utgés, J.S., MacGowan S.M., Barton G.J. LIGYSIS-web: a resource for the analysis of protein-ligand binding sites. (*Manuscript in preparation*)

Non-associated publications

Utgés J.S., Tsenkov, M.I., Dietrich, N.J.M., MacGowan, S.A., Barton, G.J. Ankyrin repeats in context with human population variation. *PLOS Comput. Biol.* **17**, e1009335 (2021). https://doi.org/10.1371/journal.pcbi.1009335.

Abstract

Fragment screening is used for hit identification in drug discovery, but it is often unclear which binding sites are functionally relevant. Here, data from 37 experiments is analysed. A method to group ligands by binding sites is introduced and sites clustered by their solvent accessibility. This identified 293 ligand sites, grouped into four clusters. C1 includes buried, conserved, missense-depleted sites and is enriched in known functional sites. C4 comprises accessible, divergent, missense-enriched sites and is depleted in functional sites.

This approach is extended to the entire PDB, resulting in the LIGYSIS dataset, accessible through a new web server. LIGYSIS-web hosts a database of 65,000 protein-ligand binding sites across 25,000 proteins. LIGYSIS sites are defined by aggregating unique relevant protein-ligand interfaces across multiple structures. Additionally, users can upload structures for analysis, results visualisation and download. Results are displayed in LIGYSIS-web, a Python Flask web application.

Finally, the human component of LIGYSIS, comprising 6800 binding sites across 2775 proteins, is used to perform the largest benchmark of ligand site prediction to date. Thirteen canonical methods and fifteen novel variants are evaluated using 14 metrics. Additionally, LIGYSIS is compared to datasets such as PDBbind or MOAD and shown to be superior, as it considers non-redundant interfaces across biological assemblies. Re-scored fpocket predictions present the highest recall (60%). The detrimental effect in performance of redundant prediction, as well as the beneficial impact of stronger pocket scoring schemes is demonstrated. To conclude, top-N+2 recall is proposed as the universal benchmark metric and authors encouraged to share their benchmark code for reproducibility.

List of Contents

List of Figures

List of Tables

List of Equations

List of Code Blocks

List of Abbreviations

#

017	Darunavir
0XR	Ethyl caffeate
3D	Three-dimensional
8PR	Paroxetine

A

A	Adenine
AA	Amino acid
ADMET	Absorption, distribution, metabolism, excretion, toxicity
ADP	Adenosine-5’-diphosphate
AFDB	AlphaFold Protein Structure Database
AMI	Adjusted mutual information
ANN	Artificial neural network
AP	Average precision
ARI	Adjusted Rand index
ASA	Accessible surface area
ATP	Adenosine-5’-triphosphate
AUC	Area under the curve

B

BGC	Glucose

BLOSUM	Block substitution matrix

C

C	Cytosine
CA	α-carbon
CC9	Curcumin
CCD	Chemical component dictionary
CHI	Calinski-Harabasz index
CI	Confidence interval
CLA	Chlorophyll A
CLR	Cholesterol
CM	Centre of mass
CMD	Conserved and missense-depleted
CME	Conserved and missense-enriched
CNN	Convolutional neural network
COOH	Carboxyl group
Cryo-EM	Cryogenic electron microscopy

D

DBI	Davies-Bouldin index
DNA	Deoxyribonucleic acid
DBSCAN	Density-based spatial clustering of applications with noise
DCA	Distance to closest ligand atom
DCC	Distance centroid to centroid
DGD	Digalactosyl diacyl glycerol
DNN	Deep neural network
DRN	Deep residual network
DS	Divergence score
DSSP	Define secondary structure of proteins

DWF	Debye-Waller factor

E

EDO	Etylene glycol
EGNN	Equivariant graph neural network
EMA	European Medicines Agency
ENA	European Nucleotide Archive
ESP	Electrostatic potential

F

FAD	Flavin-adenine dinucleotide
FBDD	Fragment-based drug discovery
FDA	Food and Drug Administration
FDR	False discovery rate
FMN	Flavin mononucleotide
FN	False negative
FP	False positive
FPR	False positive rate
FS	Fragment screening
FS	Functional score
FUC	Fucose

G

G	Guanine
GAL	Galactose
GAT	Graph attention network
GNN	Graph neural network
gnomAD	Genome aggregation database
GOL	Glycerol

GPCR	G-protein coupled receptor
GSH	Glutathione
GTP	Guanosine-5′-triphosphate

H

H	Hydrogen
HIV	Human immunodeficiency virus
HAP	HOLO4K-AlphaFold2 paired
HEM	Haem
HOH	Water
HPC	High-performance computing
HTS	High-throughput screening

I

IDR	Intrinsically disordered region
IQR	Interquartile range
I_{rel}	Relative intersection

J

JFP	N-(4-methyl-1,3-thiazol-2-yl)propanamide
JI	Jaccard index

L

LBVS	Ligand-based virtual screening
LGBM	Light gradient boosting machine
LJ	Lennard-Jones

M

MAFFT	Multiple alignment using fast Fourier transform
MAN	Mannose

MaxASA	Maximum accessible surface area
MCC	Matthews correlation coefficient
MCD	Minimum centroid distance
MDS	Multi-dimensional scaling
MES	Missense enrichment score
MHRA	Medicines and Healthcare products Regulatory Agency
ML	Machine learning
MLP	Molecular lipophilicity potential
MLP	Multi-layer perceptron
MOAD	Mother Of All Databases
mRNA	Messenger RNA
MRO	Maximum residue overlap
MSA	Multiple sequence alignment
MTD	Maximum tolerated dose
MUSCLE	Multiple sequence comparison by log-expectation

N

N	Nitrogen
NAD	Nicotinamide-adenine-dinucleotide
NAP	Nicotinamide-adenine-dinucleotide phosphate
NH2	Amino group
NI	No ions
NME	New molecular entity
N-mer	Protein peptide of *N* amino acids
NMR	Nuclear magnetic resonance
NOAEL	No observed adverse effect level
NR	Non-redundant

O

O	Oxygen
OR	Odds ratio

P

PAM	Point accepted mutation
PanDDA	Pan-dataset density analysis
PCA	Principal component analysis
PD	Pharmacodynamics
PDB	Protein Data Bank
PDBe	Protein Data Bank Europe
PDBe-KB	Protein Data Bank Europe Knowledge Base
PID	Peridinin
PK	Pharmacokinetics
PLP	Vitamin B6 phosphate
POVME	Pocket volume measurer
PR	Precision-recall curve
PTM	Post-translational modification

Q

QSAR	Quantitative structure-activity relationship

R

R	Side chain
R_g	Radius of gyration
ROC	Receiver operating characteristic
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
RRO	Relative residue overlap

RVO	Relative volume overlap
RSA	Relative solvent accessibility

S

SAH	S-adenosyl-L-homocysteine
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SAS	Solvent accessible surface
SASA	Solvent accessible surface area
SBVS	Structure-based virtual screening
SD	Standard deviation
SDP	Specificity-determining position
SES	Solvent excluded surface
SNP	Single nucleotide polymorphism
SS	Sum of squares

T

T	Thymine
TN	True negative
TP	True positive
TPR	True positive rate
TRS	Tris buffer

U

U	Uracil
U_D	Distance U
U_{max}	Maximum U statistic
UMD	Unconserved and missense depleted
UME	Unconserved and missense enriched
UPGMA	Unweighted pair-group method with arithmetic mean

UPKB UniProt Knowledgebase

U_{rel} Relative U statistic

V

VDW Van der Waals

VN Virtual node

V_R Volume ratio

VS Virtual screening

VUS Variant of uncertain significance

X

XRC X-ray crystallography

XYP Xylose

Z

ZIKV Zika virus

Chapter 1

Introduction

Preface

This Chapter introduces the fundamental concepts and methodologies essential for understanding this Thesis, along with a review of the state of the art in the field. Additionally, it provides an overview of the Thesis results in their broader scientific context.

The purpose of life

> *"Long, long ago there was a time when nothing but mere matter existed in this world. In the teeming ooze, forms of a certain something appeared, disappeared, and appeared again, and one of them eventually survived. We know it as life. The reason that life ultimately survived was because it was in its nature to multiply. Life took new forms in order to multiply, adapting to every kind of environment, and culminating in us today. Greater numbers, greater diversity, greater abundance. This is why we say that the purpose of life is to multiply."* – Hajime Isayama, *Attack on Titan* [1]

1.1 Central dogma of molecular biology

It is estimated based on geological [2], fossil [3] and phylogenetic [4] analyses that the origin of life in *our* planet Earth dates back to 3.7-4.0 billion years ago. Since then, *life* has not just survived, but adapted and evolved to give raise to a gargantuan estimated biodiversity of 8.7 million eukaryotic species [5] and upward of 1 trillion microbial species [6, 7] with the vast majority of these still to be described [8]. Despite the immense variation across species in terms of reproductive strategies, morphological, metabolic, behavioural traits or ecological niche, there is one thing *all* species have in common: nucleic acids [9]. All living species rely on nucleic acids, mostly deoxyribonucleic acid (DNA), except for some viruses [10] and viroids [11] that use ribonucleic acid (RNA), to store their genetic information. This information flows sequentially from DNA to RNA through the process of transcription and from RNA to protein through translation. This flow of molecular information is known as the *Central Dogma of Molecular Biology* [12, 13].

1.2 The genetic code

DNA is a polymer composed of two polynucleotide chains that coil around each other to form a double helix [14]. These polymer chains are formed by simpler units called nucleotides. Each nucleotide presents a common scaffold formed of a deoxyribose sugar and a phosphate and a variable nitrogen-containing nucleobase. There are four different bases: adenine (A), thymine (T), cytosine (C) and guanine (G). The information stored in DNA gets transferred to RNA through the process of transcription. RNA tends to adopt a single-stranded conformation and is also formed by nucleotides. These nucleotides differ from DNA in that they present a ribose sugar, instead of deoxyribose, and an alternative uracil (U) nucleobase instead of thymine [15].

Messenger RNA (mRNA) corresponds to the sequence of a gene and is read by the ribosomal macromolecular machinery in the process of protein synthesis, or translation. In this process, a peptide chain is formed by linking amino acids in the order specified by

the codons in the mRNA [16]. A codon is a set of three nucleotides that corresponds to one of the twenty canonical amino acids. Figure 1.1 illustrates the equivalence between these codons and the amino acid they encode, also known as the *Genetic Code* [17].

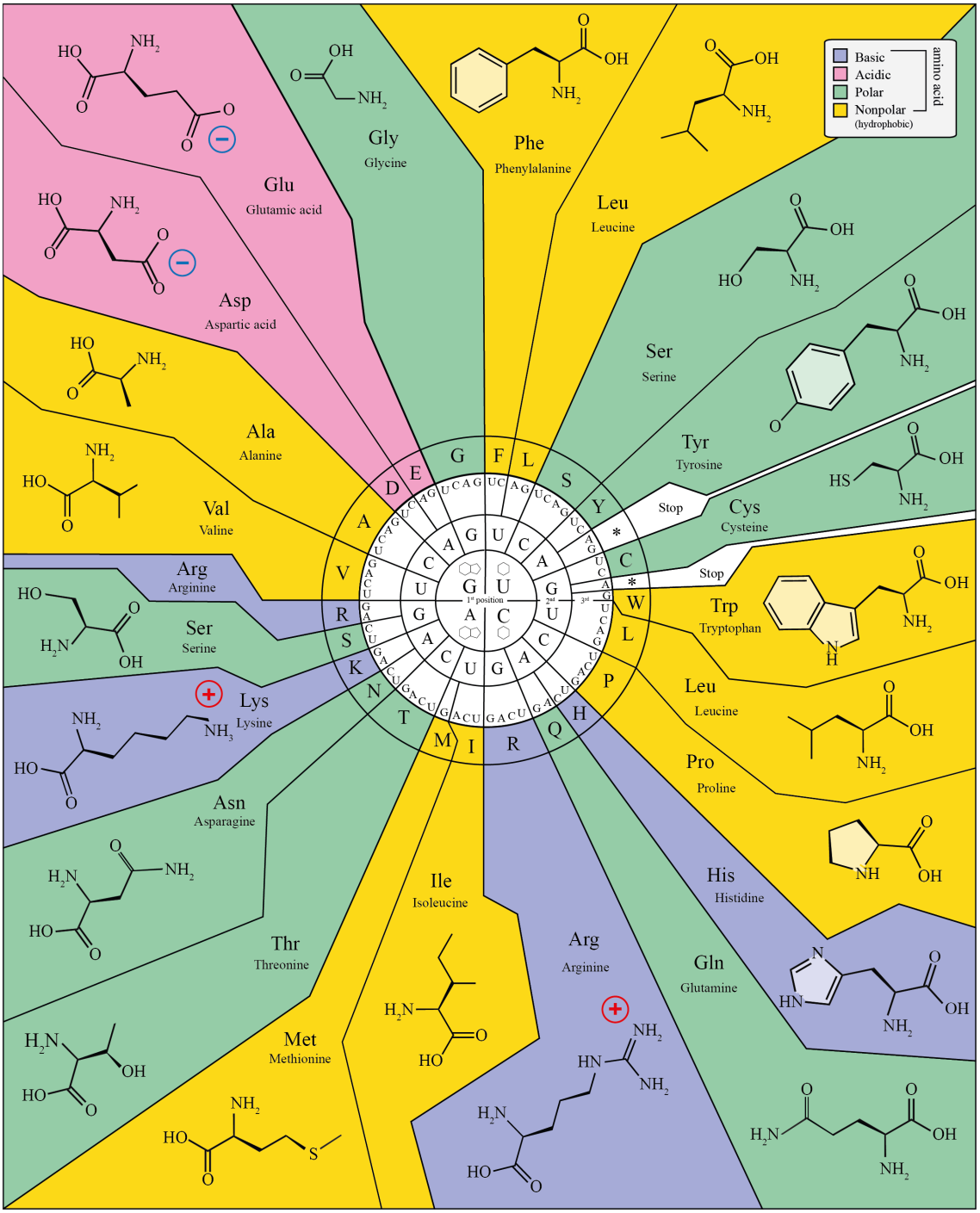

Figure 1.1. Genetic code. The genetic code illustrates the 64 codons resulting from mRNA used to synthetise proteins during translation. The chemical structure of the amino acid side chains is found next to the amino acid name. Colour indicates basic amino acids (lavender), acidic (pink), polar (green) and nonpolar (yellow). STOP codons are coloured in white. Figure adapted from Wikipedia: the free encyclopedia [18].

1.3 Proteins

Proteins are molecular machines that are involved in virtually all cellular processes including cell division, immune response or metabolism. They result from the process of translation of mRNA. Proteins are natural polymers formed of smaller monomers, called amino acids, linked to each other through peptide bonds. Proteins do not carry their function in isolation, but usually interact with other proteins, nucleic acids, ions or small molecules. Chemical compounds can be used as drugs to modulate or inhibit protein function. The amount, localisation, state and interaction with other molecules of proteins is strictly controlled by complex gene regulation networks, signalling cascades, and environment-dependent conformational changes. Any of these mechanisms failing to work correctly can result in under- or overexpression, hypo- or hypermorphism, truncated, mutated or unfolded proteins, which eventually can lead to disease.

1.3.1 Amino acid structure

There are twenty canonical amino acids that are found in all protein sequences. They receive this name because of their chemical structure, which includes both an amino and carboxylic acid functional groups. Figure 1.2 shows the general structure of an amino acid. A carbon atom is found in the centre which binds covalently to four different groups. This carbon is known as α-carbon (CA) and is attached to the amino group (–NH2), the carboxyl group (–COOH), a hydrogen atom (H) and a side chain (R) that differs across the twenty amino acid residues.

1.3.2 Amino acid properties

The different side chains of amino acids confer them distinct physicochemical properties [19]. Figure 1.3 illustrates the ten main properties, the relationship between them and which amino acids present them [20]. The three most important properties setting amino acids apart are hydrophobicity, polarity and size. Hydrophobic residues present side chains that are less soluble in water and therefore tend to be located in the interior protein core,

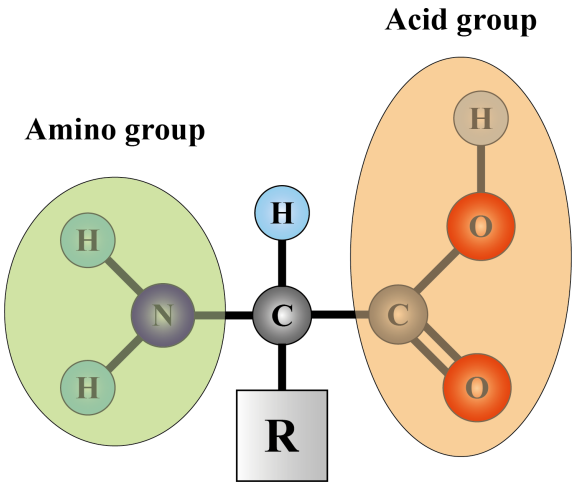

Figure 1.2. Amino acid structure. All twenty amino acids share this common structure formed by the α-carbon (CA) chemically attached to the amino (NH2) and carboxyl (COOH) groups, a hydrogen atom (H) and a side chain (R). The side chain is different and defines the amino acids.

whereas hydrophilic residues are present on the surface [21]. Polar side chains contain electronegative atoms like nitrogen (N) or oxygen (O) and favour interaction with water and other polar molecules. Size is also relevant as there is a large difference in volume between the twenty amino acids, ranging from 60 $Å^3$ (Glycine) to >200 $Å^3$ (Tryptophan). Within the categories defined by these three properties, other subsets can be found as aliphatic, aromatic, positively and negatively charged, tiny or proline. Proline has its own category because of its unique cyclical side chain which links back to the backbone [22].

The physicochemical properties of amino acids are crucial to understand the arrangement of protein atoms in three-dimensional (3D) space. Moreover, the conservation of these properties across evolutionarily related proteins is the basis for sequence comparison, analysis as well as protein structure prediction [25].

1.3.3 Substitution matrices

Similar or identical protein sequences carrying out related functions and displaying a comparable 3D structure can be found within a genome (*paralogous* sequences) and across species (*orthologous* sequences). These proteins are evolutionarily related, i.e., *homologous*, and their origin can be traced back in time to a common ancestor. The comparative analysis of related sequences within protein families provides insight into its evolution-

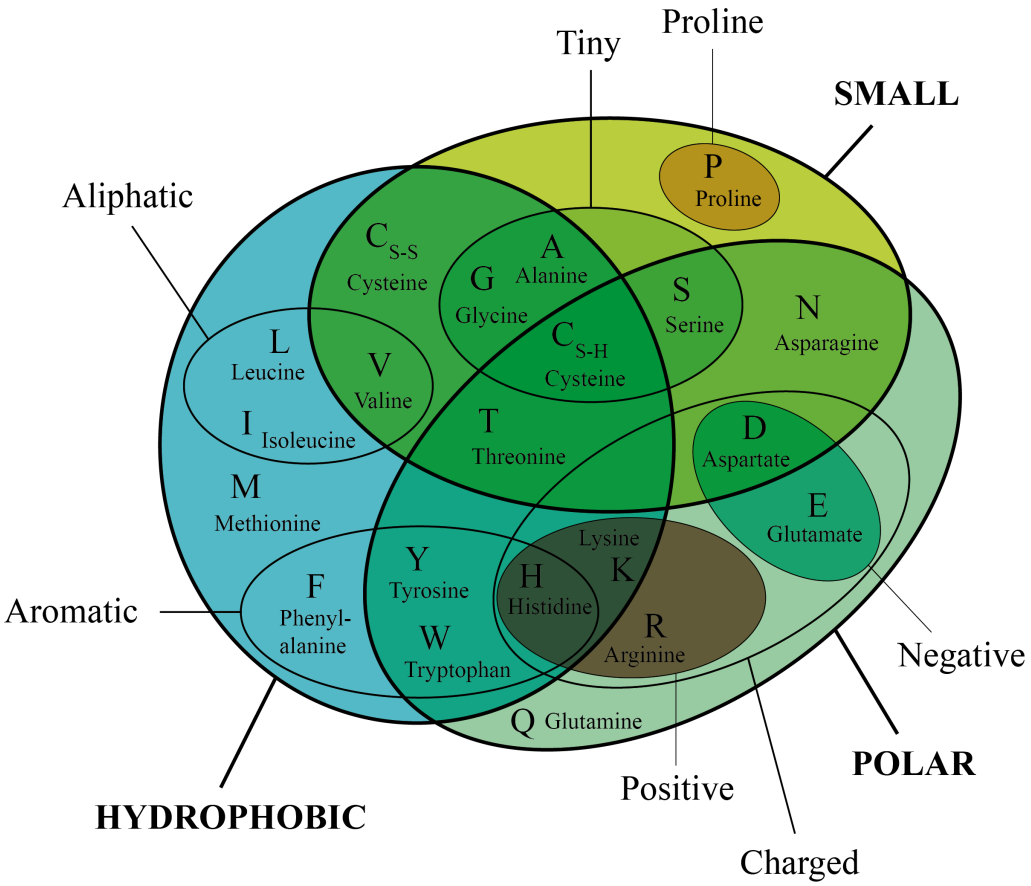

Figure 1.3. Amino acid properties. Taylor Venn diagram illustrating the different physicochemical properties of the twenty proteinogenic amino acids. Figure adapted from the Jalview website [23], which in turn adapted from Livingstone and Barton [24].

ary history [26]. Amino acid substitution matrices can be calculated by quantifying the differences between closely related sequences. These matrices indicate the likelihood of observing transitions at a given protein position between the different amino acids. Transitions between amino acids with similar physicochemical properties, e.g., aspartate $\rightarrow$ glutamate, are less likely to alter the protein structure and are therefore observed with higher frequency. The Point Accepted Mutation (PAM) [27] and Block Substitution Matrix (BLOSUM) [28] are some of the more relevant substitution matrices and serve as a scoring function for constructing alignments of multiple sequences (MSA) that are related in evolution [29].

1.3.4 Multiple sequence alignment

In a multiple sequence alignment, sequences are arranged so that homologous positions – those sharing a common ancestry – appear in the same column across proteins [30]. Through time, sequences diverge and might undergo point mutations, insertions or deletions. To handle this, aligners introduce *gaps* (–) [31]. The distribution of amino acid residues across columns in an alignment reveals conservation patterns. These patterns emerge when residues or their physicochemical properties remain invariant across sequences. Columns that present little or no variation are called *conserved* whereas columns that present a variety of amino acids with different properties are called *unconserved* or *divergent* [32]. Many methods for the alignment of multiple sequences have been developed with Clustal [33–37], MAFFT [38–40] and MUSCLE [41, 42] among the most widely used. Figure 1.4 illustrates an MSA visualised with Jalview, an interactive application for the alignment, editing and integrative analysis of sequence alignments [43].

1.3.5 Amino acid conservation

Amino acid conservation in MSAs is evidence of evolutionary constraint [27]. Throughout evolution, conserved positions have remained fixed due to their functional or structural relevance, while divergent positions accumulate substitutions resulting in variability in amino acid residues across proteins within the same family [45]. Several scores exploring different approaches to quantify conservation have been developed through the years [46]. Some of these scores consider amino acids as symbols and use their relative frequencies [47–49], or entropy [50–52] to score conservation. Others focus on their stereochemical properties [20, 22], use mutation data [53–58] or combine amino acid properties and symbol entropy [59, 60].

The score developed by Shenkin *et al.* [51] is based on Shannon's entropy (S) which is calculated with Equation 1.1 [61]. The proportion within an alignment column of each amino acid i of the $K = 20$ naturally occurring amino acids is denoted by p_i. The Shenkin score, $V_{Shenkin}$, described in Equation 1.2, measures divergence and increases as amino

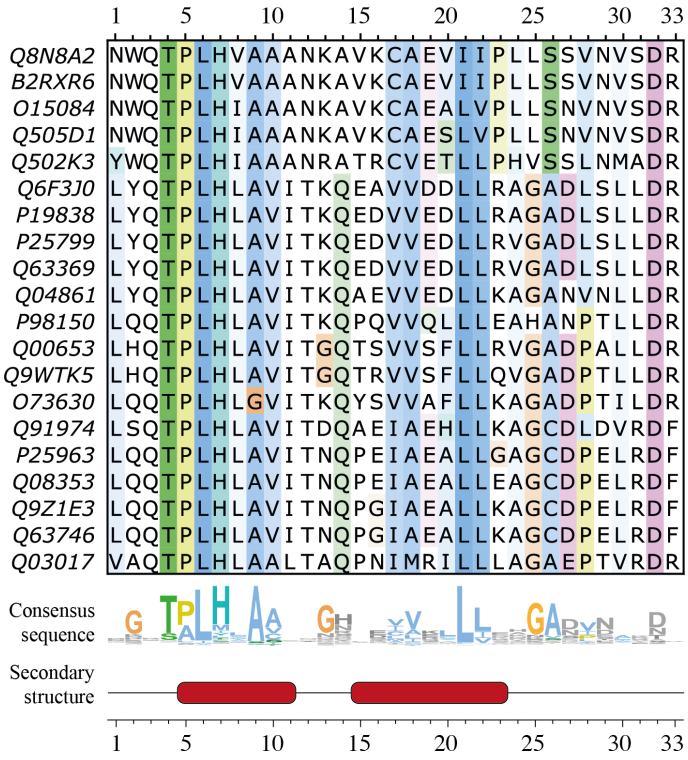

Figure 1.4. Multiple sequence alignment. Fragment of an alignment of 7407 ankyrin repeat protein sequences (IPR002110) built by Utgés *et al.* [44]. The twenty sequences displayed on this figure all have the same length and so no gaps are observed. Alignment columns are coloured in the ClustalX colour scheme [36] with hydrophobic residues in blue, polar in green, glycine in orange, proline in yellow and aromatic in cyan. Additionally, columns are shaded by their conservation, so columns in darker shades are conserved through the alignment whilst those in lighter ones are divergent. The consensus sequence recapitulates the most common residues at each position. Secondary structure assignment describes the two α-helices located at columns 5-11 and 15-23 of the MSA. Strongly conserved residues include the TPLH motif at positions 4-7 as well as Ala9, Ala10, Leu21 and Leu22 which form a series of hydrophobic interactions stabilising the helices within individual repeats as well as across them. Sequence IDs are UniProt accessions numbers. Figure obtained with Jalview [43].

acid variability grows within a column. In a fully conserved position, where all amino acids are the same, entropy is minimum ($S = 0$) and so is divergence ($V_{Shenkin} = 6$). Conversely, in a fully variable position, where all amino acids are equally represented, entropy is maximum ($S \approx 4.32$) and so is divergence ($V_{Shenkin} = 120$). Utgés *et al.* [44] defined a version of this score, $N_{Shenkin}$ (Equation 1.3), which normalises the original score by the minimum and maximum scores within the alignment and ranges 0-100.

$$S = -\sum_{i=1}^{K} p_i \log_2(p_i) \tag{1.1}$$

$$V_{Shenkin} = 2^S \times 6 \tag{1.2}$$

$$N_{Shenkin} = \frac{V_{Shenkin} - V_{Shenkin_{\text{min}}}}{V_{Shenkin_{\text{max}}} - V_{Shenkin_{\text{min}}}} \tag{1.3}$$

Beyond illuminating the evolutionary history of protein sequences, amino acid conservation patterns derived from alignments have been used to successfully predict a variety of features such as secondary structure elements [62], solvent accessibility [63], protein-protein interfaces [64], protein-ligand binding sites [65] and inter-residue contacts [66], which recently has lead to a breakthrough in the prediction of protein 3D structure [67]. There is immense power in the analysis of amino acid conservation, and in this Thesis, the normalised Shenkin divergence score is employed in a systematic manner to rank ligand binding sites on likelihood of function and highlight key residues within them.

1.3.6 Protein structure

The arrangement of protein atoms in three-dimensional space is known as protein structure. Protein structure can be defined at four different levels (Figure 1.5). The primary structure of a protein corresponds to the sequence of amino acids forming the polypeptide chain from the first residue in the amino terminus (N-term) to the last one in the carboxyl terminus (C-term) (Figure 1.5 A). Protein residues adopt local sub-structures by the formation of hydrogen bond interactions between the residue backbone atoms. These local conformations are referred to as secondary structure (Figure 1.5 B). There are two main types of secondary structure: α-helix and β-sheets [68]. The absence of any of these structures could be defined as a third type of secondary structure named coil or loop. Tertiary structure is the 3D arrangement of secondary structure elements within a single chain, resulting from the process of protein folding (Figure 1.5 C). Tertiary structure is defined by the burial of hydrophobic residues in the protein core and hydrogen bonds,

salt bridges and disulfide bonds ensuring a tight packing of residue side chains. Finally, quaternary structure results from the aggregation of two or more individual protein chains that come together to form the functional unit of the protein or multimer (Figure 1.5 D). These monomers are held together by the same interactions that stabilise tertiary structure. Quaternary structure can present different architectures depending on the number of copies involved (e.g., dimer for two, trimer for three, tetramer for four) and whether these copies are from the same sequence (homomers) or different ones (heteromers).

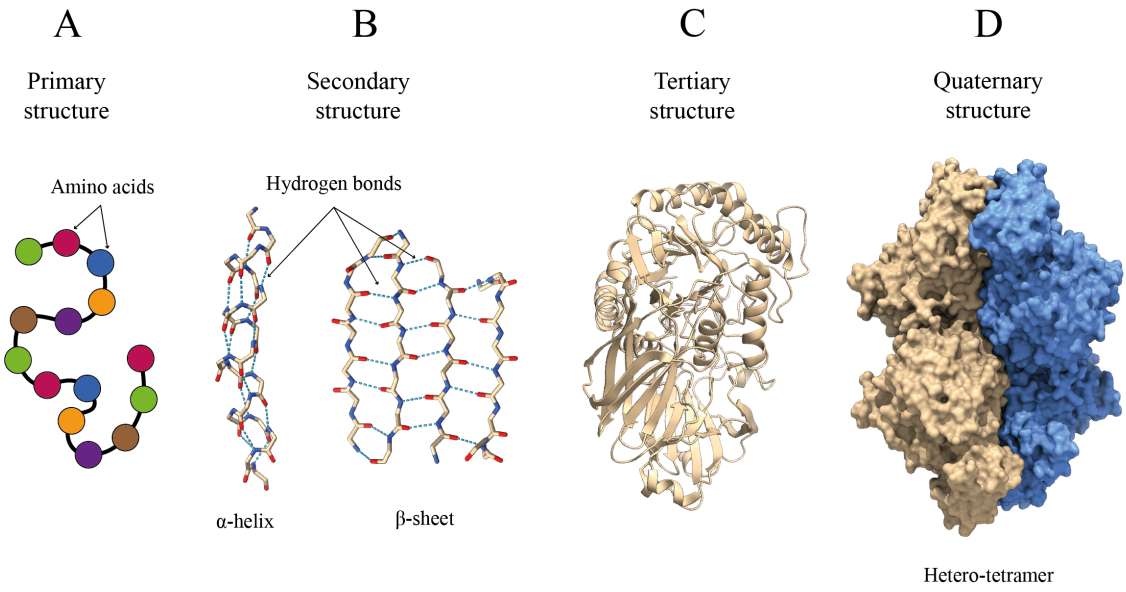

Figure 1.5. Protein structure. Four levels of protein structure: primary (**A**); secondary (**B**); tertiary (**C**); quaternary (**D**). Blue dashed cylinders illustrate hydrogen bonds holding together the secondary structures α-helix and β-sheet. Example is PDB: 8DHV [69] of β-glucuronidase of *Treponema lecithinolyticum* (A0AA82WPE8). Structure visualisation with ChimeraX [70].

1.3.7 Protein structure determination

Protein structure determination is the process of deciphering the arrangement of protein atoms in three-dimensional space. In 1958 Kendrew *et al.* [71] resolved the first protein structure for sperm whale myoglobin (P02185) using X-ray crystallography [72]. Apart from X-ray crystallography, nuclear magnetic resonance spectroscopy and more recently cryogenic electron microscopy have also been used extensively for 3D structure determination.

1.3.7.1 X-ray crystallography

The first step to resolve a protein structure using X-ray crystallography (XRC) is to obtain the protein crystal. A protein crystal is a highly ordered structure in which protein atoms are arranged in a repeating uniformly distributed pattern known as crystal lattice. Crystallising a protein can be very time consuming since the optimal conditions vary between proteins with different size, solubility or isoelectric point. Once the crystal is obtained, it is placed on an X-ray beam which will scatter the electron clouds of the atoms generating a diffraction pattern. This pattern can then be transformed to generate an electron density map revealing the position of atoms within the crystal [73, 74]. Following this, an atomic model is fitted into the electron density map to interpret the molecular structure. X-ray crystallography provides high-resolution structural information and is accordingly the most widely used method to determine protein structure accounting for ≈83% of structures deposited in the Protein Data Bank (PDB) [75].

1.3.7.2 Nuclear magnetic resonance spectroscopy

Nuclear magnetic resonance spectroscopy (NMR) is a powerful technique to determine 3D structure in solution. It was first used in 1984 by Williamson *et al.* [76] to determine the structure of proteinase inhibitor IIA from bull (P01001). NMR does not require a protein crystal, but instead a high concentration of protein in aqueous solution [77]. NMR relies on the magnetic moment or spin of certain isotopes such as ^{1}H, ^{13}C or ^{15}N. In the presence of a magnetic field, the application of radio frequency pulses to these isotopes results in a chemical shift that is diagnostic of their local electronic environment and recorded as the NMR spectrum. The chemical shifts in the spectrum are assigned to individual atoms and distance, angle and orientation restraints are derived. This information is integrated to calculate a model that is then refined to yield the final structure [78]. NMR is ideal to study the dynamics of proteins or other molecules in solution. However, this technique often results in lower structure resolution and its use is limited to smaller proteins as the spectra get more complex with increasing protein size [79].

1.3.7.3 Cryogenic electron microscopy

The use of electron microscopy to determine protein structure dates back to 1975 [80] but modern cryogenic electron microscopy (Cryo-EM) was not used to resolve a protein structure until 1990 when Henderson *et al.* [81] determined the structure of *Halobacterium halobium* bacteriorhodopsin (P02945). In Cryo-EM, proteins are rapidly frozen to very low temperatures to preserve their native state. The frozen sample is then put under an electron microscope which will generate a set of two-dimensional projections from the electron beams. These projections are later integrated into a 3D model. Cryo-EM tends to provide lower resolution than XRC or NMR. However, it is the only method that can determine the structure of large macromolecular complexes such as the spliceosome [82] or the nucleopore [83]. This resolution limitation was breached in the last decade when Bartesaghi *et al.* [84] reached a resolution of 3.2 Å for *Escherichia coli* β-galactosidase (P00722). While XRC has decades of advantage over Cryo-EM in terms of deposited structures, due to the rapid advances in the latter method, it is projected that the number of depositions between these two methods will coalesce by 2035 [85].

1.3.8 Protein structure characterisation

Beyond the determination of the arrangement of atoms in 3D space, proteins can be characterised structurally in multiple ways that offer insight into their physicochemical properties, function, stability, dynamics and interaction with other molecules. These features can then be mapped onto the molecular surface of the proteins and visually analysed (Figure 1.6).

1.3.8.1 Flexibility

The Debye-Waller factor (DWF), or B-factor, measures the attenuation of X-ray scattering caused by thermal motion [87, 88]. This attenuation is a decrease of intensity in diffraction caused by disorder. This disorder can be dynamic and result from the temperature-dependent vibration of the atoms, or static [89]. Accordingly, low values of B-factor

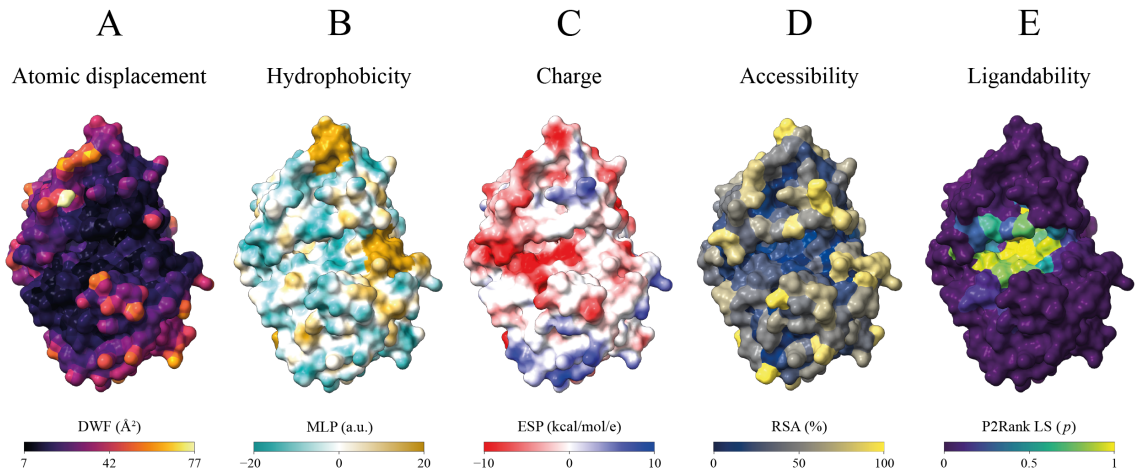

Figure 1.6. Protein structure features. Protein structure features exemplified on PDB: 4C38 [86] of bovine cAMP-dependent protein kinase catalytic subunit alpha (P00517). **(A)** Atomic displacement measured by Debye-Waller factor (DWF); **(B)** Hydrophobicity measured by molecular lipophilicity potential (MLP); **(C)** Charge measured by Coulombic electrostatic potential (ESP); **(D)** Accessibility measured by relative solvent accessibility (RSA); **(E)** Ligandability as measured by P2Rank's predicted ligandability score (LS). Structure visualisation with ChimeraX [70].

indicate rigid or well-ordered protein regions, while high values can identify flexible or dynamic regions in proteins such as loops or binding sites, as well as intrinsically disordered regions (IDR) (Figure 1.6 A). IDRs are protein regions that lack a determined three-dimensional structure and might change conformation depending on their environmental context.

1.3.8.2 Hydrophobicity

The molecular lipophilicity potential (MLP) represents the spatial distribution of lipophilicity across a molecule's surface, providing insight into the hydrophobicity or hydrophilicity of its different regions [90, 91]. High MLP values correspond to lipophilic (hydrophobic) areas and lower MLP values correlate to less lipophilic (more hydrophilic) regions (Figure 1.6 B). The analysis of protein lipophilicity is relevant for the identification of hydrophobic pockets where lipophilic ligands are likely to bind, allosteric sites, large hydrophobic patches prone to protein aggregation, as well as for protein and enzyme engineering [92]. Analysing the lipophilicity of small molecules is also important for optimising ligand design and improving drug absorption, permeability or solubility [93].

1.3.8.3 Charge

Amino acids with charged side chains, e.g., Asp, Glu, His, Lys and Arg, play an important role in the electrostatic potential of a protein, which can be calculated using Coulomb's law [94]. Electrostatic potential plays a pivotal role in the field of protein analysis as it underpins processes such as protein folding, enzyme catalysis, molecular recognition and interaction with proteins, nucleic acids and small molecules, or ligands [95] (Figure 1.6 C). Because of this, protein electrostatics analysis and fine-tuning have applications in protein design [96], protein-ligand binding affinity [97] and biocatalysis optimisation [98].

1.3.8.4 Accessibility

The surface area of a biomolecule that is accessible to solvent is known as accessible surface area (ASA) or solvent-accessible surface area (SASA). ASA was first described by Lee and Richards in 1971 [99] and is usually calculated using the "rolling ball" algorithm described by Shrake and Rupley [100]. The van der Waals (VDW) surface of a molecule is defined by the VDW radii of the atoms forming it [101]. In their algorithm, Shrake and Rupley draw a mesh of points equidistant to each atom on the molecule. These points are typically drawn at a distance of 1.4 Å, emulating the radius of a water molecule, i.e., solvent. By *rolling* this spherical probe over each atom, they established whether a mesh point was exposed to the solvent or buried and calculated the individual contribution of each atom or residue to the ASA of a protein (Figure 1.6 D). The ASA of a molecule is then the path traced by the centre of the spherical probe rolled over the VDW surface. Years later, Richards [102] defined the molecular or solvent-excluded surface (SES) which results from the trajectory of the outer edge of the sphere probe and has two components: the contact surface and the *reentrant* surface. The contact surface is the part of the VDW surface that is in direct contact with the probe. The reentrant surface is an inward-facing or concave surface resulting from the contact of the probe with multiple atoms (Figure 1.7). Conolly was the first to implement algorithms for the analytical calculation of the SES [103, 104].

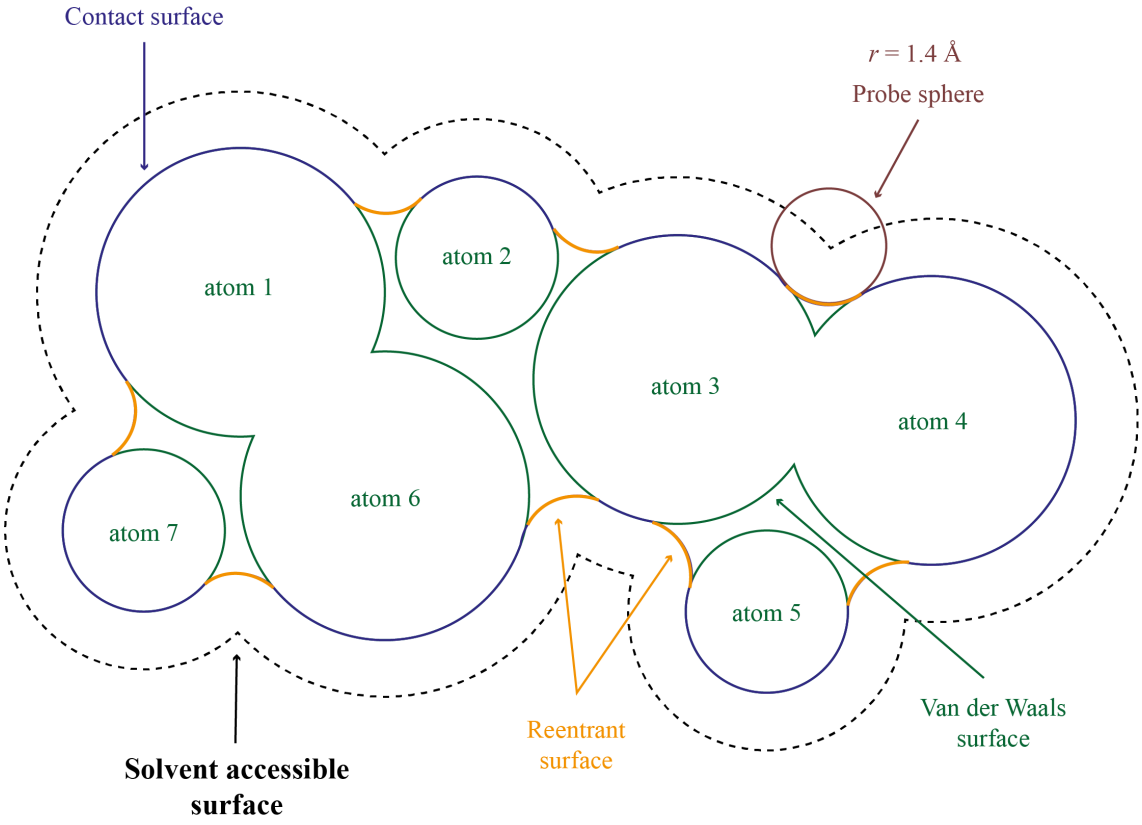

Figure 1.7. Molecular surfaces. Different definitions of the surfaces of a molecule including Van der Waals surface of atoms, solvent accessible surface area defined by Lee and Richards [99] and solvent-excluded surface by Richards [102] with its two components: reentrant and contact surfaces. Adapted from the ChimeraX website [105].

The SASA measured for a given residue in a protein structure is an absolute measure and is not directly comparable across amino acids due to the different size of their side chains. To account for these differences, SASA values are often normalised. There are multiple normalisation scales, all derived from Gly-*X*-Gly tripeptides, where *X* represents each of the twenty amino acids [106–108]. This construct is used because amino acid side chains can adopt an extended conformation and achieve their maximum ASA (MaxASA). A relative solvent accessibility (RSA) can then be obtained by dividing the ASA by the maximum allowed accessible surface area for a given residue as shown in Equation 1.4. As the name indicates, RSA is a measure relative to each side chain and can therefore be used to compare across different amino acids.

$$\text{RSA}(\%) = 100 \times \text{ASA}/\text{MaxASA} \tag{1.4}$$

The analysis of the solvent accessibility landscape of proteins provides rich insight into protein evolution, function, stability and folding. RSA can classify residues into buried and accessible to solvent. Residues with low RSA tend to be buried in the core of the protein and form a network of hydrophobic interactions that ensure the correct packing of the protein [109]. Residues with high RSA are on the surface and interact with solvent and other biomolecules. Accordingly, RSA can be used to identify active sites and hotspot residues that are key contributors to the binding interaction with ligand or protein partners [110]. Active sites tend to have intermediate RSA values (20-50%) since they need to be accessible enough to bind to their substrates but also partially protected to allow for a stable substrate binding and catalysis. Additionally, solvent accessibility is known to be correlated with evolutionary conservation [111]. Residues buried in the hydrophobic core are conserved through evolution as mutations in them have a destabilising effect and can lead to protein misfolding and aggregation.

In Chapter 2 and Chapter 3, ASA is calculated with DSSP [112] and normalised using the method of Tien *et al.* [108] to characterise the solvent accessibility profile of ligand binding sites and predict their likelihood of function.

1.3.8.5 Ligandability

Ligandability is the conceptual feature aiming to chatacterise the ability of a residue, set of residues or protein to bind a small molecule or ligand. Ligands play a critical role in protein function acting as natural co-factors, substrates, inhibitors and drugs in disease therapy (Figure 1.8). Identifying where ligands can bind to proteins is therefore of critical importance in understanding and modulating protein function. While X-ray crystallography remains the gold-standard to identify and characterise binding sites [113], over the last three decades, significant effort has been made to develop computational methods that predict binding sites from an apo three-dimensional protein structure [114]. Ligandability

can be predicred by methods like P2Rank [115] as the probability of an atom or residue having the ability to bind a ligand, and visualised on a protein surface (Figure 1.6 E).

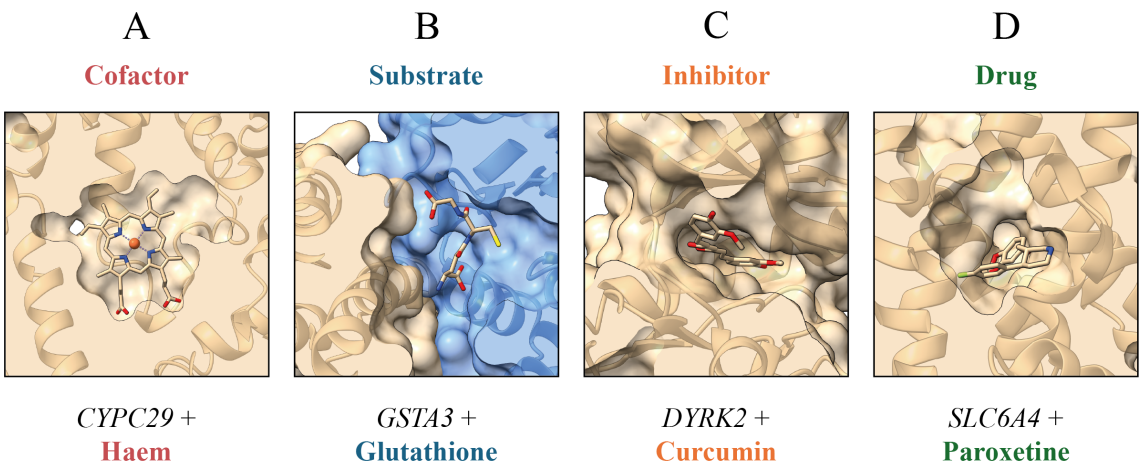

Figure 1.8. Protein-ligand complexes. Small molecule ligands interact with proteins and act as cofactors, substrates, inhibitors and drugs for therapeutic treatment. **(A)** Cytochrome P450 2C9 (P11712) interacting with haem (HEM) as a cofactor. PDB: 7RL2 [116]; **(B)** Glutathione S-transferase A3 (Q16772) interacting with its substrate glutathione (GSH). PDB: 1TDI [117]; **(C)** Dual specificity tyrosine-phosphorylation-regulated kinase 2 (Q92630) interacting with a natural inhibitor curcumin (CC9). PDB: 6HDR [118]; **(D)** Sodium-dependent serotonin transporter (P31645) binding to paroxetine (8PR), which is an antidepressant. PDB: 5I6X [119].

Ligand site prediction methods exploit a variety of techniques to suggest binding sites. Geometry-based tools like fpocket [120], Ligsite [121] and Surfnet [122] identify cavities by analysing the geometry of the molecular surface of a protein and rely on the use of grids, gaps, spheres, or tessellation [120–128]. Energy-based methods such as PocketFinder [129] rely on the interaction energy between the protein and a chemical group or probe to identify cavities [129–134]. Conservation-based methods use sequence evolutionary conservation information to find patterns in multiple sequence alignments and identify conserved key residues for ligand site identification [57, 135, 136]. Template-based methods rely on structural information from homologues and the assumption that structurally conserved proteins might bind ligands at a similar location [22, 137–141]. Combined approaches or meta-predictors integrate multiple methods or features to infer ligand binding sites, e.g., geometric features with sequence conservation [65, 142–148]. Finally, machine learning (ML) methods utilise a wide range of techniques including random forest and deep, graph, residual, or convolutional neural networks [115, 149–170].

Chapter 4 and Chapter 5 describe the largest benchmark of ligand binding site prediction to date by analysing the performance of thirteen methods using a series of metrics on a brand new reference dataset described in Chapter 3.

1.3.9 Databases

There are many databases providing relevant information for protein analysis but two of the most commonly used ones and extensively employed for the research described in this Thesis are UniProt and the Protein Data Bank (PDB).

1.3.9.1 UniProt

UniProt is a comprehensive protein sequence database including cross-references to multiple resources to provide a wealth of information about gene expression, pathogenic variation, post-translational modifications (PTMs), protein-protein interactions, domain annotations and three-dimensional structure, among others [171]. It is composed of two main components: SwissProt and TrEMBL. SwissProt is manually curated and includes 600,000 high-quality sequences often referred to as *reviewed*. TrEMBL, on the other hand, catalogues over 250 million *unreviewed* protein sequences resulting from the automatic translation of coding sequences found in the main nucleotide databases [172].

1.3.9.2 Protein Data Bank

The Protein Data Bank is a worldwide repository for the three-dimensional structures of biological macromolecules such as proteins, DNA and RNA. There are currently 230,000 3D structures deposited in the archive determined mainly through X-ray crystallography, NMR spectroscopy and Cryo-EM [173]. Through resources such as the Protein Data Bank Europe (PDBe) knowledgebase, or PDBe-KB, three-dimensional structure is linked to functional annotations including that of domains, predicted disorder, ligand binding sites or PTMs [174].

1.4 Genetic variation

Genetic variation is the difference in DNA sequence between individuals or populations of the same species. The main source of genetic variation is *de novo* mutation. Mutations are changes in a genetic sequence that usually arise during DNA replication due to errors made by the imperfect replication machinery. Mutation can also occur as a result of damage to DNA, e.g., ultraviolet radiation, or during the repair process of such damage. Genetic variation can affect a single nucleotide in the sequence, i.e., a single nucleotide polymorphism (SNP), multiple nucleotides, or larger DNA regions, even entire chromosomes, e.g., insertion, deletion, translocation, or fusion. SNPs are the only type of genetic variation described in this Thesis.

1.4.1 Types of genetic variation

1.4.1.1 Genomic location

Based on genomic location, genetic variation can be classified into *coding* variation if it affects the mRNA that codes for the protein sequence. Alternatively, *non-coding* variants are those that affect other regions that do not code for a protein product, such as introns, intergenic regions, promoters, enhancers or other regulatory elements.

1.4.1.2 Effect on coding sequence

The genetic code is *degenerate* or redundant, as there are $4 \times 4 \times 4 = 64$ codons coding for only twenty amino acids. For this reason, a change in the coding DNA sequence is not always reflected in the protein sequence. Mutations that due to the redundancy in the genetic code do not alter the protein sequence are called synonymous or silent. Conversely, nonsynonymous mutations *do* change the protein sequence and can be further classified into: missense, nonsense, stop-loss and frameshift mutations. Missense mutations are those that replace one of the twenty amino acids by a different one. They can be conservative, if the interchanged residues present similar physicochemical properties, e.g., leucine $\rightarrow$ isoleucine, or they can be non-conservative, or radical, if the exchanged

amino acids are biochemically different, e.g., lysine → threonine. Nonsense mutations replace one of the twenty amino acids by one of the three STOP codons, resulting in an early termination of the peptide chain. Stop-loss variants are the exact opposite and exchange the original STOP codon by one of the twenty amino acids, thus resulting in an abnormally elongated protein. Finally, frameshift mutations result from the insertion or deletion of nucleotides that are not a multiple of three. When this happens, the frame on which the translation machinery reads the mRNA is shifted and a completely different protein product is obtained.

While missense mutations, which simply replace one amino acid by another and can be conservative, tend to have a limited effect on protein sequence and structure, nonsense, stop-loss and frameshift mutation have more drastic consequences. Because of this, missense variants tend to be more tolerated and, along with synonymous variants, are observed at higher frequencies in the general population [175].

1.4.1.3 Impact on phenotype

Genetic variants can also be classified based on the effect they have on the phenotype, or clinical significance, which usually corresponds to an effect on the concentration, structure, function or activity of a protein [176]. Mutations that do not have a harmful effect on the protein are called neutral or benign. Since neutral variants have no noticeable effect on the *fitness* [177], i.e., the ability to leave offspring, they are not under selective pressure and consequently persist in the general population [178]. Conversely, pathogenic variants disrupt biological processes and result in disease. Disease severity will dictate the strength with which natural selection acts upon the causing variant and therefore its frequency in the population. Mutations affecting genes needed for development and survival, or essential genes, might have lethal effects and never be observed in the population [179].

It is estimated that only 2% of the more than 4 million observed human missense variants have been clinically classified as pathogenic or benign [180]. Variants of unknown significance (VUS) therefore represent the vast majority of observed missense variants and

the prediction of their effect on fitness is an ongoing challenge in human genetics [181]. Several methods exploiting different technologies have been developed over the years to tackle this challenge including SIFT [182], PolyPhen [183] and the recent AlphaMissense [184].

1.4.2 Variation is constrained

Since the sequencing of the first draft of the human genome in 2001 [185], several massively parallel methods have been developed for the high-throughput sequencing of nucleic acids [186–190]. The drastic reduction in both time and cost required to sequence DNA has enabled large-scale projects such as the 1000 Genomes Project [191] or UK Biobank [192]. The genome aggregation database (gnomAD) is a comprehensive collection of human genetic variation from over 140,000 genomes and exomes. Resources like gnomAD make it possible to carry out systematic comparative analysis to understand the distribution and constraint of genetic variation along the human genome.

In a similar way as protein sequence is constrained across species, resulting in patterns of amino acid conservation, the genomic distribution of variation within human is also restricted by factors such as protein structure and function. Several studies have demonstrated that functional elements like buried core residues, catalytic residues in enzymatic active sites and protein-protein interfaces are strongly constrained and present fewer variants than observed elsewhere in the protein [193–196]. This phenomenon is a consequence of purifying or negative natural selection acting upon the population. Variants occurring at these relevant sites are likely to impair protein function and therefore removed from the gene pool. Consequently, these positions present a lower mutational burden, or *depletion* in variation. By quantifying these evolutionary signals, functional constraint can be measured at the genic [197] and domain [198] levels and used for the functional interpretation of VUS [199]. Despite the wealth of variation data available that allows for gene- and domain-level quantification of constraint, doing so at the individual residue level remains a relative challenge still.

1.4.3 Missense enrichment score

MacGowan *et al.* approached this issue in their 2017 work [200] by aggregating variants from human paralogous residues present in the same alignment column. Residues aligning in the same column are homologous, i.e., share a common ancestor, and therefore aggregating their variants to infer residue-level constraint is a fair assumption. They developed a missense enrichmente score (MES) to numerically quantify the evolutionary constraint acting on individual residues, or positions, by leveraging the variants found not just in a protein of interest but also in their human paralogues. For each alignment column x, the number of human residues mapping to it were counted (residues_x), as well as the number of human residues mapping to all other columns (residues_{other}). Additionally, the number of variants found across all human residues aligned to the column of interest (variants_x) and all other columns (variants_{other}) were obtained. With these four quantities, the MES can be computed as showed in Equation 1.5. These four values can also be arranged in a 2×2 contingency table and the MES understood as an odds ratio (OR) expressing the likelihood of observing variants in column x relative to all other alignment columns. An OR > 1 means a column presents more variants than the average of all other columns, i.e., is *enriched* in missense variants, whilst an OR < 1 indicates *depletion* relative to the rest of the alignment. An OR = 1 indicates neutrality relative to the other columns, i.e., the number of variants found within column x follows the same distribution as the average of the rest of the alignment. Using Fisher's exact test [201] the significance of this MES (OR) can be assessed with a *p*-value.

$$\text{MES} = \frac{\text{variants}_x / \text{residues}_x}{\text{variants}_{other} / \text{residues}_{other}} \tag{1.5}$$

1.4.4 Conservation plane

Both Shenkin and MES are measures of evolutionary constraint on protein amino acids. Nevertheless, these metrics quantify it at two completely different time scales. Amino acid divergence calculated from an MSA captures the evolutionary history of a protein family

resulting of hundreds of millions of years of divergence across species originated from speciation events, large-scale genomic rearrangements and strong selective pressures, among other factors. In contrast, the missense enrichment score aims to capture the variability in *our* species emerging from migration events, genetic drift and weaker selection taking place within a much shorter time scale [202]. The stratification of conserved and divergent positions by MES yields four classifications, or four quadrants on the *conservation plane* (Figure 1.9). These are conserved positions that are missense-depleted (CMD), conserved positions, yet enriched in missense variation (CME), unconserved, or divergent, positions enriched in missense variants (UME) and unconserved and missense-depleted (UMD) positions [203].

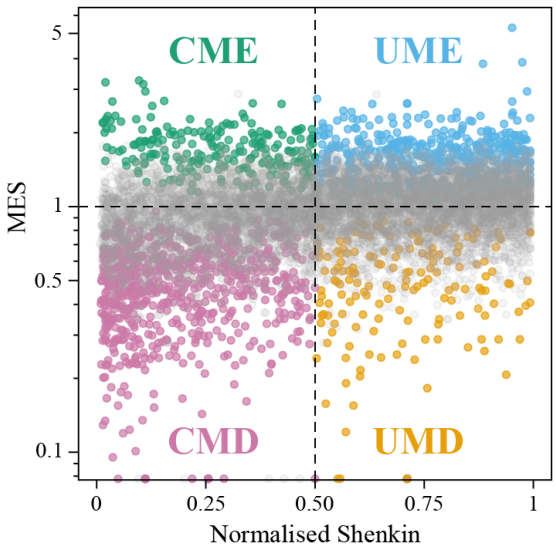

Figure 1.9. Conservation plane. The *conservation plane* arises from the comparison of within-species constraint, as measured by the missense enrichment score (MES), and across-species constraint quantified by amino acid conservation, in this case, by a normalised Shenkin divergence score. The conservation plane can be divided in four quadrants. Positions that are conserved across species and missense-depleted in human (CMD) are found in the bottom-left corner (pink). Conserved positions that are enriched in missense variation (CME) are on the top-left quadrant (green). Unconserved or divergent positions enriched in missense variants (UME) are on the top-right (blue). Finally, unconserved and missense-depleted (UMD) positions are on the bottom-right (orange). Figure adapted from MacGowan *et al.* [203].

CMD positions are the most constrained both across species (conserved) and within the human population (missense-depleted). The vast majority of them are buried in the core and are critical for protein folding, packing and stability. When they are not buried, they are highly enriched in protein-protein and protein-ligand interactions [44]. Addi-

tionally, they are enriched in ClinVar [204] pathogenic variants, further emphasising their relevance. Unconserved positions are often dismissed as they appear to be mutating freely and be under no constraint, resulting in their divergence across homologues. However, MacGowan *et al.* [203] showed that there is a subset of unconserved positions that are strongly constrained in the human population, i.e., significantly depleted in missense variation. These positions tend to be on the surface and act as specificity-determining positions (SDP) bestowing protein domains the ability to bind a wide range of substrates. Furthermore, they are enriched in pathogenic variants relative to their missense-enriched counterpart (UMEs).

1.5 Drug discovery

Drug discovery is the process of developing a new drug. It goes from the original idea conception to the market launch of a finished product and beyond. This is an extremely complex process which can take up to 12-15 years and cost more than $1 billion [205]. This high economic and time cost is caused by the high rate of failure that potential drug candidates experience during the development process, also known as *attrition*. The drug discovery and development pipeline is illustrated in Figure 1.10. This pipeline can be divided in two stages. Stage I is drug discovery and encompasses target identification and validation, hit identification and lead optimisation. Stage II corresponds to the development of the drug and includes pre-clinical, clinical trials and drug approval.

1.5.1 Target identification

Target identification is the first step in the drug discovery pipeline and one of the most critical. Drugs often fail in the clinical stages due to two main reasons: they are not safe or they do not work. Because of this, a thorough target identification and subsequent validation is vital. The goal is to identify a biomolecule to target with a drug to treat or cure a disease. The target can either be a gene, RNA or protein and the drug is usually a small molecule, peptide or a protein, e.g., antibody. An ideal target meets a series of require-

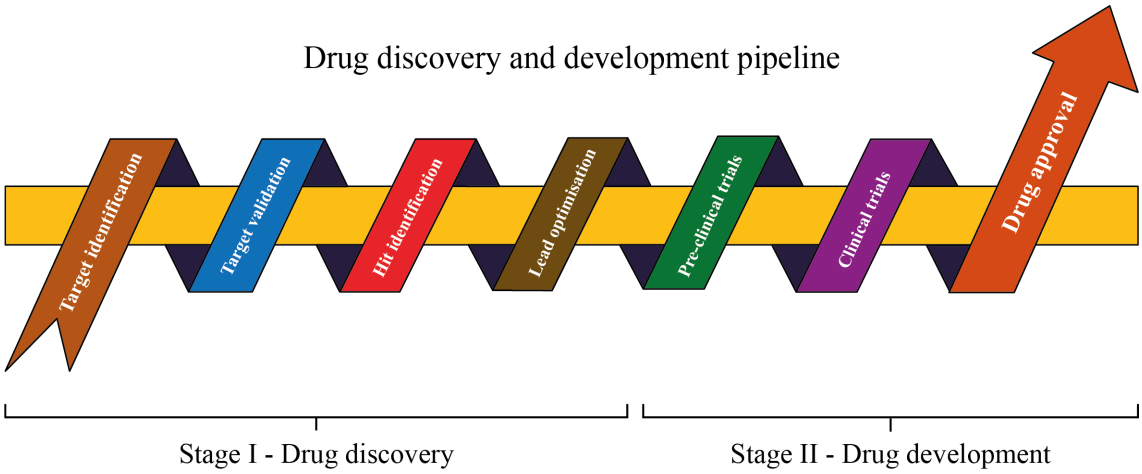

Figure 1.10. Drug discovery and development pipeline. The pipeline for discovering a drug can be divided in two stages. Stage I focuses on the discovery of a drug and includes target identification and validation, hit identification and lead optimisation. Stage II covers the development of the drug and includes pre-clinical, clinical trials, drug approval and pharmacovigilance. Figure adapted from Cui *et al.* [206].

ments: efficacy, safety and most importantly *druggability*, among others. A *druggable* target is amenable to interact with a putative drug. This interaction should trigger a biological response measurable *in vitro* and *in vivo* through biochemical or functional assays. The mining of available biomedical data from the literature, proteomics, 3D structure, genetic association studies, pathogenic variation or phenotypic screening are some of the most commonly used approaches for target identification [207].

1.5.2 Target validation

Target validation is the technical assessment of whether a target plays a critical role in a disease process and whether pharmacological modulation of the target could be effective in a particular patient population. It is predicted that a more effective target validation strategy could reduce attrition in phase II clinical trials by ≈24% lowering the cost of developing a new molecular entity (NME) by ≈30% [208]. Accordingly, the validation of a therapeutic target is a step of paramount importance within the discovery of new drugs. Some of the most frequently used approaches to validate a target include RNA interference, gene knockouts, the use of animal models and target druggability analysis, e.g., by ligand binding site prediction [209].

1.5.3 Hit identification

The next step once the target has been validated is to identify *hits*. Hits are compounds that bind to the target and elicit the desired biological activity in an assay. Their identification relies on a combination of experimental techniques, e.g., high-throughput (HTS) or fragment screening (FS), and computational techniques such as virtual screening (VS). In high-throughput screening, robotic automation is employed to evaluate large libraries of chemical compounds against a target in a biochemical or cell-based assay. HTS usually identifies a few compounds with the desired biological activity and high binding affinity to the target. FS is complementary to HTS and obtains high-quality information about the 3D structure of a protein-ligand complex by using X-ray crystallography. Lastly, provided the 3D structure of the target is known, virtual screening techniques can be used. VS encompasses a set of ligand-based (LBVS) and structure-based (SBVS) computational techniques, such as pharmacophore-mapping or protein-ligand docking, respectively. These techniques are able to identify hotspot residues relevant for ligand binding and guide the design of more effective compounds [210].

1.5.4 Lead optimisation

In this phase, identified hits are refined into promising *lead* compounds by optimising their properties before getting to pre-clinical drug candidates. This refinement aims to enhance pharmacokinetic (PK) properties such as potency, i.e., binding affinity to the target, as well as selectivity – by minimising off-target effects –, solubility, permeability, stability and toxicity. Quantitative structure-activity relationship (QSAR) studies are carried out to suggest molecules with more favourable PK properties by adding or replacing functional groups of the original hit compound. Additionally, high-throughput *in vitro* assays can be carried out to optimise the absorption (how it enters the bloodstream), distribution (how it travels within the body), metabolism (how it is broken down), excretion (how it is eliminated) and toxicity (ADMET) properties of the compounds [211].

1.5.5 Pre-clinical studies

The discovery stage concludes with the acquisition of the optimised leads and thus begins the development stage. The primary goal of pre-clinical studies is to thoroughly evaluate the safety, efficacy, pharmacokinetics and pharmacodynamics (PD) of the drug candidates before advancing to clinical trials in humans. This is achieved with a combination of *in vitro* and *in vivo* studies, including cell-bassed assays and animal models, respectively. ADMET properties are assessed to ensure a favourable pharmacological profile and toxicology studies are carried out to establish the no observed adverse effect level (NOAEL) and determine a safe starting dosage in human [212].

1.5.6 Clinical trials

Those candidates that pass through pre-clinical development will be submitted to clinical trials in voluntary human subjects. Clinical trials are divided in three phases with different goals. Phase I focuses on establishing the maximum tolerated dose (MTD) of a drug by performing strictly calculated dose escalation in a small number (20-80) of *healthy* and *diseased* individuals. Phase II will aim to establish the preliminary efficacy of the drug by comparing a *treatment* and a *placebo*, or control, group whilst closely monitoring side effects. Usually 100-300 individuals are involved in Phase II trials. Phase III confirms the safety and efficacy of the drug by involving a larger (1000-3000) and more diverse target population whilst noting potential adverse side effects. Successful completion of clinical trials results in the submission of a comprehensive report to regulatory agencies for review, marking the final step before the drug can reach the market [213].

1.5.7 Drug approval

After a drug has been approved and granted license by regulatory agencies such as the Food and Drug Administration (FDA), the European Medicines Agency (EMA) or the Medicines and Healthcare products Regulatory Agency (MHRA), it can be commercialised. Once on the market, drugs enter the post-marketing phase, also known as phase IV. In this

phase, pharmacovigilance activities are conducted to monitor long-term safety and effectiveness in larger and more diverse populations. This includes the identification of rare adverse effects and potential new therapeutic uses [214].

1.6 Fragment-based drug discovery

Fragment-based drug discovery (FBDD), or fragment screening, is a widely used technique to identify compounds binding against a specific protein target [215]. It falls within the range of tools used in the hit identification step of the drug discovery pipeline. FBDD typically uses X-ray crystallography to provide detailed information on the binding mode of small molecule fragments that bind to a protein [113]. These fragments explore the vast chemical space and usually obey the *Rule of 3*: they present low molecular weight (200-500 Da), few rotatable bonds and low hydrophobicity [216]. Hits tend to have low affinity (milimolar range) due to their small size. Nevertheless, they provide a good scaffold for optimisation and can be linked or grown to form more potent leads [217]. A typical fragment screening experiment generates a collection of three-dimensional structures with fragments bound to different regions of the protein. This is done by soaking pre-formed protein crystals in high-concentration fragment solutions, allowing the fragments to bind to the protein. After soaking, crystals are carefully washed to remove unbound fragments and cryoprotected before freezing. Once frozen, fragment-bound crystals are X-rayed, electron densities analysed and structure models obtained [218].

While many fragments group around well understood catalytic or binding sites, fragments are also observed bound to regions of the protein where the functional significance is unclear. Such sites may be functionally irrelevant or could identify previously unknown allosteric or other functionally important sites worthy of experimental investigation. Chapter 2 aims to address this issue by characterising fragment screening sites using a combination of structural, conservation and variation data, thus providing insight into the likelihood of function of such sites.

1.7 Thesis scope

Advances in the accurate prediction of protein 3D structure have resulted in a drastic reduction in the sequence-to-structure gap [219]. The UniProt knowledgebase (UPKB) catalogues 248 million protein sequences [220, 221], most of which have now structure models available through resources such as the AlphaFold Database (AFDB) [222] and other providers [223–225]. However, only a minuscule fraction of these proteins present residue-level functional annotations in UniProt – 55,000 (0.02% of UPKB) - or include biologically relevant ligands co-crystallised in the PDB [226] – 29,000 (0.01%) (Figure 1.11). The significant expense and time required for experimental validation underscores an urgent need for computational methods to characterise ligand sites systematically and highlight residues likely to be relevant to protein function.

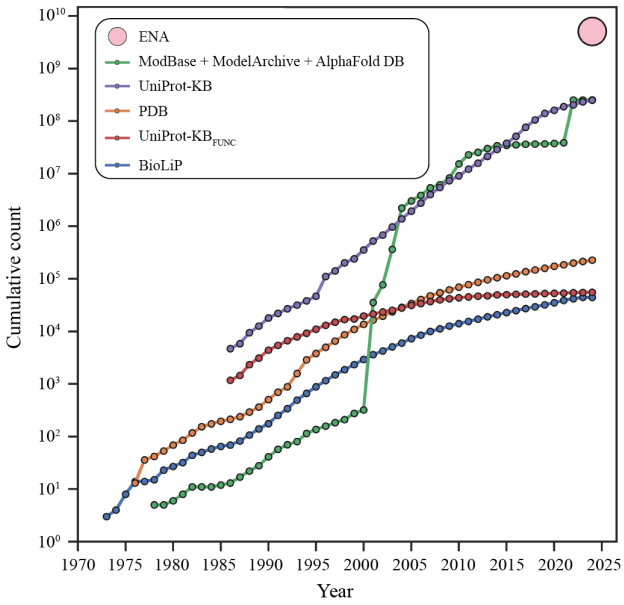

Figure 1.11. Database growth curves. Growth curves for some of the most relevant nucleotide and protein sequence and structure databases from 1970 to date. The European Nucleotide Archive (ENA) catalogues nucleotide sequences [227]. ModBase, ModelArchive and AlphaFold DB are some of the largest predicted protein structure resources. UniProt-KB$_{FUNC}$ corresponds to the subset of protein sequences with residue-level experimentally determined functional annotations in UniProt. BioLiP is a semi-curated database of biologically relevant protein-ligand complexes [228]. SWISS-MODEL was not included in this graph as growth curves could not be obtained. Likewise for ENA, for which just the number of sequences in 2024 is included. The Y-axis is in $\log_{10}$ scale.

Small molecule ligands are crucial for protein function and act as substrates, cofactors or drugs in therapy. Identifying the protein regions where these molecules bind, understanding the mode in which they do so and characterising that interface is therefore key to understanding and modulating protein function. Chapter 2 describes work for the definition, characterisation and classification of likely functional class of ligand binding sites derived from fragment screening experiments. Chapter 3 extends this approach to the entirety of the PDBe, characterising >65,000 biologically relevant protein-ligand binding sites using structural, divergence and human variation data. Additionally, a web server is introduced for users to explore this large dataset, named LIGYSIS, as well as analyse their own protein-ligand complexes. Finally, Chapter 4 and Chapter 5 describe the largest comparative performance assessment of ligand binding site prediction to date including thirteen canonical methods and fifteen novel variants defined in this work. Beyond ranking the methods by their prediction capability using several relevant metrics, this benchmark provides insight into the strengths and weaknesses of each method and paves the way for improvement in the field of ligand site prediction.

This Thesis aims to illuminate the nature of protein-ligand binding sites by analysing their structural features, evolutionary constraint, both within and across species, and using them to pinpoint those sites more likely to alter protein function if targeted. Additionally, a thorough benchmark is carried out, objectively quantifying the strengths and weaknesses of the state-of-the-art methods for ligand site prediction. Both of these contributions have potential applications in drug discovery and may help mitigate attrition in clinical trials, ultimately improving efficiency and resource allocation in drug discovery.

Chapter 2

Classification of likely functional class for ligand binding sites identified from fragment screening

Preface

This Chapter introduces a series of methods to group small molecule ligands by protein interactions and cluster ligand sites by relative solvent accessibility profile. 293 unique ligand binding sites are defined from 37 fragment screening experiments and grouped into four clusters that are differentially enriched in known functional sites. A multi-layer perceptron is developed to predict cluster labels with an accuracy of 96%, which allows functional classification of sites for proteins not in this set. Dr Stuart MacGowan conceived the idea of ligand clustering by their interactions using Jalview features. Dr Callum Ives started work on the characterisation of ligand sites integrating conservation and variation data with a set of fragment screening experiments by the Structure Genomics Consortium. *I* developed the project, extended the code, curated the dataset and performed all the analysis, the results of which are described in this Chapter.

Publications

Utgés, J.S., MacGowan, S.A., Ives, C.M., Barton, G.J. Classification of likely functional class for ligand binding sites identified from fragment screening. *Commun. Biol.* **7**, 320 (2024). https://doi.org/10.1038/s42003-024-05970-8.

2.1 Introduction

This Chapter presents a strategy to identify which fragment binding sites are most likely to be of functional importance and, therefore, to prioritise sites for further investigation. The first step is to identify binding sites from fragment screening data. Ligand binding sites are not predicted, as could be done, for instance, by P2Rank [115], fpocket [120], or molecular dynamics-based methods such as MixMD [229, 230], MDmix [231], or SILCS [232]. Instead, from a set of experimentally determined three-dimensional structures of protein-ligand complexes, ligands are defined as binding to the same site based on their interactions with the protein.

In most previous studies, the focus has been on clustering ligands by root-mean-square deviation (RMSD) [233] or Euclidean distances [234] after ligand superposition. Ligand site prediction methods, such as 3DLigandSite [137, 235] also define sites based on ligand structure superposition and RMSD. Here, an algorithm is described that defines ligand binding sites from the analysis of ligand-interacting residues on the protein. The method allows describing the extent of a fragment binding site without the need for superposition. Unsupervised methods are then applied to group the defined sites into four robust clusters based on their relative solvent accessibility profiles. This analysis suggests which sites in a set of 39 fragment screening experiments are most likely to be of functional significance through further stratification by evolutionary conservation and human population missense depletion [200, 203]. A machine learning method is subsequently developed to classify a set of interacting residues from an experimentally determined structure or a predicted ligand binding site into one of the four defined classes.

The work described in this Chapter is likely to be of interest to drug researchers performing fragment screening studies, but wider applications to ligand site classification from experimentally determined or predicted structures are also discussed.

2.2 Methods

2.2.1 Structure dataset

The Pan-Dataset Density Analysis (PanDDA) algorithm characterises a set of related crystallographic datasets of the same crystal form and identifies binding events by isomorphous difference maps [236]. Initially, 3021 three-dimensional structures determined by X-ray crystallography were selected by querying the PDBe [226] for entries containing the string "PanDDA" in their title. 1542 of the structures included bound ligands for 39 different proteins. Four proteins, which formed part of multi-protein complexes including additional ligands were excluded to leave a set of single-protein-ligand complexes from 35 different proteins and a total of 1450 structures. The structures had resolutions from 0.9-3.3 Å, with a mean resolution of ≈1.5 Å. Preferred biological assemblies, as defined by PISA [237], were downloaded from the PDBe via ProIntVar [238].

2.2.2 Binding site definition

Ligand binding site definition or prediction approaches are usually based on the spatial superposition of ligand molecules and their clustering on metrics such as Euclidean distance or RMSD [137, 233–235]. These methods rely on structural superimposition and can be computationally expensive when dealing with large numbers of structures. Here, binding sites are defined from protein-ligand interactions without the need for superposition (Figure 2.1).

Only non-ion ligands of interest were used for the binding site definition. These do not include water molecules (HOH), nor other by-products of the experimental conditions, such as ethylene glycol (EDO), glycerol (GOL) or tris (TRS). Ligand contacts were de-

termined with Arpeggio [239]. For a given ligand, a binding fingerprint is defined as the UniProt sequence numbers of the residues the ligand interacts with. For a pair of ligands L_A and L_B, with interaction fingerprints A and B, their relative intersection, I_{rel}, is defined (Equation 2.1) by dividing the intersection of sets A and B by the maximum possible intersection between the two sets, given by the minimum fingerprint length (Equation 2.2).

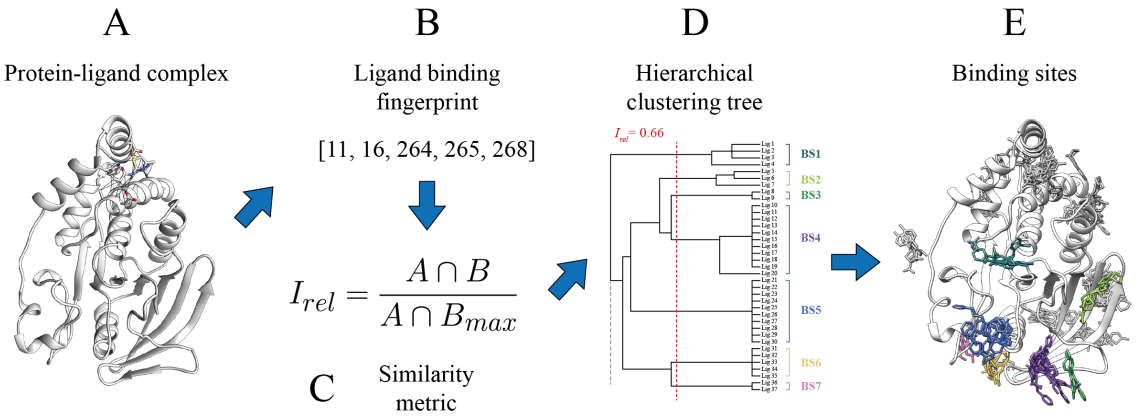

Figure 2.1. Ligand binding site definition algorithm. The method defines ligand binding sites from a set of three-dimensional structures portraying the complex of a protein of interest bound to ligands. **(A)** Protein-ligand complex exemplified by Tyrosine-protein phosphatase non-receptor type 1, PTP-1B, (P18031) bound to N-(4-methyl-1,3-thiazol-2-yl)propanamide (JFP). PDB: 5QDJ [240]; **(B)** Ligand binding fingerprint, consisting of the UniProt sequence numbers of ligand-interacting residues; **(C)** Formula of the similarity metric employed: relative intersection, I_{rel}; **(D)** Hierarchical clustering tree resulting from a similarity matrix, cut at a threshold to determine distinct clusters of ligands. This subtree of the full tree highlights 7 out of 18 binding sites defined on PTP-1B. The dashed line indicates that the tree continues beyond what is shown here; **(E)** Superposition of ligands binding to the protein, coloured by ligand cluster. Only ligands found in groups (sites) 1-7 are coloured by their membership and the rest are coloured in grey. Cartoon PDB: 5QDJ.

$$I_{rel} = \frac{A \cap B}{A \cap B_{max}} \tag{2.1}$$

$$A \cap B_{max} = \min(\text{len}(A), \text{len}(B)) \tag{2.2}$$

I_{rel} serves as a similarity metric, ranging from 0 to 1, suitable to perform hierarchical clustering on the ligands. Single-linkage hierarchical clustering was conducted with the OC software [241]. After exploring various I_{rel} threshold values for cutting the tree, I_{rel} = 0.66 was selected as the optimal threshold for this dataset. As I_{rel} is a similarity metric, this threshold indicates that a ligand shares at least two-thirds of its binding residues with

at least one other ligand within the same cluster. A total of 293 ligand binding sites across 37 protein domains were defined this way. For each protein, all structures were multiply aligned by STAMP [242]. Ligand binding sites were visualised in UCSF Chimera [243].

2.2.3 Multiple sequence alignments

Two of the 35 proteins included fragment screening experiments targeting multiple domains, or protein products, resulting in 39 protein-fragments sets. A representative sequence was selected for each of the 39 sets of structures. These sequences were used to search SwissProt [244] for homologues and generate multiple sequence alignments. The search was performed with jackHMMER [245] using default parameters and 5 iterations. Evolutionary divergence within the alignments was quantified with $N_{Shenkin}$, a normalised version of the Shenkin divergence score [51] defined by Utgés *et al.* [44].

2.2.4 Human variants and enrichment

VarAlign [238] was used to retrieve genetic variants from gnomAD v2.1 [246] found in the human sequences within the multiple sequence alignment generated for each target protein. gnomAD contains exomes and genomes from 141,456 unrelated individuals with no known phenotypic conditions and is therefore a reasonable representation of the general *healthy* population. Variants found in the human sequences within the alignments were mapped to individual alignment columns and missense enrichment scores (MES) were calculated. MES represents the enrichment in missense variants of an alignment column relative to the average of the other columns in the alignment [200, 203]. 95% confidence intervals (CI) and *p*-values were used to assess the significance of these ratios [247]. MES was also calculated for the defined ligand binding sites. The MES of a binding site represents the enrichment in missense variants of a binding site relative to the rest of protein residues. Alignment columns as well as binding sites were classified as enriched (MES $>$ 0), depleted (MES $<$ 0) or neutral (MES = 0). Enrichment was not calculated for two of the 39 proteins since no human homologues were identified.

2.2.5 Binding site clustering

Secondary structures were defined with DSSP [112] via ProIntVar, and relative solvent accessibility (RSA) was calculated with the method of Tien *et al.* [108]. The defined binding sites were grouped according to the pattern of RSA as follows and summarised in Figure 2.2.

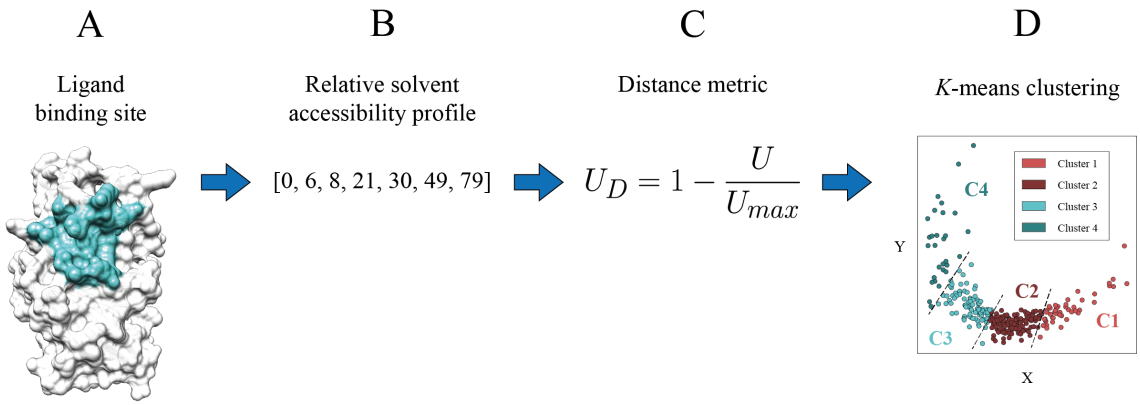

Figure 2.2. Binding site clustering algorithm. The method here clusters ligand binding sites defined across different proteins based on their solvent accessibility profiles. **(A)** Defined ligand binding site for PTP-1B (P18031) on PDB: 5QDU [240]; **(B)** Relative solvent accessibility profile of a binding site, represented by the RSA of the site residues; **(C)** Formula of the distance metric employed: distance U, U_D; **(D)** Multidimensional scaling (MDS) representation of binding sites coloured according to the four clusters determined by the *K*-means algorithm. Dashed lines represent the cluster limits.

Given two binding sites, A and B, with RSA profiles r_A and r_B formed by n_A and n_B amino acids residues, respectively, U_A and U_B are calculated with Equation 2.3. The Mann-Whitney U statistic [248], as implemented in SciPy [249], was chosen as it has a maximum theoretical value (U_{max}) (Equation 2.4). A relative U value, U_{rel}, ranging from 0 to 1 is obtained by dividing the U value by U_{max}. The more similar r_A and r_B are, the bigger U and U_{rel} are. Thus, U_{rel} is a similarity score. Subtracting U_{rel} from 1 gives the U distance, U_D, (Equation 2.5). U_D is indicative of how different r_A and r_B are and can be used to cluster binding sites by their RSA profiles.

$$U_A = R_A - \frac{n_A(n_A+1)}{2}, \quad U_B = R_B - \frac{n_B(n_B+1)}{2} \tag{2.3}$$

$$U_A + U_B = n_A n_B, \quad U = \min(U_A, U_B) \rightarrow U_{max} = \frac{n_A n_B}{2} \tag{2.4}$$

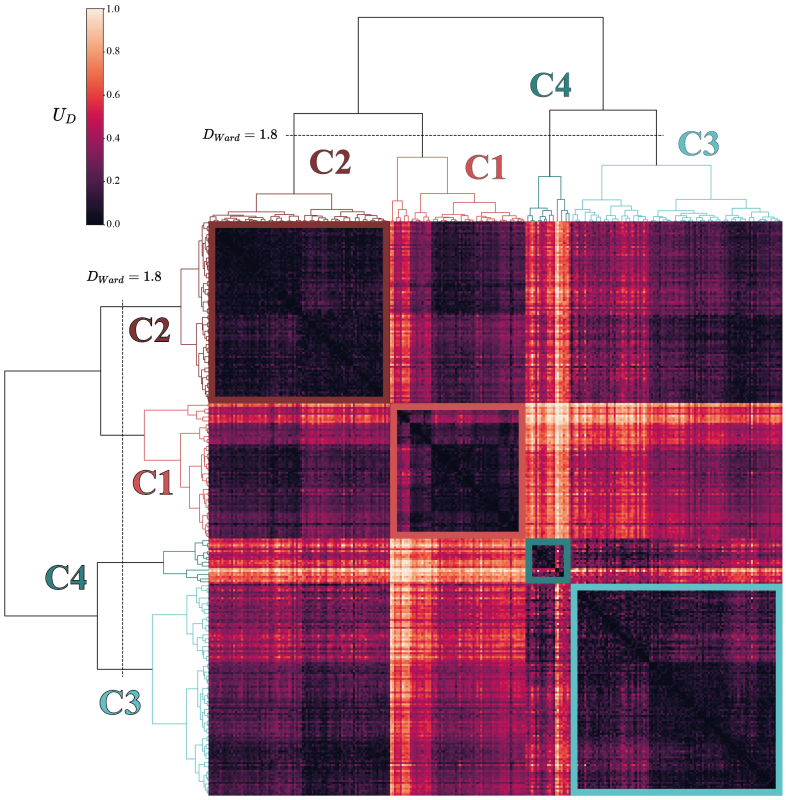

Figure 2.3. Binding sites Ward clustering. Cluster map of the U distance, U_D, matrix of the 293 defined binding sites clustered by the Ward hierarchical clustering method implemented in SciPy. The tree is cut at $D_{Ward} = 1.8$, giving four clear clusters. These clusters are labelled so they correspond to the ones obtained with K-means. Clusters in the heat map are represented by dark squares around the diagonal. U_D is a distance; therefore, clusters include sites that are similar to each other, and present lower distances (dark colour).

$$U_{rel} = \frac{U}{U_{max}} \rightarrow U_D = 1 - U_{rel} \tag{2.5}$$

After calculating pairwise distances between the RSA profiles of the defined binding sites, K-means clustering [250] was performed. Several clustering algorithms were tried to realise this task, including some hierarchical, or connectivity-based, such as single and complete-linkage [251], unweighted average linkage clustering (UPGMA) [252] or Ward linkage [253], as well as centroid-based, such as K-means. Overall, the clusters obtained by the different methods were similar. Ward linkage and K-means resulted in the most similar clusters, displaying an average similarity between clusters of 85% (Figure 2.3).

Finally, multidimensional scaling (MDS) [254] with $N = 2$ dimensions was performed to visualise the clusters. K-means was selected as it presented better contained clusters,

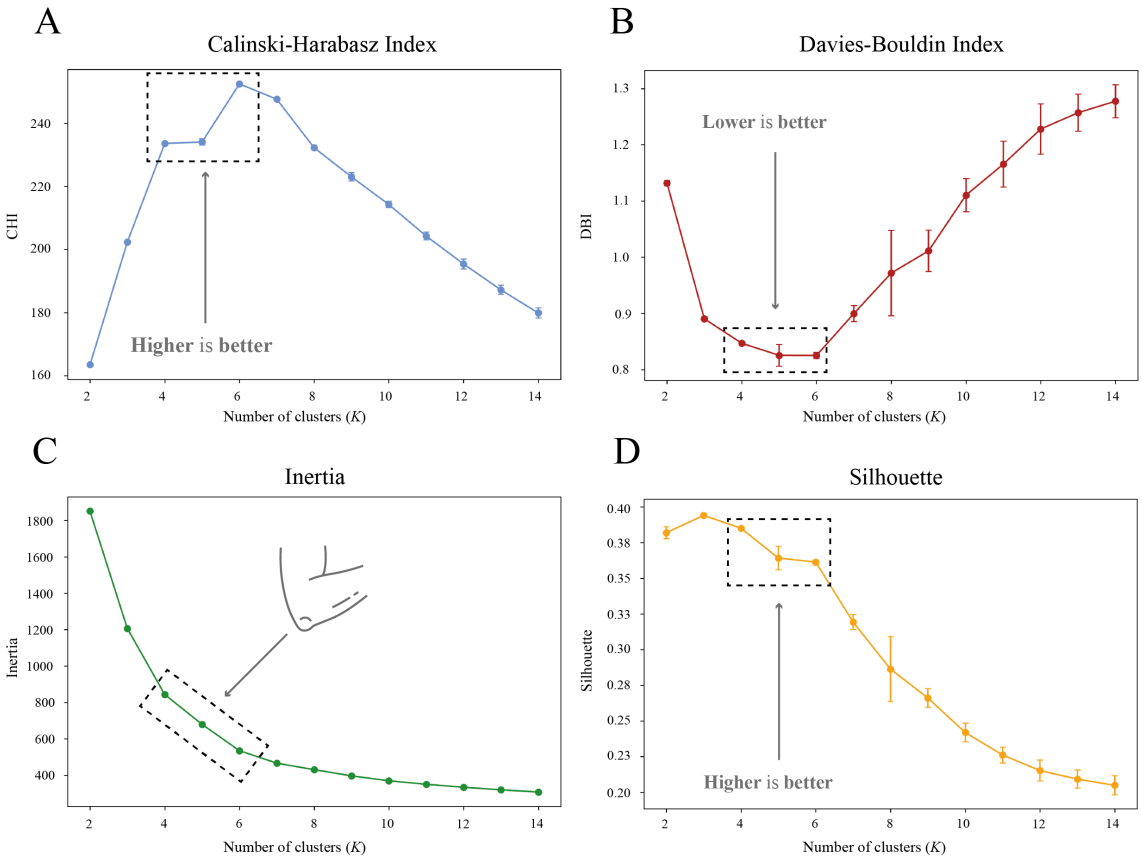

Figure 2.4. *K*-means clustering robustness. Cluster analysis to assess the quality of the *K*-means clustering. For each $K \in [2, 14]$, clustering was bootstrapped 1000 times with different initial random states. Error bars indicate 1 SD. **(A)** Calinski-Harabasz Index (CHI); **(B)** Davies-Bouldin Index (DBI); **(C)** Inertia; **(D)** Silhouette. All methods agreed the optimal clustering of this dataset lies in $K \in [4, 6]$.

i.e., less overlapping between members of different clusters. The silhouette [255] and elbow [256] methods, as well as Calinski-Harabasz index (CHI) [257] and Davies-Bouldin index (DBI) [258] were used for finding the optimal K (Figure 2.4), in conjunction with the MDS, trees resulting from hierarchical clustering algorithms, and the visual representation of the RSA profiles, to decide on a final number of $K = 4$ clusters: C1, C2, C3, and C4. Clustering was repeated 1000 times with different random states and 289/293 (98.6%) sites were always present in the same cluster, thus suggesting cluster robustness.

2.2.6 Binding site cluster prediction

Two different predictive models were developed with the aim of classifying binding sites into the defined RSA clusters obtained with *K*-means, as described above. The first uses

the K-nearest neighbour (KNN) algorithm as implemented in Scikit-learn [259], with K = 3. The input for this KNN model are the rows of the U_D matrix, containing the distances between pairs of binding site RSA profiles.

The second model is a multi-layer perceptron (MLP) [260], a type of artificial neural network (ANN), constructed with Keras, [261] with a single hidden fully connected layer between the input layer of 11 neurons, and the output layer of 4 neurons, one for each cluster label. RSA profiles present different lengths depending on the size (number of amino acids) of the binding site. As this input is not suitable for the neural network, binding sites were encoded as an 11-element vector. The first element of the vector encodes the size of the binding site relative to the maximum site size of 40 residues within this dataset. The other 10 elements represent the proportion of residues forming the binding site with an RSA within a 10% interval: [0, 10), [10, 20), …, and [90, 100].

2.2.6.1 MLP ablation

A thorough hyperparameter optimisation was carried out by examining the effect that a series of hyperparameter changes had on the prediction accuracy relative to the current ML setup, labelled as *current*. Sixty-four single-hyperparameter changes were performed, one at a time. For each variation, 100 models were trained with different seeds and the average validation accuracies compared to the current MLP. Sixty-four pairwise *t*-tests were conducted to compare the accuracy means, and Benjamini-Hochberg correction [262] applied. FDR and increment in accuracy, Δ_{acc} (Equation 2.6) were used to describe the results, where $acc_{current}$ is the average validation accuracy of the current ML setup across the 100 seeds, and $acc_{variant}$ is the average accuracy across 100 seeds of each one of the 64 variant models. $\Delta_{acc} < 0$ represents a decrease in performance relative to the current ML architecture, whereas $\Delta_{acc} > 0$ corresponds to a higher accuracy. The results of this ablation study are described below and graphically represented in Figure 2.5 and Table 2.1.

$$\Delta_{acc}(\%) = acc_{variant} - acc_{current} \tag{2.6}$$

Number of layers Removing the single hidden layer resulted in a significant decrease in accuracy, $\Delta_{acc} = -11\%$ (FDR < 0.05). The addition of more layers did not improve accuracy: 2-layer $\Delta_{acc} = -1\%$ (FDR < 0.05), 10-layer $\Delta_{acc} = -8.9\%$ (FDR < 0.05), or was not statistically different from the current setup baseline: 5-layer $\Delta_{acc} = -0.15\%$ (FDR = 0.42).

Neurons per layer The addition of neurons $N_{\text{neurons}} = [11, 20, 25, 50, 100]$ in the single hidden layer did not improve the current accuracy (FDR > 0.05). The removal of neurons did not have an effect of performance for $N_{\text{neurons}} = [4, 5, 6, 7, 8, 9]$ (FDR > 0.05), or had a significant negative effect for 1 neuron, $\Delta_{acc} = -15\%$ (FDR < 0.05), 2 neurons $\Delta_{acc} = -4\%$ (FDR < 0.05), and 3 neurons, $\Delta_{acc} = -1\%$ (FDR < 0.05). This result suggests that 5 neurons on a single hidden layer might be enough to achieve a comparable accuracy to the current model.

Activation function The usage of different activation functions either negatively affected the accuracy of the MLP ($\Delta_{acc} < 0$) or had no effect (FDR > 0.05).

Loss function Different loss functions resulted in a substantial loss of accuracy: $\Delta_{acc} \approx -50\%$ (FDR < 0.05). This was expected as they are not appropriate for a multi-label classifier, unlike sparse categorical cross-entropy.

Weight initialiser Most weight initialisers were tested and either negatively affected the accuracy of the MLP ($\Delta_{acc} < 0$) or had no effect (FDR > 0.05). However, `RandomNormal`, `RandomUniform`, and `TruncatedNormal` did improve the accuracy, but by less than 1%, $\Delta_{acc} < +1\%$ (FDR < 0.05).

Optimiser Regarding optimisers, they either severely negatively affected accuracy with a $\Delta_{acc} \approx -30\%$ (FDR < 0.05), had no significant effect (FDR > 0.05), or very slightly improved accuracy, such as `RMSProp` $\Delta_{acc} < +1\%$ (FDR < 0.05).

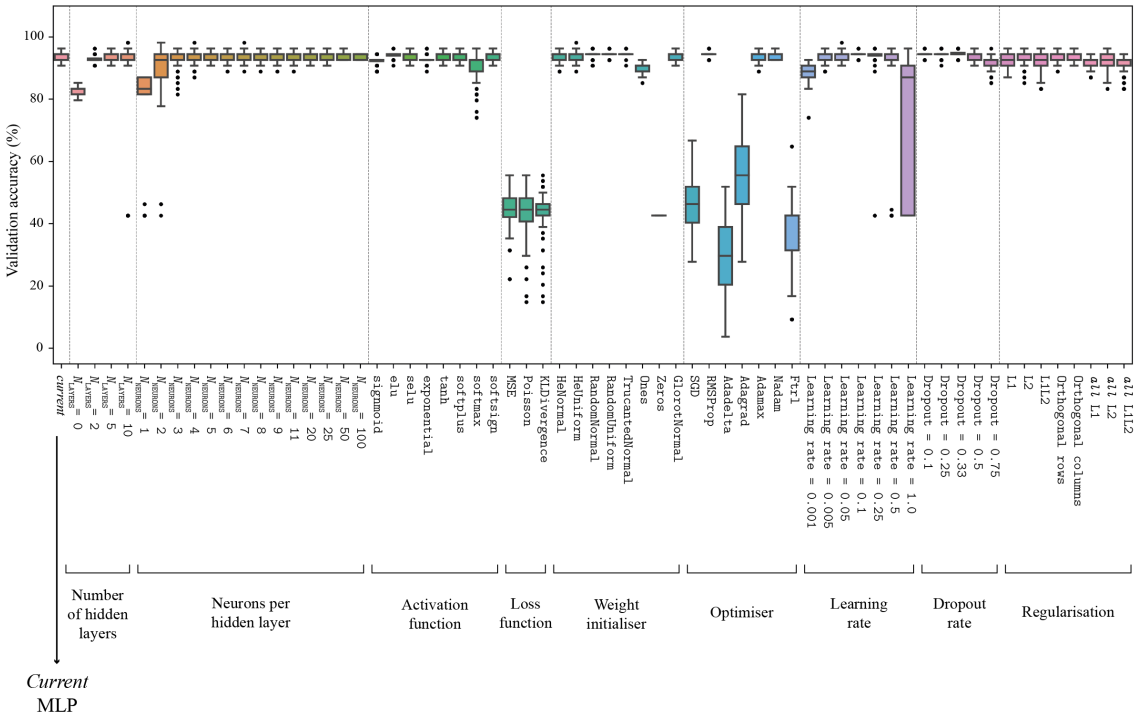

Figure 2.5. MLP ablation study. Ablation study performed on the MLP. Sixty-four single hyperparameter changes were conducted one at a time to explore the hyperparameter space and the effect they have on the prediction accuracy relative to the current machine learning setup, labelled as *current*. Box and whiskers represent the distribution of validation accuracy across 100 random seeds. The box contains the central 50% of the data, i.e., Q1 – median (Q2) – Q3 also known as interquartile range (IQR). Whiskers extend to 1.5 × IQR, and beyond them are the outliers. Dashed lines mark the separation between different hyperparameters: number of layers, neurons, activation, loss functions, weight initialisers, optimisers, learning, dropout rates and regularisation techniques.

Learning rate Extreme learning rates of 0.001 (too small) and 1.0 (too big) negatively affected prediction: $\Delta_{acc} < -5\%$ (FDR < 0.05). Intermediate rates had either no significant effect (FDR > 0.05) or a small effect: $\|\Delta_{acc}\| < 1\%$.

Dropout rate Regarding dropout rates, a rate of 75% negatively affected prediction with $\Delta_{acc} < -2\%$ (FDR < 0.05). Lower dropout rates: 0.1, 0.25, and 0.33 did improve the accuracy, but the effect size was very small, $\Delta_{acc} < +1\%$ (FDR < 0.05). This result agreed with the effect of the removal of neurons per layer and showed that fewer neurons on a single hidden layer might be enough to achieve a comparable accuracy to the current model, since dropping them out has no effect.

Regularisation Overall, implementing kernel, bias or activity regularisation techniques did not improve prediction accuracy, but worsened it: $\Delta_{acc} \ni [-2.56, -0.46]$ (FDR < 0.05).

Model	**Accuracy** (%)	Δ_{acc} (%)	**FDR**
current	93.9	-	-
$N_{\texttt{LAYERS}}$ = 0	82.9	−11.0	0
$N_{\texttt{LAYERS}}$ = 2	92.9	−1.0	0
$N_{\texttt{LAYERS}}$ = 5	93.8	−0.1	0.42
$N_{\texttt{LAYERS}}$ = 10	85.0	−8.9	0
$N_{\texttt{NEURONS}}$ = 1	79.0	−14.9	0
$N_{\texttt{NEURONS}}$ = 2	89.8	−4.2	0
$N_{\texttt{NEURONS}}$ = 3	92.9	−1.1	0
$N_{\texttt{NEURONS}}$ = 4	93.7	−0.3	0.24
$N_{\texttt{NEURONS}}$ = 5	93.8	−0.1	0.54
$N_{\texttt{NEURONS}}$ = 6	93.7	−0.2	0.26
$N_{\texttt{NEURONS}}$ = 7	93.6	−0.3	0.08
$N_{\texttt{NEURONS}}$ = 8	93.6	−0.3	0.08
$N_{\texttt{NEURONS}}$ = 9	93.8	−0.1	0.49
$N_{\texttt{NEURONS}}$ = 11	94.0	+0.1	0.65
$N_{\texttt{NEURONS}}$ = 20	94.0	+0.1	0.68
$N_{\texttt{NEURONS}}$ = 25	94.0	−0.0	0.83
$N_{\texttt{NEURONS}}$ = 50	93.9	−0.0	0.91
$N_{\texttt{NEURONS}}$ = 100	93.8	−0.1	0.49
`sigmoid`	92.4	−1.5	0
`elu`	94.1	+0.2	0.38
`selu`	93.7	−0.2	0.15
`exponential`	92.8	−1.1	0
`tanh`	93.8	−0.1	0.41

Table 2.1 (continued)

Model	**Accuracy** (%)	Δ_{acc} (%)	**FDR**
`softplus`	93.3	−0.6	0
`softmax`	90.0	−3.9	0
`softsign`	93.5	−0.4	0.02
`MSE`	44.1	−49.8	0
`Poisson`	43.6	−50.3	0
`KLDivergence`	43.9	−50.0	0
`HeNormal`	93.7	−0.2	0.37
`HeUniform`	93.7	−0.2	0.21
`RandomNormal`	94.4	+0.5	0.02
`RandomUniform`	94.4	+0.5	0.01
`TruncatedNormal`	94.5	+0.6	0
`Ones`	90.0	−3.9	0
`Zeros`	42.6	−51.3	0
`GlorotNormal`	93.9	−0.0	0.79
`SGD`	46.7	−47.2	0
`RMSProp`	94.5	+0.6	0
`Adadelta`	29.4	−64.5	0
`Adagrad`	55.1	−38.8	0
`Adamax`	94.1	+0.2	0.6
`Nadam`	94.1	+0.2	0.48
`Ftrl`	34.7	−59.2	0
`Learning rate = 0.001`	89.0	−4.9	0
`Learning rate = 0.005`	93.5	−0.4	0.04
`Learning rate = 0.05`	93.6	−0.3	0.06
`Learning rate = 0.1`	94.3	+0.4	0.05
`Learning rate = 0.25`	93.1	−0.8	0.27

Table 2.1 (continued)

Model	**Accuracy** (%)	Δ_{acc} (%)	**FDR**
`Learning rate = 0.5`	88.2	−5.7	0
`Learning rate = 1.0`	69.6	−24.3	0
`Dropout rate = 0.1`	94.5	+0.6	0
`Dropout rate = 0.25`	94.6	+0.7	0
`Dropout rate = 0.33`	94.7	+0.8	0
`Dropout rate = 0.5`	94.1	+0.2	0.54
`Dropout rate = 0.75`	91.9	−2.0	0
`L1`	92.5	−1.4	0
`L2`	93.5	−0.4	0.04
`L1L2`	92.0	−1.9	0
`Orthogonal rows`	93.5	−0.4	0.02
`Orthogonal columns`	93.9	−0.0	0.62
`all` `L1`	91.8	−2.1	0
`all` `L2`	92.2	−1.7	0
`all` `L1L2`	91.4	−2.5	0

Table 2.1. MLP ablation study. Sixty-four single hyperparameter changes were conducted one at a time to explore the hyperparameter space and the effect they have on the prediction accuracy relative to the current ML setup, labelled as *current*. Accuracy (%) is the validation accuracy average across 100 random seeds. Δ_{acc} (%) represents the difference in accuracy between the variant MLP model and the current setup. Negative values result from a decrease in performance, whereas positive ones mean an improvement in classification accuracy. FDR was employed to assess the significance of these differences.

2.2.6.2 Performance evaluation

The complete dataset ($N = 293$) was split into a blind test set (1/11 = 27) and a training set (10/11 = 266). Ten repeats of a stratified 10-fold cross-validation were performed to assess the robustness of the ANN and compare it with the KNN model, as well as a baseline of the same models trained on randomly shuffled data, and completely random label assignment

($p = 0.25$). The reliability of the ANN predictions was assessed by means of a confidence score calculated as in Cuff and Barton [263], which represents how certain the MLP is of each individual prediction (Equation 2.7). The score is based on the difference between the top first and second probabilities assigned by the network to each of the classes, p_1, and p_2, respectively. For example, if the output of the network were P = [0.95, 0.02, 0.03, 0.0]. The probabilities would be sorted, so $p_1 = 0.95$, $p_2 = 0.03$ and a confidence score of 9 would be obtained.

$$\text{confidence score} = \lfloor 10 \times (p_1 - p_2) \rfloor \tag{2.7}$$

The KNN is based on distances to all training data and therefore consistently gives higher classification accuracy than the ANN model, where sites are represented by their binned RSA profile, and are thus completely unaware of other sites, and their distances to them (Figure 2.6 A). Both methods are significantly better than random. The average cross-validation accuracy across all repeats is 98, 90, 33, 31 and 24% for KNN and ANN models, their randomly trained versions and completely random label assignment, respectively. The baseline accuracy of the randomly trained models is higher than 25% since the dataset is unbalanced, with classes C1 and C2 overrepresented.

Figure 2.6 B shows the confidence of the ANN predictions across the 10 repeats of the 10-fold cross-validation. The overall accuracy is 90%. Those predictions with a confidence score greater or equal to 5 present an accuracy of 97% and cover 75% of all predictions. Finally, Figure 2.6 C shows the same two-dimensional representation of the K-means clusters identified in Figure 2.2 D and demonstrates that those binding sites with lower prediction confidence are mostly located at the borders between clusters. Sites that switch cluster labels depending on the seed are also located in these regions.

Once the model hyperparameters were optimised, 50 models were trained on 10/11 of the data (N = 266) for the 10 different seeds used to initialise the models. From a final pool of 500 models, the one presenting the highest validation accuracy and lowest validation loss was chosen, with a validation accuracy of 96%. This model, as well as KNN were used to predict on the blind test set. There was no significant difference in

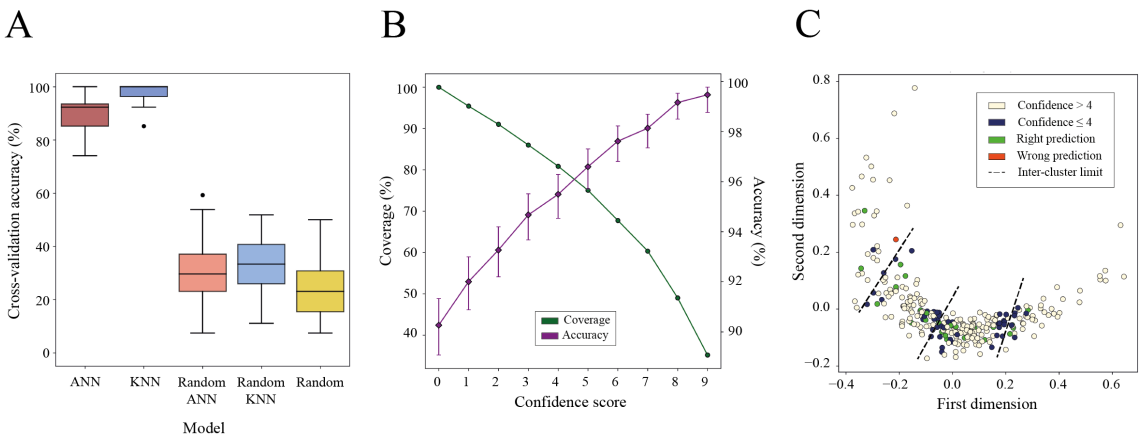

Figure 2.6. MLP cross-validation and blind test results. (A) Average accuracy of the 10-repeat 10-fold (N = 100) cross-validation of the KNN and ANN predictive models compared to a baseline of the same models trained on randomly shuffled data, as well as complete random prediction (p = 0.25). Boxes represent the central 50% of the data. Whiskers extend to 1.5 × IQR and beyond them are the outliers; **(B)** Cross-validation accuracy and proportion of binding sites against cumulative confidence score from the trained ANN. Sites with a confidence score greater or equal to 5 present an average accuracy of 97% and represent 75% of all predictions. Predictions are for the 2660 cross-validation data points: 10 different repeats of 10 distinct splits of 26-27 binding sites each. Accuracy error bars indicate 95% CI of the proportion [264]; **(C)** MDS representation of the 293 binding sites. Training data are coloured according to the average confidence of their prediction in the cross-validation. Test data are coloured according to whether they were correctly predicted or not. Dashed lines indicate the limits of K-means clusters.

performance of the ANN and KNN models. Accuracies were 96% (26/27), 95% CI = [82, 99] and 100% (27/27), 95% CI = [88, 100], for ANN and KNN, respectively. The adjusted Rand index (ARI) [265, 266], as well as adjusted mutual information (AMI) [267, 268] were calculated. ARI_{ANN} = 0.93, 95% CI = [0.81, 1.0] [269], AMI_{ANN} = 0.93, 95% CI = [0.82, 1.0]. ARI_{KNN} = 1.0, AMI_{KNN} = 1.0. 95% CI of AMI was calculated by bootstrap resampling (N = 10,000). All three metrics agreed on the high performance of the MLP. Figure 2.6 C illustrates how the binding site, which label was wrongly predicted by the ANN model, is located on the limits between adjacent clusters C3 and C4. This result agrees with the K-means clustering reliability and confidence score analysis of the cross-validation, where the same inter-cluster regions are highlighted due to their lower clustering reliability and low confidence prediction. This suggests that the core of the clusters is stable, and that the ANN confidence score may be used to identify binding sites that are at the borders of clusters.

2.2.7 Site function classification

Ligand binding sites were divided into two groups – *known function* and *unknown function* – by searching UniProt [270] for feature annotations indicative of function, e.g., metal, substrate binding, active site, etc. via the UniProt proteins API [271]. Seventeen out of the 35 proteins presented at least one UniProt annotated residue in one binding site. Manual curation using protein homology within the proteins in the dataset added 9 more functionally annotated proteins. This gave a total of 44 sites from 26 proteins classified as of known function. All other sites were classified as of unknown function.

2.2.8 Statistics and reproducibility

All data analyses were carried out primarily with the following Python libraries: NumPy [272], Pandas [273, 274] and SciPy. Keras and Scikit-learn were used for machine learning, with Matplotlib [275] and Seaborn [276] for plotting. All statistical tests performed were two-tailed and used a significance level $\alpha = 0.05$. Sample sizes and measures of significance are reported and described in the text, figures and legends.

2.2.9 Data and code availability

The software developed to carry out this analysis as well as the main summary result tables are available in the following repository: `https://github.com/bartongroup/FRAGSYS` [277].

2.3 Results

2.3.1 Defined binding sites

The focus here is on human proteins, allowing the additional information from human population variation data to be explored. Two of the 39 protein domains – products of the SARS-CoV-2 replicase polyprotein 1ab (P0DTD1) – were removed, since they did

not include any human homologues. The remaining 37 protein domains accounted for 1309 three-dimensional structures that included interactions with 1601 ligands of interest, of which 998 were unique. 293 ligand binding sites were defined across these domains, formed by 2664 unique ligand binding residues. The total number of binding sites per domain ranges from 1 to 24, with 33/37 domains presenting more than one defined binding site. The median number of sites per domain is seven.

Figure 2.7 illustrates three examples of the 37 domains for which ligand binding sites were defined by the algorithm presented in this Chapter. The grouping of the ligands into the defined sites reflects the similarity between the interaction fingerprints of the ligands with the protein.

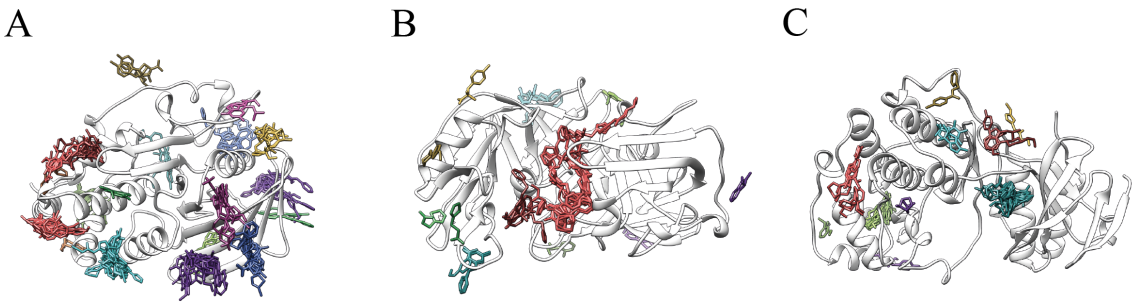

Figure 2.7. Ligand clusters defined by the binding site definition algorithm. A single protein cartoon representation, coloured in white, is shown for simplicity for each example. Ligands are coloured according to the site they bind to. **(A)** There were 110 structures depicting human tyrosine-protein phosphatase non-receptor type 1 (P18031) binding 143 ligand molecules, 104 of which were unique. 18 binding sites were defined. Cartoon PDB: 5QDU [240]; **(B)** The 68 ligands, 30 of which were unique, found across 50 structures of the chestnut blight fungus endothiapepsin (P11838) were classified into 12 distinct binding sites. Cartoon PDB: 5R1Y [278]; **(C)** For mouse mitogen-activated protein kinase 14 (P47811), 52 structures portrayed the interaction with 53 ligand molecules, 50 of which were unique, which clustered into 10 ligand binding sites. Cartoon PDB: 1LEW [279].

Figure 2.8 shows that the 293 defined binding sites are diverse in size (number of amino acids), solvent accessibility, evolutionary divergence and missense depletion. The binding site size ranges from 2-40 residues with a median of 9, while the median site RSA ranges from 4-80%, with a median of 30%. For evolutionary divergence, the average site $N_{Shenkin}$ spreads from 0-80, peaking at 40. Lastly, MES spans −0.75 to 1.0, peaking at neutrality (MES $\approx$ 0).

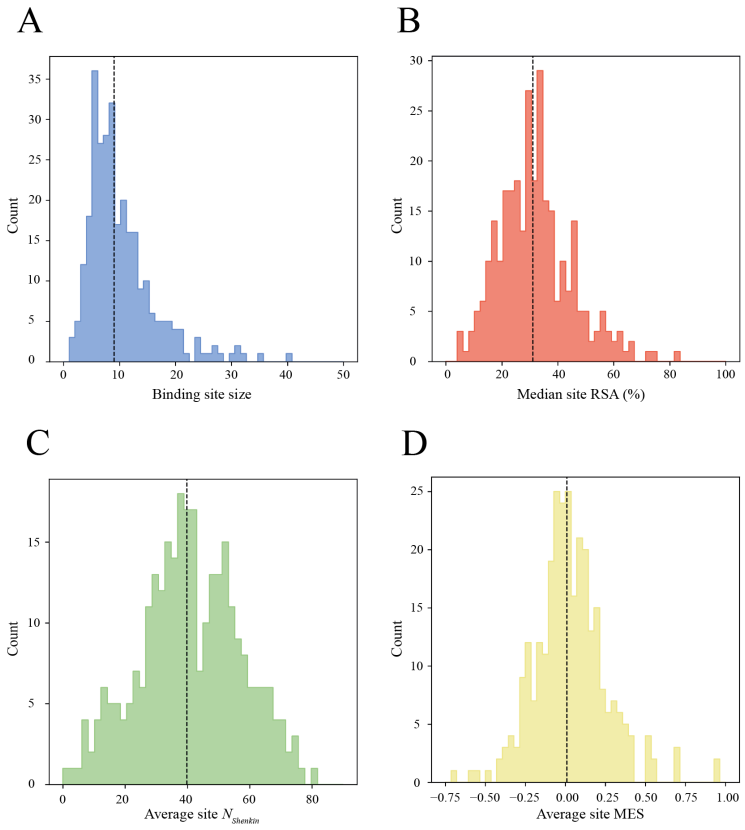

Figure 2.8. Variation in binding site features. Distribution of size **(A)**, median RSA **(B)**, average $N_{Shenkin}$ **(C)** and MES **(D)** across the 293 binding sites defined across the 37 fragment screening datasets. Black dashed lines indicate the median of each distribution.

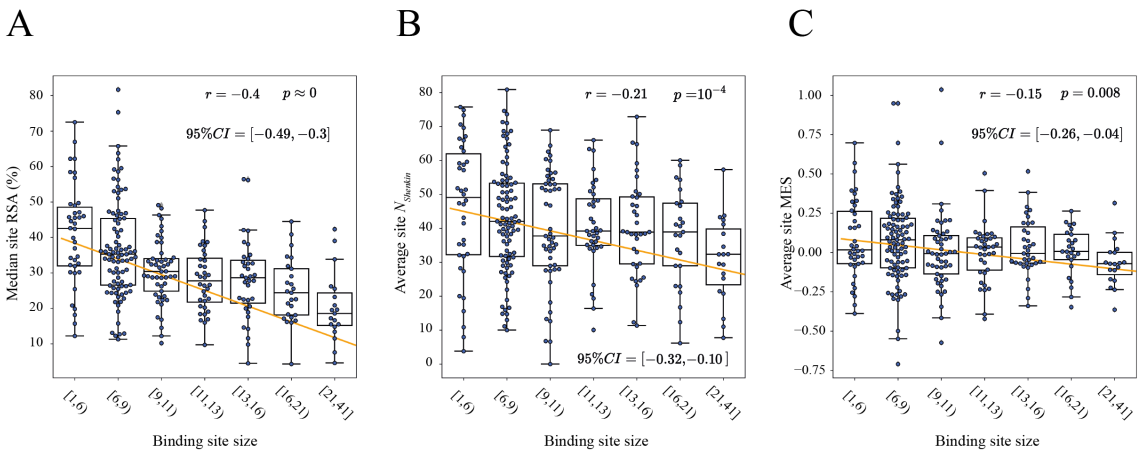

Figure 2.9. Relation between different binding site properties. A regression line was fitted to all data points prior to binning ($N = 293$ binding sites). Pearson's correlation coefficient r [280], associated p-value and 95% CI of r [281] are also shown. Data points were grouped into bins according to different binding site size intervals, represented by box and swarm plots. **(A)** Median site RSA (%) *vs* binding site size, in amino acids; **(B)** Average $N_{Shenkin}$ *vs* binding site size; **(C)** Average site MES *vs* site size. Boxes represent the IQR, whiskers extend to 1.5 × IQR and outliers are found beyond.

Despite the diversity among sites, some general trends were observed. Figure 2.9 A shows that larger binding sites tend to be less accessible to solvent ($r = -0.4$, $p \approx 0$). Figure 2.9 B illustrates that larger sites are less divergent across homologues ($r = -0.21$, $p = 10^{-4}$) while Figure 2.9 C presents how larger sites show lower enrichment in neutral missense variants within the human population, i.e., are on average more depleted in missense variants than sites of a smaller size ($r = -0.15$, $p = 0.008$). Correlations between MES and $N_{Shenkin}$, and RSA and $N_{Shenkin}$ were not significant (95% CI $r \ni 0$).

2.3.2 RSA-based binding site clustering

Figure 2.10 depicts the four clusters defined by *K*-means and the RSA profiles of the sites within them, while Figure 2.11 illustrates six binding sites from each cluster to highlight the range of binding site size and different accessibility within and across clusters. Cluster 1 includes 46 sites, 127 sites are found on C2, 91 in C3 and 29 in C4.

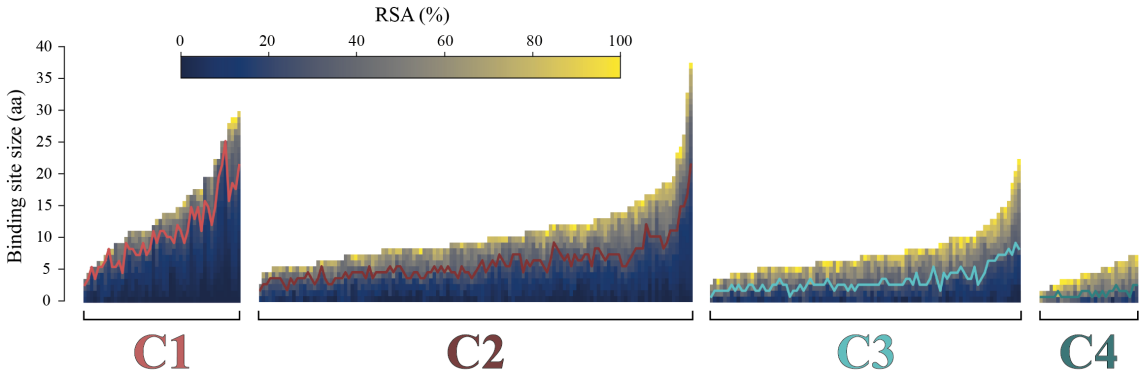

Figure 2.10. Profiles of RSA-based binding site clusters. RSA profiles of the 293 binding sites that were grouped in four clusters (C1-C4) by *K*-means based on the difference between their RSA profiles (U_D). $N_{C1} = 46$; $N_{C2} = 127$; $N_{C3} = 91$; $N_{C4} = 29$. Each binding site is represented by a vector, plotted as a bar here. The elements of the vector represent the residues that form the binding site and are sorted according to their RSA, so buried residues are at the beginning of the vector (bottom), and more accessible residues towards the end (top). Each element of the vector, or section of the bar, is coloured by RSA using the matplotlib *cividis* colour palette. Within each cluster, binding sites are sorted by their number of amino acids. Over each cluster, a line is drawn at RSA = 25% to illustrate the different burial proportion.

The proportion of residues with an RSA < 25% in Figure 2.10 follows a different profile in each cluster, which is confirmed in Figure 2.12 A. C1 is the most buried with a proportion of residues with RSA < 25% of 0.68, ($p_{buried} = 0.68$), followed by C2 with p_{buried}

= 0.47, then C3, p_{buried} = 0.30 and lastly C4 (p_{buried} = 0.10). Figure 2.12 B displays the difference in binding site size between the clusters. There is variation within clusters in site size, but certain patterns are still apparent. C1 includes the largest sites, with an average size of $\bar{s}$ = 15 residues, followed by C2 with $\bar{s}$ = 11, then C3 with $\bar{s}$ = 8 and finally C4 with $\bar{s}$ = 5. Figure 2.12 C shows the two-dimensional MDS representation of the binding sites, also shown in Figure 2.2 D and Figure 2.6 C in the Methods section of this Chapter. While there is some overlap between neighbouring clusters, C1 and C4 are significantly different. Sites near the cluster borders are those that switch groups depending on the random initialisation of the clustering. To summarise, C1 includes on average the largest, most buried sites, whereas C4 includes the smallest and most accessible. C2 and C3 are not as different as C1 and C4, but still differ in size and burial proportion with C2 including larger and overall, less accessible sites than C3.

These results support U_D as a metric that effectively quantifies the difference in solvent exposure profile across binding sites resulting in four clusters which differ not only in RSA, but also in binding site size. This effect might be explained by the negative correlation between solvent accessibility and binding site size shown in Figure 2.9.

Figure 2.12 D shows the RSA distribution across the four clusters. C1 exhibits a strikingly different distribution to the rest of clusters, peaking at RSA ≈ 5%, indicating a high density of buried residues. C2 still presents an excess of buried residues relative to C3-C4, though not as high as C1. C4 displays the most different distribution to C1, peaking at RSA ≈ 50-70%. This definition agrees with Figure 2.10 and Figure 2.12 A.

To further characterise the defined clusters, the distributions of the $N_{Shenkin}$ divergence score and MES of the residues found in the clusters were analysed. Regarding evolutionary divergence (Figure 2.12 E), C1 presents a different distribution to the rest of the clusters, with a peak at $N_{Shenkin} \approx 5$, i.e., most of the residues forming the sites within this cluster are highly conserved. The other clusters present flatter distributions with increasing proportion of divergent residues ($N_{Shenkin} > 25$) p_{C2} = 0.55, p_{C3} = 0.67, and p_{C4} = 0.69. $N_{Shenkin}$ is a divergence score ranging from 0 to 100, therefore residues with $N_{Shenkin} \leq 25$ – p_{C1} = 0.58, p_{C2} = 0.45, p_{C3} = 0.33 and p_{C4} = 0.31 – represent stronger conservation, or

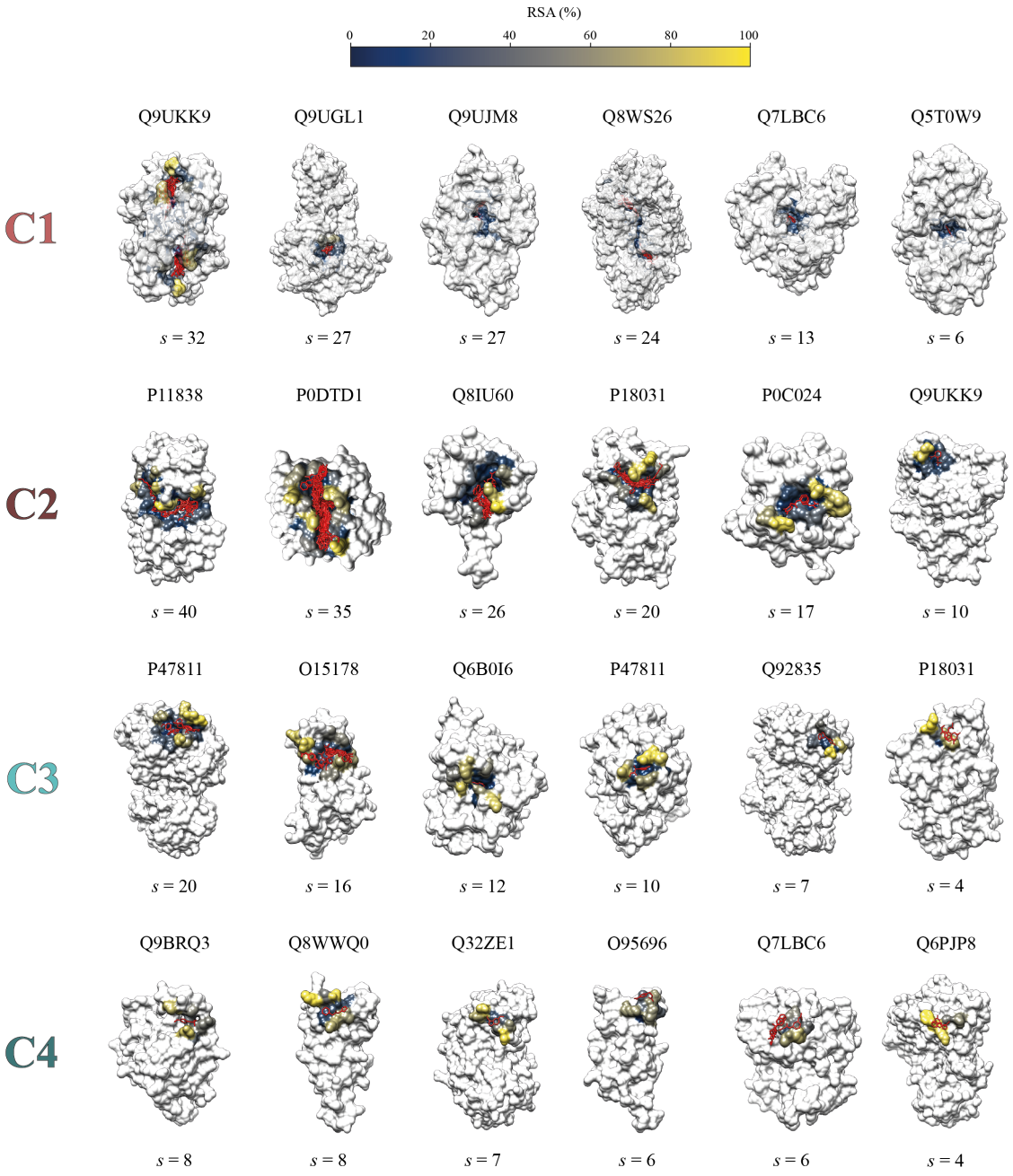

Figure 2.11. Examples of RSA-based binding site clusters. Six examples of binding sites are shown in structure for each cluster. Examples were selected to represent the range of binding site sizes within each cluster. Binding site residues are coloured by their RSA using the *cividis* colour scheme. The rest of the protein is coloured in white. Ligands binding to the site in question are coloured in red. Identifiers are UniProt accession numbers and PDB codes are provided for protein representative structures. C1 – Q9UKK9 - PDB: 5QJL [282], Q9UGL1 - PDB: 5FZ0 [283], Q9UJM8 - PDB: 5QIB [284], Q8WS26 - PDB: 5QPM [285], Q7LBC6 - PDB: 5RAN [286], Q5T0W9 - PDB: 5QHN [287]; C2 – P11838 - PDB: 5R1Y [278], P0DTD1 - PDB: 5S4B [288], Q8IU60 - PDB: 5QP9 [289], P18031 - PDB: 5QDU [240], P0C024 - PDB: 5QGI [290], Q9UKK9 - PDB: 5QJL; C3 – P47811 - PDB: 5RA5 [291], O15178 - PDB: 5QSA [292], Q6B0I6 - PDB: 5PHL [236], P47811 - PDB: 5RA5, Q92835 - PDB: 5RY0 [293], P18031 - PDB: 5QDU; C4 – Q9BRQ3 - PDB: 5RA5 [294], Q8WWQ0 - PDB: 5RJU [295], Q32ZE1 - PDB: 5RHI [296], O95696 - PDB: 5PNY [236], Q7LBC6 - PDB: 5RAN, Q6PJP8 - PDB: 5Q1Z [297].

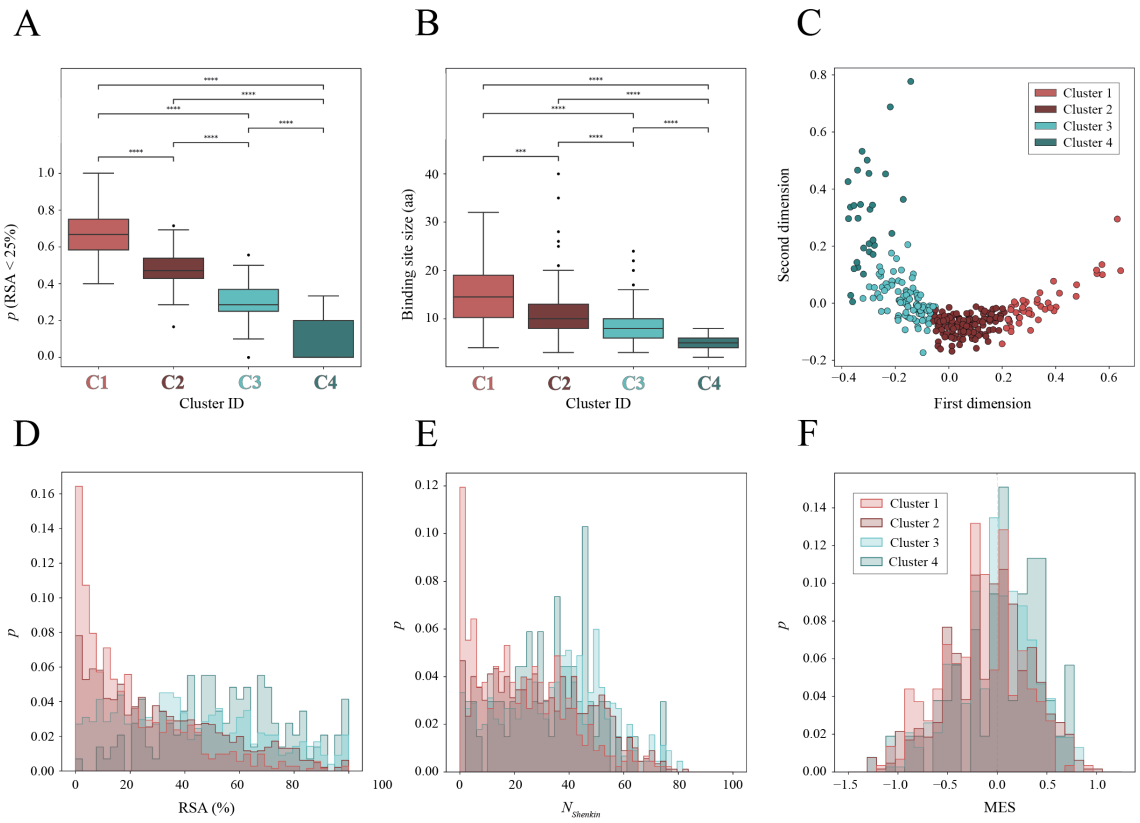

Figure 2.12. Binding site cluster features. **(A)** Box plot of the proportion of residues with RSA < 25% per binding site across the four clusters defined by *K*-means clustering; **(B)** Box plot of the binding site size, in amino acids, across clusters. Pairwise Mann-Whitney-Wilcoxon tests were performed to assess the differences between the clusters. Boxes represent the IQR and whiskers extend to 1.5 × IQR. Outliers are found beyond. *p*-value annotation legend – *ns*: $p > 0.05$, *: $0.01 < p \leq 0.05$, **: $10^{-2} < p \leq 10^{-3}$, ***: $10^{-4} < p \leq 10^{-3}$, ****: $p \leq 10^{-4}$; **(C)** MDS representation of the 293 binding sites on 2 dimensions. Data points represent binding sites and are coloured based on the cluster they group in; **(D)** Histogram of RSA (%) of the residues found within the ligand binding sites in each cluster; **(E)** Histogram of $N_{Shenkin}$ within cluster residues; **(F)** MES histogram plots for the four clusters defined.

lower divergence, than $N_{Shenkin} > 25$. This agrees with the pattern observed on the RSA distributions (Figure 2.12 D), as buried residues tend to be evolutionarily conserved [298]. In terms of missense depletion (Figure 2.12 F), the distribution of C1 is slightly shifted to the left, towards more negative values, i.e., more depleted residues, with $\overline{\text{MES}}_{C1} = -0.17$. The distributions of C2-C4 are not statistically different, but present increasing average missense enrichment scores: $\overline{\text{MES}}_{C2} = -0.07$, $\overline{\text{MES}}_{C3} = -0.02$, and $\overline{\text{MES}}_{C4} = +0.06$. Once again, this pattern agrees with the ones observed with site size, solvent accessibility and evolutionary divergence. Sites that are more buried tend to be larger, more conserved across homologues and depleted in missense variation in human.

2.3.3 Clusters predict differential functional enrichment

A key goal of the work presented in this Chapter was to identify which sites from a fragment screening experiment are most likely to be functional and so worth investigating further. Figure 2.13 shows the relative enrichment in functional sites across the four defined clusters. C1 is the most enriched in functional sites, with 17/46 sites being classed as of *known function*: OR = 4.46, $p \approx 0$. C2 was next with 21/127 (OR = 1.15, $p = 0.75$). C3 with 6/91 is depleted relative to the other clusters (OR = 0.33, $p = 0.01$) and finally C4 with 0/29: OR = 0.16, $p = 0.04$. RSA-based defined ligand binding site clusters are significantly and differentially enriched in functional sites. Based on their enrichment, a binding site found in C1 is ≈4, ≈14, and ≈28-fold more likely to be functional than a site in C2, C3 and C4, respectively.

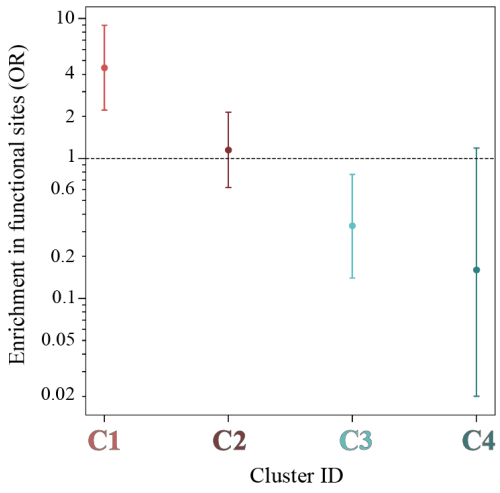

Figure 2.13. Binding site cluster enrichment in known functional sites. This enrichment score is an odds ratio (OR). Error bars indicate 95% CI of the OR. The Y-axis is in $\log_{10}$ scale. A pseudo-count of 1 was added to each cell of the contingency table to be able to calculate the score.

Functional definitions in UniProt tend to lag behind the literature, and therefore a literature search found support for 12 sites in C1 that are without UniProt annotations (Table 2.2) with two examples discussed below. No literature support was found for the remaining seventeen sites in C1 suggesting they may be novel, functionally important sites. Table 2.3 shows the full list of C1 sites that are predicted to be functionally important with 2/17 examples discussed below.

2.3.4 Example C1 site functional predictions supported by literature but not annotated in UniProt

2.3.4.1 Zika virus NS3

The Zika virus (ZIKV) genome polyprotein (Q32ZE1) is 3419 amino acids long and codes for three structural proteins: capsid (C), envelope (E) and membrane (M), as well as seven non-structural proteins: NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5. NS3 is a critical serine proteinase for viral polyprotein processing and genomic regulation. It includes a protease domain at the N-terminus and a helicase domain on the C-terminus. The helicase is responsible for RNA unwinding during replication, thus representing a promising drug target against ZIKV [299].

There are 10 sites in NS3 identified from 17 structures, with 17 unique ligands, and all are functionally unannotated in UniProt. The analysis here shows Binding Site 7 (BS7) to lie in Cluster 1 and therefore it is likely to be functional.

The site is located between domains I-III, involving residues from η2, α3 on domain I and α10, α11 on domain III as defined by Tian *et al.* [300] (Figure 2.14). Mottin *et al.* [301] predicted four RNA binding sites on NS3. One of them, the RNA exit crevice, is located between domains I–III, and involves α3 and α10 residues. Raubenolt *et al.* [302] probed four different allosteric sites on this protein. One of them, D3, was manually curated, included α11 and α12, and overlapped with BS7. Later, Durgam and Guruprasad [303] stated that four of the ten residues forming this site: Ala264, Thr265, Lys537 and Asp540 bind to RNA when in complex with NS3. These results strongly suggest that this region plays an important role in RNA binding to NS3 and therefore represents a favourable site to target to modulate function. Moreover, the site is on average missense-depleted: MES = −0.28. Ala264 ($N_{Shenkin}$ = 18, MES = −0.79), Thr267 ($N_{Shenkin}$ = 53, MES = −0.55) and Ser293 ($N_{Shenkin}$ = 72, MES = −0.48) are the three key positions out of the 10 forming this binding site, as they are all constrained within the human orthologs of this protein. Ala264 is conserved across homologues, whereas Thr267 and Ser293 are divergent, yet missense-depleted, and consequently could be relevant for binding specificity.

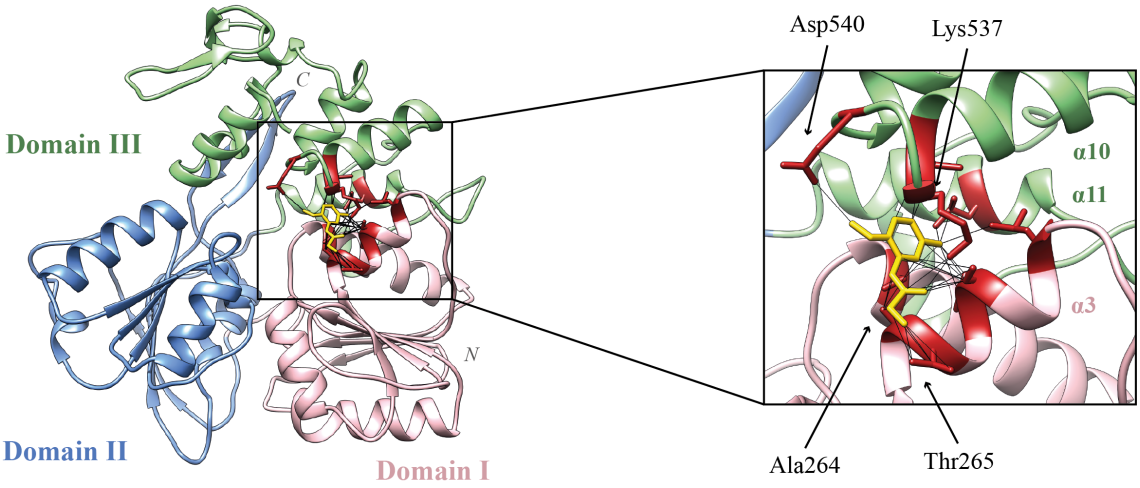

Figure 2.14. Binding Site 7 of Zika virus NS3. Non-structural protein NS3 of Zika virus (Q32ZE1) binding to N-(2-methoxy-5-methylphenyl)glycinamide (NY7) in BS7. PDB: 5RHG [304]. Domains I, II and III are coloured in pink, blue and green, respectively. Binding Site 7, which is in Cluster 1, is highlighted. The other 9 binding sites, which fall in C2 (3), C3 (3) and C4 (3), are not shown. Ligand binding residues are coloured in red and NY7 in yellow. Protein-ligand interactions are represented by black lines.

2.3.4.2 SARS-CoV-2 NSP13

The Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) replicase polyprotein 1ab (P0DTD1) is 7096 amino acids long and codes for 16 non-structural proteins [305]. NSP13 is a helicase that unwinds double-stranded RNA in the 5'-3' direction to provide a single-stranded template for viral RNA amplification [306]. NSP13 also has NTPase activity, which provides the energy for the RNA unwinding [307]. NSP13 plays a fundamental role in the replication and transcription of the SARS-CoV-2 genome and is thought to be a promising drug target against SARS-CoV-2 viral infection [308]. NSP13 has five domains. Two "RecA like" subdomains 1A and 2A, in charge of nucleotide binding and hydrolysis, as well as three other domains: an N-terminal zinc-binding domain, the helical "stalk" domain, and a beta-barrel 1B domain [309]. It is the most conserved protein across coronaviruses, with a sequence identity of >99% [310].

Twenty-four sites were defined on the surface of NSP13. Two binding sites were identified as C1: BS6 and BS16 (Figure 2.15). Visual inspection showed the two sites to be adjacent with a total of 16 residues. Three fragments bind to the site, which is located in the nucleotide and RNA binding interface of NSP13 between the 1B and 2A domains. This is

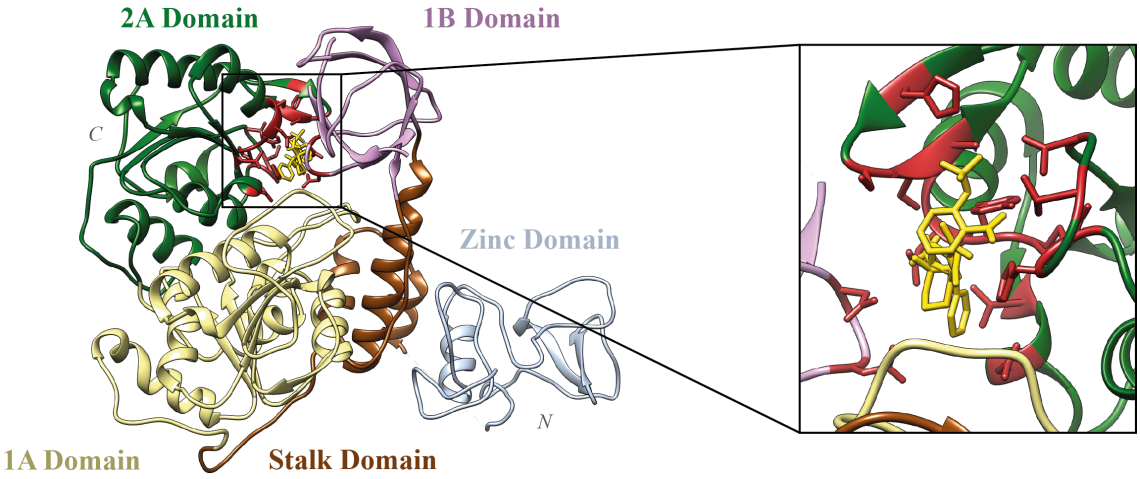

Figure 2.15. Binding Site 6+16 of SARS-CoV-2 NSP13. Non-structural protein NSP13 of SARS-CoV-2 (P0DTD1) binding to 3 ligands in BS6+16 (Cartoon PDB: 5RMH) [311]. 1A, 1B, 2A, stalk and zinc domains are coloured in yellow, pink, green, brown and grey, respectively. Ligand binding residues are coloured in red and ligands in yellow. Interactions are not shown for simplicity.

the region where the 5' end of the RNA binds [312]. This pocket is determined to be highly druggable and drugs binding to it might be effective against other coronaviruses, due to the pocket's high amino acid conservation [311]. This agrees with the results presented here, as this site has an average $N_{Shenkin} = 32$ and MES = −0.18. Of the 16 positions in this site, four show high conservation across homologues and missense depletion in human: Pro514 ($N_{Shenkin} = 30$, MES = −0.56), Asp534 ($N_{Shenkin} = 9$, MES = −0.56), Thr552 ($N_{Shenkin} = 48$, MES = −1.87) and His554 ($N_{Shenkin} = 36$, MES = −0.85). Thr552 shows highest conservation across species and lowest missense enrichment (−1.87) and so is most likely to have a key function in this protein family.

2.3.5 Examples of potentially novel C1 cluster functional predictions

2.3.5.1 Human tenascin

Human tenascin, TN, (P24821) is a hexameric extracellular matrix glycoprotein implicated in a variety of functions including cell migration, cell attachment, matrix assembly and proinflammatory cytokine synthesis [313]. TN is known to interact with viruses and play a role in viral infections, e.g., human immunodeficiency virus subtype 1 (HIV-1), and has been reported as a biomarker for disease severity [314]. It also plays a key role

in wound healing [315], is involved in diverse cardiovascular diseases [316] and in breast cancer [317]. For these reasons, there is considerable effort invested into a better understanding of the function of TN and targeting it for therapeutic effect.

The fragment screening data for TN includes 11 structures with 11 unique ligands bound. These ligands were grouped into four binding sites, none of which are annotated in UniProt. One of the four binding sites is in C1 and thus predicted to be of functional importance. The site is found on the fibrinogen C-terminal domain of the protein, which functions as a molecular recognition unit that interacts with either proteins or carbohydrates (Figure 2.16). This site shows high conservation across species ($N_{Shenkin}$ = 15), as well as missense-depletion in human (MES = −0.33). Accordingly, this site is likely to be of relevance for TN function. Among the 15 residues within the site, Val2012 ($N_{Shenkin}$ = 5, MES = −1.0), Gly2046 ($N_{Shenkin}$ = 0, MES = −0.67), Phe2047 ($N_{Shenkin}$ = 0, MES = −0.67), Trp2055 ($N_{Shenkin}$ = 0, MES = −0.54) and Gly2057 ($N_{Shenkin}$ = 0, MES = −0.83) are the most critical interacting residues and highly conserved across homologues.

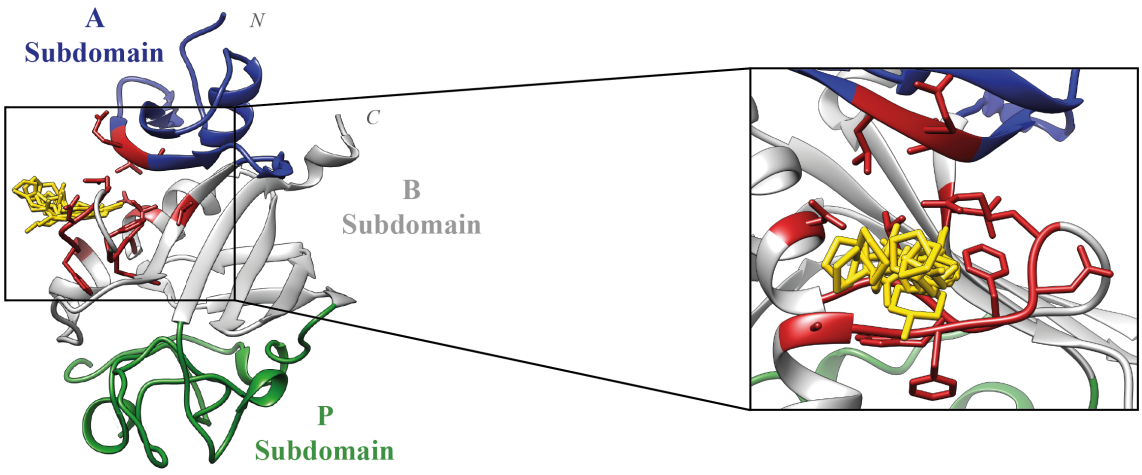

Figure 2.16. Binding Site 0 of human tenascin. Human tenascin, TN, (P24821) binding to 8 ligands in BS0. Cartoon PDB: 5R60 [318]. A, B and P subdomains as defined by Yee *et al.* [319] are coloured in blue, grey and green, respectively.

2.3.5.2 Human 5-aminolevulinate synthase

ALAS2 is a gene located on the X chromosome that codes for the human mitochondrial erythroid-specific 5-aminolevulinate synthase, ALAS-E, (P22557). This dimeric enzyme

carries out the first and rate-limiting step of the haem synthesis pathway: the pyridoxal 5'-phosphate (PLP)-dependent condensation of succinyl-CoA and glycine to form amino-laevulinic acid [320]. Across eukaryotes, these enzymes have developed extensions surrounding the catalytic core on both the N and C-termini [321]. The N-terminal extensions include the mitochondrial targeting sequence [322], whereas the C-terminal extension (C-ext) plays an autoinhibitory role by regulating substrate binding and product release [323]. Mutations affecting C-ext can result in gain-of-function, such as X-linked protoporphyria [324], as well as loss-of-function disorders, e.g., X-linked sideroblastic anaemia [325]. Accordingly, ALAS-E is a potential therapeutic target for the treatment of such diseases.

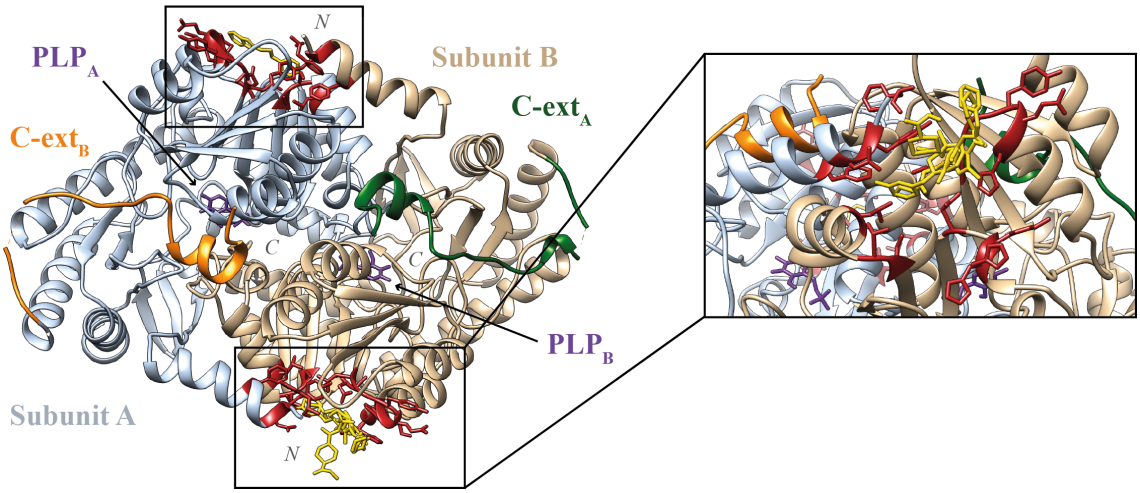

Figure 2.17. Binding Site 1 of human ALAS-E. Human erythroid-specific mitochondrial 5-amino-levulinate synthase, ALAS-E, (P22557) binding to 7 ligands in BS1. (Cartoon PDB: 5QR0 [326]). Subunits A, B, C-terminal extensions A, B, as well as PLP cofactors are coloured in grey, beige, green, orange and purple, respectively. Ligand binding residues in red, and ligands in yellow.

Thirty-three unique ligands across 25 ALAS-E structures were grouped into ten binding sites, of which only one is annotated as functional. Three sites were classified as C1. Two of these are known to be on the interface between subunits, form key interactions to maintain the assembly and are close to the PLP binding site [323]. However, one (BS1) is not mentioned in the literature. This site is located on a deep pocket at the N-terminal region of the protein structure (Figure 2.17). Residues in this site are strongly conserved and depleted in missense variation: $N_{Shenkin} = 29$, MES = −0.13. Together, this suggests the site has a functional role in the protein, perhaps as an allosteric regulator or

through interaction with a partner such as succinate-CoA ligase, SCS-α, (Q96I99) [327]. Out of the 16 residues forming the site, Lys381 ($N_{Shenkin}$ = 38, MES = −0.94) is the most missense-depleted position in the site and should be considered when growing or optimising a fragment binding to this site.

2.4 Discussion

In this Chapter, a method was presented to identify binding sites from fragment screening data and group the sites into four robust clusters by an RSA profile metric. 29/46 sites in Cluster 1 have functional support from the literature (UniProt 17/46; literature search 12/46 – Table 2.2). Seventeen further sites have similar profiles, but no evidence of their functional significance was found in the literature. Two examples from this set are shown that have compelling support for functional significance from conservation and missense depletion scores of functional significance. Furthermore, all novel potentially functional sites are listed in Table 2.3 as a resource for further experimentation on these proteins.

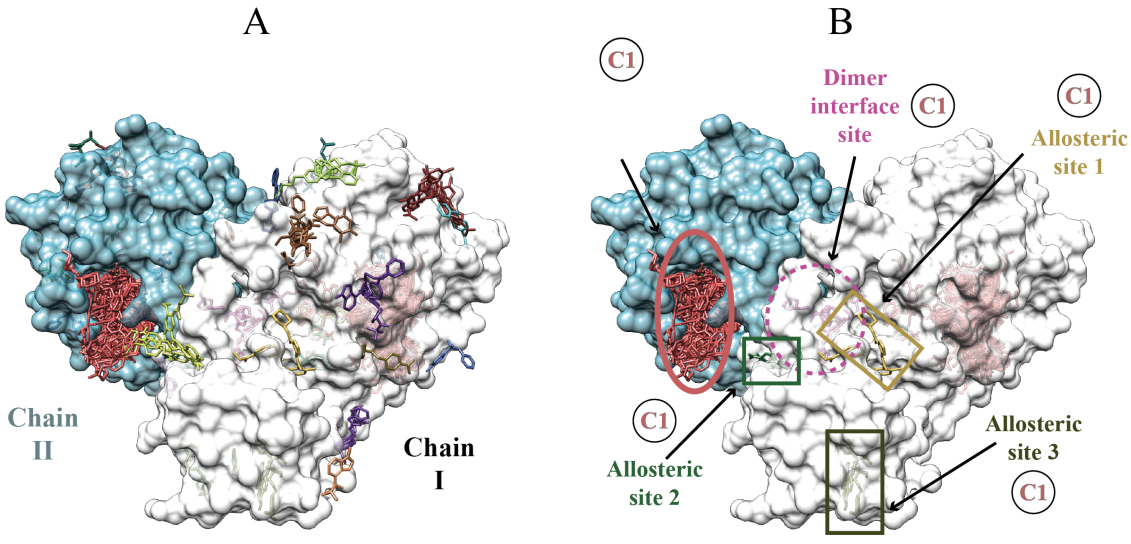

Figure 2.18. SARS-CoV-2 MPro fragment screening. **(A)** Twenty-five defined ligand binding sites on the SARS-CoV-2 main protease, MPro, (P0DTD1) from 971 ligands from 511 structures; **(B)** Five of the 9 C1 sites included the known MPro active site and four known potential allosteric sites [328, 329]. Surface PDB: 5R7Y [328].

As a case study, this method was applied to the SARS-CoV-2 main protease, MPro (P0DTD1). Twenty-five sites were defined from 511 structures, from which 8 were classed

as C1, 12 as C2, 3 as C3 and only 2 as C4. Of the 8 C1 sites, one corresponds to the active site and three to allosteric sites 1, 2 and 3 [329]. A further C1 site is located at the dimer interface and known to be a potential allosteric site [328] (Figure 2.18). The remaining three C1 sites may be important, but each binds only a single ligand and their function is currently unclear.

This Chapter focuses on a small set of proteins heavily studied by fragment screening methods. However, the method described here can be applied to classify any observed or predicted ligand binding site. Accordingly, future work should seek to classify all known ligand binding sites in the PDBe and provide tools to predict the likely functional class of sites inferred by tools such as P2Rank [115] or GRaSP [153, 330] from AlphaFold2 [67, 222] or other 3D structure models.

It is natural to focus on sites that are most likely to be of functional significance and by extension, possible targets to modulate function. However, binding sites identified here that are predicted to be least likely to have function may also be valuable as favourable locations for tagging proteins for degradation [331], phosphorylation [332], dephosphorylation [333] or other modulation [334, 335].

UniProt ID	RSA (%)	$N_{Shenkin}$	MES	*p*	# aas	# ligs	UniProt residue numbers	Literature support
Q32ZE1	17.4	38.4	−0.21	0.02	10	1	1762, 1763, 1765, 1766, 1769, 1791, 1991, 2034, 2035, 2038	RNA binding [303], RNA exit site [301], D3 site [302]
Q9Y2J2	14.6	38.2	+0.01	0.84	15	1	117, 118, 119, 203, 206, 207, 210, 231, 232, 235, 236, 253, 282, 283, 286	GPC binding [336]
Q9Y2J2	13.4	43.3	+0.02	0.7	21	4	154, 161, 162, 163, 164, 185, 186, 189, 208, 212, 217, 295, 297, 298, 299, 300, 301, 315, 375, 376, 379	Calmodulin binding [336]
Q8WS26	16.2	28.9	−0.22	0.26	19	2	105, 106, 107, 108, 109, 112, 151, 154, 155, 158, 159, 162, 170, 171, 173, 174, 175, 176, 179	IPP, DMAPP binding [337, 338]
Q8WS26	22.1	31	+0.18	0.58	8	2	308, 312, 315, 316, 320, 324, 384, 423	IPP binding [338]

Table 2.2 (continued)

UniProt ID	RSA (%)	$N_{Shenkin}$	MES	*p*	# aas	# ligs	UniProt residue numbers	Literature support
P18031	20.8	33.9	+0.05	0.48	14	1	1, 2, 3, 4, 6, 10, 19, 242, 243, 244, 245, 246, 247, 271	Conformational change [240], Cluster II [339]
P47811	17.1	55	+0.08	0	19	10	191, 192, 197, 198, 232, 236, 242, 246, 249, 250, 251, 252, 255, 259, 291, 292, 293, 294, 296	MAP insert motif, Trp197 pocket [291, 340]
Q6B0I6	15.8	41.8	+0.12	0.43	12	5	193, 224, 225, 227, 228, 239, 240, 241, 242, 243, 277, 279	Cryptic binding site [341]
P0DTD1	12.9	34.3	−0.13	0.45	12	2	5501, 5503, 5809, 5810, 5811, 5838, 5839, 5840, 5841, 5856, 5858, 5878	RNA binding [311]
P0DTD1	22.3	51.5	−0.04	0.87	9	1	5806, 5809, 5810, 5811, 5839, 5874, 5876, 5878, 5879	RNA binding [311]

Table 2.2 (continued)

UniProt ID	RSA (%)	$N_{Shenkin}$	MES	*p*	# aas	# ligs	UniProt residue numbers	Literature support
P22557	16	47.8	−0.09	0.61	16	10	148, 152, 155, 267, 268, 271, 272, 409, 413, 506, 570, 572, 573, 574, 575, 576	Dimerisation interface [323]
P22557	12.7	53.1	+0.08	0.61	7	2	271, 293, 294, 295, 296, 297, 575	Conformational change, PLP binding, succinyl-CoA inhibition [323]

Table 2.2. Literature supported C1 sites. These are 12 C1 sites with no functional annotations in UniProt, therefore labelled as *unknown function*, for which literature has been found that support their functional relevance. UniProt ID indicates the protein UniProt accession. RSA is the median site RSA (%). $N_{Shenkin}$ is the average normalised Shenkin score for the site. MES is the average missense enrichment score for the site. *p* is the *p*-value associated to the site MES. # aas is the number of residues forming the site. # ligs is the number of ligands binding to the site. UniProt residue numbers is a list of the UniProt residue numbers of the residues forming the site. Literature support contains a brief description of the literature-reported site function and references.

UniProt ID	RSA (%)	$N_{Shenkin}$	MES	*p*	# aas	# ligs	UniProt residue numbers
Q5T0W9	22.4	36.2	−0.24	0.08	12	10	149, 150, 151, 177, 233, 234, 235, 236, 270, 273, 274, 277
Q5T0W9	9.7	38.6	−0.05	0.79	12	2	125, 126, 127, 129, 229, 255, 256, 257, 272, 275, 276, 279
Q8WVM7	19.8	57.7	−0.23	0.62	5	1	285, 288, 322, 325, 326
Q15047	18.1	12.4	+0.08	0.78	18	2	295, 296, 297, 298, 300, 301, 302, 324, 328, 329, 330, 332, 333, 357, 389, 392, 393, 394
Q8WS26	19.5	57.3	−0.11	0.57	21	26	84, 87, 88, 89, 90, 214, 217, 218, 221, 222, 225, 268, 269, 273, 277, 281, 285, 290, 295, 299, 303
Q9UGL1	28.7	31.3	−0.09	0.66	10	1	53, 57, 506, 582, 583, 606, 607, 609, 610, 613
Q9UGL1	16.6	34	−0.01	1	12	3	658, 659, 662, 663, 666, 667, 670, 701, 736, 737, 738, 741
P15379	18.3	19.4	+0.09	0.63	11	1	23, 24, 40, 41, 50, 146, 148, 162, 163, 164, 165
Q9UJM8	24.3	42.8	−0.11	0.86	6	1	5, 11, 323, 327, 328, 331
Q6B0I6	21.9	36.6	−0.15	0.68	4	1	50, 209, 265, 285
Q6B0I6	12.2	26	−0.06	0.84	7	1	44, 199, 275, 276, 297, 300, 303

Table 2.3 (continued)

UniProt ID	RSA (%)	$N_{Shenkin}$	MES	*p*	# aas	# ligs	UniProt residue numbers
Q9UKK9	9.8	29.6	−0.05	0.73	15	1	65, 66, 67, 69, 75, 77, 124, 125, 145, 146, 147, 175, 200, 205, 206
Q92835	16.5	33.7	−0.05	0.78	19	46	615, 616, 617, 618, 620, 621, 622, 624, 625, 630, 631, 632, 633, 634, 635, 636, 637, 638, 674
Q92835	12.2	39.4	+0.02	0.92	12	1	560, 561, 562, 570, 571, 572, 573, 574, 578, 817, 839, 840
Q96HY7	11.6	38.5	+0.07	0.75	14	1	57, 58, 60, 61, 64, 105, 106, 107, 121, 122, 125, 126, 147, 151
P22557	17.5	40.6	+0.04	0.72	16	7	143, 145, 146, 149, 348, 349, 350, 351, 352, 353, 380, 381, 383, 402, 403, 406
P24821	14.2	24.4	−0.29	0	15	8	2010, 2011, 2012, 2025, 2045, 2046, 2047, 2048, 2049, 2050, 2054, 2055, 2056, 2057, 2060

Table 2.3. Novel C1 sites. These are 17 C1 sites with no functional annotations in UniProt, therefore labelled as *unknown function*, without any literature support. These sites therefore represent novel predicted functional sites. UniProt ID indicates the protein UniProt accession. RSA (%) is the median site RSA. $N_{Shenkin}$ is the average normalised Shenkin score for the site. MES is the average missense enrichment score for the site. *p* is the *p*-value associated to the site MES. # aas is the number of residues forming the site. # ligs is the number of ligands binding to the site. UniProt residue numbers is a list of the UniProt residue numbers of the residues forming the site.

Chapter 3

LIGYSIS-web: a resource for the analysis of protein-ligand binding sites

Preface

This Chapter refines and extends the ligand binding site definition and characterisation approach introduced in Chapter 2 to the entire PDB. LIGYSIS is a ligand binding site analysis dataset comprising biologically relevant protein-ligand interactions from 30,000 proteins with experimentally determined structures across species. Additionally, the LIGYSIS web server is presented: LIGYSIS-web. LIGYSIS-web is a resource to explore the LIGYSIS database of ligand binding sites, as well as to analyse custom user structure sets and visualise them in an interactive and dynamic way. Dr Stuart MacGowan deployed the web server as a publicly accessible service within the Dundee Resource for Sequence Analysis and Structure Prediction, making essential changes to the codebase for production. He was also responsible for the user submission and results interfaces and implemented job handling by integrating LIGYSIS-web with Slivka. All other aspects of the work presented in this Chapter were carried out by *me*.

Publications

Utgés, J.S., MacGowan, S.M., Barton, G.J. LIGYSIS-web: a resource for the analysis of protein-ligand binding sites. (*Manuscript in preparation*)

3.1 Introduction

Chapter 2 proposed a novel method for defining ligand binding sites from multiple protein-ligand complexes derived from fragment screening. Sites were then categorised by solvent accessibility into four clusters, each showing differential enrichment in known functional sites. These clusters provide a basis for ranking sites by likelihood of functional significance.

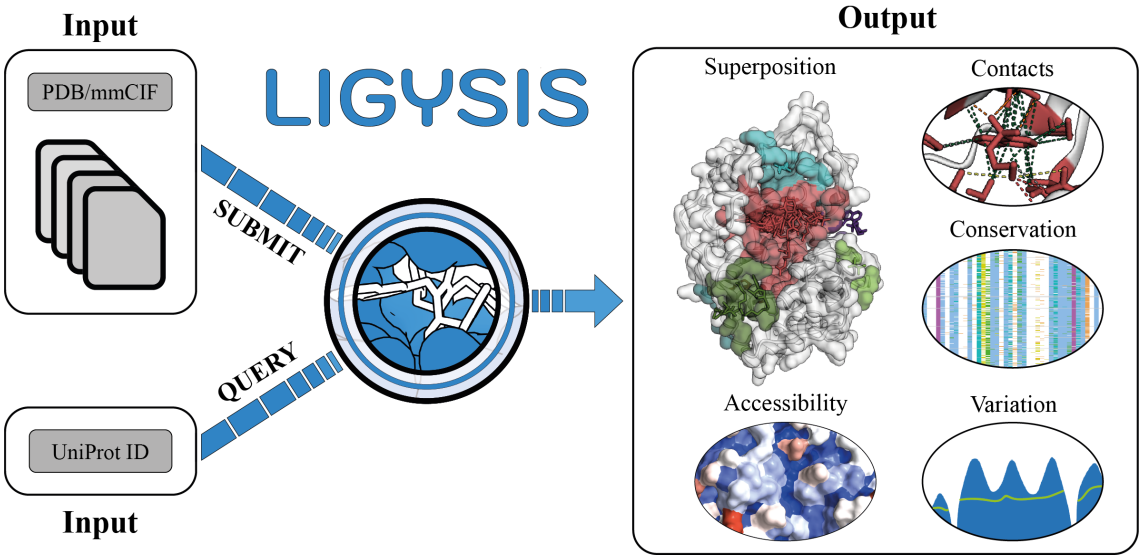

Figure 3.1. LIGYSIS-web. LIGYSIS-web is a resource for the analysis of protein-ligand binding sites. Users can query the pre-computed LIGYSIS database of 64,782 protein-ligand binding sites across 25,003 proteins in UniProt, or submit their own set of protein-ligand complexes for analysis. Results can then be explored dynamically in the LIGYSIS web server, a Python Flask web application. These results include protein-ligand contacts information, evolutionary divergence scores from multiple sequence alignments of homologous proteins, human missense enrichment scores and solvent accessibility, providing an integrated view of the likelihood of function of the defined binding sites and individual residues within them.

In this Chapter, this method is applied to the entire PDBe, resulting in the LIGYSIS dataset, which Utgés and Barton [342] employed to benchmark protein-ligand binding site prediction methods. This will be discussed in detail in Chapter 4 and Chapter 5. Addition-

ally, this Chapter introduces the ligand site analysis web server *LIGYSIS-web*, a resource for the analysis of protein-ligand binding sites (Figure 3.1). LIGYSIS-web hosts the LIGYSIS dataset, comprising 64,782 ligand binding sites, defined from 435,038 biologically relevant ligands, across 25,003 proteins with protein-ligand complexes deposited in the PDBe. Furthermore, users can submit their own structures to the LIGYSIS web server for analysis, visualise the results in a dynamic manner and download the results for further analysis. LIGYSIS-web can be found at: `https://www.compbio.dundee.ac.uk/ligysis/` [343].

3.2 Methods

3.2.1 Derivation of the LIGYSIS dataset

There are 248 million proteins in the UniProtKB database [344], ≈65,000 of which (0.02%) of which have at least one experimentally determined three-dimensional structure deposited in the Protein Data Bank (PDB) [345]. 28,997 of these proteins (45% – 0.01% of UPKB) present at least one structure in complex with a biologically relevant ligand as defined by BioLiP [228]. This corresponds to 29,657 different *structural segments* as defined on the PDBe-KB. These segments represent a UniProt sequence region with one or more structurally overlapping chains [346]. A protein might present multiple domains which are usually represented by different structural segments. Protein chains mapping to a given UniProt accession identifier were obtained from the PDBe aggregated API endpoint: `uniprot/superposition/` [174, 347]. Transformation matrices to superpose protein chains for a structural segment were downloaded from the PDBe FTP site [348, 349] and used to structurally align all chains mapping to each segment with BioPython [350].

Preferred biological assemblies, as defined by PISA [237], were downloaded from PDBe via ProIntVar [238]. Protein-ligand contacts were determined with pdbe-arpeggio [239]. Figure 3.2 illustrates the ligand site definition approach used to obtain the new dataset presented here: LIGYSIS. This approach is an extension of the one used in Chap-

ter 2. For a pair of ligands, L_A, L_B, fingerprints A, B are defined as sets containing the UniProt residue numbers of the amino acids interacting with each ligand. PDB residues were cross-referenced to UniProt using the SIFTS mapping present in the macromolecular crystallographic information files (mmCIF) located under the `_atom_site.pdbx_sifts_xref_db` fields [351, 352]. Relative intersection, I_{rel}, (Equation 2.1) is a similarity metric that quantifies how similar these fingerprints are [353]. Subtracting I_{rel} from 1 gives a distance, D (Equation 3.1), which takes the value of 0 when A and B share all the binding residues and 1 when they share none. For a given protein segment, interacting with M biologically meaningful ligands across N chains, ligand fingerprints were clustered using average linkage with SciPy [249] and ligand sites obtained by cutting the tree at $D = 0.5$. pdbe-arpeggio currently does not support multiple occupancy atoms, and so ligands with atom occupancy $\neq 1$ were not considered in this analysis. This affected 6256 structures (6% of LIGYSIS dataset). 64,782 sites were defined from 435,038 ligands across 26,260 structural segments. These segments mapped to 25,003 different proteins, which were represented by 104,456 structures. The original LIGYSIS pipeline is summarised in Figure 3.3.

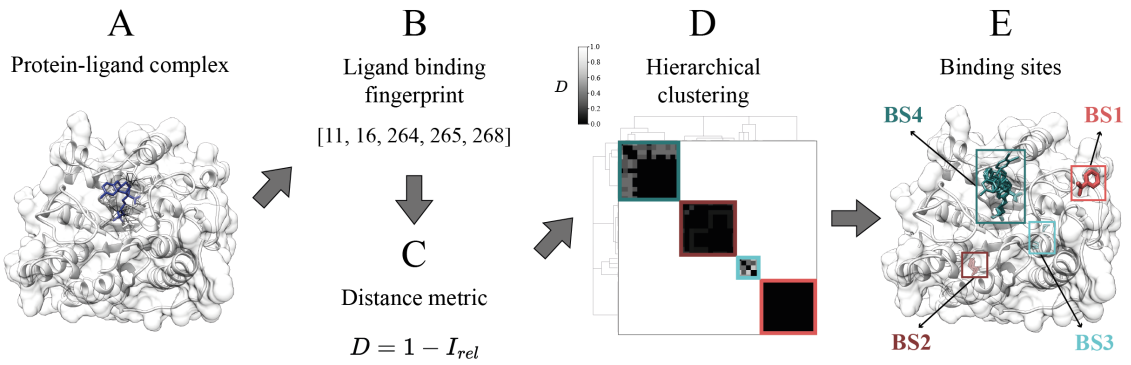

Figure 3.2. LIGYSIS ligand binding site definition algorithm. For a given protein-ligand interaction complex **(A)**, a ligand binding fingerprint **(B)** was obtained as the set of unique UniProt sequence residue numbers interacting with the ligand as defined by pdbe-arpeggio; **(C)** Fingerprints from different ligands binding to the same protein were compared and a distance calculated; **(D)** This distance was employed to perform hierarchical clustering, which grouped the different ligands in distinct clusters binding to the same region of the protein or binding site; **(E)**. The example here is human arginase-2, mitochondrial (P78540) represented by PDB: 4IXV [354] with XA1 bound.

$$D = 1 - I_{rel} \tag{3.1}$$

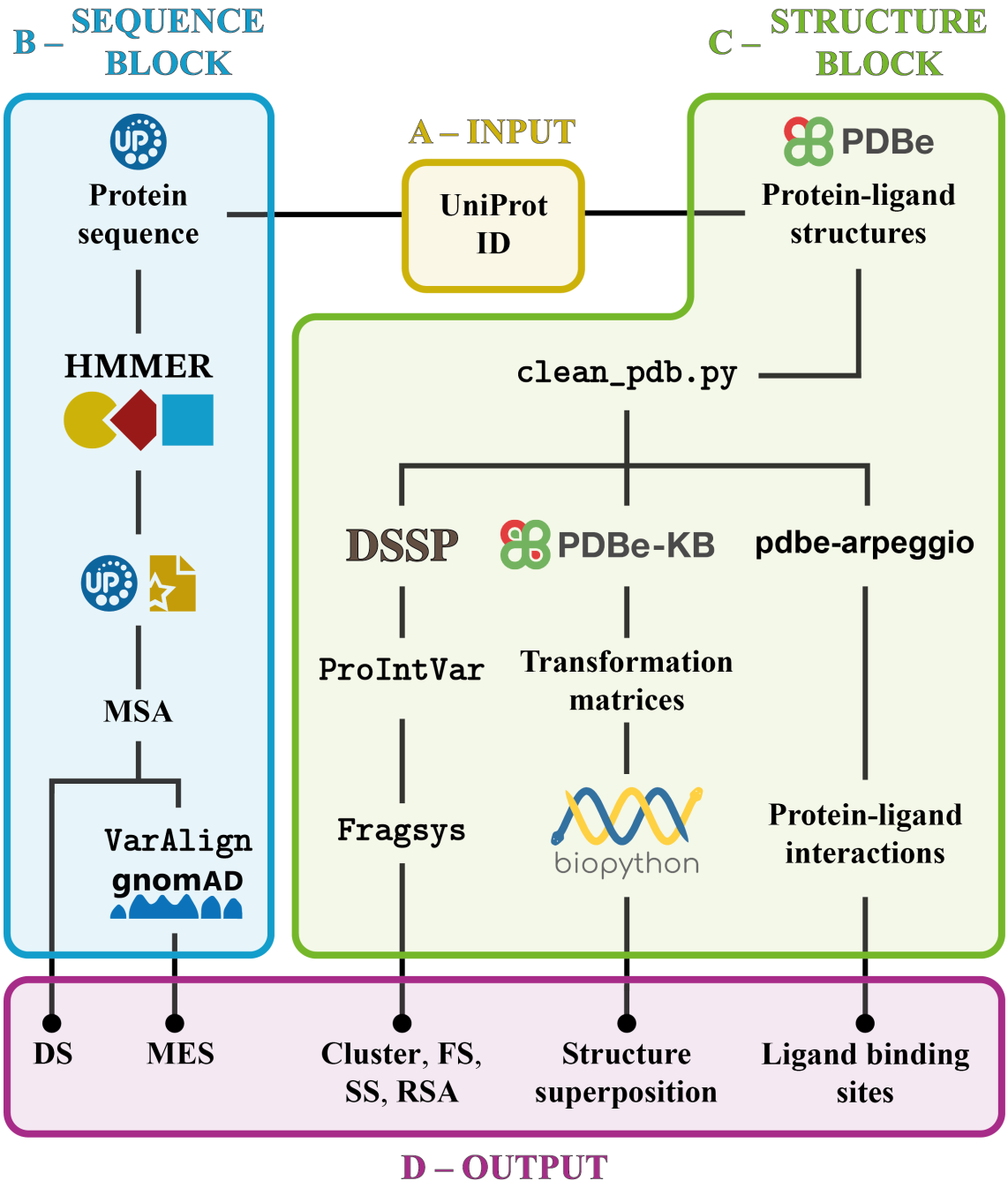

Figure 3.3. LIGYSIS original pipeline. Flow diagram of the LIGYSIS original pipeline. The only required input is a UniProt ID **(A)**. The pipeline can be divided into two parts: sequence and structure blocks. The sequence block extracts the protein sequence from UniProt, searches for homologues on SwissProt with HMMER and builds and MSA. Divergence scores (DS) are computed from this MSA. Missense enrichment scores (MES) are also calculated after extracting variants from gnomAD with VarAlign **(B)**. The structure block retrieves ligand-binding structures from the PDBe, runs DSSP via ProIntVar to obtain secondary structure (SS) and relative solvent accessibility (RSA) data. The MLP described in Chapter 2 (FRAGSYS) is employed to obtain an RSA-cluster label and functional score (FS). Structural superposition is carried out with BioPython using PDBe-KB transformation matrices and binding sites defined from protein-ligand contacts calculated with pdbe-arpeggio **(C)**. All these results are merged in the output result tables **(D)**.

3.2.2 Alignments and variants

The canonical sequence for each UniProt accession was used to perform a homologue sequence search in SwissProt [244]. jackHMMER [245] was employed with three iterations to generate a multiple sequence alignment (MSA). Amino acid divergence was quantified with the normalised Shenkin [51] divergence score $N_{Shenkin}$ [44]. Genetic missense variants mapping to human sequences in the MSA were retrieved from gnomAD [246] using VarAlign [200]. Missense enrichment scores (MES), i.e., odds ratio (OR), were calculated for alignment columns [203] and 95% confidence intervals and *p*-values used to evaluate their significance [247].

3.2.3 RSA-based clustering and score

Accessible surface area was calculated by DSSP [112] via ProIntVar [238] and normalised [108] to relative solvent accessibility (RSA). The Keras [261] multilayer perceptron (MLP) [260] described in Chapter 2 [353] was employed to predict site RSA-based cluster labels: C1 – C4. These clusters are differentially enriched in functional sites annotated in UniProt [271] ($OR_{C1} \approx 28 \times OR_{C4}$).

For a binding site i, a functional score FS_i is calculated with Equation 3.2 as the dot product of the P_i and F vectors. In this equation, p_{i_j} represents the probability of site i belonging in Cluster j and f_j denotes the proportion of known annotated functional sites within Cluster j. The probabilities p_{i_j} are derived from the vector P_i (Equation 3.3), which is returned by the MLP and provides the probabilities of a site i belonging to each class. The vector F (Equation 3.4) contains the proportions of functional sites in each cluster, which were determined through hierarchical clustering and functional classification of the human subset of the LIGYSIS dataset, composed of 13,000 sites across 3500 proteins. Both the functional label and score serve as metrics that indicate the likelihood of a binding site being functional and can be used to rank binding sites within a protein.

$$FS_i = P_i \cdot F = \sum_{j=1}^{4} p_{i_j} f_j \tag{3.2}$$

$$P_i = [\, p_{i_1},\, p_{i_2},\, p_{i_3},\, p_{i_4} \,] \tag{3.3}$$

$$F = [\, f_1,\, f_2,\, f_3,\, f_4 \,] = [\, 0.52,\, 0.18,\, 0.05,\, 0.04 \,] \tag{3.4}$$

3.2.4 LIGYSIS customised pipeline

The LIGYSIS customised pipeline is used for the analysis of user-submitted structures to the server. The customised pipeline does not rely on the PDBe-KB for mapping chains to a protein, nor transformation matrices. Instead, STAMP [242] is used to superpose the uploaded structures, which can be in PDB (*.ent*, *.pdb*) or mmCIF (*.cif*) format (Figure 3.4). However, all structures in the set must be in the same format, e.g., either all PDB or all mmCIF. Structures must present the same number of protein chains, either all monomers, dimers, trimers, etc. The LIGYSIS web server currently supports only homomeric protein-ligand complexes, i.e., complexes between N copies of a single protein sequence and any number of ligands. For structures mapping to a protein in UniProt, it is required to submit the corresponding UniProt identifier, so residues across different structures can be mapped to the same reference. If the submitted structures represent a protein not in UniProt, this field can be left blank. The residue numbering employed will be from the structure, so structures are expected to present the same numbering scheme.

3.2.5 Server architecture

The LIGYSIS web server is implemented using the Python Flask framework [355], with dynamic content rendered via Jinja templates [356] and data transferred through Flask routes. The frontend is structured using HTML for content and layout and a combination of Bootstrap [357] and plain CSS [358] for styling. JavaScript [359] enhances interactivity by integrating a 3Dmol.js structure viewer [360, 361] for molecular visualisation, interactive tables and Chart.js graphs [362]. jQuery [363] and AJAX [364] requests facilitate asynch-

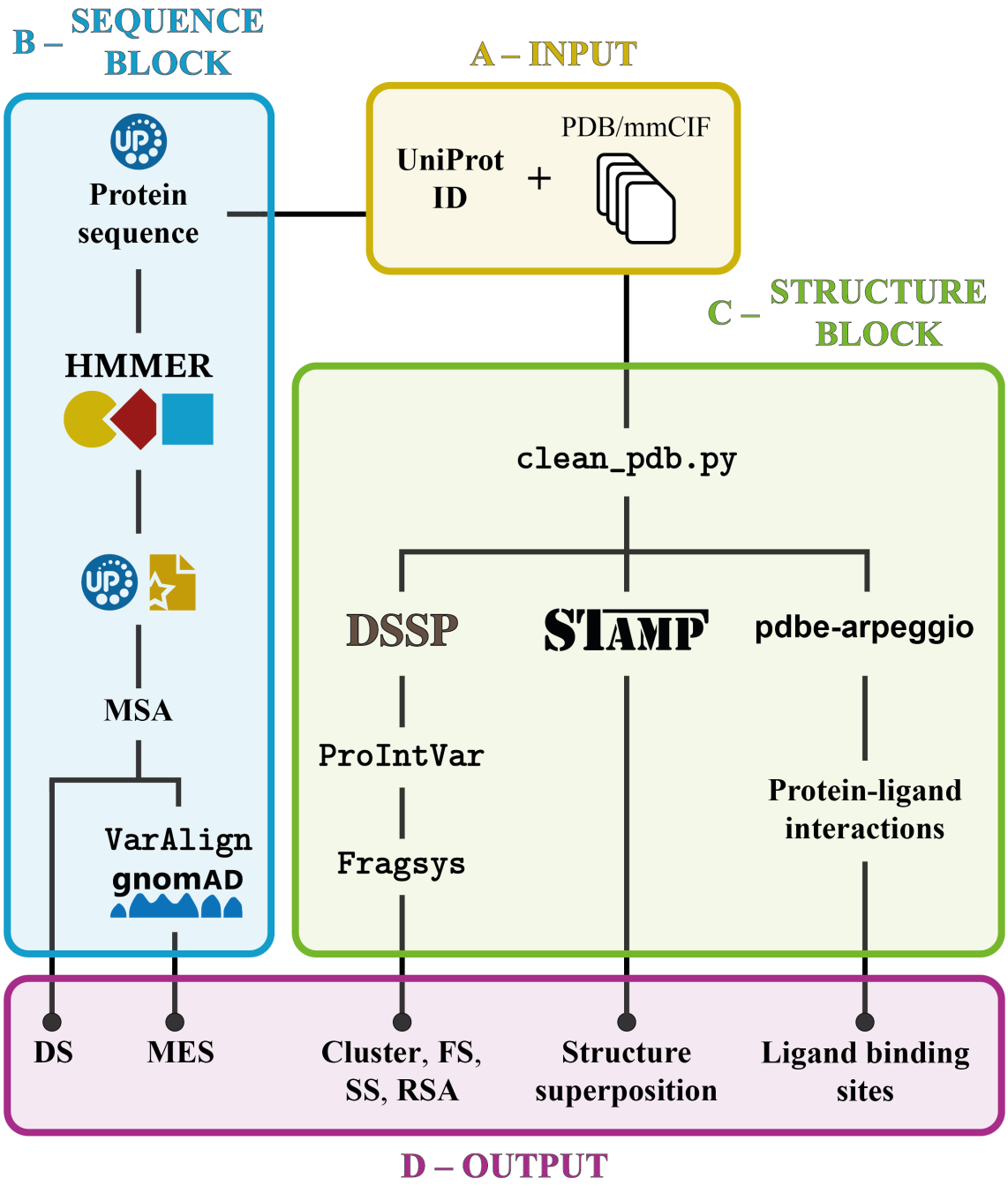

Figure 3.4. LIGYSIS customised pipeline. Flow diagram of the LIGYSIS customised pipeline. The input needed is a UniProt ID and a set of PDB/mmCIF structures **(A)**. The pipeline can be divided into two parts: sequence and structure blocks. The sequence block extracts the protein sequence from UniProt, searches for homologues on SwissProt with HMMER and builds and MSA. Divergence scores (DS) are computed from this MSA. Missense enrichment scores (MES) are also calculated after extracting variants from gnomAD with VarAlign **(B)**. The structure block runs DSSP via ProIntVar to obtain secondary structure (SS) and relative solvent accessibility (RSA) data. The MLP described in Chapter 2 (FRAGSYS) is employed to obtain an RSA-cluster label and functional score (FS). Structural superposition is carried out with STAMP and binding sites defined from protein-ligand contacts calculated with pdbe-arpeggio **(C)**. All these results are merged in the output result tables **(D)**.

ronous updates, enabling dynamic data exchange between the client and server. Illustrative examples of each of these components can be found in Code Block 3.1 (HTML), Code Block 3.2 (JavaScript), Code Block 3.3 (CSS) and Code Block 3.4 (Python). The LIGYSIS web server can be accessed through: https://www.compbio.dundee.ac.uk/ligysis/ [343].

```html
<div
  class="col-auto justify-content-end"
  style="padding-left: 0px; padding-right: 0px;">
  <button
    id="saveAllArpeggioDataButton"
    style="border: 1px solid black; color: black; display: flex;
    align-items: center; border-radius: 5px; padding: 5px 10px;"
    onclick="saveAllAssembliesContactData()">
    <img
      src="/static/images/download.svg"
      style="width: 1.5rem; height: 1.5rem; margin-right: 8px;">
     <b>ALL</b> Assemblies Contacts
  </button>
</div>
```

Code Block 3.1. HTML `saveAllArpeggioDataButton` download button. This HTML element contains a button (`saveAllArpeggioDataButton`) that downloads the pdbe-arpeggio protein-ligand contacts for all biological assemblies mapping to a protein segment. The JavaScript `saveAllAssembliesContactData` function is called upon clicking. A combination of Bootstrap classes and custom CSS is used for styling.

```javascript
function saveImage(canvasId, filename) {
  var canvas = document.getElementById(canvasId);
  var link = document.createElement('a');
  link.href = canvas.toDataURL('image/png', 1);
  link.download = `${filename}.png`;
  link.click();
}
```

Code Block 3.2. JavaScript `saveImage` function. `saveImage` JavaScript function as implemented in LIGYSIS-web. This function takes an HTML canvas element identifier (`canvasId`) and a file name (`filename`) and saves a screenshot as a PNG file. It is used to save images from the Chart.js graphs as well as the 3Dmol.js viewer.

```css
.spinner-overlay {
    position: absolute;
    top: 0;
    left: 0;
    width: 100%;
    height: 100%;
    background: rgba(255, 255, 255, 0.8);
    display: flex;
    justify-content: center;
    align-items: center;
    z-index: 1;
    box-sizing: border-box;
}
```

Code Block 3.3. CSS `spinner-overlay` class. Attributes of the `spinner-overlay` CSS class. This class is used to overlay a transparent white background to sit behind the spinner wheel that is shown while LIGYSIS-web is loading structures, reading files or performing calculations.

```python
@app.route('/get-contacts', methods = ['POST'])
def get_contacts():
    data = request.json
    active_model = data['modelData']
    prot_id = data['proteinId']
    seg_id = data['segmentId']
    (...)
    response_data = {
        'contacts': json_cons,
        'ligands': struc_ligs_data,
        'protein': struc_prot_data,
    }
    return jsonify(response_data)
```

Code Block 3.4. Python Flask `/get-contacts` route. Simplified `/get-contacts` Python Flask route implemented in LIGYSIS-web. This route retrieves the PDB code (`modelData`) for a given structure depicting the interaction between one or multiple ligands and a protein mapping to a UniProt accession (`proteinId`) and a PDBe-KB segment (`segmentId`). Then, it reads a series of files, including pdbe-arpeggio output, processes them and returns the data in a suitable format to draw the relevant protein-ligand interactions in the 3Dmol.js viewer.

User job submission is handled through Slivka-bio v0.8.3 [365, 366] and jobs run on the School of Life Sciences, University of Dundee HPC infrastructure. Once the job

execution has finished, result files are served to the client and displayed in the same way as the LIGYSIS database entries.

3.2.6 Data Availability

The code for the LIGYSIS-web Python Flask application can be found in this GitHub repository: https://github.com/bartongroup/LIGYSIS-web [367]. The code for the LIGYSIS pipeline, employed to generate the LIGYSIS dataset, which LIGYSIS-web explores can be found here: https://github.com/bartongroup/LIGYSIS [368]. The code for the LIGYSIS pipeline adapted to handle user-submitted jobs can be found in its repository: https://github.com/bartongroup/LIGYSIS-custom [369]. Source code for Slivka-bio can be found on the project repository: https://github.com/bartongroup/slivka-bio [238]. LIGYSIS-web is available at: https://www.compbio.dundee.ac.uk/ligysis/ [343].

3.3 LIGYSIS-web

3.3.1 LIGYSIS-web results page

The LIGYSIS web server can be used in two modes: (1) to explore the LIGYSIS dataset and (2) to submit a set of structures for analysis in the LIGYSIS customised pipeline. The LIGYSIS dataset comprises ≈25,000 proteins with deposited structures of proteins bound to biologically relevant ligands on the PDBe. Results can be explored by searching for a UniProt accession, entry, or name of the protein of interest. User job results can be accessed through the link provided when submitting the job or through the jobs table in the user session tab, and explored just like pre-computed entries.

Figure 3.5 illustrates the LIGYSIS-web results page. This page is divided into three panels: *Binding Sites*, *Structure* and *Binding Residues* panels. At the top of the panels, general information about the target protein can be found: UniProt accession (link to UniProt), entry, protein names and number of chains, ligands and binding sites.

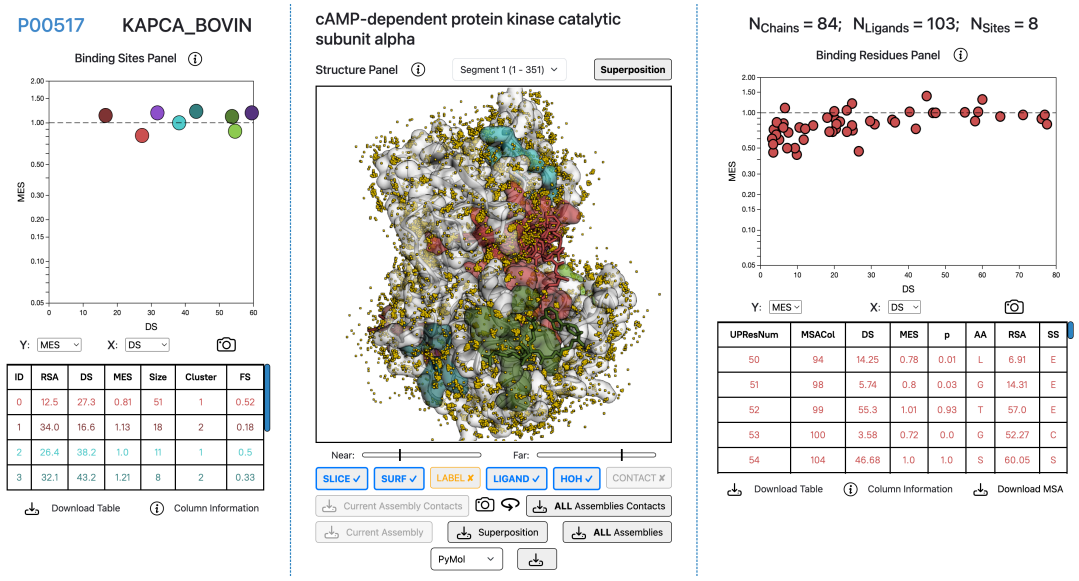

Figure 3.5. LIGYSIS-web results page. The results page is divided into three panels: *Binding Sites* (left), *Structure* (centre), and *Binding Residues* (right) panels. In this example, 103 ligands across 84 structures have been clustered into 8 different ligand binding sites for bovine cAMP-dependent protein kinase catalytic subunit alpha, PKA C-alpha (P00517). This is the only segment of PKA C-alpha and covers its whole sequence (1-351). Chart.js scatter points and table rows represent binding sites and their average features on the Binding Sites Panel, whereas they represent individual amino acid residues on the Binding Residues Panel. Both panels interact with the 3Dmol.js central structure viewer through hover and click events. Below this viewer, function buttons can be found to hide/show slab controls, surfaces, labels, ligands, water molecules and protein-ligand contacts. Structures and contact data can also be downloaded from these buttons. Surfaces, ligands and water molecules are displayed in this screenshot. Superposition representative PDB: 1SVH [370], chain: A.

3.3.1.1 Binding Sites panel

Figure 3.6 depicts the Binding Sites Panel of the results page. This panel is formed by a dynamic Chart.js canvas and a table. Both are displaying the mean binding site features, calculated from averaging the features of the residues forming the site. These features are relative solvent accessibility (RSA), normalised Shenkin divergence score (DS), missense enrichment score (MES), the size of the site, i.e., number of amino acid residues, the RSA-derived cluster label and its associated functional score (FS). The table rows can be sorted by any of these variables and the axes of the chart dynamically changed. MES uses a logarithmic ($\log_{10}$) scale, as it is an odds ratio. The chart as well as the table are linked to the structure viewer by hover and click events. Hovering on a data point/row will temporarily display the side chains of the site residues, whereas clicking on it would

fix them in the view until another site is clicked or the current one unclicked. Refer to Figure 3.6 legend for more details on how to interpret the MES *vs* DS graph.

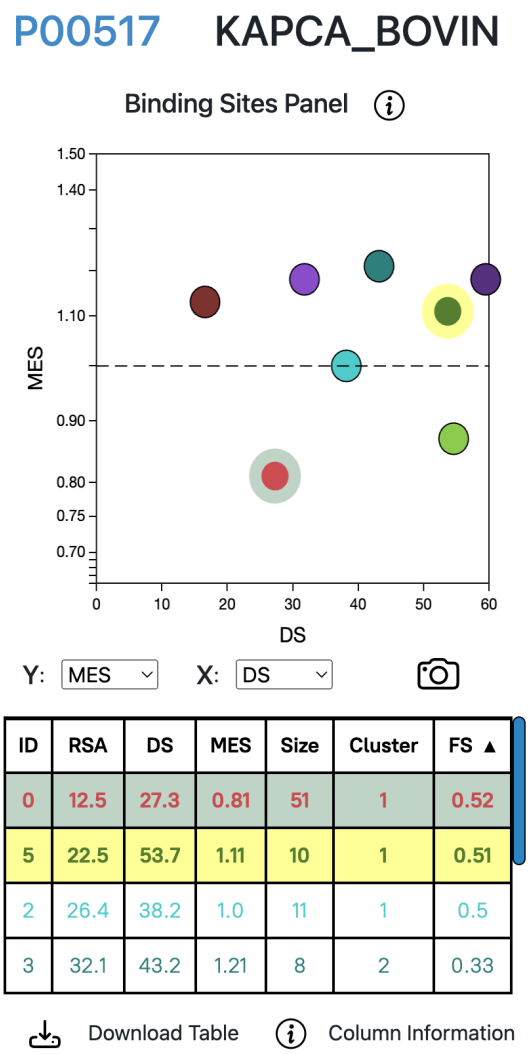

Figure 3.6. LIGYSIS-web results page Binding Sites Panel. UniProt accession identifier and entry name for cAMP-dependent protein kinase catalytic subunit alpha (P00517) at the top. Below, Chart.js scatter of MES ($\log_{10}$ scale) *vs* divergence score (DS) for the 8 binding sites defined for PKA C-alpha. The dashed line indicates neutrality, i.e., missense variation within the site is no different than in the rest of the protein. Binding Site 2 (cyan) is an example of this. In contrast, Binding Site 0 (pastel red), the known active site of the kinase, is depleted in missense variation (MES = 0.81). Additionally, it is buried (RSA = 12%), conserved across homologues (DS = 27/100) and presents a high functional score (FS = 0.52). Binding sites with low divergence and missense enrichment scores (lower-left quadrant) are most likely to be functional, whereas those at the top-right are least likely to have biological effect when bound to a ligand, e.g., BS6 (purple). Binding sites table is sorted by functional score (FS). Binding Site 0 is clicked (green highlight) and Binding Site 5 is being hovered over (yellow highlight).

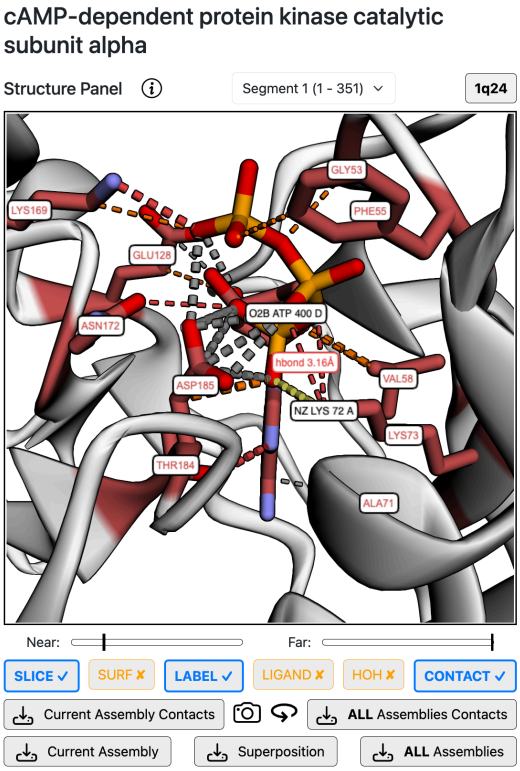

Figure 3.7. LIGYSIS-web results page Structure Panel. Structure Panel of the LIGYSIS-web results page. Preferred biological assembly of PKA C-alpha (P00517) PDB: 1Q24 [371] with displayed interactions with adenosine tri-phosphate (ATP). Water molecules and ligand-binding residue labels are displayed. Interaction label between NZ atom of Lys73, chain: A, and O2B of ATP 400, chain: D, is displayed on hover. This is a hydrogen bond and the distance between the atoms is of 3.16 Å. Residues in contact with multiple ligand atoms, as Lys73, which forms hydrogen bond and ionic interactions with three atoms, or Asp185, which interacts with four ATP atoms, are likely to be more relevant for the binding mode than residues with a single contribution to the binding interface, as Ala71. Non-carbon atoms are coloured based on the Jmol colouring scheme [372].

3.3.1.2 Structure panel

Figure 3.7 depicts the central Structure Panel of the results page. This panel is divided into three parts: the main 3Dmol.js viewer, in the centre, segment and structure selectors above it, and the structure buttons below it. The Segment selector is a drop-down menu showing the segment identifier and its protein sequence coordinates. The Structure selector is a drop-up menu that allows the user to swap between the Superposition view (default) and individual biological assemblies. The Superposition view consists of a white cartoon representation of a segment representative chain and the superposed ligands coloured by binding site. These are hidden to start with and can be displayed by clicking

on the "LIGAND" button and hidden if clicked again. The "SURF" button will display the protein chain surface, which will be white for non-ligand binding residues and coloured for binding residues. The same applies to water molecules and the "HOH" button. To see the labels of a clicked binding site, click on the "LABEL" button, and click again to hide them. The "CONTACT" button is disabled for the superposition view but can be clicked when exploring an assembly or structure. The relevant ligands, as well as the residues interacting with them will be coloured based on their site, and dashed cylinders depicting the protein-ligand interactions calculated by pdbe-arpeggio will be displayed. They are coloured based on the Arpeggio colour scheme, e.g., green for hydrophobic, red for polar and yellow for ionic [239]. The width of these cylinders is representative of the distance between the atoms. Thicker cylinders denote a clash between the Van der Waals (VDW) radii of the atoms (closer) whilst thinner cylinders indicate interaction between the VDW radii of the atoms. The view can also be sliced between two planes to focus on a region of interest. The slab or slice controls are displayed with the "SLICE" button. Hiding them does not reset the slab, just hides the controls. Clicking on the "SLICE" button again will show the controls again and allow for further slab adjustment. Clicking on the circular arrow generates a spin animation on the Y axis. Clicking on any ligand atom on the structure viewer links to the new PDBe-KB ligand pages [373] providing enhanced ligand annotations and a holistic view of small molecules for their biological context.

3.3.1.3 Binding Residues panel

Figure 3.8 illustrates the the Binding Residues Panel. This panel is similar to the Binding Sites Panel and also comprises a Chart.js graph and an interactive table. However, data points and table rows correspond to individual binding site residues and not the site as a whole, as they did in the Binding Sites Panel. The table displays the UniProt residue number (UPResNum), the column in the multiple sequence alignment (MSACol), the $N_{Shenkin}$ divergence score (DS), missense enrichment score (MES) and associated p-value, amino acid name (AA), relative solvent accessibility (RSA) and secondary structure (SS). SS corresponds to the original 8-state DSSP classification: 3_{10} helix (G), α-helix (H), π-helix

(I), β-sheet (E), β-bridge (B), helix turn (T), bend (S) and coil (C) [112]. Variables on the chart axes can also be changed, and table rows sorted by column. Currently, only hover events are supported for this panel, clicking chart points or table rows has no effect. Refer to Figure 3.8 legend for more details on how to interpret the MES *vs* DS graph.

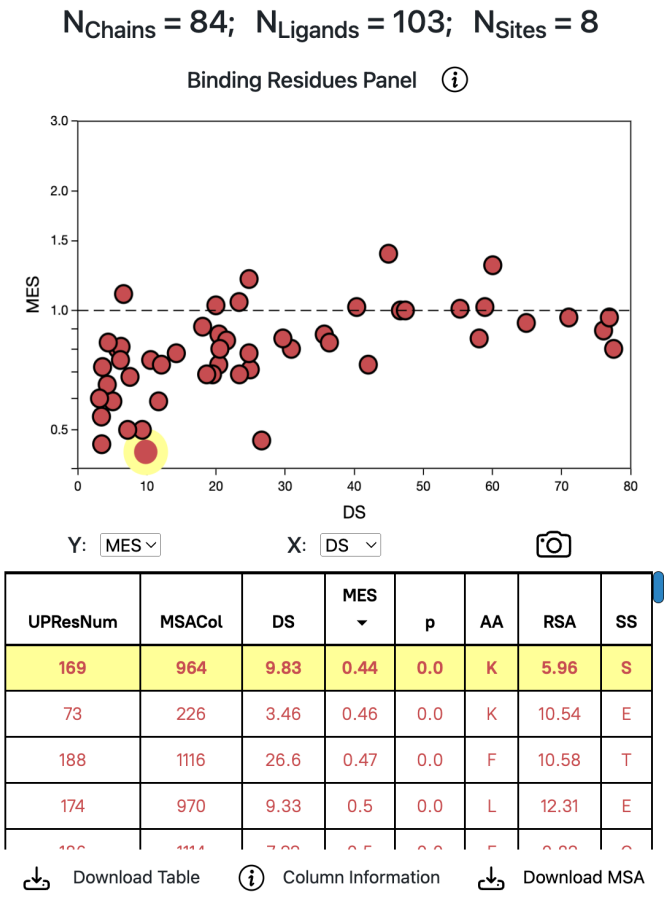

Figure 3.8. LIGYSIS-web results page Binding Residues Panel. Missense enrichment score (MES) *vs* normalised Shenkin divergence score (DS) for the 51 residues of Binding Site 0 of PKA C-alpha (P00517). This site corresponds to the ATP binding site of PKA C-alpha. Most residues on this site present MES < 1 (below the dashed line), indicating depletion in human missense variation resulting from selective pressure or constraint. Missense-depleted residues that are also conserved are most likely to be functional and cause an effect on the protein function if targeted. An example is Lys169, which is buried (RSA = 6%), conserved across homologues (DS = 10) and depleted in missense variation (MES = 0.44, $p \approx 0$). This residue is known to be functionally relevant as it interacts with ATP, which agrees with the results displayed on the table. Residues are sorted by missense enrichment on the table (lowest at the top) and Lys169, which is the most depleted, is being hovered over (yellow highlight).

3.3.2 Data export

A "Download Table" button can be found on the Binding Site and Binding Residues panels. Clicking on this button will download the binding site or binding site residues tables, respectively, in CSV (*.csv*) format. The multiple sequence alignment from which the divergence scores are derived can also be downloaded in Stockholm format (*.sto*) by clicking on the "Download MSA" button. Additionally, (pdbe-arpeggio) protein-ligand contacts can be downloaded in tabular format by clicking on the "Download Current Assembly Contacts" and "Download ALL Assemblies Contacts" for the assembly currently being explored, or all of them. The first option is not available when exploring the Superposition view, since contacts are calculated on individual assemblies, and the second option will download a zipped folder of CSV files.

The ligand superposition view, i.e., representative chain with bound ligands across structures, can be saved to ChimeraX (*.cxc*) or PyMOL (*.pml*) script files by clicking on "Download Superposition" and then selecting the preferred viewer. The same can be done for individual ("Download Current Assembly") or all assemblies ("Download ALL Assemblies"). Figure 3.9 A-C exemplifies this with the ligand superposition for human PKA C-alpha and Figure 3.9 D-F illustrates contacts between PKA C-alpha (P00517) and M77 on PDB: 1Q8W [374]. Screenshots of both graphs as well as 3Dmol.js view can be saved to PNG (*.png*) files by clicking on the camera icon on the corresponding panel.

3.4 LIGYSIS-web analysis of bovine rhodopsin

G-protein coupled receptors (GPCR) comprise the largest protein receptor family in the human genome, with ≈800 members [375]. GPCRs have a regulatory role in most physiological processes including the visual, gustatory and smell sense, immune and nervous system activity as well as in disease [376]. Consequently, GPCRs represent an important target in drug therapy, covering ≈35% of drugs approved by the Food and Drug Administration (FDA) [377, 378]. GPCRs share a conserved seven-transmembrane helix fold connected by three intra- and three extra-cellular loops [379] and bind to a variety of en-

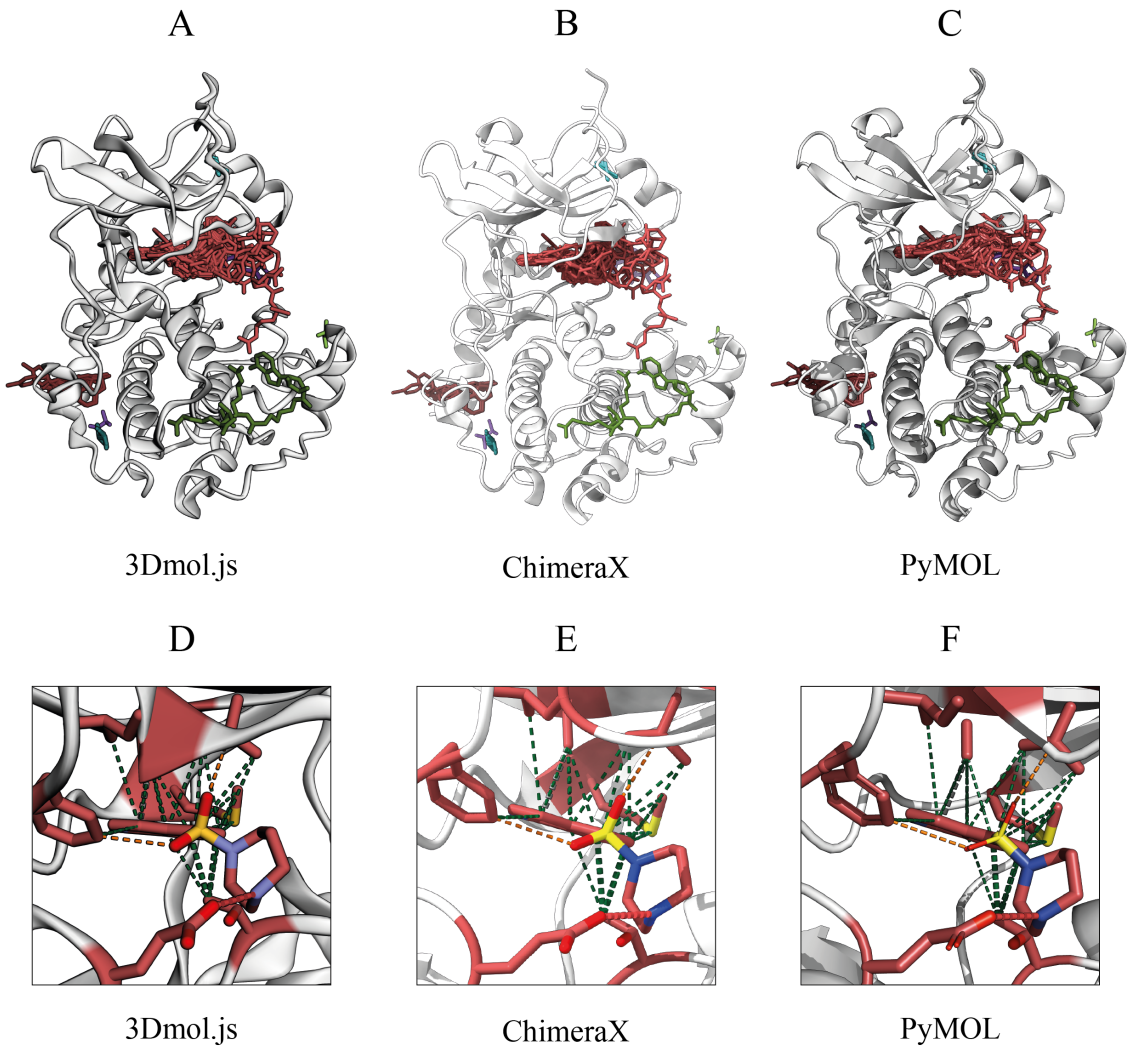

Figure 3.9. LIGYSIS-web supports ChimeraX and PyMOL. LIGYSIS-web employs 3Dmol.js for visualisation. Additionally, it supports download of both superposition view as well as individual assemblies in the common structure viewers ChimeraX and PyMOL. In this superposition example **(A-C)**, PDB: 1SVH [370], chain: A is the representative chain for cAMP-dependent protein kinase catalytic subunit alpha (P00517), for which 8 binding sites are defined from 103 ligands across 84 structures. Ligands are coloured by their binding site; PKA C-alpha interacting with M77 on PDB: 1Q8W [374] visualised with 3Dmol.js **(D)**, ChimeraX **(E)** and PyMOL **(F)**. Non-carbon atoms are coloured based on the Jmol colouring scheme [372] in **(D-F)**.

dogenous ligands including peptides, ions, lipids or neurotransmitters [380, 381]. Once activated by external stimuli, e.g., ligand binding, GPCRs employ G-proteins to interact with downstream effectors, thus triggering intracellular signalling cascades [382]. To accommodate for the wide variety of substrates that GPCRs bind, specificity-determining positions (SDP) are required at the protein-ligand interface. These sites are unconserved across homologues and depleted in missense variation in human (UMD) [203].

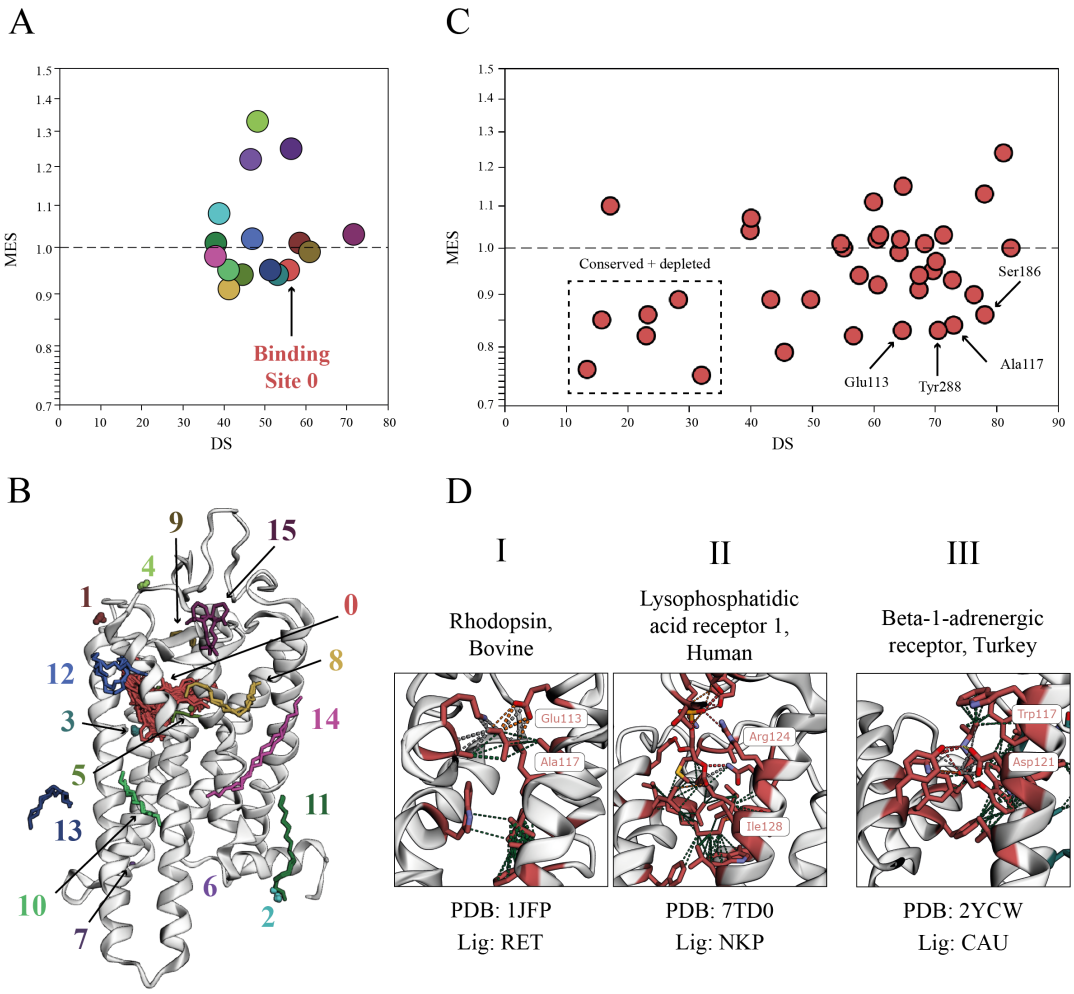

Figure 3.10. LIGYSIS analysis of bovine rhodopsin. **(A)** Scatter of average site missense enrichment score (MES) *vs* divergence (DS) for the 16 ligand binding sites defined for bovine rhodopsin (P06299). Binding Site 0 is unconserved (DS = 56) and depleted in missense variation (MES = 0.95); **(B)** Superposition view of 145 ligands of interest across 49 structures of rhodopsin visualised with 3Dmol.js; **(C)** MES *vs* DS for the 38 residues in Binding Site 0. Residues within the dashed rectangle are conserved across homologues and missense-depleted in human, indicating common function and constraint across members of the family. However, Glu113, Ala117, Ser186 and Tyr288 in bovine rhodopsin are divergent within the family yet constrained in human, suggesting a substrate specificity determining role; **(D)** Examples of multiple GPCRs binding to endogenous substrates substrates (I-II) and inverse agonist (III). (I) Bovine rhodopsin binding to retinal (RET) – PDB: 1JFP [383]. (II) Human lysophosphatidic acid receptor 1 (Q92633) binding to oleoyl lysophosphatidic acid (NKP) - PDB: 7TD0 [384]. (III) Turkey beta-1-adrenergic receptor (P07700) binding to carazolol (CAU) – PDB: 2YCW [385]. Sequence divergence across members of the family is required to accommodate for a variety of substrates. However, each of these residues is missense-depleted and forms ligand-specific interactions within each GPCR. Jmol colouring scheme [372] is applied to non-carbon atoms.

Figure 3.10 illustrates how SDPs can easily be identified in LIGYSIS-web for the example of bovine rhodopsin, a GPCR. Figure 3.10 A shows the average site missense enrichment *vs* divergence for the 16 binding sites defined in LIGYSIS from 145 ligands of interest across 45 structures (Figure 3.10 B). Binding Site 0 corresponds to the orthosteric GPCR site and is on average unconserved (DS $>$ 50) and missense-depleted (MES $<$ 1). Four of the 38 residues within this site are significantly depleted in missense variation ($p <$ 0.05): Glu113, Ala117, Ser186 and Tyr288 (Figure 3.10 C). Figure 3.10 D exemplifies the specificity role of these positions by showing how different GPCRs bind to their substrates with different amino acids at these positions.

3.5 Discussion

The LIGYSIS web server is a free and open resource accessible to all users without any login requirement for the analysis of protein-ligand binding sites. It hosts the LIGYSIS dataset, an integrative protein-ligand complex dataset including 65,000 biologically relevant binding sites across 25,000 proteins with structures on the PDBe. Additionally, users can upload their structures for analysis, results visualisation and download. LIGYSIS defines binding sites by clustering protein-ligand interactions and characterises them by evolutionary divergence, missense variation and solvent accessibility, thus offering insight into the likelihood of function of sites as well as individual residues. These results are dynamically displayed on LIGYSIS-web, a Python Flask web application using Chart.js for dynamic graph rendering and 3Dmol.js for structure visualisation. The server can be accessed through this link: `https://www.compbio.dundee.ac.uk/ligysis/`.

As with most web resources, LIGYSIS is under continuous development, subject to user needs and feature requests. These might include implementing a search by ligand functionality, overcoming the current limitation of multi-occupancy ligand atoms, the implementation of heteromeric protein-ligand complexes or analysing predicted ligand binding sites by methods as P2Rank [115], fpocket [120] or IF-SitePred [168].

Chapter 4

Comparative evaluation of methods for the prediction of protein-ligand binding sites

Preface

This Chapter describes the largest benchmark of ligand binding site prediction methods to date, comparing thirteen original methods using 14 informative metrics and the LIGYSIS dataset as a reference. LIGYSIS, introduced in Chapter 3, is compared to widely used training and test sets and the advantages of using LIGYSIS over these other datasets are shown. A specific metric referred to as top-N+2 recall is proposed as a more robust metric for ligand site prediction and a recommendation made for open-source sharing of both method and benchmark code. The work in this Chapter was solely carried out by *me*.

Publications

Utgés, J.S. and Barton, G.J. Comparative evaluation of methods for the prediction of protein-ligand binding sites. *J. Cheminform.* **16**, 126 (2024). https://doi.org/10.1186/s13321-024-00923-z.

Method	Source	Review	Install	Docs	Model	Included
VN-EGNN	✓	✓	✓	✓	✓	✓
IF-SitePred	✓	✓	✓	✓	✓	✓
GrASP	✓	✓	✓	✓	✓	✓
RefinePocket	✓	✓	?	✗	✓	✗
EquiPocket	✓	✗	?	✗	✓	✗
GLPocket	✓	✓	?	✗	✓	✗
SiteRadar	✗	✓	✗	✗	✗	✗
NodeCoder	✓	✗	?	✓	✗	✗
DeepPocket	✓	✓	✓	✓	✓	✓
RecurPocket	✓	✗	?	✗	✓	✗
PointSite	✓	✓	✗	✓	✓	✗
DeepSurf	✓	✓	✗	✓	✓	✗
PUResNet	✓	✓	✓	✓	✓	✓
Kalasanty	✓	✓	✗	✓	✓	✗
BiteNet	✗	✓	✗	✓	✗	✗
GRaSP	✓	✓	✓	✗	✓	✗
P2Rank	✓	✓	✓	✓	✓	✓
PRANK	✓	✓	✓	✓	✓	✓
DeepSite	✗	✓	✗	✗	✗	✗

Table 4.1. Method selection criteria. These are the criteria employed to select machine learning-based methods for this benchmark. Nineteen machine learning-based methods were considered and seven, for which all requirements were met, were selected. Source: whether the method is open source and code is publicly accessible; Review: whether the method has been published after peer-review; Install: whether installation of the method was successful; Docs: whether the method is sufficiently documented to install it and run it on an example input; Model: whether the method provides pre-trained model weights; Included: whether the method was included in this analysis. Check marks (✓) indicate meeting the requirement and crosses (✗) the opposite. Question marks (**?**) indicate uncertainty – installation was not attempted for some methods as they already did not meet other requirements. Methods in bold font are the ones included in this benchmark.

4.1 Introduction

In this Chapter, thirteen ligand binding site prediction tools are compared to and tested on the LIGYSIS reference dataset, introduced in Chapter 3. LIGYSIS identifies human protein-ligand binding sites from biologically relevant ligands, defined by BioLiP [228], across protein structures determined by X-ray crystallography. The methods assessed in this Chapter include geometry-based fpocket [120], Ligsite [121] and Surfnet [122], energy-based PocketFinder [129] and machine learning methods, exemplified by PRANK [149], P2Rank [115, 150], DeepPocket [160], PUResNet [156, 170], GrASP [167], IF-SitePred [168] and VN-EGNN [169]. Open source, peer-reviewed and easy-to-install methods were prioritised (Table 4.1). This set of methods represents the most complete and relevant set of ligand binding site prediction tools benchmarked to date and is representative of the state-of-the-art within the field.

Table 4.2 and Table 4.3 summarise the methods evaluated in this work, which were executed with their standard settings. VN-EGNN [169] combines virtual nodes with equivariant graph neural networks. Virtual nodes, represented by ESM-2 embeddings [386] are passed through a series of message-passing layers until they reach their final coordinates, which represent the centroid of predicted pockets. Pocket residues are not reported. IF-SitePred [168] represents protein residues with ESM-IF1 embeddings [387] and employs 40 different light gradient boosting machine (LGBM) models [388] to classify residues as ligand-binding if all forty models return a $p > 0.5$. It later utilises PyMOL [389] to place a series of cloud points which are clustered using DBSCAN [390] and a threshold of 1.7 Å. Pocket centroids are obtained by averaging the clustered points' coordinates, scored and ranked based on the number of cloud points. Like VN-EGNN, IF-SitePred does not report pocket residues. GrASP [167] employs graph attention networks to perform semantic segmentation on all surface protein atoms, represented by 17 atom, residue and bond-level features, scoring which are likely part of a binding site. Atoms with a score $>$ 0.3 are clustered into binding sites using average linkage and a threshold of 15 Å. Pocket scores are calculated as the sum of squares of binding site atom scores. PUResNet [156]

combines deep residual and convolutional neural networks to predict ligand binding sites using an 18-element vector of atom-level features and one-hot encoding to represent grid voxels. Voxels with a score > 0.34 are clustered into binding sites using DBSCAN and a threshold of 5.5 Å [170]. Pockets are represented by their residues, but neither pocket centroid, nor score or ranking are reported. Similarly to PUResNet, DeepPocket [160] exploits convolutional neural networks on grid voxels represented by 14 atom-level features to re-score ($\text{DeepPocket}_{\text{RESC}}$) and additionally extract new pocket shapes ($\text{DeepPocket}_{\text{SEG}}$) from fpocket candidates. P2Rank [115] relies on solvent accessible surface (SAS) points placed over the protein surface, represented by 35 atom and residue-level features, and a random forest classifier to score them based on their likelihood of binding a ligand. SAS points with a score > 0.35 are clustered into sites using single linkage and a threshold of 3 Å. $\text{P2Rank}_{\text{CONS}}$ [152] works in the same manner but considers an extra feature: amino acid conservation as measured by Jensen-Shannon divergence [391]. Both report residue and pocket level scores, as well as pocket centroid and rank. PocketFinder [129] uses the Lennard-Jones [392] transformation on a 1 Å grid surrounding the protein surface to predict protein cavities. PocketFinder does not report pocket centroid, score or rank. Finally, geometry-based methods: fpocket [120], Ligsite [121] and Surfnet [122] rely on the geometry of the molecular surface to find cavities. fpocket is the only one of these three methods that reports pocket centroid, score, rank and residues. Additionally, fpocket reports multiple pocket features including surface area, volume, hydrophobicity, charge and druggability.

This Chapter compares these thirteen methods to each other and to the LIGYSIS reference dataset according to a range of metrics including the number of ligand sites, their size, shape, proximity and overlap. Additionally, this Chapter identifies the strengths and weaknesses of prediction assessment metrics and leads to guidance for developing ligand binding site prediction tools or using them to understand protein function and in drug development. This work represents the first independent ligand site prediction benchmark for over a decade, since Schmidtke *et al.* [393] and Chen *et al.* [394] and the largest to date in terms of dataset size (2775), methods compared (13) and metrics employed (14).

Method	Approach	Features	# Features	P centroid	P residues	P score	P rank	R score
VN-EGNN	EGNN + VN	ESM-2 embeddings	1280	✓	✗	✓	✓	✗
IF-SitePred	LGBM	ESM-IF1 embeddings	512	✓	✗	✓	✓	✗
GrASP	GAT - GNN	Atom + residue + bond	17	✓	✓	✓	✓	✓
PUResNet	DRN + 3D-CNN	Atom + one-hot encoding	18	✗	✓	✗	✗	✗
DeepPocket	fpocket + 3D-CNN	Atom	14	✓	✓	✓	✓	✗
$P2Rank_{CONS}$	Random forest	Atom + residue	36	✓	✓	✓	✓	✓
P2Rank	Random forest	Atom + residue	35	✓	✓	✓	✓	✓
$fpocket_{PRANK}$	fpocket + Random forest	Atom + residue	34	✗	✓	✓	✓	✗
fpocket	α-spheres	–	–	✗	✓	✓	✓	✗
$PocketFinder^{+}$	LJ potential	–	–	✗	✗	✗	✗	✓

Ligsite^{+}	Cubic grid	–	–	✗	✗	✗	✗	✓
Surfnet^{+}	Gap regions	–	–	✗	✗	✗	✗	✓

Table 4.2. Ligand binding site prediction methods summary (I). All these methods were used with their default settings. Check marks (✓) indicate that a method provides a given output and crosses (✗) the contrary. Dashes (–) indicate a field is not applicable for a given method, e.g., features for non-machine learning-based methods. Approach: the techniques applied by the method; Features/# Features: the features and their number if the method is machine learning-based; P centroid/P residues/P score/P rank/R score: whether the method reports the pocket centroid, pocket residues, pocket score, pocket ranking and residue *ligandability* score, respectively. For example, P2Rank uses a random forest classifier on SAS points represented by 35 atom and residue features. EGNN + VN: equivariant graph neural network + virtual nodes; LGBM: Light gradient boosting machine; GAT: graph attention network; GNN: graph neural network; DRN: deep residual network; 3D-CNN: three-dimensional convolutional neural network; LJ potential: Lennard-Jones potential.

Method	**R score threshold**	**Cluster**	**Algorithm**	**Threshold** (Å)
VN-EGNN	–	–	–	–
IF-SitePred	0.50 (*all* 40)	Cloud points	DBSCAN	1.7
GrASP	0.30	Atoms	Average	15
PUResNet	0.34	Atoms	DBSCAN	5.5
DeepPocket	–	–	–	–
$P2Rank_{CONS}$	0.35	SAS points	Single	3
P2Rank	0.35	SAS points	Single	3
fpocket	–	α-spheres	Multiple	1.7, 4.5, 2.5
$fpocket_{PRANK}$	–	–	–	–
$PocketFinder^{+}$	–	Grid points	**?**	*search*
$Ligsite^{+}$	–	Grid points	**?**	*search*
$Surfnet^{+}$	–	Grid points	**?**	*search*

Table 4.3. Ligand binding site prediction methods summary (II). All these methods were used with their default settings. Information about the clustering strategies employed by the methods. R score threshold: whether the method uses a residue ligandability threshold; Cluster: the instances they cluster to define the distinct pockets; Algorithm: the clustering algorithm used; Threshold: the distance threshold employed (Å). For example, GrASP utilises average linkage clustering on atoms with predicted ligandability score > 0.30 and a threshold of 15 Å. A dash (–) indicates that the category is not applicable, i.e., VN-EGNN does not employ clustering in their prediction of ligand binding sites. Question marks (**?**) indicate variables for which values were not be found. "*search*" represents an iterative process to find optimal clustering thresholds.

4.2 Methods

4.2.1 LIGYSIS reference dataset

Chapter 3 describes the LIGYSIS pipeline and ligand binding site definition approach, which groups small molecule ligands across multiple biological assemblies of the same protein. In this Chapter, the human subset of the LIGYSIS dataset is employed as a reference dataset for the benchmark of a relevant selection of tools for the prediction of ligand binding sites.

There are 20,423 human reviewed proteins in UniProt [270]. 7640 of these proteins (37.4%) present experimentally determined three-dimensional structures deposited in the Protein Data Bank (PDB) [345]. 5455 of these (71.4%) present at least one ligand-binding structure. Non-biologically relevant ligands were removed in accordance with BioLiP [228], leaving 3513 proteins and 4037 structural segments. A structural segment is defined by the PDBe-KB as a protein region with structural coverage that maps to a contiguous section of their corresponding UniProt sequence. A protein can have multiple segments. For example, each domain of a multi-domain protein, for which there are independent structures, would correspond to a segment. Transformation matrices were obtained from the PDBe-KB [349] and used to structurally align with BioPython [350] a total of 64,498 protein chains across 33,715 structures for the 4037 segments. These matrices result in high-quality structural alignments between different protein chains across PDB structures.

Figure 4.1 illustrates the superposition process of the original LIGYSIS pipeline by highlighting pancreatic alpha-amylase (P04746) and peroxisome proliferator-activated receptor gamma (P37231). Figure 4.2 illustrates the little variation amongst the Cα atoms of the ligand binding residues. The human subset of the LIGYSIS dataset includes 8244 ligand binding sites, i.e., sets of UniProt residues. From here on, this subset of the LIGYSIS dataset will be referred as *LIGYSIS* for brevity.

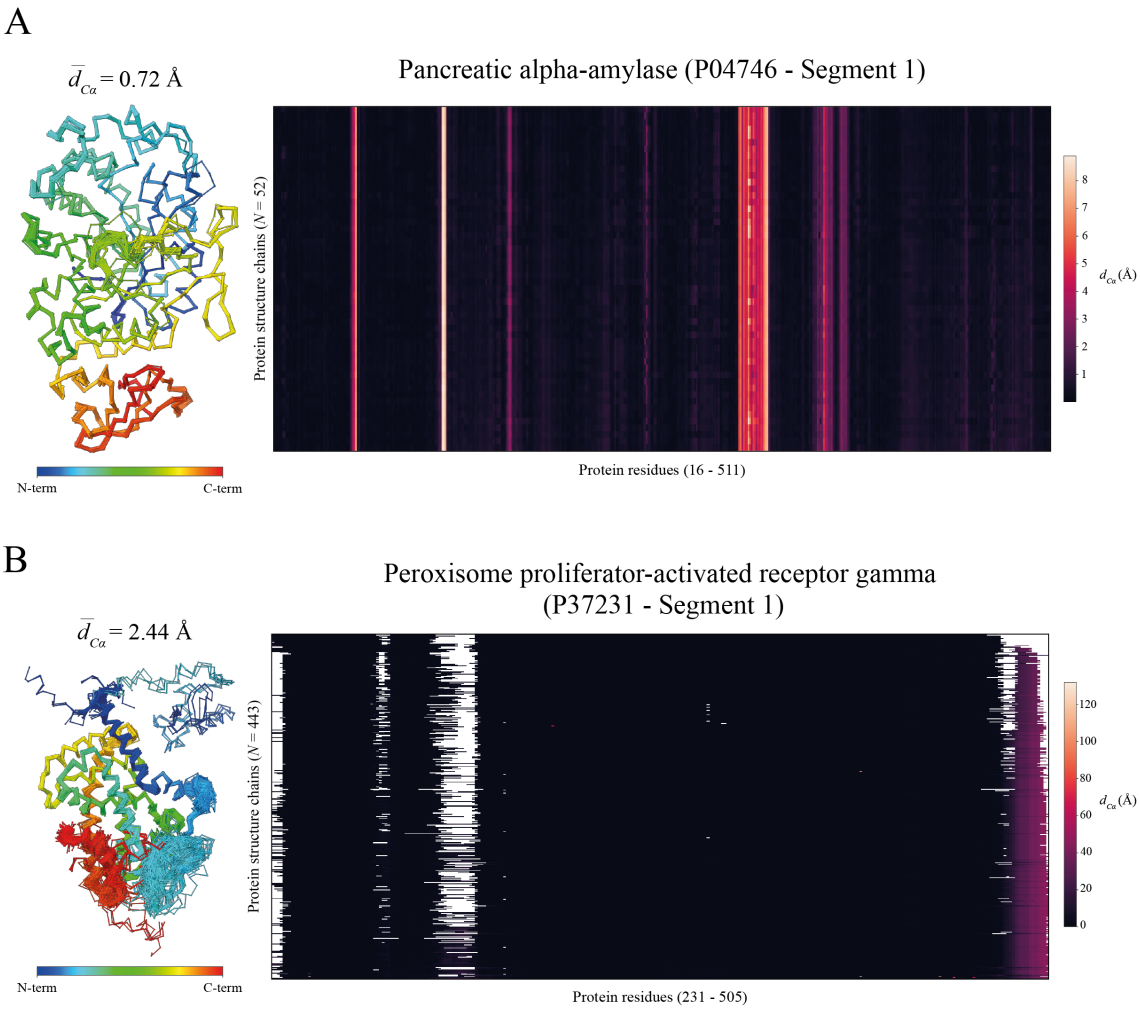

Figure 4.1. Protein chains superposition. PDBe-KB transformation matrices were utilised to structurally align protein chains. For each example, superposed chain trace (Cα atoms) are shown in sticks and coloured using the rainbow scheme from N- to C-terminus and average distance across residues from the aligned chains to the PDBe-KB-defined representative chain is reported as $\overline{d}_{C\alpha}$ (Å). Superposition is visualised also with a heatmap. Protein chain residues are on the X axis and aligned protein chains on the Y axis. Protein chains are sorted by the average distance to the representative chain, so more dissimilar chains are on the bottom. Heatmap cells are coloured based on their $\overline{d}_{C\alpha}$ using the *rocket* colour scheme. White cells represent residues present in the representative chain but not the aligned one, i.e., discontinuities, chain breaks or residues not present in the input sequence. Residues with very high $\overline{d}_{C\alpha}$ (>20 Å) represent alternative locations that were not transformed correctly. **(A)** Pancreatic alpha-amylase (P04746) with 52 superposed chains; **(B)** Peroxisome proliferator-activated receptor gamma (P37231) with 443 chains.

4.2.2 Comparison of datasets

Training and test datasets were downloaded for all machine learning-based methods reviewed in this Chapter. Datasets were compared to the LIGYSIS reference set, in terms of number of sites per protein, ligand-interacting chains, chain length, site size (number of

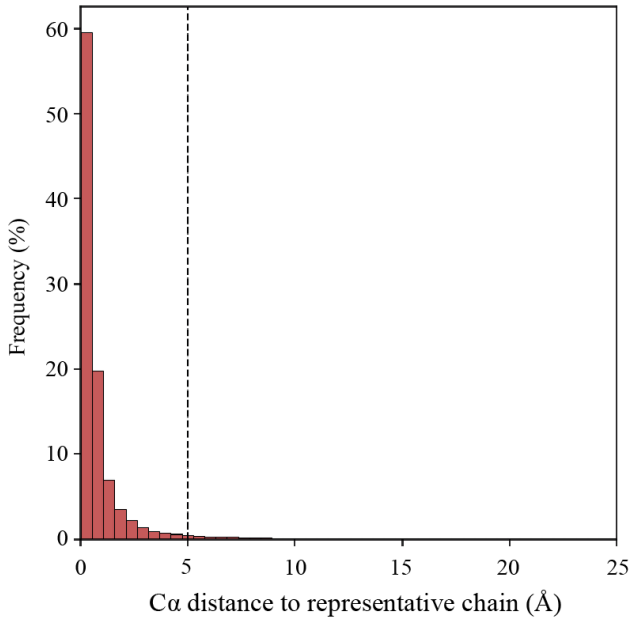

Figure 4.2. Distance to representative chain for ligand binding residues. This histogram represents the distribution of the average Cα distance across transformed chains to the representative chain for 74,536 ligand binding residues across the 2478 segments that present more than one chain. The black dash line indicates 5 Å. 95% of ligand binding residues are within 5 Å of the representative structure in average across chains. This demonstrates that the variation in the Cα trace for ligand binding residues across different structures of the same protein is very small.

amino acids), ligand composition, size and diversity. Ligand diversity was quantified by Shannon's entropy [61] (Equation 4.1) where p_i represents the proportion of each ligand i of the R ligands observed in a dataset. Ligand data was extracted from the Chemical Component Dictionary (CCD) [395]. An overlap (%) was calculated for each dataset as the proportion of LIGYSIS binding sites that were covered by at least one ligand in a test dataset. A straightforward approach was adopted by calculating the intersection of ligand codes between LIGYSIS and each dataset. Ligand codes were defined as a string of PDB ID + "_" + ligand ID, e.g., "6GXT_GTP" corresponds to the guanosine-5'-triphosphate (GTP) of the PDB entry with ID: 6GX7 [396].

$$H' = -\sum_{i=1}^{R} p_i \ln(p_i) \tag{4.1}$$

4.2.3 Training datasets

VN-EGNN is trained on a subset [156] of the sc-PDB (v2017) [397–400] (sc-PDB$_{SUB}$). sc-PDB is a comprehensive database of pharmacological protein-ligand complexes. The database is composed of proteins in complex with buried, biologically relevant synthetic or natural ligands deposited in the PDB. sc-PDB contains unique non-repeating protein-ligand pairs, meaning that only one ligand is considered per PDB entry. Smith *et al.* [167] enriched this dataset with 9000 extra ligands resulting in a version of sc-PDB referred to here as sc-PDB$_{RICH}$, which GrASP trained on. This dataset is not publicly accessible and therefore was not considered in this analysis. DeepPocket used the full sc-PDB set to train on, sc-PDB$_{FULL}$. IF-SitePred used a sequence identity-filtered version of the non-redundant subset of the binding Mother Of All Databases (MOAD) [401–404], which considers only protein family leaders, here referred to as bMOAD$_{SUB}$. The binding MOAD is a large collection of crystal structures with clearly identified biologically relevant ligands with binding data extracted from the literature. PRANK and P2Rank used the CHEN11 dataset to train, which aimed to cover all SCOP [405–407] families of ligand binding proteins in a non-redundant manner [394]. CHEN11 not only considers the ligands in each structure but is enriched with ligands binding to homologous structures. Finally, P2Rank utilised the JOINED dataset for validation. JOINED is a combined dataset formed by other smaller datasets: ASTEX [408], UB48 [143], DT198 [409] and MP210 [146], which represent diverse collections of protein-ligand complexes, including bound/unbound states, drug-target complexes and other ligand site predictor benchmark sets.

4.2.4 Test datasets

The majority of ligand binding site predictors published since 2018 have been using two datasets that were first presented by Krivák *et al.* [115]: COACH420 and HOLO4K, or subsets of them. COACH420 is comprised by a set of 420 single-chain structures binding a mix of drug-like molecules and naturally occurring ligands which is disjunct with the CHEN11 and JOINED datasets. COACH420 is a modified version of the original COACH

test set [138, 141]. HOLO4K is a larger set, $N \approx 4000$, based on the list by Schmidtke *et al.* [393], which includes a mix of single- and multi-chain complexes, also disjunct with P2Rank training (CHEN11) and validation (JOINED) datasets. PRANK employed the small datasets comprising the JOINED set for testing. VN-EGNN, DeepPocket and GrASP use the Mlig and Mlig+ subsets of the COACH and HOLO4K datasets, which include strictly biologically relevant ligands as defined by the binding MOAD. IF-SitePred tested on the HOLO4K-AlphaFold2 Paired (HAP) and HAP-small sets. HAP is a subset of the HOLO4K dataset which presents high-quality models in the AlphaFold database [222]. HAP-small is a smaller subset of HAP that only contains proteins with sequence identity lower than 25% to proteins in the P2Rank training set. VN-EGNN uses the refined version of PDBbind (v2020), referred here as $\text{PDBbind}_{\text{REF}}$, as a third test set. Like binding MOAD, the PDBbind database provides a comprehensive collection of experimentally measured binding affinity data for macromolecular complexes [410–415]. Specifically, the refined set includes protein-ligand complexes for which binding data was obtained from the literature and met certain experimental quality thresholds. Lastly, SC6K is a dataset presented by Aggarwal *et al.* [160] containing 6000 protein-ligand pairs from PDB entries submitted from 01/01/2018 – 28/02/2020.

4.2.5 Protein chain alignment

For each protein chain, atomic coordinates were translated to be centred at the origin, $O = (0, 0, 0)$, and rotated using a rotation matrix R. The two principal components of the coordinate space pc_1 and pc_2 were obtained using principal component analysis (PCA) [416]. A third component, $pc_\perp$, was obtained with the cross-product of the other two, to ensure orthogonality. A rotation matrix P was constructed from these vectors (Equation 4.2). By placing the main component pc_1 on the second row of P, the Y axis was fixed as the major axis, representing the height of the protein chain. The second largest axis is the X axis, representing the width of the protein, and lastly the depth is represented by the smaller magnitude of the Z axis. The final rotation matrix R was obtained by multiplying

P by the negative identity matrix NI (Equation 4.3 and Equation 4.4). This was done to maintain the left-handedness of the protein chains whilst ensuring a consistent alignment on the major axes.

$$pc_{\perp} = pc_1 \times pc_2 \quad \rightarrow \quad P = \begin{bmatrix} pc_2 \\ pc_1 \\ pc_{\perp} \end{bmatrix} \tag{4.2}$$

$$NI = -1 \cdot I_3 = -1 \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{4.3}$$

$$R = P \cdot NI \tag{4.4}$$

4.2.6 Protein chain characterisation

For a protein chain with N amino acid residues, the centre of mass (CM) was calculated by averaging the coordinates, r_i, of all atoms (Equation 4.5), and from it, the radius of gyration, R_g, was derived (Equation 4.6) [417]. Since protein chains were already aligned on the Y axis and centred on the origin, O, the dimensions of the protein chain weree obtained as the magnitude of the PCA components or *eigenvectors*, i.e., the *eigenvalues*. The dimensions represent width, height and depth for the X, Y and Z axes, respectively.

$$\text{CM} = \frac{1}{n}\sum_{i=1}^{n} r_i \rightarrow \text{CM} = \text{O} = (0,0,0) \tag{4.5}$$

$$R_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \text{CM})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \text{O})^2} \rightarrow R_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n} r_i^2} \tag{4.6}$$

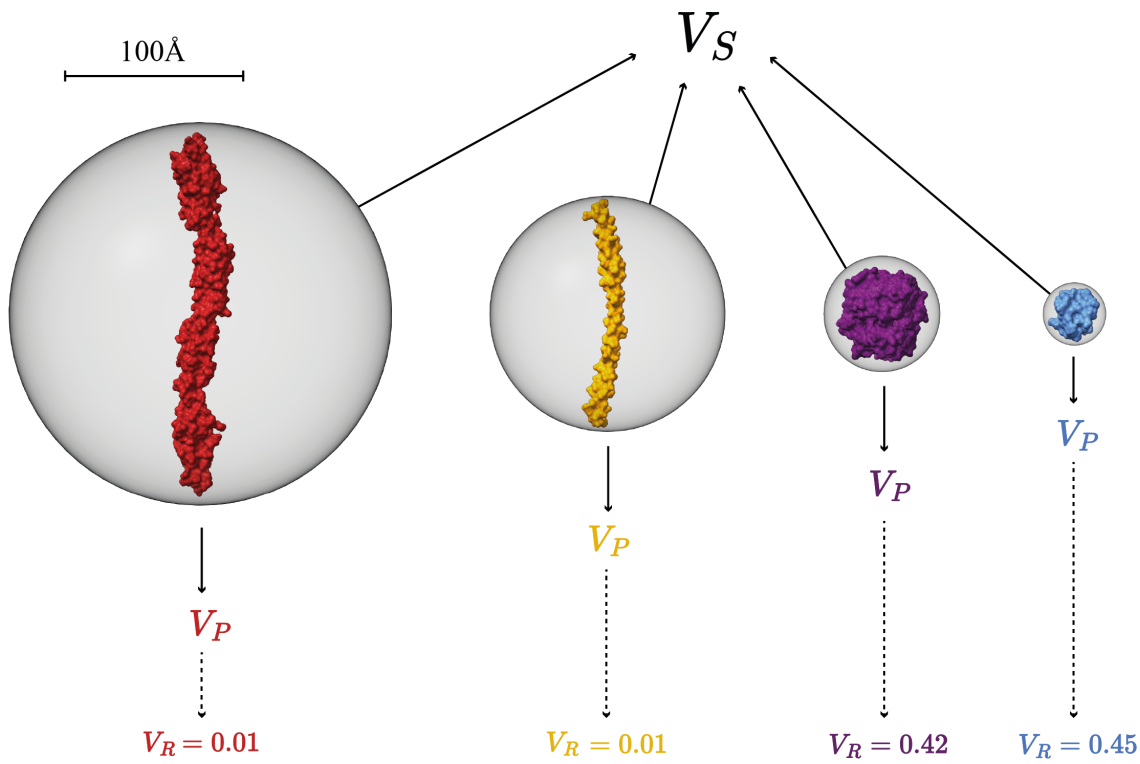

Figure 4.3. Protein chain shape and size classification approach. The volume of the sphere enclosing the protein chain as well as the protein chain volumes were calculated and their ratio obtained (V_R). Globular proteins present more spherical shapes and therefore occupy a higher portion of the sphere volume, resulting in higher volume ratios. Non-globular, elongated or fibrous proteins, on the other hand, do not, and present lower volume ratios. After extensive visual examination, a threshold was established at V_R = 0.08, and so proteins classified in these two groups. Proteins were also classified as "tiny" if their chain was $\leq$ 100 amino acids. Examples for each class are from left to right: Q9Y5G1 – PDB: 6MER [418], chain: A; P05412 – PDB: 5T01 [419], chain: A; Q12884 – PDB: 6Y0F [420], chain: A; Q8NE86 – PDB: 5KUJ [421]. chain: A.

Protein chain volumes were calculated using ProteinVolume [422]. A sphere enclosing the protein and centred on the protein centre of mass was obtained. The radius of this sphere was the maximum Euclidean distance [423] between the protein atoms and the CM (Equation 4.7). The volume of the sphere was calculated using Equation 4.8. Proteins were classified into four different groups based on their shape and size. Protein chains with $\leq$ 100 amino acids were classified as "tiny". Regarding the shape, protein chains were classified into "elongated" if their protein to sphere volume ratio (V_R) $\leq$ 0.08 (Equation 4.9), i.e., the protein volume contained no more than 8% of the sphere volume. This threshold was derived empirically by the visual inspection of all 3448 protein chains on the LIGYSIS set. Otherwise, proteins were considered globular (Figure 4.3). In this

manner, protein chains were classified into *globular* ($N = 2104$; 61%), *elongated* $N = 670$; 19%), *elongated tiny* ($N = 341$; 10%) and *globular tiny* ($N = 333$; 10%).

$$R = \max \|r_i - \mathrm{CM}\| \tag{4.7}$$

$$V_S = \frac{4}{3}\pi R^3 \tag{4.8}$$

$$V_R = \frac{V_P}{V_S} \tag{4.9}$$

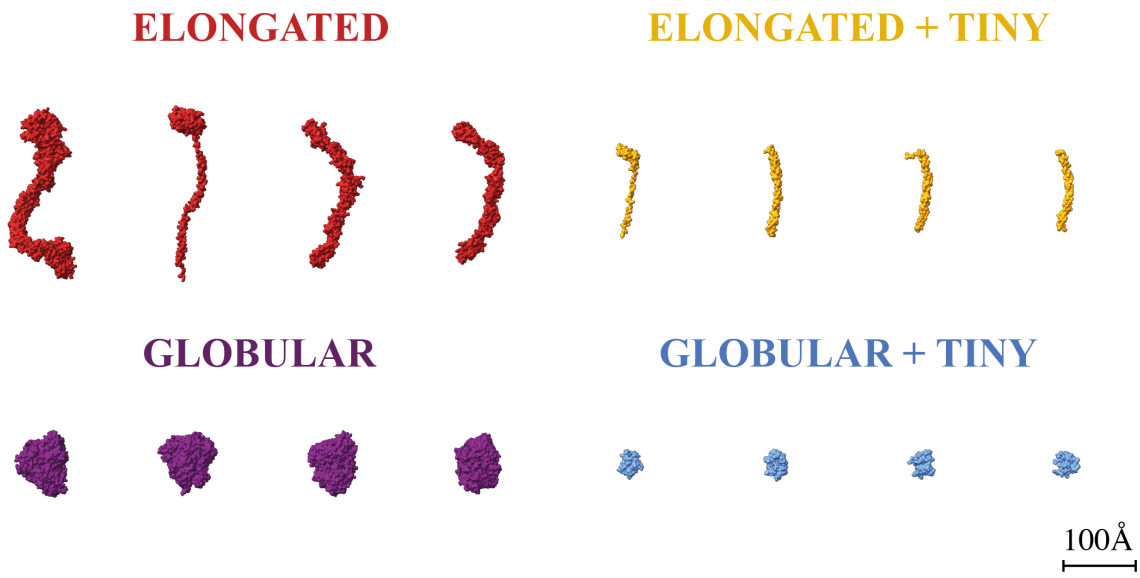

Figure 4.4. Protein shape class examples. Four examples of each protein chain group to illustrate the outcome of the approach. Elongated: Q8NEZ3 – PDB: 8FGW [424], chain: C; P02679 – PDB: 3GHG [425], chain: C; Q14126 – PDB: 7A7D [426], chain: A; Q08554 – PDB: 5IRY [427], chain: A. Elongated + tiny: Q9H2S9 – PDB: 2MA7 [428], chain: A; Q9BV73 – PDB: 6OQA [429], chain: H; Q8IYW5 – PDB: 5YDK [430], chain: F; P60880 – PDB: 3RK2 [431], chain: G. Globular: O43451 – PDB: 3TOP [432], chain: A; P21399 – PDB: 2B3Y [433], chain: A; Q9UI17 – PDB: 5L46 [434], chain: B; P27487 – PDB: 3VJM [435], chain: A. Globular + tiny: Q9UN19 – PDB: 1FAO [436], chain: A; Q12923 – PDB: 1D5G [437], chain: A; P42566 – PDB: 1C07 [438], chain: A; P42566 PDB: 1EH2 [439], chain: A.

4.2.7 Ligand binding site prediction

The representative chain as defined in the PDBe-KB was selected for each segment in the LIGYSIS dataset. Structures were cleaned using the `clean_pdb.py` script [440]. Eleven different ligand binding site prediction tools were used to predict on the 3448 representative chains of the LIGYSIS dataset: VN-EGNN [169], IF-SitePred [168], GrASP [167], PUResNet [156, 170], DeepPocket [160], P2Rank [115, 150], PRANK [149], fpocket [120, 123], PocketFinder^{+} [129], Ligsite^{+} [121] and Surfnet^{+} [122]. Conservation scores for P2Rank were obtained from PrankWeb [441] and used for further prediction. This variant of P2Rank employing amino acid conservation is referred to as P2Rank$_{\text{CONS}}$ [152, 161]. When running DeepPocket, the *-r* threshold was removed and so all fpocket candidates were passed to the CNN-based segmentation module for pocket shape estimation. fpocket predictions re-scored by DeepPocket are referred as DeepPocket$_{\text{RESC}}$, whereas pockets extracted by the segmentation module of DeepPocket are referred as DeepPocket$_{\text{SEG}}$. PRANK was also used to re-score fpocket (fpocket$_{\text{PRANK}}$). This combination had already been explored in previous studies [115, 149, 150, 442]. Re-implementations of Capra *et al.* [145] were used for PocketFinder, Ligsite and Surfnet, indicated by the "+" superscript. VN-EGNN, IF-SitePred, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+} do not provide a list of residues for each pocket, but a list of centroids and their scores for the first two, and a list of grid points for each predicted pocket for the last three. For VN-EGNN, residues within 6 Å of the virtual nodes were considered pocket residues. No residues were found within this threshold for 429 predicted pockets (≈3%). For IF-SitePred, residues within 6 Å of the clustered cloud points that resulted on a predicted pocket centroid were considered as pocket residues. Pocket residues were obtained in a similar manner for PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}, by taking those residues within 6 Å of the pocket grid points. In total, thirteen methods were considered in this analysis: VN-EGNN, IF-SitePred, GrASP, PUResNet, DeepPocket$_{\text{RESC}}$, DeepPocket$_{\text{SEG}}$, P2Rank$_{\text{CONS}}$, P2Rank, fpocket$_{\text{PRANK}}$, fpocket, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}.

Seven of the methods provide residue *ligandability* scores. P2Rank and P2Rank$_{CONS}$ report calibrated probabilities of residues being ligand-binding. Similarly, GrASP predicts the likelihood for any given heavy atom to be part of a binding site. A residue-level score was obtained for GrASP by taking the maximum score of the residue atoms. For IF-SitePred, a residue ligandability score (LS) was computed by averaging the 40 predicted probabilities of a residue being ligand-binding (Equation 4.10). Though calculated in a different way, these three scores range 0-1, represent the likelihood of a residue binding a ligand, i.e., ligandability, and can therefore be compared. PocketFinder^{+}, Ligsite^{+} and Surfnet^{+} also provide residue scores, which maximum value can be > 1.

$$LS = \frac{1}{40}\sum_{i=1}^{40} p_i \tag{4.10}$$

VN-EGNN, PUResNet, DeepPocket$_{RESC}$, DeepPocket$_{SEG}$, fpocket$_{PRANK}$ and fpocket do not report residue-level scores. However, binary labels represent whether a residue is part of a pocket (label: 1) or not (label: 0), in the same manner as for all other methods. Throughout this Chapter, the terms "site" and "pocket" are used indistinctly. Methods are sorted in chronological order across all figures, tables and legends.

4.2.8 Binding site characterisation

Radius of gyration was calculated for pockets as it was done for whole proteins (Equation 4.6). Distance between pockets was calculated as the Euclidean distance [423] between their centroids and overlap between pocket residues with the Jaccard index (JI) (Equation 4.11) [443, 444]. POVME 2.0 was employed for pocket volume calculation [445–447]. A single inclusion region defined by the smallest rectangular prism containing all pocket atoms was used. The prism was centred on the pocket centroid and its dimensions determined by the distance between the two farthest coordinates on each axis. No exclusion regions were used. Points outside the convex hull were deleted. A contiguous-points region was defined as a 5 Å-radius sphere on the pocket centroid (Figure 4.5).

$$JI(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.11}$$

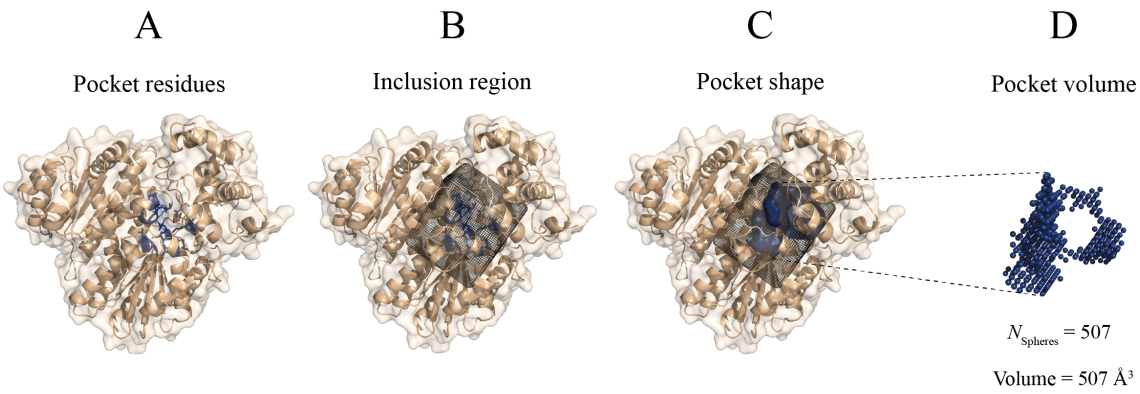

Figure 4.5. Pocket volume calculation algorithm. (A) PUResNet predicted pocket for PDB: 4PX2 [448]. Pocket residues are coloured in blue and have their side chains displayed; **(B)** An inclusion region is determined by the pocket atoms; **(C)** POVME 2.0 calculates the pocket shape within the inclusion region; **(D)** The pocket shape is defined by a series of unit-volume (1 Å^3) spheres. The pocket volume is calculated as the number of spheres within the pocket. Structure visualisation with PyMOL v2.5.2 [389].

4.2.8.1 Determination of DCC threshold

Most methods employ distance to closest ligand atom (DCA) and a threshold of 4 Å to consider a prediction as correct. Because of the way the LIGYSIS dataset has been curated, it is easier to use DCC, since binding sites result of the clustering of multiple ligands, and not just a single ligand binding a protein. Despite DCC being much stricter than DCA, the same threshold of 4 Å is used for both metrics when benchmarking methods [156, 160, 169]. Figure 4.6 A shows the relation between DCC and pocket residue overlap for the *best* prediction for each method and each observed pocket. The *best* prediction is that with the minimum Euclidean distance to the observed pocket centroid. Across all methods, there were more than 15,000 predicted pockets with DCC $>$ 4 Å and residue overlap $\geq$ 0.5. Setting the DCC threshold at 4 Å results in the wrong labelling of these predictions as "false positives". For this reason, a more meaningful DCC threshold was empirically established through the visual inspection of predicted-observed pocket pairs. Figure 4.6 B suggests that this threshold lies between 10-15 Å, where the proportion of pockets with $I_{rel} \geq 0.5$ gradually decreases until reaching 0.

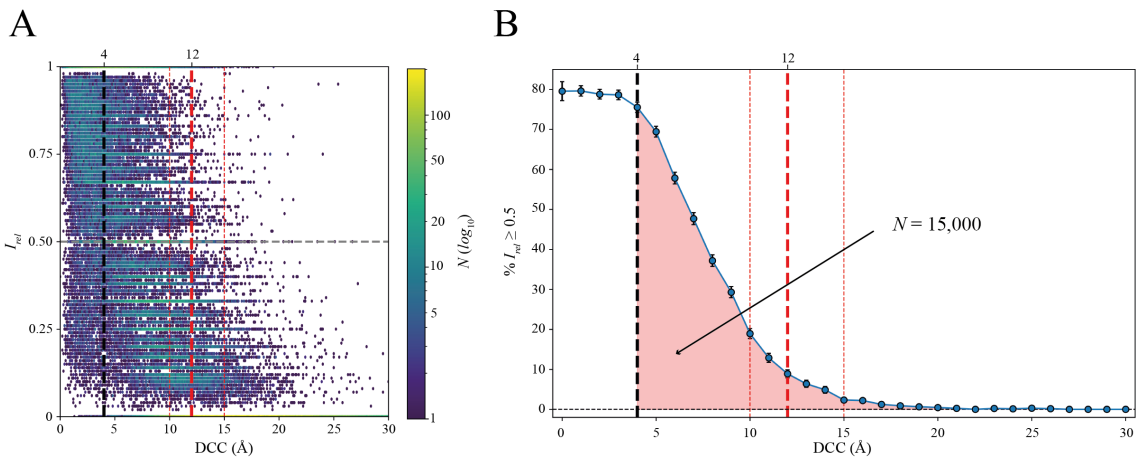

Figure 4.6. ***I_{rel}* vs DCC. (A)** Hexagonal binned plot of I_{rel} (Y) *vs* DCC (X). Data points are grouped into hexagonal bins, which are coloured by the number of data points within each bin using the *viridis* colour palette. The colour bar axis is in $\log_{10}$ scale. The dashed lines indicate the literature consensus DCC = 4 Å threshold (black) and an arbitrary I_{rel} threshold of 0.5 (gray), i.e., coverage of half of the observed ligand-binding residues by the predicted pocket. The dashed red lines delimit the likely location of a potentially more informative DCC threshold; **(B)** Cumulative proportion of predicted pockets with $I_{rel} \geq 0.5$ for each DCC 1-Å interval. The commonly used threshold of DCC = 4 Å labels >15,000 predictions with $I_{rel} \geq 0.5$ as false. Error bars indicate 95% CI of the proportion.

A hard threshold was set at D = 20 Å and a decision made, so that based purely on distance, pockets with DCC > 20 Å would not be considered as correct predictions. For each DCC interval of 1 Å, the pocket with the highest and lowest I_{rel} were inspected (Figure 4.7 and Figure 4.8). This initial visual inspection further supported the hypothesis that a more meaningful DCC threshold was between 10-14 Å. For the next step, only predicted-observed pocket pairs with minimal overlap ($I_{rel} < 0.25$) were considered. Starting at DCC = 10 Å, and using unit (1 Å) intervals, the 100 farthest pocket pairs were inspected for each interval. The percentage of correct predictions was therefore calculated as the number of pockets labelled as "correct" upon visual inspection (%). For D = 10 Å, 94% of pockets were correct (Figure 4.9), 86% for D = 11 Å (Figure 4.10), 85% for D = 12 Å (Figure 4.11) and 66% for D = 13 Å. Due to the considerable drop of correct pockets at D = 13 Å, the final distance threshold was set at D = 12Å. Accordingly, predictions were considered as true positives if DCC ≤ 12 Å (Equation 4.12).

$$\textit{True positive} \iff \text{DCC} \leq 12\ \text{Å} \tag{4.12}$$

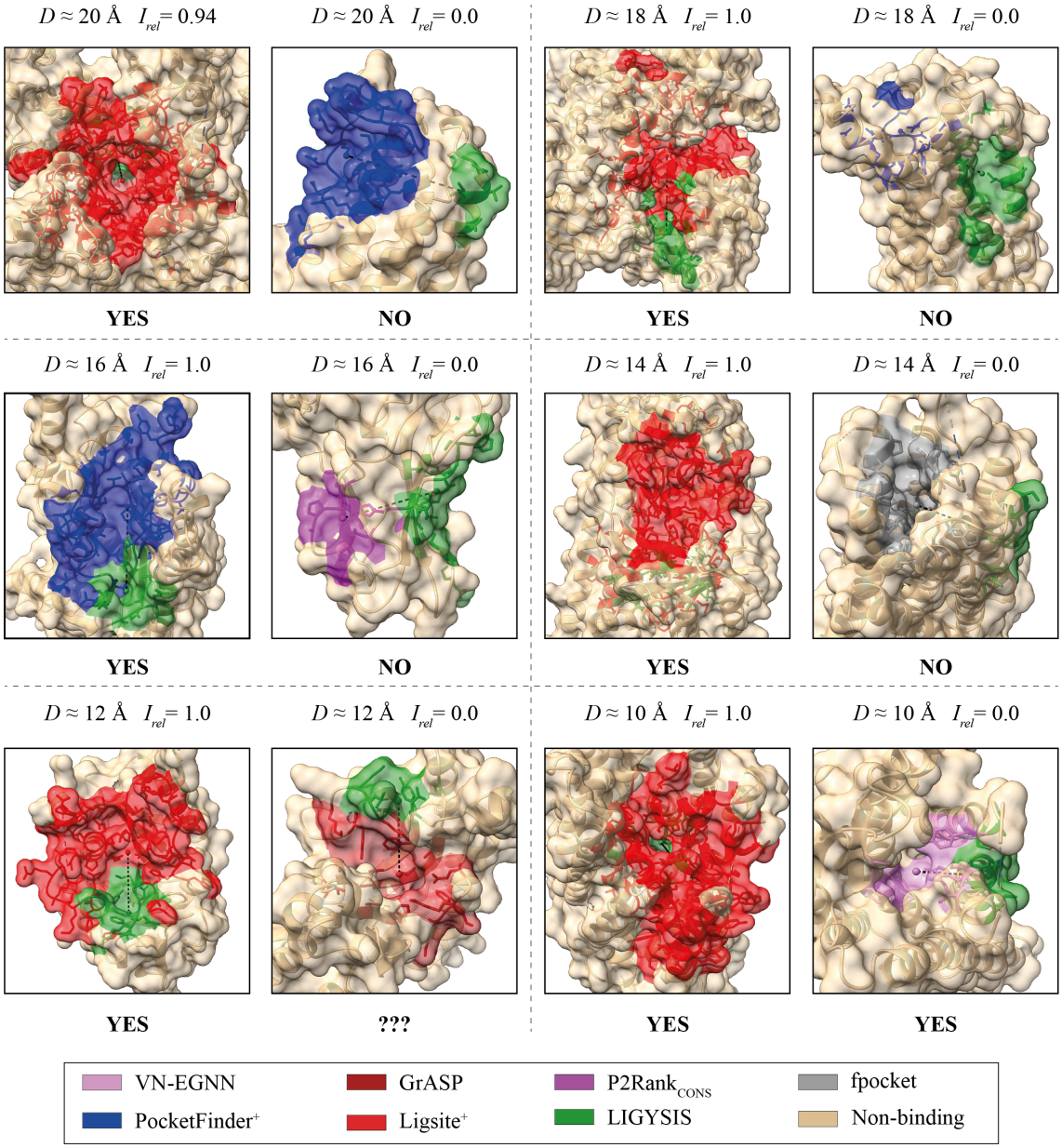

Figure 4.7. Determination of DCC threshold (I). Highest and lowest-residue overlap predictions for each 2-Å DCC interval. Observed LIGYSIS sites are coloured in green, predicted pockets in other colours. *D* represents DCC and I_{rel} the relative intersection between predicted and observed pocket residues, i.e., proportion of observed site residues covered by predicted pocket residues. "YES" or "NO" labels indicate whether a prediction was considered correct upon visual inspection. "???" at DCC = 12 Å illustrates the inflection point between 10-12 Å, where it is no longer clear whether predicted pockets within this DCC interval and $I_{rel} \approx 0$ agree with the observed pockets. To facilitate the visualisation of the observed pocket, this one was coloured after the predicted one. Otherwise, for cases where $I_{rel} = 1$, only the predicted pocket would be shown. Despite 1-Å intervals were inspected, only representatives of 2-Å intervals are shown here for simplicity. Examples from left to right and top to bottom: O75417 – PDB: 5A9J [449], chain: D; P01574 – PDB: 1AU1 [450], chain: A; Q01118 – PDB: 7TJ8 [451], chain: A; Q92847 – PDB: 7W2Z [452], chain: R; Q9BY49 – PDB: 1YXM [453], chain: A; O15178 – PDB: 8FMU [454], chain: B; Q14534 – PDB: 6C6P [455], chain: A; P21728 – PDB: 8IRR [456], chain: R; Q9P2W7 – PDB: 1V84 [457], chain: A; Q9P0M2 – PDB: 5JJ2 [458], chain: A; Q8N695 – PDB: 7SL9 [459], chain: A; P41145 – PDB: 6VI4 [460], chain: A.

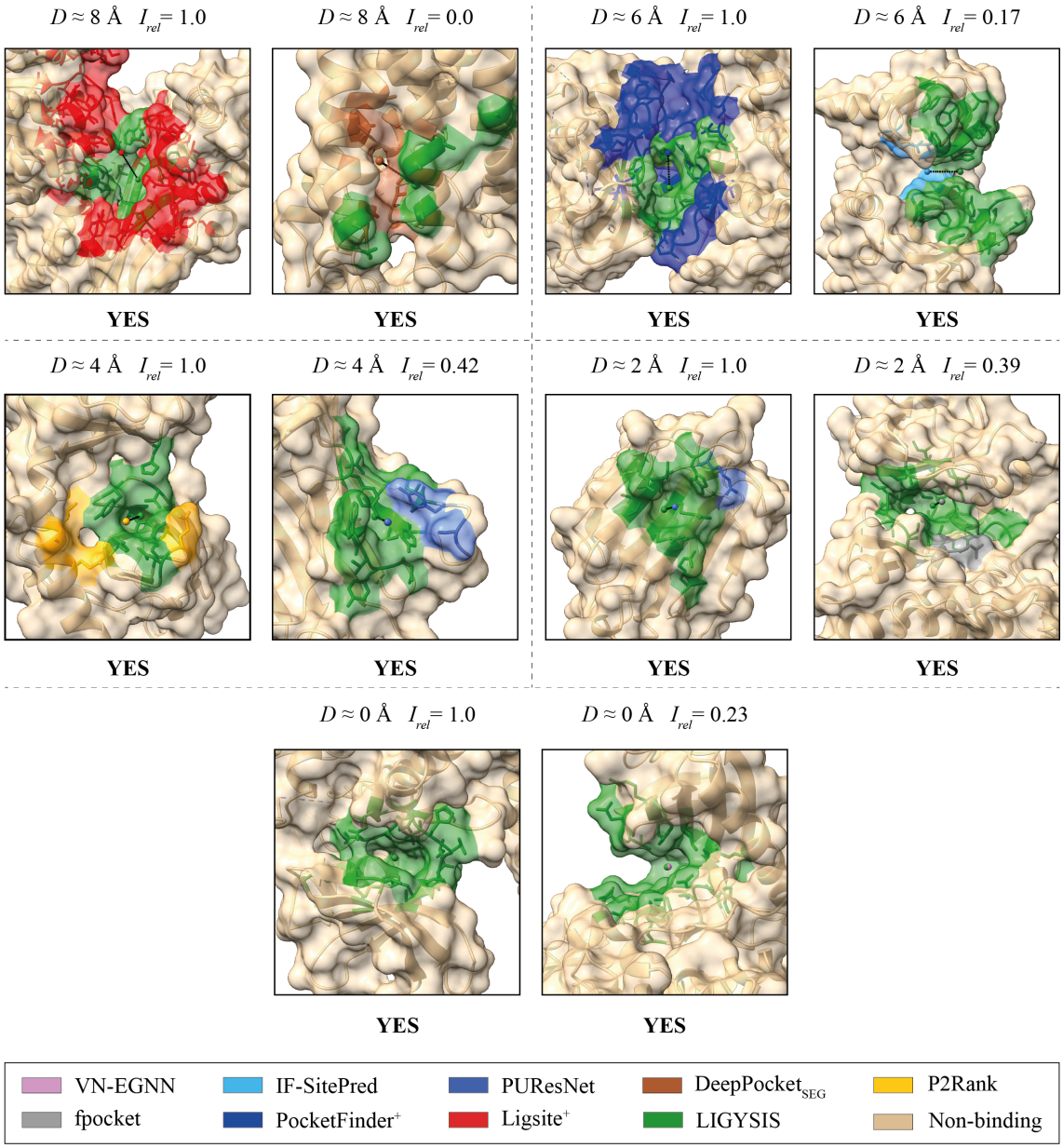

Figure 4.8. Determination of DCC threshold (II). Highest and lowest-residue overlap predictions for each 2-Å DCC interval. Observed LIGYSIS sites are coloured in green and predicted pockets in other colours. *D* represents DCC and I_{rel} the relative intersection between predicted and observed pocket residues, i.e., proportion of observed site residues covered by predicted pocket residues. "YES" indicates that a prediction was considered correct upon visual inspection. To facilitate the visualisation of the observed pocket, this one was coloured after the predicted one. Otherwise, for cases where I_{rel} = 1 only the predicted pocket would be shown. Despite 1-Å intervals were inspected, only representatives of 2-Å intervals are shown here for simplicity. Examples from left to right and top to bottom: Q9UDR5 – PDB: 5L78 [461], chain: B; P16473 – PDB: 7XW7 [462], chain: R; Q09013 – PDB: 2VD5 [463], chain: B; Q15047 – PDB: 6BHD [464], chain: A; Q9BXT4 – PDB: 5M9N [465], chain: A; Q9UGM1 – PDB: 4UY2 [466], chain: B; Q9UHV8 – PDB: 5XG7 [467], chain: A; Q15303 – PDB: 3BCE [468], chain: A; Q8IVW4 – PDB: 3ZDU [469], chain: A; P49760 – PDB: 6KHE [470], chain: A.

$D = 10$ Å and $I_{rel} < 0.25 \longrightarrow p = 94/100 \rightarrow 94\%$

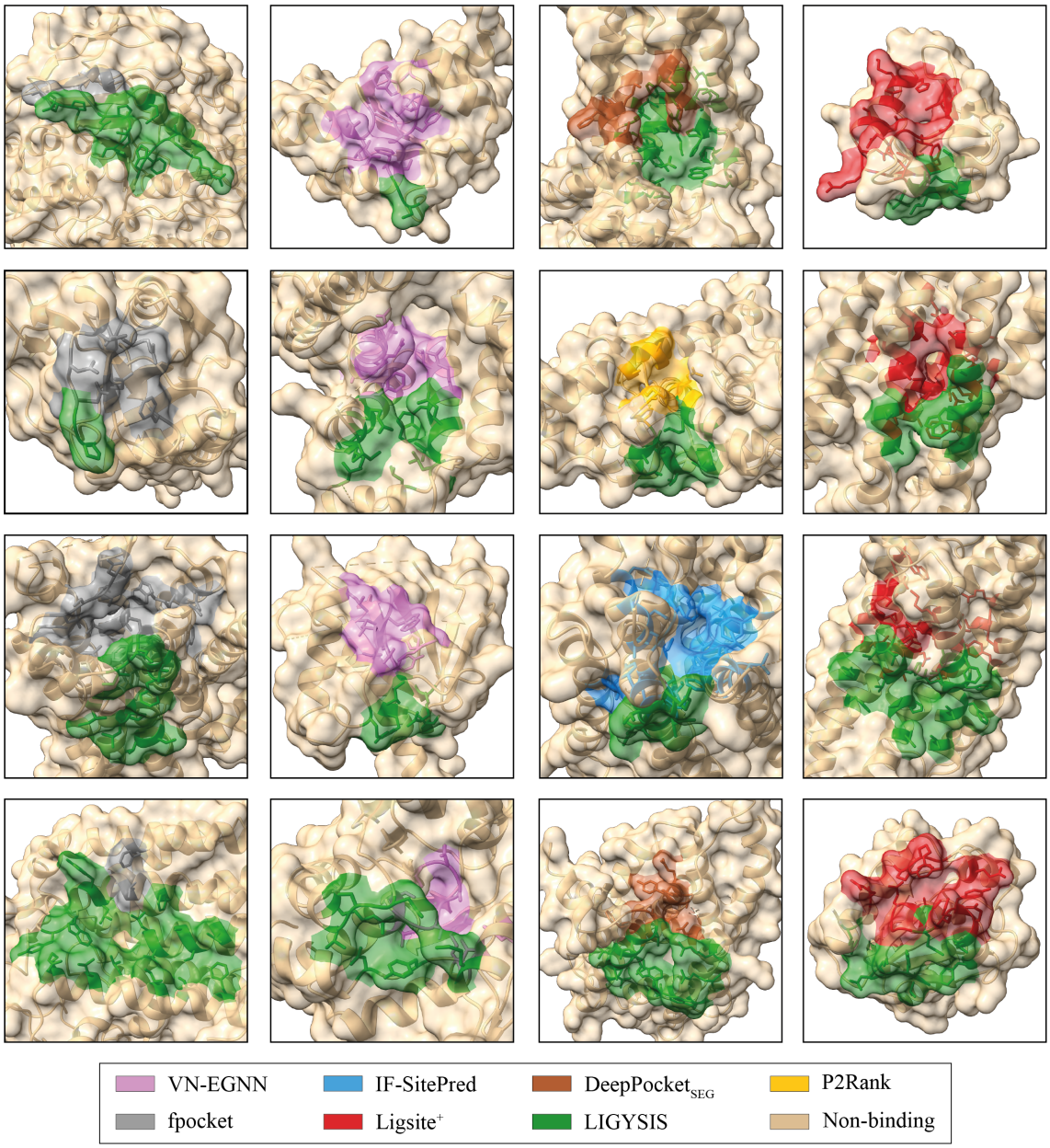

Figure 4.9. Predicted-observed pocket pairs at DCC = 10 Å and $I_{rel} < 0.25$. 94/100 visually inspected examples were considered as correct predictions on the basis that the predicted and observed pockets were adjacent, i.e., their surface area is in contact, and it is therefore easy to imagine a ligand that would bind to this region. LIGYSIS observed sites are coloured in green and predicted pockets in other colours. Examples from left to right and top to bottom: P22303 – PDB: 5HQ3 [471], chain: B; O76003 – PDB: 2YAN [472], chain: A; P13498 – PDB: 8WEJ [473], chain: A; Q9HD26 – PDB: 2LOB [474], chain: A; P35247 – PDB: 5OXS [475], chain: C; Q96TC7 – PDB: 7CC7 [476], chain: A; P10828 – PDB: 1NQ1 [477], chain: A; Q9BXJ8 – PDB: 7F3U [478], chain: B; P21728 – PDB: 8IRR [456], chain: R; O00213 – PDB: 3D8E [479], chain: C; P21728 – PDB: 8IRR, chain: B; P48546 – PDB: 7DTY [480], chain: R; Q9H3H5 – PDB: 6BW6 [481], chain: B; O95278 – PDB: 4RKK [482], chain: A; Q96BI3 – PDB: 5FN5 [483], chain: C; Q93096 – PDB: 5BX1 [484], chain: A.

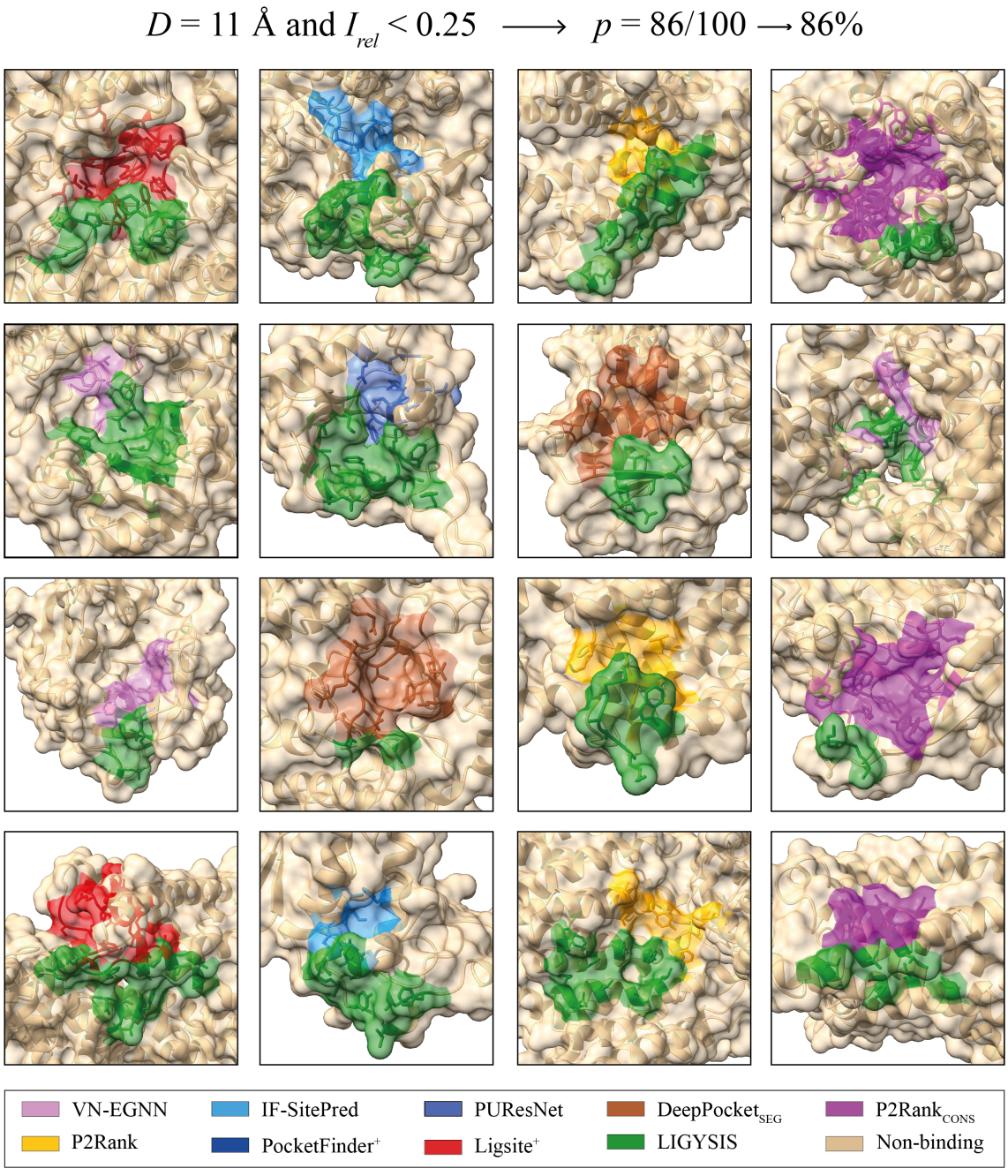

Figure 4.10. Predicted-observed pocket pairs at DCC = 11 Å and I_{rel} < 0.25. 86/100 visually inspected examples were considered as correct predictions. LIGYSIS observed sites are coloured in green and predicted pockets in other colours. Examples from left to right and top to bottom: O43451 – PDB: 3TOP [432], chain: A; P04180 – PDB: 5TXF [485], chain: B; P49190 – PDB: 7F16 [486], chain: R; P21728 – PDB: 8IRR [456], chain: R; Q16394 – PDB: 7SCH [487], chain: A; Q8IU60 – PDB: 5MP0 [488], chain: D; P56192 – PDB: 5Y6L [489], chain: A; Q5VSL9 – PDB: 7K36 [490], chain: I; P20160 – PDB: 1FY1 [491], chain: A; P06280 – PDB: 4NXS [492], chain: B; Q9H490 – PDB: 7WLD [493], chain: U; O95631 – PDB: 7NDG [494], chain: G; Q00975 – PDB: 7MIX [495], chain: A; O15496 – PDB: 5G3M [496], chain: B; Q8TCJ2 – PDB: 6S7T [497], chain: A; P41145 – PDB: 6VI4 [460], chain: A.

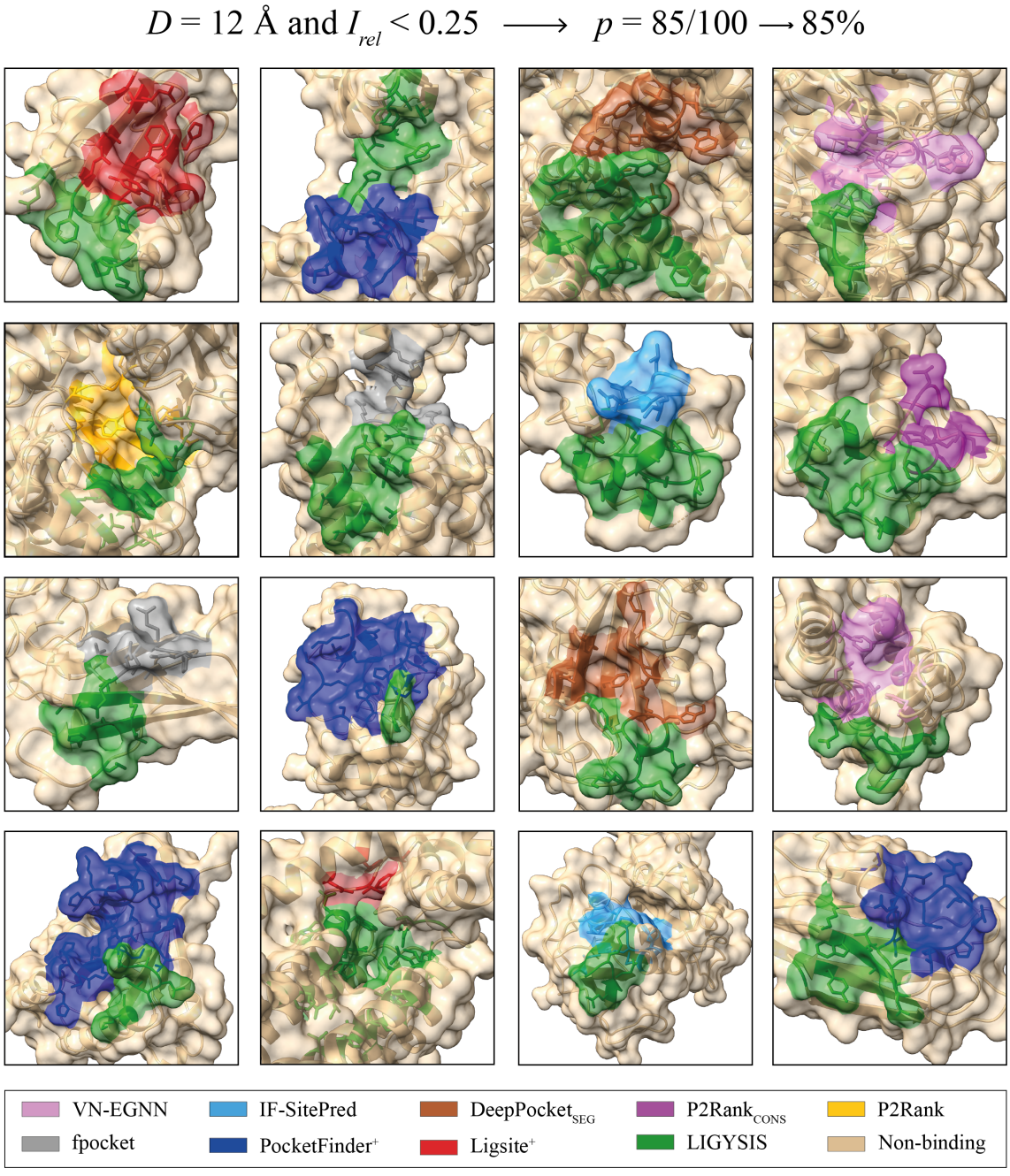

Figure 4.11. Predicted-observed pocket pairs at DCC = 12 Å and I_{rel} < 0.25. 85/100 visually inspected examples were considered as correct predictions. LIGYSIS observed sites are coloured in green and predicted pockets in other colours. Examples from left to right and top to bottom: P50616 – PDB: 2Z15 [498], chain: C; Q9BXT4 – PDB: 5M9N [465], chain: B; Q9P0X4 – PDB: 7WLK [499], chain: A; P08648 – PDB: 3VI3 [500], chain: A; P22102 – PDB: 2QK4 [501], chain: B; P48546 – PDB: 7DTY [480], chain: R; Q99496 – PDB: 4S3O [502], chain: B; O15496 – PDB: 5G3M [496], chain: B; P16471 – PDB: 3MZG [503], chain: B; P35247 – PDB: 5OXS [475], chain: C; Q00688 – PDB: 2MPH [504], chain: A; Q6PL18 – PDB: 7M98 [505], chain: A; Q9H082 – PDB: 6ZAY [506], chain: A; P00156 – PDB: 5XTE [507], chain: J; P00374 – PDB: 1DRF [508], chain: A; P02746 – PDB: 2JG9 [509], chain: F.

4.2.9 Prediction evaluation

LIGYSIS binding sites consist of sets of UniProt residue numbers to which ligands bind across the multiple structures of a protein. The thirteen ligand binding site predictors benchmarked in this Chapter predict only on the representative chains for each protein. These representative structures are defined in the PDBe-KB based on three criteria: data quality, sequence coverage and resolution [346]. Despite this, representative chains might still be missing some residues present in other structures. To compare LIGYSIS binding sites to predicted sites on the representative chains, UniProt sequence mappings are needed for each residue in the LIGYSIS-defined sites. For this reason, LIGYSIS entries with ligand-binding residues missing UniProt residue mappings on the protein's representative chain were discarded, resulting in a set of 3048 human proteins, including 3448 segments. After predicting on these 3448 LIGYSIS chains, only chains where all residues across all predicted sites presented UniProt residue mapping were kept. This resulted in a final set of 2775 protein chains that was employed to assess the performance of the methods.

The performance of ligand binding site prediction methods can be evaluated at two different levels: *residue* level and *pocket* level. Prediction at the residue level involves the discrimination of those residues that are likely to interact with a ligand, whereas the aim of pocket-level prediction is to define distinct regions on a protein, i.e., pockets, where *a* ligand is likely to bind. This region can either be defined by a centroid, a group of cloud/grid points, a set of residues or a combination of these. Some methods are *residue*-centric, which predict first at the residue-level, use a threshold to select high-probability ligand-binding residues, and then cluster them into pockets. Residue-centric methods include IF-SitePred or GrASP. Other (*pocket*-centric) methods directly predict the location or shape of the pocket without the need of predicting at the residue level first. Some of these methods can use their pocket-level prediction to report residue ligandability scores, e.g., P2Rank$_{CONS}$, P2Rank, PocketFinder^{+}, Ligsite^{+} or Surfnet^{+}. Others, such as VN-EGNN, PUResNet, DeepPocket or fpocket do not report residue ligandability scores.

4.2.9.1 Residue-level predictions

GrASP, P2Rank$_{\text{CONS}}$, P2Rank, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+} all offer residue ligandability scores. Additionally, a ligandability score was derived for IF-SitePred using Equation 4.10. Prediction at the residue level is a binary classification problem: binding (label: 1) or non-binding (label: 0). Given a ligandability threshold t_{LS}, a residue i with a ligandability score LS_i is classified as "positive" if $LS_i > t_{LS}$. Conversely, the residue is classed as "negative" if $LS_i \leq t_{LS}$. Further stratification results from comparing predictions to the LIGYSIS reference dataset.

- True positive (TP): residue classified as positive that binds a ligand according to the reference.
- False positive (FP): residue classified as positive that does not bind a ligand in the reference.
- True negative (TN): residue classified as negative that does not bind a ligand.
- False negative (FN): residue classified as negative but is known to bind a ligand according to the reference.

With these four classes, true positive rate (TPR) (Equation 4.13), false positive rate (FPR) (Equation 4.14), precision (Equation 4.15) and recall (Equation 4.16) can be calculated and receiver operating characteristic (ROC) and precision-recall (PR) curves plotted. ROC and PR curves were obtained for each of the LIGYSIS protein chains. Mean ROC and PR curves were obtained by taking the mean TPR and FPR (ROC curve) and mean precision and recall (PR curve) from the single-protein curves at each score interval. These average curves are representative of the variation across proteins for these metrics. Mean area under the curve (AUC) for ROC and average precision (AP) were calculated by averaging the areas and precisions across curves. Baselines for these are 50% and the proportion of true binding residues (10%), respectively. ROC and AUC can't be calculated for VN-EGNN, PUResNet, DeepPocket, fpocket$_{\text{PRANK}}$ and fpocket as these methods do not provide residue ligandability scores.

$$\text{TPR (\%)} = 100 \times \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.13}$$

$$\text{FPR (\%)} = 100 \times \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{4.14}$$

$$\text{Precision (\%)} = 100 \times \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.15}$$

$$\text{Recall (\%)} = 100 \times \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.16}$$

Pocket binary labels (0: not a pocket residue; 1: pocket residue) can also determine TP, FP, TN and FN for each residue in a protein chain P_i. VN-EGNN, IF-SitePred, PocketFinder$^+$, Ligsite$^+$ and Surfnet$^+$ do not report pocket residues. For these methods, residues within 6 Å of the pocket centroid, cloud points and grid points (3×), respectively, were labelled as pocket residues (label: 1). All other residues in P_i were labelled as non-binding (label: 0). An F1 score was computed from all residues in P_i, which combines precision and recall into a unified metric, capturing the accuracy and completeness of predictions at the residue level (Equation 4.17). The Matthews correlation coefficient (MCC) [510] (Equation 4.18) was also calculated. The median F1 score and MCC across dataset proteins were reported for each method.

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.17}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{4.18}$$

4.2.9.2 Pocket-level predictions

Ligand binding site prediction at the pocket level is a multi-instance prediction problem. There are no *negatives* predicted at the pocket level of ligand binding site prediction, only *positives*. A positive is a predicted pocket, which will be true (TP) or false (FP) depending on whether it is observed in the reference data. False negatives are those pockets observed in the reference data that are not predicted. They are the pockets the method fails to predict, and therefore, are not scored. A true negative would be a "non-pocket" that is *not* predicted. This can't be quantified easily and even if it was, it would not be scored by the method, as it is not predicted. For this reason, in this context, neither TPR, nor FPR can be calculated. Consequently, ROC/AUC can't be utilised to assess ligand binding site prediction at the pocket level. False negatives are known, but not scored, and therefore PR/AUC is not an option either. What *can* be calculated is the recall given a certain criterion. In this case, because of the nature of the LIGYSIS dataset, where defined sites result from the clustering of multiple ligands, the distance between the predicted pocket centroid and the observed binding site (DCC) was chosen.

For each observed binding site in the LIGYSIS reference, the "best" prediction for each method was chosen. This is defined as the prediction with the minimum Euclidean distance to the observed pocket centroid or DCC. Once the observed-predicted pairs were obtained, only those with DCC $\leq$ 12 Å were considered as correct predictions. A threshold of 12 Å was chosen as 4 Å is too strict a threshold when using DCC. A threshold of 4 Å works well for the distance to closest ligand atom (DCA) but does not for DCC. The top-N and N+2 ranking predictions were considered to calculate success rate, or recall (Equation 4.19), and maximum recall was calculated by considering all predictions, regardless of their score or rank. N represents the number of observed sites for a given protein.

$$\text{Success rate (\%)} = 100 \times \frac{\text{observed sites with predicted site DCC} \leq 12 \text{ Å}}{\text{observed sites}} \tag{4.19}$$

Additionally, instead of conventional ROC, ROC_{100} [511, 512] was used to measure the predictive performance of the methods. To do this, for each method, all predictions across dataset proteins were ranked based on pocket score and cumulative true positives were plotted against cumulative false positives until 100 false positives were reached. In a similar way, a precision curve can be calculated by taking the top-N, in this case N = 1000, predictions. This curve measures how precision changes as more predictions with lower scores are considered. This is indicative of how informative pocket scores are.

Precision and recall are key measures for evaluating the performance of ligand binding site prediction methods. However, these indicators are calculated and interpreted slightly differently depending on the context a prediction is analysed, i.e., pocket *vs* residue level, as well as the metric employed, e.g., F1 score, MCC, ROC or PR curves. At the residue level, the prediction is a binary classification task, where each residue is classified as binding (label: 1) or non-binding (label: 0). In this case, precision reflects the proportion of residues predicted as binding that are true, i.e., observed in the reference data. Recall measures the proportion of true binding residues that are correctly identified. For the calculation of F1 and MCC, a residue is labelled "positive" or "negative" depending on whether it is part of a predicted pocket. However, for ROC and PR curves, the positive and negative labels are derived based on a ligandability threshold, t_{LS}. Prediction at the pocket level represents a multi-instance prediction task. Precision indicates the proportion of predicted pockets that are observed in the reference data whilst recall represents the proportion of true binding pockets that are correctly predicted. It is important to keep this in mind to correctly interpret precision and recall across different contexts.

To measure the similarity in shape and residue membership between predicted and observed pockets, relative residue overlap (RRO) and relative volume overlap (RVO) were employed. For an observed-predicted pocket pair, RRO represents the proportion of observed ligand-binding residues (R_o) that are covered by the predicted pocket residues (R_p) (Equation 4.20). The POVME output was used for the calculation of RVO (Figure 4.12). POVME defines the volume of a pocket as a series of equidistantly spaced spheres of unit-volume. As predictions by the different methods were on the same coordinate reference,

these pocket volume spheres were already aligned, and the volume overlap was calculated as the proportion of spheres in the observed pocket (V_o) that overlap with the predicted pocket spheres (V_p) (Equation 4.21).

$$\text{RRO (\%)} = 100 \times \frac{|R_p \cap R_o|}{R_o} \tag{4.20}$$

$$\text{RVO (\%)} = 100 \times \frac{|V_p \cap V_o|}{V_o} \tag{4.21}$$

4.2.10 Statistics and reproducibility

VN-EGNN was installed from their repository [514] and run locally. Likewise, for IF-SitePred [515]. GrASP was obtained from their repository [516] and predictions generated using their Google Colab Notebook [517]. PUResNet predictions were obtained through the PUResNet v2.0 web server [518]. DeepPocket was installed from GitHub [519] and executed locally. P2Rank v2.4.2 [520] was used to run all predictions as well as PRANK re-scoring. fpocket v4.0 was installed via Conda [521]. For PocketFinder, Ligsite and Surfnet, the ConCavity v0.1 "+" re-implementations were employed [522].

Other recent methods including RefinePocket [166], EquiPocket [165], GLPocket [164], SiteRadar [163], NodeCoder [162], RecurPocket [159], PointSite [158], DeepSurf [157], Kalasanty [155], BiteNet [154], GRaSP [153] or DeepSite [151] were not included in this analysis due to technical reasons. Peer-reviewed, open-source methods with publicly accessible code, clear installation instructions, well defined dependencies and accessible command line interfaces were prioritised for this benchmark (Table 4.1). This set of thirteen methods is representative of the state-of-the-art within the field.

ChimeraX v1.7.1 [70] was used for structural visualisation in all figures unless otherwise stated, in which case PyMOL v2.5.2 was employed [389]. Performed statistical tests were two-tailed and $\alpha = 0.05$. Sample sizes and measures of significance are reported in text, figures and legends.

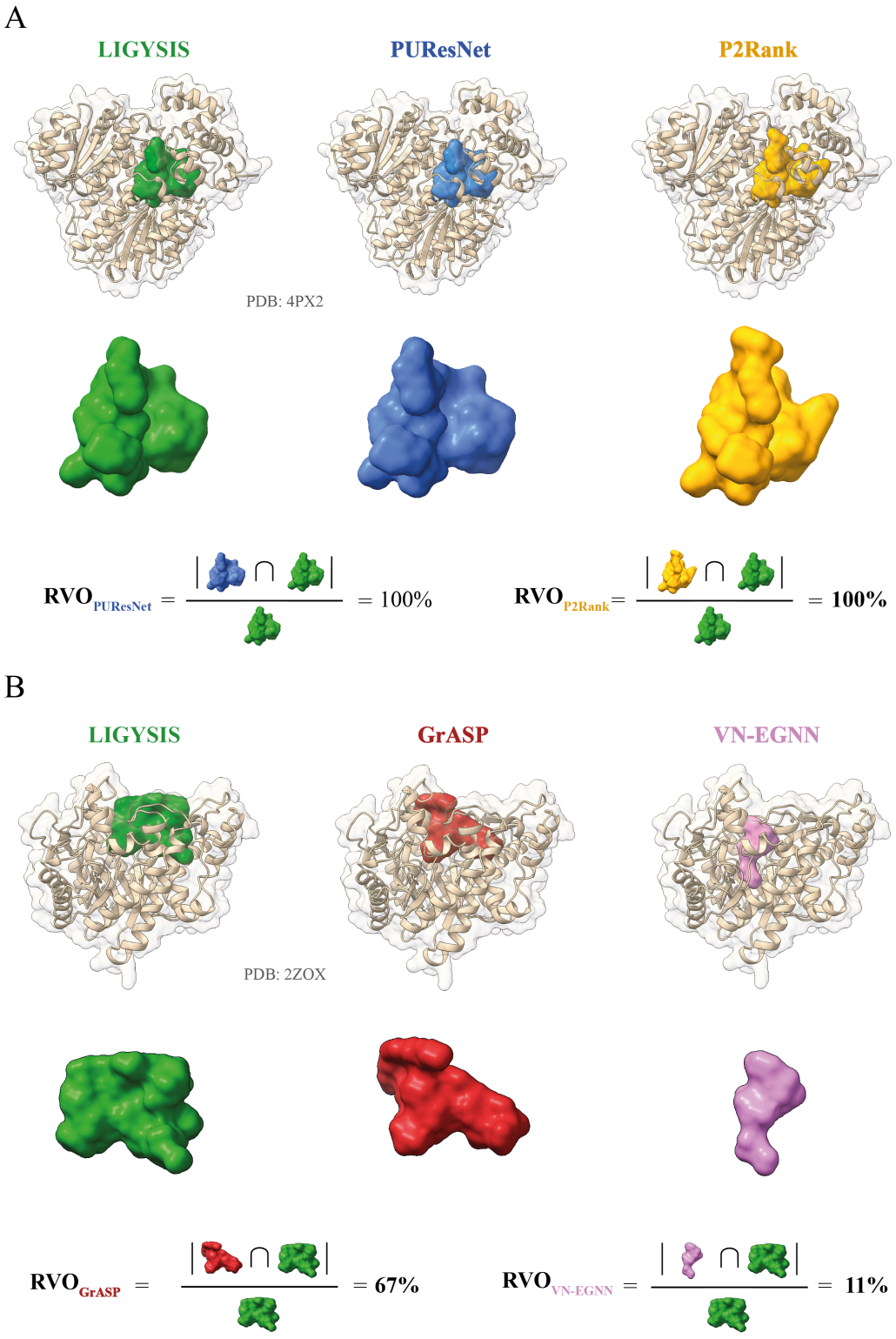

Figure 4.12. Relative Volume Overlap (RVO) calculation. (A) Example of two very accurate predictions by PUResNet and P2Rank on PDB: 4PX2 [448]. Pocket volumes were calculated with POVME 2.0 and represented by coloured surfaces. These volumes result from the addition of unit-volume spheres on a grid. To obtain the RVO, the intersection of these spheres between predicted and observed site was divided by the number of observed pocket spheres. Both predictions cover the entirety of the observed pocket volume; **(B)** GrASP and VN-EGNN predictions of a site on PDB: 2ZOX [513]. Th volumes of these predicted sites overlap less with the observed site: RVO = 67% for GrASP and RVO = 11% for VN-EGNN.

4.2.11 Data and code availability

The main results tables and files necessary to replicate the analysis described in this Chapter can be found here: https://doi.org/10.5281/zenodo.13121414 [523]. Software developed to carry out this analysis is found in this GitHub repository: https://github.com/bartongroup/LBS-comparison [524].

4.3 Results

4.3.1 The LIGYSIS dataset

The human subset of the LIGYSIS dataset includes protein-ligand complexes for 3448 human proteins. For each protein, biologically relevant protein-ligand interactions – in accordance with BioLiP [228] – were considered across the PISA-defined [237] biological assemblies of the multiple entries deposited in the PDBe [345]. Ligands were clustered using their protein interaction fingerprint to identify ligand binding sites as described in Chapter 3. The full LIGYSIS dataset includes ≈30,000 proteins with known ligand-bound complexes. Here, the human subset of LIGYSIS is employed as a manageable set to run all prediction methods on and referred to as *LIGYSIS* for brevity.

The LIGYSIS dataset differs from previous train and test sets for ligand binding sites by considering biological units, aggregating multiple structures of the same protein and removing redundant protein-ligand interfaces. The asymmetric unit is the smallest portion of a crystal structure that can reproduce the complete unit cell through a series of symmetry operations. The asymmetric unit often does not correspond to the biological assembly, or unit, and relying on it can lead to artificial crystal contacts or redundant protein-ligand interfaces. The biological unit is the biologically relevant and functional macromolecular assembly for a given structure and might be formed by one, multiple copies or a portion of the asymmetric unit [525]. LIGYSIS consistently considers biological units, which is key in any analysis that delves into molecular interactions at residue or atomistic level. An example of this is illustrated in Figure 4.13 with PDB: 1JQY [526], part of the HOLO4K

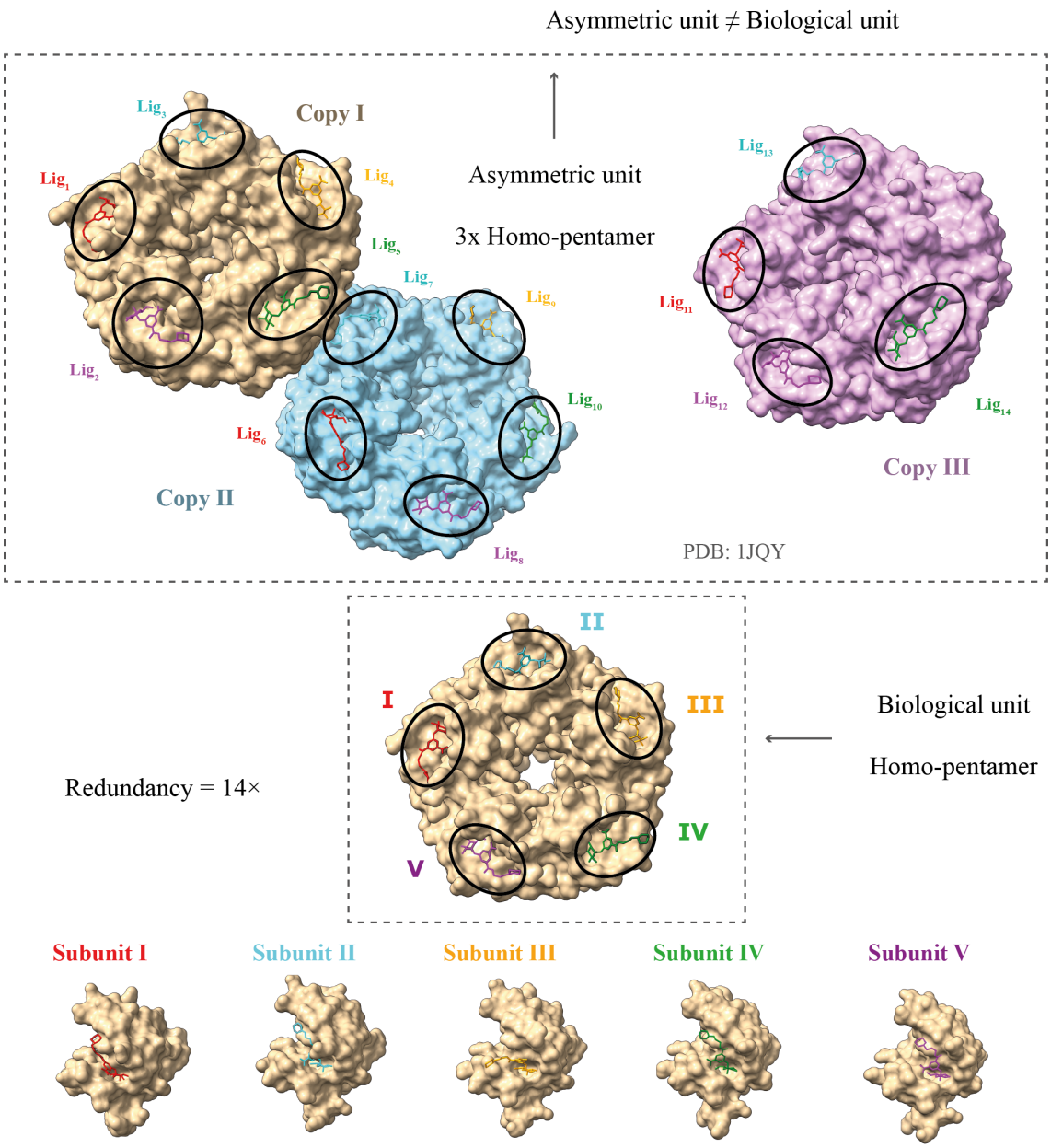

Figure 4.13. Redundancy in protein-ligand interfaces (I). For PDB: 1JQY, the asymmetric unit comprises three copies of a homo-pentamer, whereas the biologically functional assembly is a single pentamer. An A32 ligand molecule binds to each copy, except for one, of each of the three pentamers. This results in the same protein-ligand interface repeated 14 times, i.e., 14× redundancy. The dashed rectangles indicate the asymmetric and biological units.

dataset, where the asymmetric unit is formed by three copies of a homo-pentamer, whereas the biological unit comprises a single pentamer. In this structure, 14 molecules of BMSC-0010 (A32) interact with 14 copies of *Escherichia coli* heat-labile enterotoxin B chain (P32890). This protein-ligand interface is the same repeated 14 times.

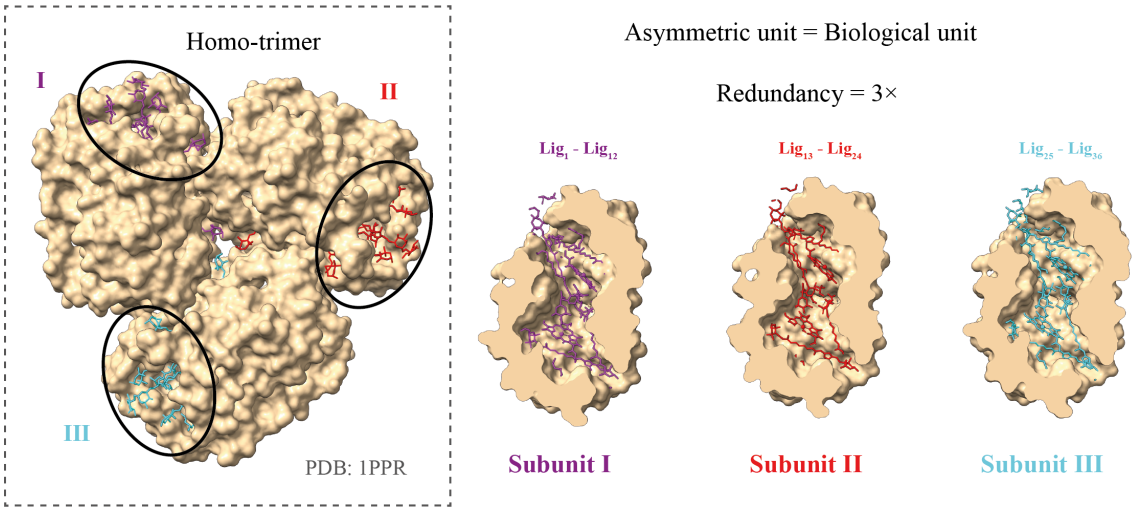

Figure 4.14. Redundancy in protein-ligand interfaces (II). For PDB: 1PPR both the asymmetric and biological units are a homo-trimer. Different molecules of the same ligands are binding to the same interfaces across the three copies of the trimer, i.e., 3× redundancy. A dashed rectangle indicates the asymmetric (and biological) unit.

Protein-ligand interface redundancy can also be an issue when the asymmetric unit equals the biological assembly (Figure 4.14). In PDB: 1PPR [527], also in HOLO4K, molecules of chlorophyll A (CLA), peridinin (PID) and digalactosyl diacyl glycerol (DGD) bind to the three copies of a peridinin-chlorophyll a-binding protein 1, chloroplastic, PCP, (P80484) trimer, resulting in a redundancy of 3×. To account for this, LIGYSIS considers unique non-redundant protein-ligand interfaces by retrieving the UniProt sequence numbers of the residues the ligands interact with, so 1/14 interfaces were retrieved for PDB: 1JQY and 12/36 for PDB: 1PPR.

Figure 4.15 shows the comparison between PDB: 4GQQ [528], part of the PDBbind dataset, and the LIGYSIS entry for human pancreatic alpha-amylase (P04746), which representative structure is also 4GQQ. The entry in PDBbind represents a single protein-ligand complex, whereas LIGYSIS clusters 195 ligands across 51 structures to define 13 unique ligand binding sites. LIGYSIS aggregates all biologically relevant protein-ligand interactions for a protein in a non-redundant manner, thus representing the most complete and integrative protein-ligand binding dataset up to date. For this reason, LIGYSIS is proposed as a new benchmark dataset for the prediction of ligand binding sites and used in this Chapter to evaluate a set of thirteen ligand binding site prediction tools.

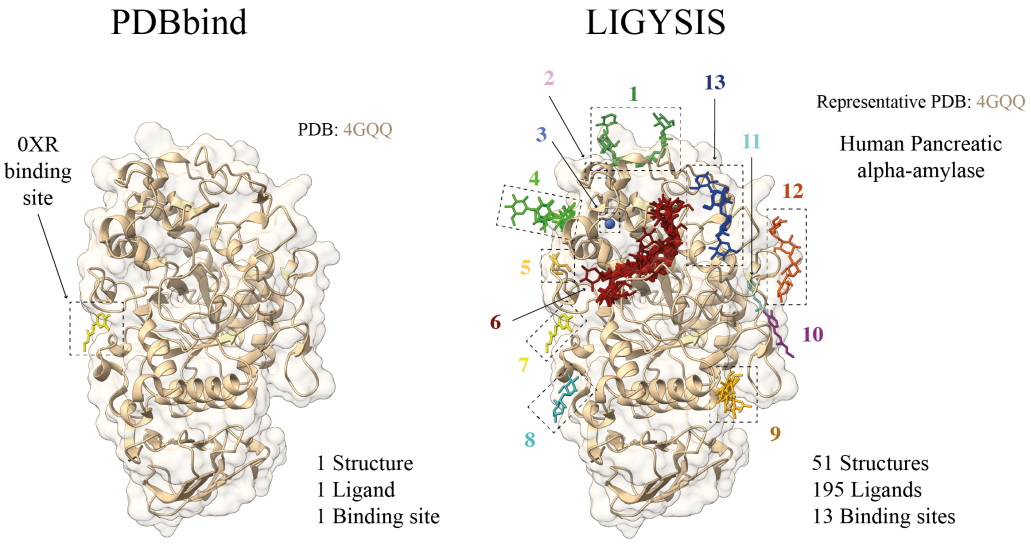

Figure 4.15. Comparison of PDBbind and LIGYSIS. PDBbind provides complexes between a protein and the most biologically relevant ligand in a structure. For PDB: 4GQQ, this is ethyl caffeate (0XR). LIGYSIS considers all unique biologically relevant protein-ligand interactions across all structures for a given protein. For human pancreatic alpha-amylase (P04746), which representative structure is 4GQQ, 13 ligand binding sites were defined from 195 ligands across 51 structures. LIGYSIS provides a better representation of the ligand-binding capabilities of a protein than a single protein-ligand complex and constitutes therefore a better ground truth for benchmarking ligand binding site prediction tools.

4.3.2 Comparison of datasets

To assess the scope and limitations of the methods surveyed in this Chapter, their training and test sets were compared to LIGYSIS by number of sites per protein, number of interacting protein chains per ligand site, ligand size, ligand site size and ligand composition. sc-PDB$_{FULL}$ represents the full sc-PDB, used for training by DeepPocket; bMOAD$_{SUB}$ the subset of binding MOAD used for training by IF-SitePred; PDBbind$_{REF}$ the reference subset of PDBbind, which VN-EGNN uses for testing. Only original datasets were considered in this analysis e.g., HOLO4K, but not HOLO4K$_{Mlig}$ nor HOLO4K$_{Mlig+}$, HAP or HAP-small. The same goes for Mlig and Mlig+ versions of COACH420, sc-PDB$_{SUB}$ and sc-PDB$_{RICH}$. *ALL** represents all methods in this work except for fpocket, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}. Table 4.4 summarises the size of the datasets, which methods employ them and their overlap with LIGYSIS, which differs from all other sets since it considers biologically relevant ions, comprising ≈ 40% of its sites. For additional reference, LIGYSIS$_{NI}$, a subset of LIGYSIS without ions, is included in this analysis.

Dataset	Type	# Structures	# Sites	# Ligands	Overlap (%)	Methods
LIGYSIS	NEW	3448	8244	**65,116**	–	–
$LIGYSIS_{NI}$	NEW	2275	4572	38,595	–	–
$sc\text{-}PDB_{FULL}$	TRAIN	**17,594**	**17,594**	17,594	**801 (9.7)**	VN-EGNN, GrASP, PUResNet, DeepPocket
$bMOAD_{SUB}$	TRAIN	5899	11,184	11,184	606 (7.6)	IF-SitePred
CHEN11	TRAIN	**244**	**479**	**479**	**40 (0.5)**	PRANK, P2Rank
$PDBbind_{REF}$	TEST	5316	5316	5316	310 (3.8)	VN-EGNN
SC6K	TEST	6147	6147	6147	259 (3.1)	DeepPocket
HOLO4K	TEST	4009	10,175	10,175	207 (2.5)	*ALL**
COACH420	TEST	413	624	624	41 (0.5)	VN-EGNN, GrASP, DeepPocket, PUResNet, P2Rank
JOINED	TEST	557	752	752	110 (1.3)	PRANK

Table 4.4. Datasets summary statistics. # Structures, # Sites and # Ligands represent the number of PDB structures, ligand sites and total number of ligands for each dataset. For LIGYSIS and $LIGYSIS_{NI}$, 3448 and 2775, are the number of structural segments, each represented by a single chain. For each segment, biologically relevant ligands across structures were considered: N = 23,321 (LIGYSIS) and N = 19,012 ($LIGYSIS_{NI}$). The number of ligands is not equal to the number of sites for LIGYSIS and $LIGYSIS_{NI}$, as ligands from multiple structures of the same protein are aggregated into unique sites. Overlap is the number of LIGYSIS binding sites represented by at least one protein-ligand complex for a given dataset. Percentage relative to LIGYSIS also reported. Methods represents the ligand site predictors that use these datasets for training or test. For # Structures, # Sites and # Ligands, highest values are coloured in bold blue font and lowest in orange. This is the other way around for Overlap (%).

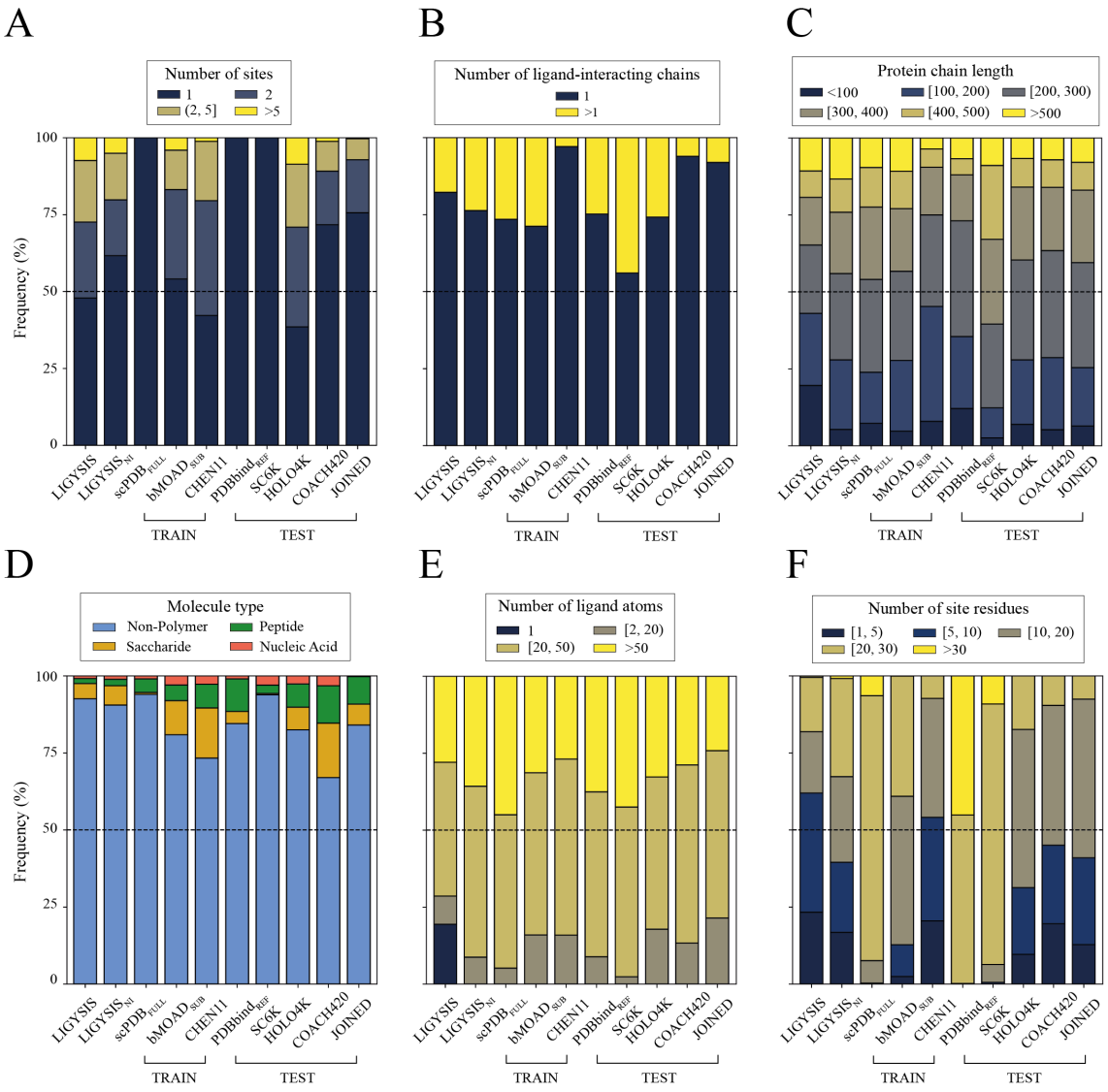

Figure 4.16. Comparison of datasets (I). Panels A, B, C, E and F plot the frequencies (%) of binned intervals of a discrete variable, coloured using the *cividis* palette. Interval ranges were selected to facilitate data interpretation. **(A)** Number of ligand binding sites per dataset entry; **(B)** Number of ligand-interacting protein chains. This represents whether the ligand interacts with a single protein chain or more; **(C)** Length of ligand-interacting protein chains (number of amino acids); **(D)** Ligand molecule type frequency as described in the CCD [395]; **(E)** Number of ligand atoms; **(F)** Binding site size, i.e., number of ligand-interacting protein residues. Dashed lines are drawn at frequency = 50%. $\text{LIGYSIS}_{\text{NI}}$ – a subset of LIGYSIS with no ions (NI) – is included in this analysis, as most training and test datasets do not consider ions.

Figure 4.16 A shows the number of binding sites per entry across datasets. $\text{sc-PDB}_{\text{FULL}}$, $\text{PDBbind}_{\text{REF}}$ and SC6K only consider the most relevant ligand for each entry. COACH420 and JOINED mostly present single-ligand entries (≈70%). The datasets $\text{bMOAD}_{\text{SUB}}$ (46%) and CHEN11 (58%) present more similar distributions to LIGYSIS, where 54% of the protein chains present more than one binding site. This percentage decreases for

$LIGYSIS_{NI}$ as ion sites are removed: 38%. HOLO4K presents the highest proportion of multi-ligand entries: 62%. Both HOLO4K and COACH420 are based on asymmetric units and not biological assemblies. For HOLO4K, 1811 structures (40%) present different numbers of chains between the asymmetric and biological units. This is even more frequent in COACH420: 234 (56%). Moreover, multimeric complexes might present the same protein-ligand interface repeated across the copies of the complex (Figure 4.13 and Figure 4.14). Considering predictions of these interfaces as independent could lead to the overestimation of the performance of a predictor. Regarding the number of chains interacting with a given ligand (Figure 4.16 B), CHEN11 and COACH420 present the smaller fraction of multimeric protein-ligand interactions: 3% and 6%, respectively, whereas SC6K presents the highest (44%). The rest of the methods range between 20-30%. There are no striking differences in the size of the ligand-interacting proteins, represented by the number of residues (Figure 4.16 C). Figure 4.16 D represents the ligand type composition of the datasets. Non-polymer ligands dominate all datasets (>66%) and the proportion of peptides and nucleic acids differ across datasets, with JOINED and LIGYSIS presenting fewer ligands of these types (0.9% and 1.6%). $sc\text{-}PDB_{FULL}$ and SC6K are depleted in saccharides (<1%). Figure 4.16 E depicts the difference in the number of atoms of the ligands in each dataset. LIGYSIS is, as expected, different due to its ion ligand content. However, there is no difference between $LIGYSIS_{NI}$ and the other datasets. Figure 4.16 F conveys the difference in the number of ligand-interacting residues. LIGYSIS has the largest proportion of small sites, 1-10 residues (56%). This is directly related to its prominent ion component. This frequency decreases to 36% when ions are removed in $LIGYSIS_{NI}$. CHEN11, COACH420, JOINED and HOLO4K are more similar to $LIGYSIS_{NI}$, whereas $sc\text{-}PDB_{FULL}$, $PDBbind_{REF}$ and SC6K are clearly different and present almost exclusively large sites, larger than 20 amino acids (>90%).

Figure 4.17 explores the ligand diversity on each dataset by showing the top-5 most frequent ligands per dataset, the percentage of the total number of ligands they represent, as well as Shannon's entropy, *H'*. Shannon's entropy is a measure of diversity. Larger numbers indicate a more evenly spread distribution of a larger number of different molecules,

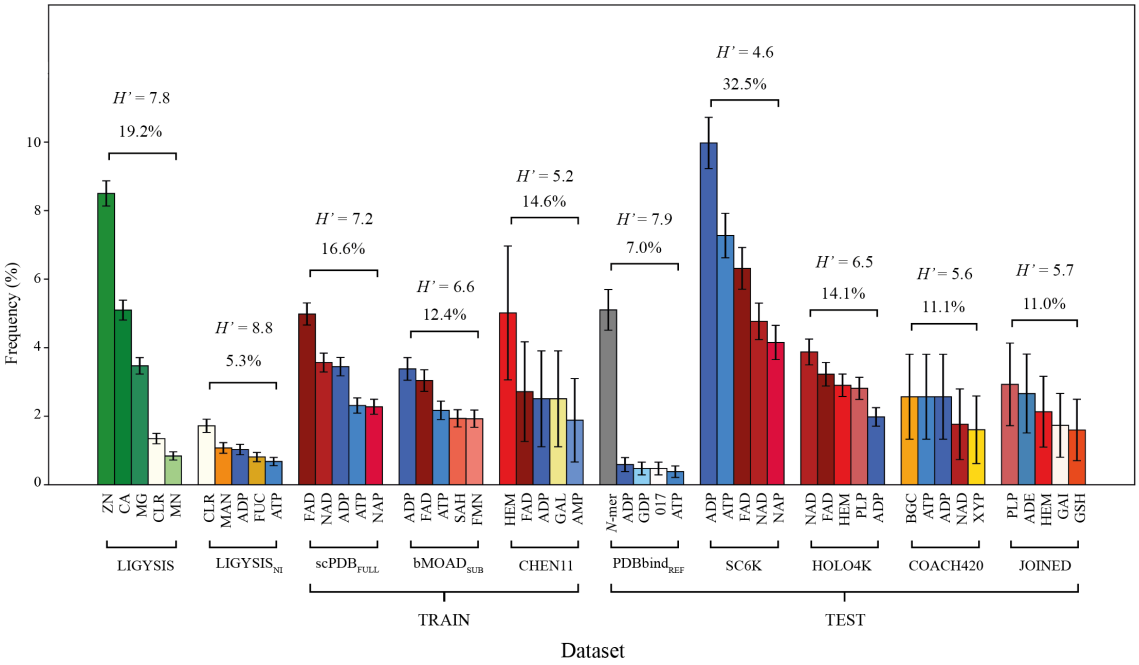

Figure 4.17. Comparison of datasets (II). Five most frequent ligands per dataset. Error bars represent 95% confidence interval of the proportion [264]. Ligands of similar type are coloured in shades of the same colour: greens for ions, reds for co-factors, blues for energy-carrier molecules, yellows for sugars, grey for peptides and white for other non-polymeric ligands. Above the bars, Shannon's entropy and the proportion of all ligands in each set covered by these top-5 can be found. Both are measures of ligand diversity within each dataset. $\text{LIGYSIS}_{\text{NI}}$, a subset of LIGYSIS without ions, is included in this analysis, as most training and test datasets do not consider ions. 017: darunavir; ADE: adenine; BGC: glucose; CLR: cholesterol; GAI: guanidine; GSH: glutathione; MAN: mannose; FUC: fucose; NAP: nicotinamide-adenine-dinucleotide phosphate; SAH: S-Adenosyl-L-homocysteine; FMN: flavin mononucleotide; GAL: galactose; *N*-mer: protein peptides of *N* amino acids; PLP: vitamin B6 phosphate; XYP: xylose.

whereas small numbers indicate higher frequency of a few ligands. While four out of the top-5 ligands of LIGYSIS are ions – Zn^{+2} (ZN), Ca^{+2} (CA), Mg^{+2} (MG), Mn^{+2} (MN) – and represent 19.2% of all ligands, its diverse composition is comparable to that of $\text{PDBbind}_{\text{REF}}$. Removing ions, $\text{LIGYSIS}_{\text{NI}}$ becomes the most diverse dataset with $H' = 8.8$ and its top-5 ligands covering only 5.3% of all ligands in the set. SC6K is the least diverse with its top-5 most frequent ligands covering 33% of all ligands. All datasets, except for LIGYSIS, $\text{LIGYSIS}_{\text{NI}}$ and $\text{PDBbind}_{\text{REF}}$, are dominated by co-factor ligands, such as flavin-adenine dinucleotide (FAD), nicotinamide-adenine dinucleotide (NAD) and haem (HEM), or energy carrier molecules such as adenine tri-, di- and monophosphate (ATP, ADP, AMP). Short peptides (<10 aas) are the most common ligands in $\text{PDBbind}_{\text{REF}}$ (5%), and energy carriers represent <2% of the top-5 ligands. Cholesterol (CLR), mannose

(MAN) and fucose (FUC) are some of the most common ligands in $LIGYSIS_{NI}$. LIGYSIS (including ions) was utilised for the pocket characterisation and performance evaluation analyses.

4.3.3 Binding pocket characterisation

After removing backbone-only atom chains and those missing UniProt residue mappings, the final LIGYSIS set included 2775 protein chains. Not all methods predicted pockets on all protein chains. VN-EGNN, GrASP, fpocket, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+} predicted in >99% of the chains, $P2Rank_{CONS}$ on 93%, followed by P2Rank on 86%, PUResNet and $DeepPocket_{SEG}$ (85%) and finally IF-SitePred, which only predicted pockets on 75% of the chains. PUResNet, DeepPocket, $P2Rank_{CONS}$ and P2Rank often fail to predict on smaller proteins (<100 aas), as well as non-globular or elongated proteins, representing 60-80% of proteins with no predicted pockets. However, for IF-SitePred, larger globular proteins represent ≈50% of all proteins where this method does not predict pockets (Figure 4.18).

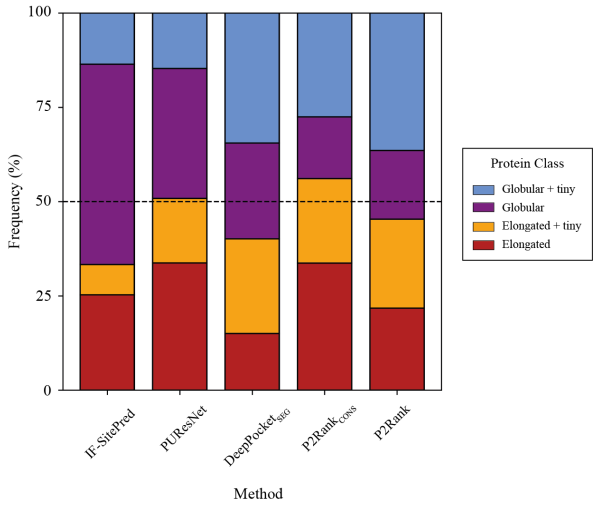

Figure 4.18. Where methods do not predict any sites. IF-SitePred does not predict any ligand binding sites on 700 of the 2775 protein chains in the LIGYSIS set (25%), PUResNet on 415 (15%), $DeepPocket_{SEG}$ (426; 15%), $P2Rank_{CONS}$ (196; 7%) and P2Rank (373; 13%). All methods struggle to predict on elongated proteins, regardless of their size, as well as on tiny globular proteins. Globular proteins comprise the most common group amongst the proteins with no predictions for IF-SitePred (53%). The dashed line indicates frequency of 50%.

Method	**Coverage** (%)	**# Total Pockets**	**# Pockets per protein**	R_g (Å)	**MCD** (Å)	**MRO**
LIGYSIS (ref)	2775	6882	1, 1, 27	5.9	14.1	0
(**d**) VN-EGNN	2764 (99.6%)	13,582 (×2.0)	1, 5, 7	5.9	**1.1**	**0.85**
(**d**) IF-SitePred	**2075 (74.8%)**	44,948 (×6.5)	1, **20**, 129	5.9	3.4	0.55
(**d**) GrASP	2771 (99.9%)	4694 (×0.7)	1, 1, 12	7.9	21.4	0
(**d**) PUResNet	2360 (85.1%)	2621 (×0.4)	1, 1, 4	8.1	27	0
(**d**) $\text{DeepPocket}_{\text{SEG}}$	2349 (84.7%)	21,718 (×3.2)	1, 6, 196	7.7	4.6	0.4
(**d**) $\text{P2Rank}_{\text{CONS}}$	2759 (92.9%)	12,412(×1.8)	1, 3, 57	7.1	13.9	0.05
(**d**) P2Rank	2402 (86.6%)	10,180 (×1.5)	1, 3, 85	7.1	13.8	0.05
(**d**) fpocket	2759 (99.4%)	**57,859 (×8.4)**	1, 17, **349**	6.3	9.7	0.15
(**d**) PocketFinder^{+}	2775 (100%)	8913 (×1.3)	1, 3, 23	8.6	18.7	0.05
(**d**) Ligsite^{+}	2775 (100%)	6903 (×1.0)	1, 2, 12	**9.1**	16.7	0.09
(**d**) Surfnet^{+}	2775 (100%)	9043 (×1.3)	1, 3, 40	8.4	17.2	0.07

Table 4.5. Ligand site characterisation. LIGYSIS is not a ligand site predictor, but a reference dataset derived from experimentally determined protein-ligand complexes. These predictions result from the default prediction by the methods, indicated by **(d)** preceding method names. Coverage is the number of chains where methods predict at least one pocket. Percentage is relative to number of LIGYSIS chains. Total number of pockets and ratio of predicted pockets per reference site in parenthesis, e.g., for each LIGYSIS site, fpocket predicts 8.4 pockets on average; Minimum, median and maximum number of predicted pockets per chain; Median pocket radius of gyration, R_g, (Å); Minimum centroid distance (MCD) (Å) measures how close predicted pockets are; Maximum residue overlap (MRO) measures residue overlap between pockets, e.g., the median overlap between VN-EGNN predicted pockets is 85%. Bold font indicates the most extreme values within each column.

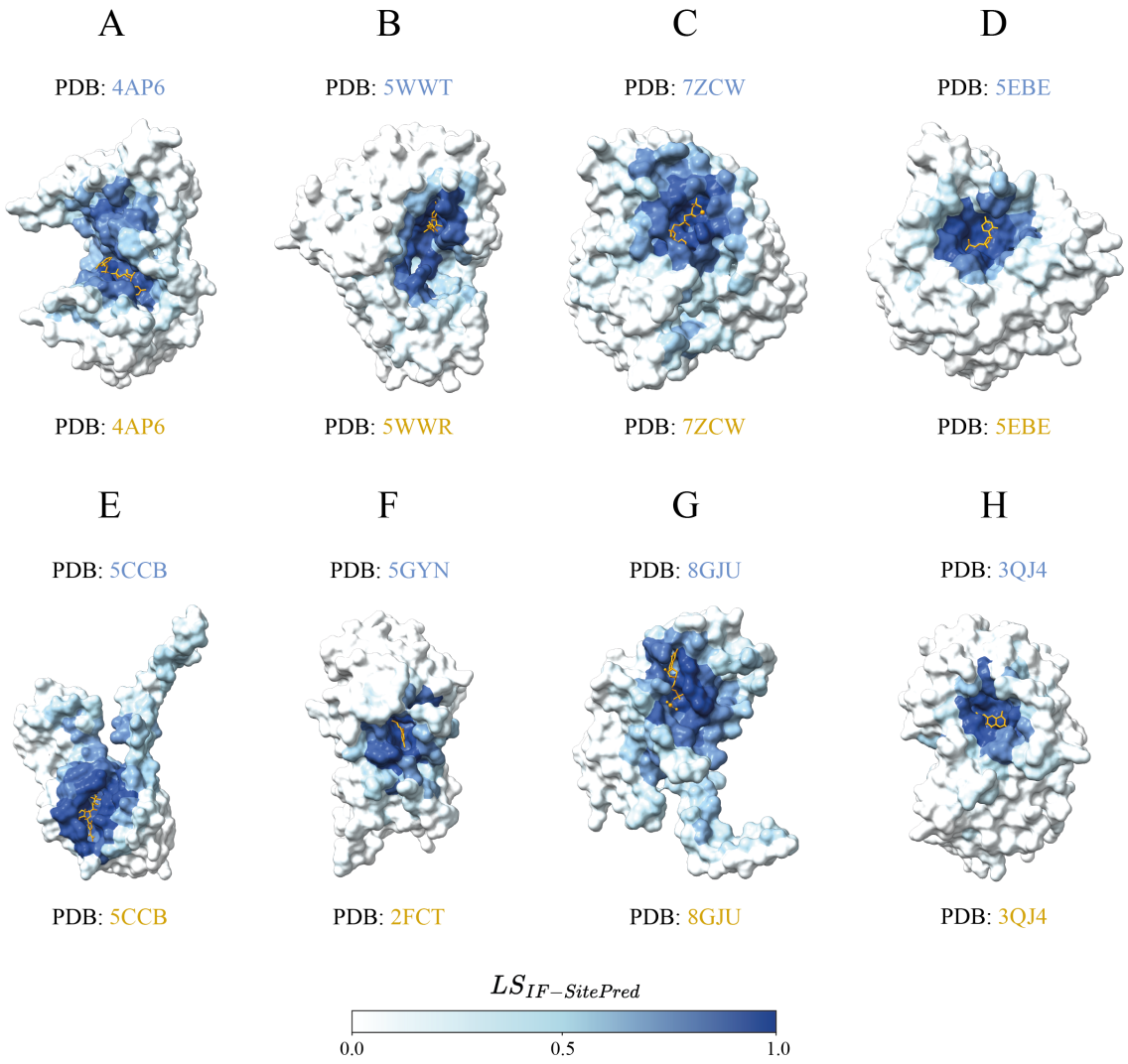

Figure 4.19. IF-SitePred "missed" predictions. Eight examples of human protein chains where IF-SitePred does not report any predicted ligand binding sites. Predictions are made on ligand-stripped chains. Ligand molecules, in orange, are superposed to illustrate how the ligandability scores calculated from the 40 IF-SitePred preddiction models recapitulate the observed binding site. These are protein representative chains and ligand molecules might not be observed in the same PDB entry; **(A)** GDP-fucose protein O-fucosyltransferase 2, Q9Y2G5, with GFB superimposed (PDB: 4AP6) [529]; **(B)** tRNA (cytosine(72)-C(5))-methyltransferase NSUN6, Q8TEA1, (PDB: 5WWT) [530] with superposed SFG (PDB: 5WWR) [530]; **(C)** Tubulin beta-2B chain, Q9BVA1, with G2P (PDB: 7ZCW) [531]; **(D)** Cyclic GMP-AMP phosphodiesterase SMPDL3A, Q92484, with C5P (PDB: 5EBE) [532]; **(E)** tRNA (adenine(58)-N(1))-methyltransferase catalytic subunit, Q96FX7, with SAH (PDB: 5CCB) [533]; **(F)** Chronophin, Q96GD0, (PDB: 5GYN) [534] with PLP (PDB: 2FCT) [535]; **(G)** Mitochondrial Methylmalonic aciduria type A protein, Q8IVH4, with GDP (PDB: 8GJU) [536]; **(H)** Renalase, Q5VYX0, (PDB: 3QJ4) with FAD [537]. Residues are coloured based on the ligandability score calculated by averaging the probabilities predicted by each of the 40 IF-SitePred prediction models. This is a score ranging 0-1 which is indicative of the likelihood of a given residue binding a ligand. Clear pockets can be observed formed by residues with high ligandability scores (darker blue colour), which agree with the sites where ligands bind.

Predicted residue ligandability scores for P2Rank$_{CONS}$, P2Rank and IF-SitePred (which is novel from this work), were examined for proteins with no predicted pockets. Figure 4.19 illustrates eight examples of proteins where residues with IF-SitePred high ligandability scores (Equation 4.10) cluster in space into clear binding sites that are not reported by this method. This suggests that IF-SitePred is too strict in selecting only those residues predicted as ligand-binding by *all* 40 models. The cloud point selection clustering approach and threshold in this method may also play a role in this.

Table 4.5 summarises the ligand site characterisation analysis. fpocket predicts the most sites out of all the methods, with 57,859, followed by IF-SitePred (44,948), DeepPocket$_{SEG}$ (21,718), VN-EGNN (13,582), P2Rank (12,412), P2Rank$_{CONS}$ (10,180), Surfnet^{+} (9043), PocketFinder^{+} (8913), Ligsite^{+} (6903), GrASP (4694) and PUResNet, which predicts fewest sites (2621). LIGYSIS defines 6882 binding sites from experimental data. Relative to LIGYSIS, the prediction methods have ratios of predicted/defined sites ranging from 8.4 (fpocket) to 0.4 (PUResNet) with P2Rank$_{CONS}$ in the middle, predicting 1.5 pockets per observed reference site. IF-SitePred, DeepPocket$_{SEG}$, P2Rank and fpocket predict more pockets on larger protein chains, whereas the rest of methods do not (Figure 4.20). This effect is most clear with fpocket, which predicts 350 pockets for chain A of PDB: 7SUD [538], a structure of the DNA-dependent protein kinase catalytic subunit, DNPK1, (P78527) with 3736 amino acid residues. In contrast, VN-EGNN, which initially places $K = 8$ virtual nodes, results in a maximum of 8 predicted pockets, regardless of protein chain size, and PUResNet predicts a single pocket in 90% of the proteins.

Figure 4.21 and Figure 4.22 represent how the eleven sets of unique ligand site predictors compare to each other, as well as to LIGYSIS, which *defines* ligand sites from experimentally determined biologically relevant protein-ligand complexes. There are eleven unique sets of predictions since DeepPocket$_{RESC}$ and fpocket$_{PRANK}$ do not predict their own pockets, but instead re-score and re-rank original fpocket predictions. DeepPocket$_{SEG}$ predictions are different as new pocket shapes are extracted by its CNN segmentation module. Figure 4.21 A shows how PUResNet, VN-EGNN and GrASP differ from the other methods with a maximum of 4, 7 and 12 predicted pockets, respectively. PocketFinder^{+},

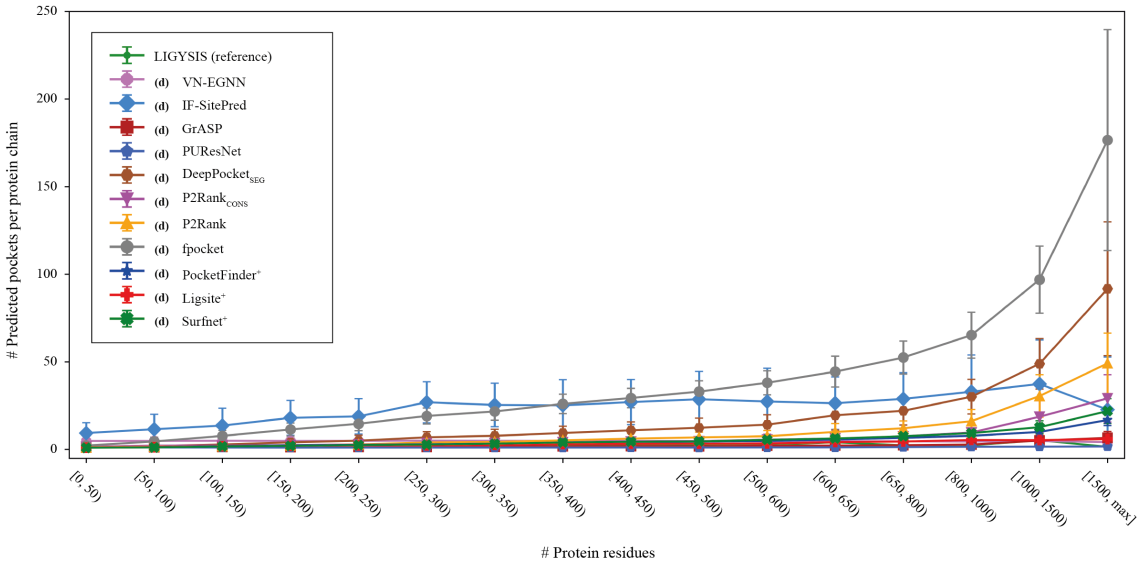

Figure 4.20. Number of pockets *vs* protein size. Number of defined (LIGYSIS) and predicted (other methods) sites against protein chain size, i.e., number of amino acid residues. The number of protein residues was discretised into intervals of 50 until 650, and larger intervals until the maximum of ≈3800. Error bars represent one standard deviation (SD).

Ligsite$^+$ and Surfnet$^+$ present narrow distributions like LIGYSIS and with medians of 1-3 pockets per protein. P2Rank$_{CONS}$ and P2Rank also present a median of 3 pockets per protein, but display wider distributions as they can predict up to 60 and 80 pockets per protein, respectively. Overall, P2Rank$_{CONS}$ predicts fewer pockets than P2Rank. DeepPocket$_{SEG}$, fpocket and IF-SitePred follow, with a median of 6, 17 and 20 pockets. The difference in number of pockets between DeepPocket$_{SEG}$ and DeepPocket$_{RESC}$ or fpocket is due to the fact that 60% of fpocket candidates were not extracted by the CNN segmentation module implemented in DeepPocket.

Figure 4.21 B shows the distribution of pocket radius of gyration, R_g. VN-EGNN and IF-SitePred differ from the rest of methods with narrow distributions and medians around 6 Å. These two methods do not report pocket residues. Instead, they were obtained using a distance threshold of 6 Å from the centroid, for VN-EGNN, and cloud points, for IF-SitePred. This is reflected by examining the percentage of pockets with R_g $>$ 10 Å, which is 0% and 0.1% for VN-EGNN and IF-SitePred. This is a striking difference compared to the LIGYSIS reference and other methods: 1.8% (fpocket), 4.8% (P2Rank), 5.7% (DeepPocket$_{SEG}$), 6.4% (GrASP), 6.5% (P2Rank$_{CONS}$), 11.6% (PUResNet), 12.6%

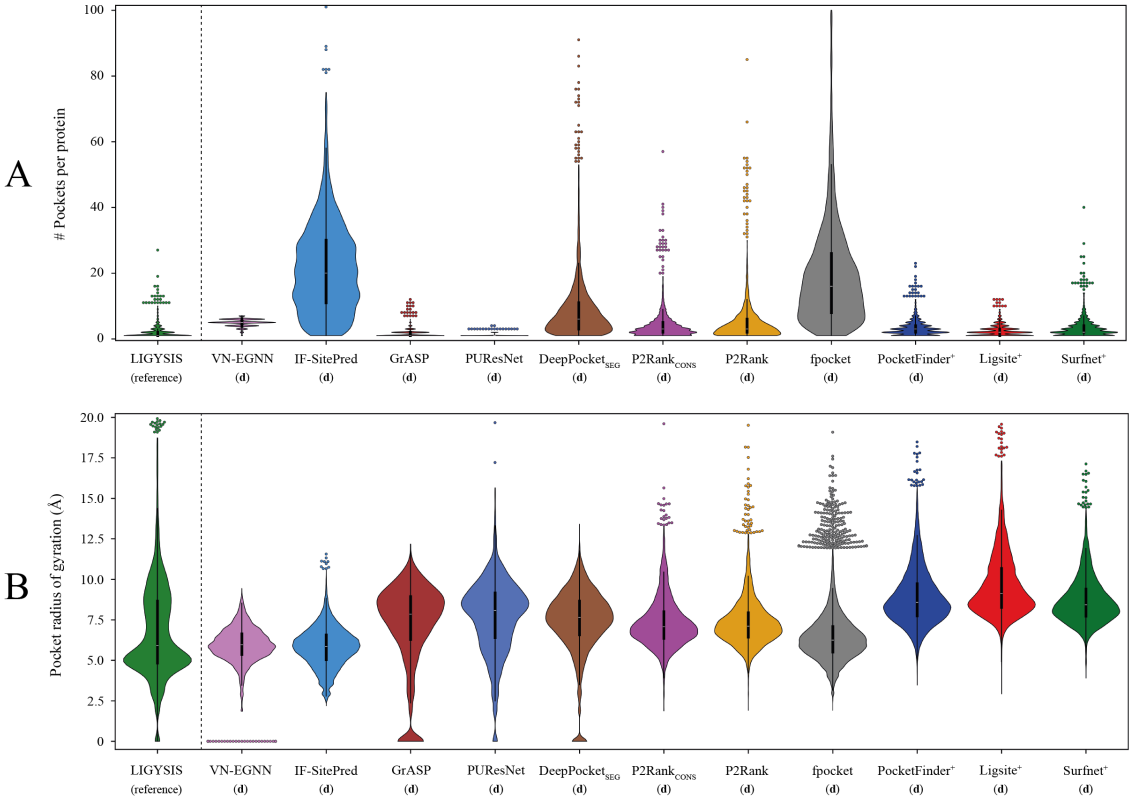

Figure 4.21. Binding pocket characterisation (I). Violin plots show the main distributions and swarm plots are used to show outliers. Data points farther than four standard deviations (SD) from the mean were considered outliers. The limit of the Y axis is the maximum non-outlier value plus a buffer value. This way, only the most extreme outliers are hidden, which maximises visual interpretation of the data whilst minimising the number of data points not shown. Within the violin plots are box plots representing the underlying distribution. A line represents the median and the box contains the interquartile range (IQR) and whiskers extend to 1.5 × IQR. **(A)** Number of pockets per protein; **(B)** Pocket radius of gyration, R_g, (Å).

(LIGYSIS), 16% (Surfnet$^+$), 21.4% (PocketFinder$^+$) and 33.5% (Ligsite$^+$). The latter three predict the sites with largest median $R_g \approx 9$ Å. VN-EGNN, GrASP, PUResNet and DeepPocket$_{\text{SEG}}$ predict sites with $R_g = 0$ Å. This is rather infrequent (7.8% GrASP) and $<$3% for the other three. These examples correspond to singletons, i.e., pockets formed by only one amino acid.

Figure 4.22 A illustrates how close predicted sites are to each other within a protein chain. Pairwise distances between the centroids of all ligand site pairs for a protein were calculated, and for each site, the minimum distance was taken. Sites predicted by VN-EGNN, IF-SitePred and DeepPocketSEG are very close to each other, with median distances ($\tilde{d}$) of 1.1, 3.4 and 4.6 Å, respectively. fpocket follows with $\tilde{d} = 9.7$ Å. The rest of the methods and LIGYSIS (reference) present median distances ranging between 13-18 Å

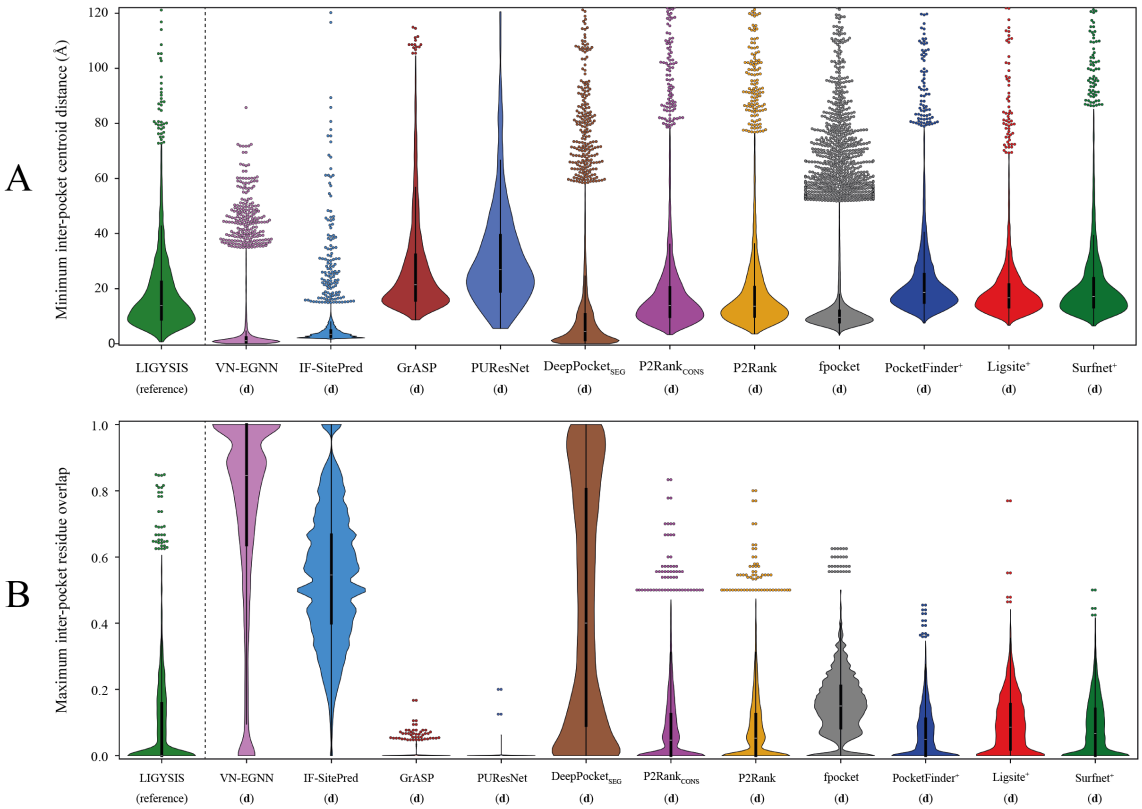

Figure 4.22. Binding pocket characterisation (II). Violin plots show the main distributions and swarm plots are used to show outliers. Data points farther than four standard deviations (SD) from the mean were considered outliers. The limit of the Y axis is the maximum non-outlier value plus a buffer value. This way, only the most extreme outliers are hidden, which maximises visual interpretation of the data whilst minimising the number of data points not shown. Within the violin plots are box plots representing the underlying distribution. A line represents the median and the box contains the interquartile range (IQR) and whiskers extend to 1.5 × IQR. **(A)** Minimum inter-pocket centroid distance (MCD) (Å). This is a measure of how close predicted pockets are to each other within a protein; **(B)** Maximum inter-pocket residue overlap (MRO). Residue overlap was calculated as Jaccard index. This is a measure of how much the pockets overlap in terms of binding residues.

(LIGYSIS, P2Rank$_{CONS}$, P2Rank, PocketFinder$^+$, Ligsite$^+$, Surfnet$^+$), and finally GrASP ($\tilde{d}$ = 21.7 Å) and PUResNet ($\tilde{d}$ = 27 Å). Both versions of P2Rank present the most similar distribution to what is observed in the LIGYSIS reference.

Figure 4.22 B depicts the overlap existing between residues that form the predicted pockets within a protein. All pairwise overlaps were calculated between pockets in a chain, and for each pocket, the maximum was taken. This is a measure of how much predicted pockets overlap with each other. This measure is directly related to how close pockets are, and so VN-EGNN, IF-SitePred and DeepPocket$_{SEG}$ present very high overlaps $\tilde{o}$ = 0.85, $\tilde{o}$ = 0.55 and $\tilde{o}$ = 0.4, respectively. fpocket follows with $\tilde{o}$ = 0.15, Ligsite$^+$ ($\tilde{o}$ =

0.09), Surfnet^{+} ($\tilde{o} = 0.07$), P2Rank$_{CONS}$, P2Rank and PocketFinder^{+} ($\tilde{o} = 0.05$), and finally LIGYSIS, GrASP and PUResNet with $\tilde{o} = 0$. GrASP is the only method of the thirteen presented here that clusters atoms directly, and as a result, overlap between pockets is minimal. Other methods cluster cloud points (IF-SitePred), SAS points (P2Ranks), voxels (PUResNet, DeepPocket), α-spheres (fpocket), or grid points (PocketFinder^{+}, Ligsite^{+}, Surfnet^{+}) but not residues, resulting consequently in higher overlapping.

Proximity in space between predicted sites, as well as residue overlap, are indicators of redundant ligand binding site prediction, i.e., duplicate predictions of a single observed ligand site. This is especially prevalent in predictions by VN-EGNN, IF-SitePred and DeepPocket$_{SEG}$. This phenomenon could negatively impact the precision and recall of these methods. Accordingly, correcting for redundancy should have a significant positive impact on the performance of these methods. In contrast, GrASP and PUResNet which predict a small number of pockets show low proximity and overlap of predicted sites, and so redundancy is not an issue.

4.3.4 Evaluation of predictive performance

4.3.4.1 Pocket level evaluation

The ideal ligand binding site predictor would have a high precision, i.e., most of the predictions it makes are correct, whilst maintaining a high recall, i.e., recapitulating most of the observed sites. Moreover, the ideal predictor returns predictions that are non-redundant, i.e., it does not predict the same pocket multiple times. Additionally, pockets are ranked in a systematic manner according to a strong pocket scoring scheme that captures well the nature of existing ligand binding sites, therefore ranking the predicted pockets from more likely (high score, top) to least likely (low score, bottom). An ideal predictor would also perform well at the residue level. This means it is able to capture the likelihood of a residue to bind a ligand. This can be done by means of a residue ligandability score, which might also highlight key residues, the more ligandable within a binding site. Ligand site prediction methods were benchmarked with these criteria in mind.

Method	Recall$_{\text{top-}N}$ (%)	Recall$_{\text{top-}N+2}$ (%)	Recall$_{\text{max}}$ (%)	Precision$_{1K}$ (%)	# TP$_{100\ FP}$	RRO (%)	RVO (%)
(**d**) VN-EGNN	27.5 (#11)	40.9 (#12)	49.3 (#10)	**92.5 (#1)**	**1301 (#1)**	**32.8 (#12)**	**27.6 (#11)**
(**d**) IF-SitePred	**19.8 (#12)**	**25.7 (#13)**	52.1 (#6)	91.0 (#2)	961 (#3)	46.5 (#11)	40.4 (#9)
(**d**) GrASP	48.0 (#2)	49.9 (#5)	50.0 (#8)	**92.5 (#1)**	1017 (#2)	54.5 (#7)	59.8 (#6)
(**d**) PUResNet	40.6 (#6)	41.1 (#11)	**41.1 (#12)**	81.6 (#6)	534 (#8)	61.0 (#4)	63.9 (#4)
(**d**) DeepPocket$_{\text{SEG}}$	35.4 (#10)	43.8 (#10)	56.5 (#5)	82.6 (#4)	670 (#5)	57.5 (#5)	60.3 (#5)
(**d**) DeepPocket$_{\text{RESC}}$	46.6 (#4)	58.1 (#2)	89.3 (#2)	81.7 (#5)	637 (#6)	53.1 (#9)	38.2 (#10)
(**d**) P2Rank$_{\text{CONS}}$	**48.8 (#1)**	53.9 (#3)	57.0 (#4)	90.7 (#3)	932 (#4)	56.4 (#6)	43.8 (#8)
(**d**) P2Rank	46.7 (#3)	51.9 (#4)	57.0 (#3)	79.2 (#7)	586 (#7)	54.4 (#8)	58.2 (#7)
(**d**) fpocket$_{\text{PRANK}}$	**48.8 (#1)**	**60.4 (#1)**	**91.3 (#1)**	81.7 (#5)	526 (#9)	52.6 (#10)	38.2 (#10)
(**d**) fpocket	38.8 (#8)	46.5 (#8)	**91.3 (#1)**	47.3 (#9)	94 (#11)	52.6 (#10)	38.2 (#10)
(**d**) PocketFinder^{+}	39.2 (#7)	47.8 (#7)	50.5 (#7)	42.0 (#10)	64 (#12)	72.3 (#2)	75.9 (#2)
(**d**) Ligsite^{+}	41.3 (#5)	48.4 (#6)	49.7 (#9)	52.3 (#8)	115 (#10)	**77.6 (#1)**	**77.0 (#1)**
(**d**) Surfnet^{+}	37.7 (#9)	45.8 (#9)	48.9 (#11)	**39.5 (#11)**	**61 (#13)**	71.7 (#3)	72.0 (#3)

Table 4.6. Pocket level evaluation. This table illustrates the performance of default methods indicated by (**d**) preceding method names. Recall considering top-*N*, *N*+2 and *all* predictions (max) regardless of rank, i.e., maximum recall. Precision for the top-1000 scored predictions. Number of TP reached for the first 100 FP (# TP$_{100\ FP}$). Mean relative residue overlap (RRO) for those sites correctly predicted and relative volume overlap (RVO) for sites that have a volume, i.e., are pockets or cavities, and not fully exposed sites, which do not have a volume. RRO and RVO represent the overlap in residues and volume relative to the observed site. See Section 4.2.9.2 for definitions of RRO and RVO. Bold font indicates the best (blue) and worst (orange) performing methods for each metric.

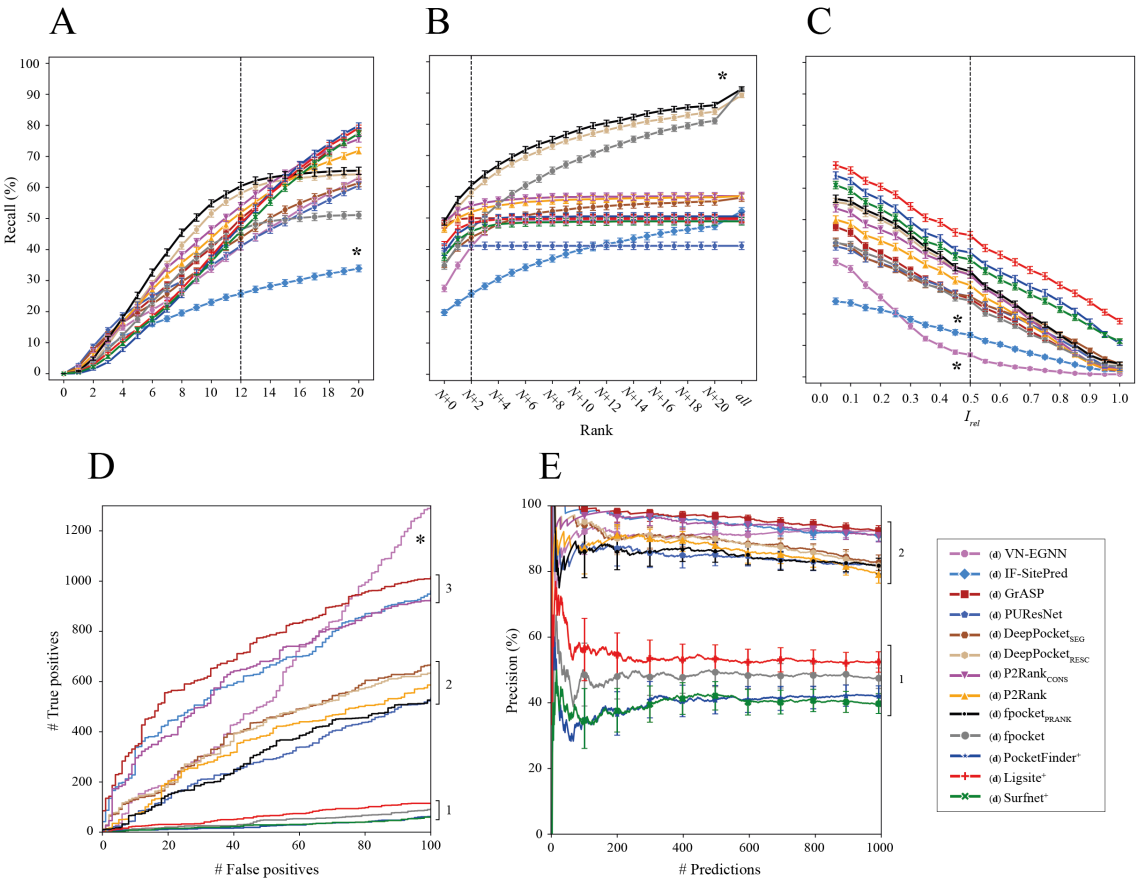

Figure 4.23. Ligand binding site prediction benchmark at the pocket level. These curves correspond to the default predictions of the thirteen methods, indicated by (**d**) preceding their names. **(A)** Recall, percentage of observed sites that are correctly predicted by a method within the top-N+2 predictions according to a DCC = 12 Å threshold; **(B)** Recall using DCC = 12 Å but considering increasing rank thresholds, i.e., top-N, N+1, N+2, etc. *all* represents the maximum recall of a method, obtained by considering all predictions, regardless of their rank or score; **(C)** Recall curve for top-N+2 predictions using I_{rel} as a criterion; **(D)** ROC_{100} curve (cumulative # TP against cumulative FP until 100 FP are reached); **(E)** Precision curve for the top-1000 predictions of each method across the LIGYSIS dataset, $Precision_{1K}$. Error bars represent 95% CI of the recall **(A-C)** and precision **(E)**, which is 100 × proportion. Numbers at the right of the panels indicate groups or blocks of methods that perform similarly for each metric. Stars (*) indicate outlier methods, or methods that perform very differently than the rest.

Figure 4.23 A illustrates the recall curve for top-N+2 pockets for each method, where N is the number of observed sites in a given protein. Reported recall is obtained using a threshold DCC = 12 Å (see Section 4.2.8.1 for details). Re-scored fpocket predictions by PRANK ($fpocket_{PRANK}$) and DeepPocket ($DeepPocket_{RESC}$) yield the highest recall with 60.4% and 58.1%, closely followed by $P2Rank_{CONS}$ (53.9%) and P2Rank (51.9%). The rest of the methods present recall <50% with PUResNet, VN-EGNN and IF-SitePred presenting the lowest recall of 41.1%, 40.9% and 25.7%, respectively (Table 4.6).

Figure 4.23 B shows the recall curve considering different top-N+X predictions. Most methods reach a plateau by top-N+5, as they do not predict many pockets. However, methods that predict more pockets per protein, such as IF-SitePred or fpocket, fpocket$_{\text{PRANK}}$, DeepPocket$_{\text{SEG}}$ and DeepPocket$_{\text{RESC}}$, which take fpocket predictions as a base, increase their recall as more predictions are considered. fpocket, fpocket$_{\text{PRANK}}$ and DeepPocket$_{\text{RESC}}$ reach a maximum recall of ≈90% when *all* predictions are considered, regardless of score or rank. The rest of the methods present maximum recall of ≈50-60%.

Figure 4.23 C depicts the recall curve when residue overlap is used instead of DCC as a criterion. In this case, Ligsite^{+}, PocketFinder^{+} and Surfnet^{+} came on top with recall ≈45% at $I_{rel} \geq 0.5$. This is explained by their prediction of massive cavities, that while often fully contain or overlap with the observed pocket, do not meet the DCC criterion, as their centroids are farther than 12 Å from the observed site.

Figure 4.23 D represents the cumulative number of TP against FP when predictions across the proteins in the reference dataset are sorted by score. This shows how effective the scoring scheme of each method is in ranking their predictions to reflect the nature of ligand binding sites. At 100 FP the # TP fall into three different blocks and one outlier: Ligsite^{+}, fpocket, Surfnet^{+} and PocketFinder^{+} are at the bottom with # TP$_{100\ \text{FP}} \ni (60, 120)$. Secondly, fpocket$_{\text{PRANK}}$, DeepPocket$_{\text{SEG}}$, DeepPocket$_{\text{RESC}}$, P2Rank and PUResNet follow with # TP$_{100\ \text{FP}} \ni (530, 670)$. Re-scoring fpocket predictions with PRANK or DeepPocket results in up to +500 TP. GrASP, IF-SitePred and P2Rank$_{\text{CONS}}$ present a high # TP$_{100\ \text{FP}}$ ranging 900-1000. Finally, VN-EGNN sits at the top with # TP$_{100\ \text{FP}}$ = 1301. However, this number might not be representative, as the # TP could be inflated due to the redundancy in the predictions by VN-EGNN. This is the same for IF-SitePred and DeepPocket$_{\text{SEG}}$. Redundant correct predictions of the same pocket count as multiple TP, whereas they should only count as 1 TP. Newer methods, e.g., GrASP, P2Rank$_{\text{CONS}}$, VN-EGNN and IF-SitePred, despite redundancy in prediction for the latter two, are better at ranking their predicted pockets, presenting up to 900 more TP for 100 FP than earlier methods. This means their scoring schemes are significantly better at capturing the essence of a ligand binding site. Including evolutionary conservation in P2Rank (P2Rank$_{\text{CONS}}$) results in an

increase of +346 TP relative to default P2Rank, indicating that the fewer predicted pockets (and their scores) are a more faithful representation of the observed LIGYSIS dataset.

Figure 4.23 E provides insight into the precision of the methods by examining how this metric changes as more predictions are considered. In the same manner as for Figure 4.23 D, predictions across proteins in the LIGYSIS dataset were sorted and cumulative precision was plotted for the top-1000 scoring predictions. Methods group into two clear blocks. Newer (machine learning-based) methods VN-EGNN, GrASP, IF-SitePred, $P2Rank_{CONS}$, $DeepPocket_{SEG}$, $fpocket_{PRANK}$, $DeepPocket_{RESC}$, PUResNet and P2Rank are highly precise ($Precision_{1K}$ = 80-95%). Earlier (geometry/energy-based) methods $Ligsite^{+}$, fpocket, $Pocket\text{-}Finder^{+}$ and $Surfnet^{+}$ present lower $Precision_{1K}$ of 40-50%. $fpocket_{PRANK}$ and $DeepPocket_{RESC}$ take fpocket (geometry-based) predictions as a starting point and achieve much higher # TP_{100FP} (+500) as well as $Precision_{1K}$ (+30%). This is further evidence that performance can be boosted with a solid scoring scheme and agrees with previous studies [115, 149, 150, 442].

Table 4.6 summarises these results and shows the mean relative residue overlap (RRO) and relative volume overlap (RVO), which measure how well predicted sites align with those observed in the LIGYSIS reference in terms of their shape. VN-EGNN and IF-SitePred present the smallest RRO and RVO, but it is important to note that these methods do not report pocket residues and so residues were taken within 6 Å of their centroid and pocket spheres, respectively. $PocketFinder^{+}$, $Ligsite^{+}$ and $Surfnet^{+}$ present unusually high average RRO and RVO (>70%). This is a consequence of the massive size of their predicted cavities, that rather than overlap with the observed site, fully contain and are much larger than it. This might not be convenient in the context of pocket finding for drug discovery, where more clearly defined drug-like sites might be of interest. GrASP, PUResNet and $DeepPocket_{SEG}$ present high values of RRO ≈ 60% and RVO ≈60% whilst presenting a size distribution more like LIGYSIS (Figure 4.21 B) and provide the best representation of the observed sites regarding shape and residue membership similarity.

4.3.4.2 Residue level evaluation

Ligand binding site prediction tools can also be evaluated at the residue level. F1 score as well as Matthews correlation coefficient (MCC) were utilised to do so. For each protein chain, F1 and MCC were calculated, distributions graphed and means reported (Table 4.7). Binary labels were employed to calculate these scores – 1 if the residue is found in a pocket and 0 otherwise – and compared to the ground truth, i.e., whether a residue binds a ligand in the LIGYSIS reference. For VN-EGNN, IF-SitePred, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}, which do not report pocket residues, pocket residues were obtained by considering those residues within 6 Å of the pocket centroid, cloud, and grid points, respectively. DeepPocket$_{\text{RESC}}$ was not considered for this analysis since its predictions correspond to re-scored and re-ranked fpocket candidates.

Figure 4.24 A-B illustrate the distributions of the F1 score and MCC for each method on the 2775 protein chains of the final LIGYSIS set. Both metrics agree that PUResNet (F1 = 0.41, MCC = 0.39), GrASP (F1 = 0.39, MCC = 0.33) and P2Rank$_{\text{CONS}}$ (F1 = 0.36, MCC = 0.30) are the top-3 performing methods in this task of binary classification into pocket (label: 1) and non-pocket residues (label: 0). fpocket presents the lowest F1 = 0.23 and MCC = 0.12 since it predicts many unobserved pockets, and therefore ligand-binding residues, that count as FP here.

IF-SitePred does not originally report a residue ligandability score beyond a binary label (0, 1). Nevertheless, in this Chapter, a score was computed by utilising the scores returned by the 40 prediction models of IF-SitePred. These scores range 0-1 and can be averaged as probabilities (Equation 4.10). This is here referenced as IF-SitePred ligandability score. For IF-SitePred, GrASP, P2Rank$_{\text{CONS}}$, P2Rank, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}, which report a residue level score (beyond a binary label), ROC and PR curves were plotted (Figure 4.24 C) and mean area under the curve (AUC) and average precision (AP) reported. This was not possible for VN-EGNN, PUResNet, DeepPocket$_{\text{SEG}}$, DeepPocket$_{\text{RESC}}$, fpocket$_{\text{PRANK}}$ and fpocket as they do not report residue ligandability scores.

Method	F1	MCC	AUC (%)	AP (%)
(**d**) VN-EGNN	0.29 (#8)	0.26 (#4)	–	–
(**d**) IF-SitePred	0.29 (#9)	0.24 (#6)	76 (#2)	45 (#2)
(**d**) GrASP	0.39 (#2)	0.34 (#2)	70 (#4)	42 (#3)
(**d**) PUResNet	**0.41 (#1)**	**0.39 (#1)**	–	–
(**d**) DeepPocket$_{SEG}$	0.27 (#10)	0.21 (#9)	–	–
(**d**) P2Rank$_{CONS}$	0.36 (#3)	0.30 (#3)	**76 (#1)**	**46 (#1)**
(**d**) P2Rank	0.31 (#4)	0.26 (#5)	74 (#3)	42 (#3)
(**d**) fpocket	**0.23 (#11)**	**0.12 (#11)**	–	–
(**d**) PocketFinder^{+}	0.31 (#5)	0.22 (#7)	68 (#6)	**34 (#6)**
(**d**) Ligsite^{+}	0.31 (#6)	0.21 (#8)	70 (#5)	38 (#4)
(**d**) Surfnet^{+}	0.29 (#7)	0.20 (#10)	**68 (#7)**	35 (#5)

Table 4.7. Residue level evaluation. These results come from default predictions, indicated by (**d**) preceding method names. DeepPocket$_{RESC}$ was not considered in this analysis as their predictions are re-scored and re-ranked fpocket's. Ligand binding site prediction benchmark at the residue level was calculated from 2775 protein chains in the LIGYSIS set. Mean F1 score, mean Matthews correlation coefficient (MCC), mean ROC area under the curve (AUC) and mean precision recall (PR) curve average precision (AP). Numbers following a hash (#) indicate how methods rank for each metric. Bold font indicates the best (blue) and worst (orange) performing methods. Pocket binary labels (0, 1) were employed for the calculation of F1 and MCC and obtained from predicted pockets. Residue ligandability scores were employed to calculate ROC/AUC and PR/AP. Reported AUC and AP are means resulting from the average across the 2775 LIGYSIS chains. This was not possible for VN-EGNN, PUResNet, DeepPocket$_{SEG}$ and fpocket as these methods do not provide such scores, indicated by a dash (–).

Figure 4.25 illustrates the variation in ROC/AUC for each method across the 2775 chains in the LIGYSIS set. P2Rank$_{CONS}$ and IF-SitePred, with the ligandability score calculated in this work (Equation 4.10), present the highest mean AUC = 76%, closely followed by P2Rank (AUC = 74%). Surfnet^{+} presents the lowest AUC = 68%. Figure 4.24 D shows the mean PR curves, which agree with ROC AUC and highlight P2Rank$_{CONS}$ as the method with the highest average precision = 46%, followed by IF-SitePred (with Equation 4.10 scoring) with AP = 45% and PocketFinder^{+} the lowest with (AP = 34%). Figure 4.26 displays the variability across LIGYSIS proteins for PR curve and AP.

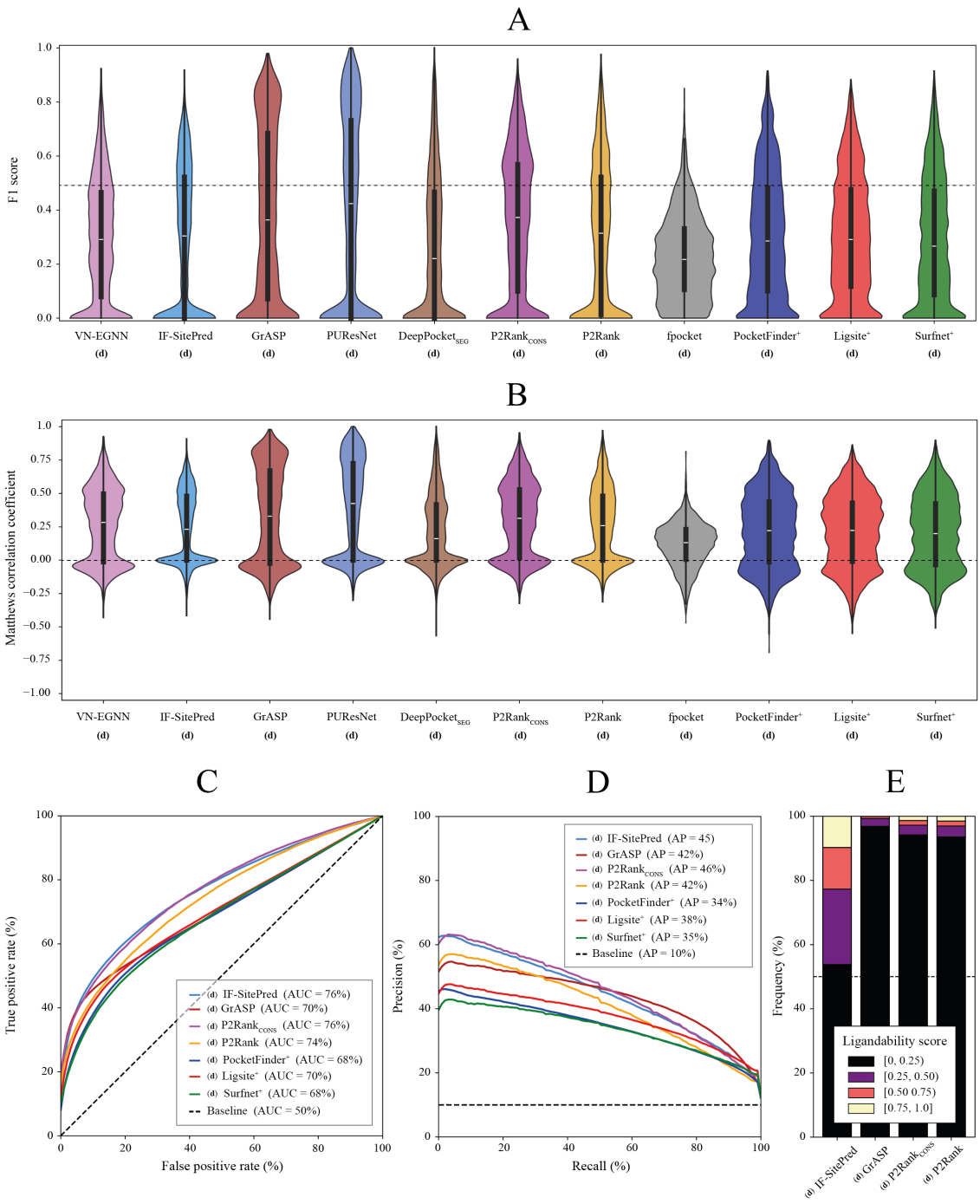

Figure 4.24. Ligand binding site prediction benchmark at the residue level. DeepPocket$_{RESC}$ predictions were not included in F1 and MCC analyses as these are re-scored and re-ranked fpocket predictions and the results would be the same as for fpocket. **(A)** F1 score distributions; **(B)** MCC distributions. In both panels, each data point corresponds to the score obtained from all residues in a protein chain. $N = 2775$ chains; **(C)** Mean ROC curve for methods that report a residue score. A dashed line represents the baseline, 1 FP for each TP, i.e., diagonal and AUC = 50%; **(D)** Mean PR curves. A dashed line represents the baseline, i.e., percentage of observed binding residues (true positives + false negatives) = 10%; **(E)** Distribution of residue ligandability scores for IF-SitePred, GrASP, P2Rank$_{CONS}$ and P2Rank. PocketFinder$^+$, Ligsite$^+$ and Surfnet$^+$ were not included as their scores do not range 0-1, and a small number of scores can reach values > 25. These are results from predictions generated from the original methods with default parameters (**d**).

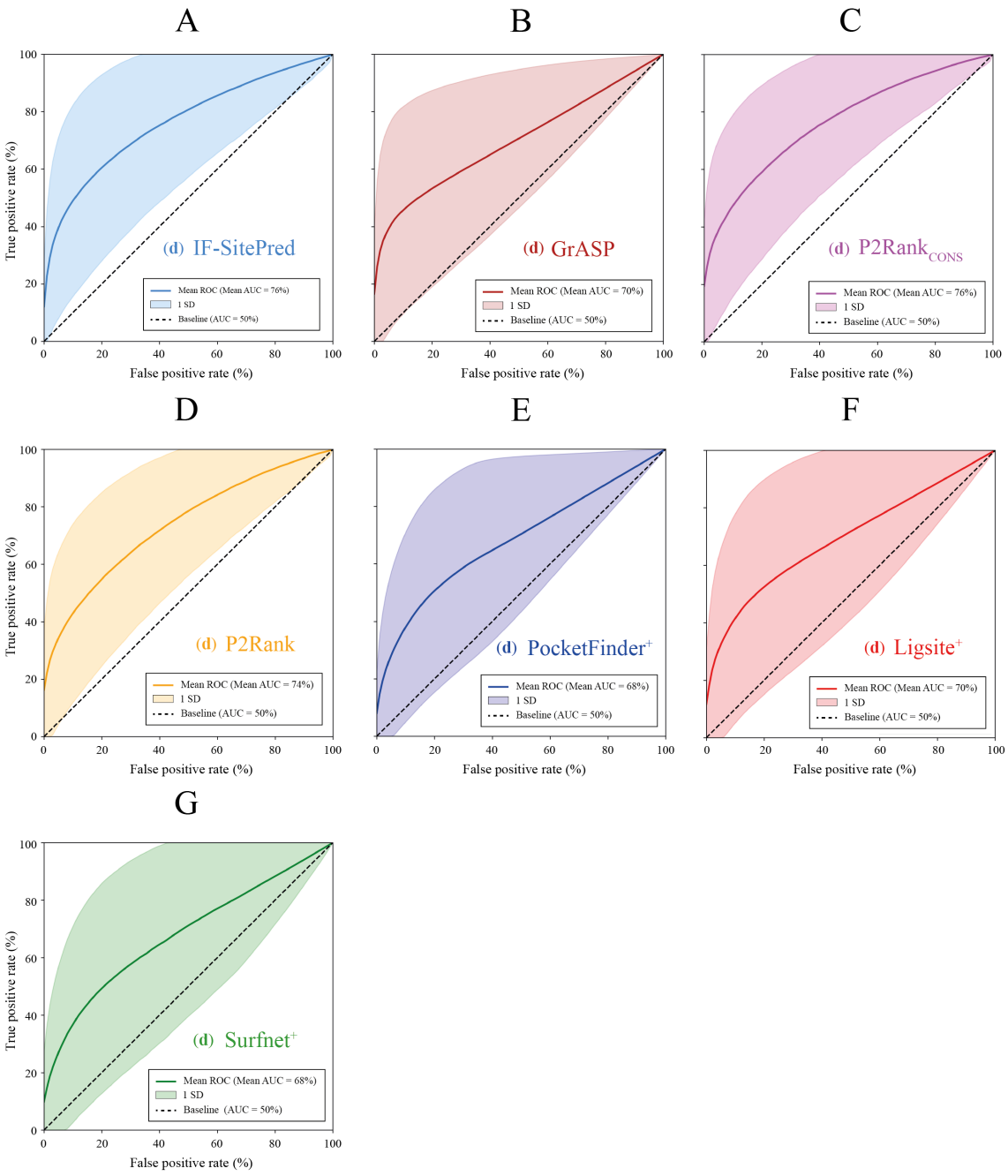

Figure 4.25. Variation in ROC curve and AUC across LIGYSIS proteins. For each of the methods that report, or for which residue ligandability scores were computed, a ROC curve was obtained for each of the 2775 protein chains in the LIGYSIS set and AUC calculated. Plotted curves represent the mean ROC curve for each method. These were obtained by averaging the TPR for each FPR interval across proteins. The shaded area represents one standard deviation (1 SD) from the mean ROC curve. Reported AUC is the mean AUC calculated by averaging the AUC for the 2775 ROC curves obtained. Baseline AUC is random chance (AUC = 50%). **(A)** IF-SitePred; **(B)** GrASP; **(C)** P2Rank$_{CONS}$; **(D)** P2Rank; **(E)** PocketFinder^{+}; **(F)** Ligsite^{+}; **(G)** Surfnet^{+}. These results originate from default methods, indicated by **(d)** preceding method names.

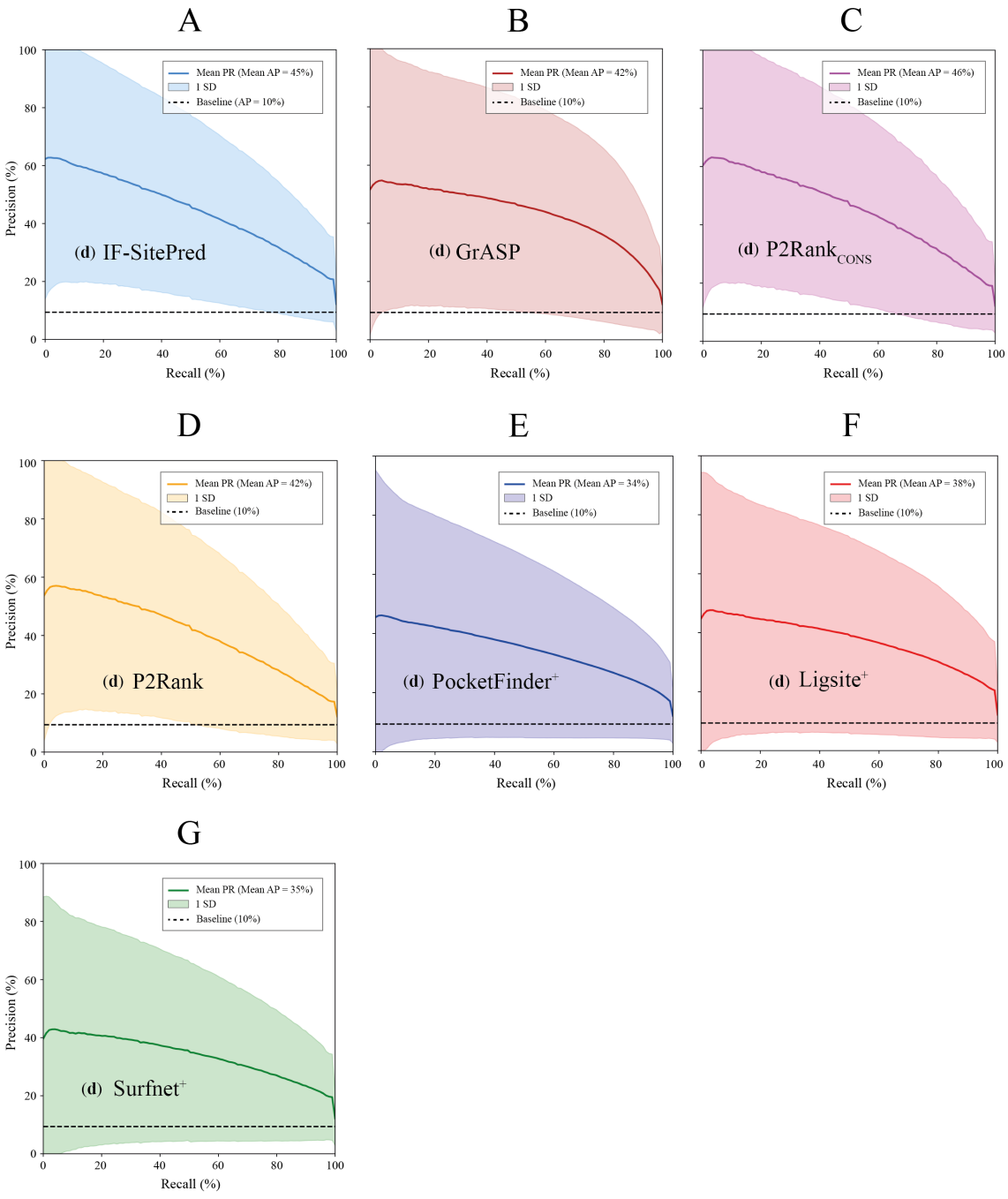

Figure 4.26. Variation in PR curve and AP across LIGYSIS proteins. For each of the methods that report, or for which residue ligandability scores were computed, a precision-recall (PR) curve was obtained for each of the 2775 protein chains in the LIGYSIS set and average precision (AP) calculated. Plotted curves represent the mean PR curve for each method. These were obtained by averaging the precision for each recall interval across proteins. The shaded area represents one standard deviation (1 SD) from the mean PR curve. Reported AP is the mean AP calculated by averaging the AP for the 2775 PR curves obtained. Baseline AP is the percentage of observed ligand-binding residues (AP = 10%). **(A)** IF-SitePred; **(B)** GrASP; **(C)** P2Rank$_{CONS}$; **(D)** P2Rank; **(E)** PocketFinder^{+}; **(F)** Ligsite^{+}; **(G)** Surfnet^{+}. These results originate from default methods, indicated by (**d**) preceding method names.

Figure 4.24 E shows IF-SitePred presenting a different residue ligandability score distribution to GrASP, P2Rank$_{CONS}$ and P2Rank. The IF-SitePred ligandability score, resulting from averaging the scores from the 40 IF-SitePred models, is the most "generous" with ≈20% of the residues presenting a score > 0.5, in contrast with GrASP, which residue scoring is very strict $P(LS \geq 0.5) = 0.75\%$ and P2Ranks (≈3%). This difference, combined with the mean ROC and PR curves, further supports the use of the IF-SitePred ligandability score proposed in this Chapter to define the predicted binding sites for this method. It also suggests that GrASP might benefit from a less strict residue level scoring scheme. PocketFinder^{+}, Ligsite^{+} and Surfnet^{+} were not included in this analysis as their scores do not range 0-1 and very high scores (>25) can be obtained.

4.4 Discussion

This Chapter describes the most complete comparative analysis of ligand binding site prediction methods to date, spanning three decades of methods development. Firstly, predictions from the thirteen methods, as well as observed sites from the new reference dataset introduced here, LIGYSIS, were compared in terms of the number of proteins methods predict on, the number of predicted sites per protein, their size, distance and overlap between the sites. This analysis provides insight into how the different methods work and hints at potential limitations or room for improvement, e.g., the prediction of a fixed number of sites per protein, or considerable proximity and overlap between predictions. Secondly, predictions from thirteen canonical ligand binding site prediction methods were objectively evaluated using the LIGYSIS set. This evaluation considered prediction at the residue level by F1 score, MCC, ROC/AUC and PR/AP, as well as the pocket level by recall for top-N, N+2 and *all* predictions, Precision$_{1K}$, # TP$_{100\,FP}$, RRO and RVO. This is the first independent ligand site prediction benchmark since Schmidtke *et al.* [393] in 2010 and Chen *et al.* [394] shortly after in 2011, and the largest to date both in terms of reference dataset size (2775), methods compared (10) and metrics employed (14).

Recall (% of observed sites that are correctly predicted) is more informative than precision (% of predictions that are correct), particularly, recall considering top-N+2 ranked predictions. In most cases, not all the existing binding pockets are observed with a ligand bound. In other words, the reference data are incomplete, with 33-50% of existing sites yet to be observed with ligands bound in a structure, as conjectured by Krivák and Hoksza [115]. Considering only the top-N predicted pockets assumes that there are exactly N real pockets for a given protein, which might not be the case. A method could predict a *real* pocket that is yet to be observed and rank it before other predicted *and* observed pockets. By considering the top-N+2 pockets, the noise in the reference data is controlled for, to some extent, and a more accurate representation of the method performance is obtained. In a context of discovery, where the true ligand binding sites of the target are unknown, it is more useful to have multiple predictions that might or might not correspond to real sites (lower precision), rather than a single or few predictions that are very precise but are missing other likely sites (lower recall). Most methods do well in predicting the most obvious (orthosteric) site. This site, however, might not be available for therapeutic targeting and it is convenient to predict other sites that could modulate function acting as allosteric sites. Precision is a metric that provides valuable insight and is covered in this work. However, it must always be contextualised with recall. This Chapter shows that the most precise methods do not correspond to higher recalling methods. A method predicting the most obvious site, that could be identified by eye, might be 90% precise, but present a lower recall, e.g., 30%. This said, methods predicting fewer pockets with higher precision might prove more advantageous when users aim to study a particular region of interest in a protein, a few high-priority sites are needed for experimental validation or false predictions are costly in downstream analysis.

Some methods define "success rate" as the precision of the top-1 or top-3 scoring predictions, which is not a very representative performance assessment metric. For this reason, in this Chapter, *I* strongly encourage method developers not only to share the code of their approach, but also the code of their benchmarking analysis. Furthermore, the definition of success rate must be standardised as recall, as some methods use recall,

whereas others use precision, both under the name of *success rate*. This can be confusing when comparing the results from different analyses. Moreover, due to the inherent noise in the reference data, i.e., not all existing pockets are known, recall considering top-N+2 is more informative than taking top-1, top-3 or top-N predictions. In any case, success rate must be clearly defined in a given publication, so readers can fully understand the implications of the metric employed in a benchmark.

It is clear from the results described in this Chapter that a DCC threshold of 4 Å is too conservative and a more flexible DCC threshold of 10-12 Å should be used for comparable performance with DCA = 4 Å. According to this work and the LIGYSIS reference, most predictions with DCC 4-12 Å overlap with or are adjacent to observed sites and should be considered as correct predictions. The reason for this is the inherent noise in the ground truth, i.e., a ligand binding to a cavity might not be representative of all ligands that could bind to it. For most proteins, not all existing ligand sites are characterised and as different ligands can bind to the same region, it is unrealistic to use such a small DCC threshold. The results in this Chapter show several examples of correct predictions of observed cavities with DCC $>$ 4 Å.

Re-scoring of original fpocket predictions by PRANK ($\text{fpocket}_{\text{PRANK}}$) and DeepPocket ($\text{DeepPocket}_{\text{RESC}}$) present the highest recall considering top-N+2 predictions (≈60%). $\text{P2Rank}_{\text{CONS}}$ and P2Rank follow closely with ≈53% recall and IF-SitePred presents the lowest recall (25.7%). fpocket and methods that re-score its predictions predict the most pockets per protein, reaching a maximum recall between 80-90%. $\text{P2Rank}_{\text{CONS}}$, P2Rank and $\text{DeepPocket}_{\text{SEG}}$ follow with a maximum recall of ≈60%. The rest of the methods range 40-55%. This indicates that while there are still some pockets left unpredicted by fpocket (10-20%), the maximum recall of this method is 20-30% higher than any other method. However, when considering top-N+2 pockets, fpocket recalls only 47% of the observed pockets. $\text{fpocket}_{\text{PRANK}}$ and $\text{DeepPocket}_{\text{RESC}}$ gain $>$10% in recall by simply re-scoring those predictions. This highlights the paramount importance of a robust scoring scheme, which captures well the nature of binding sites and places those with a higher probability of being real binding sites at the top of the ranking.

Precision$_{1K}$ and # TP$_{100\,FP}$ show that newer methods like VN-EGNN, IF-SitePred and GrASP are the most precise methods. However, because of redundancy in predictions (VN-EGNN, IF-SitePred), or low number of predicted pockets per protein (VN-EGNN and GrASP) are limited in their recall. Their high precision indicates that their models learn and capture well the nature of ligand binding sites and so they represent a great avenue to pursue in the field of ligand binding site prediction.

The usefulness of residue-level metrics like F1 score and MCC is limited, as methods that predict the easiest sites, e.g., PUResNet (high precision, low recall) perform better on these metrics, while methods that predict more pockets, e.g., fpocket and re-scored versions (lower precision, higher recall) obtain worse results. Pocket-level metrics, such as top-N+2 recall, are more representative of the ability to predict ligand binding sites.

While datasets like PDBbind, bMOAD or PLINDER [539] are extremely useful to train, validate and test deep learning models for rigid body docking [540], flexible pocket docking [541] or pocket-conditioned ligand generation [542], they might not be ideal as a test set for ligand binding site prediction. LIGYSIS analyses all unique, biologically relevant protein-ligand interfaces, including ions, across the biological assembly from multiple experimentally determined structures of a given protein. It then clusters these ligands based on their interactions with the protein, resulting in the observed binding sites. Beyond considering biological assemblies and unique protein-ligand interfaces, the greatest innovation in LIGYSIS is leveraging the extensive structural data on the PDBe-KB to aggregate ligand interactions across different structures of the same protein, thus capturing better the ligand-binding capabilities of a protein than just taking a single protein-ligand complex. In doing so, LIGYSIS represents the most complete and non-redundant protein-ligand complex dataset for the prediction of protein-ligand binding sites to date.

The benchmark is performed on LIGYSIS including ion binding sites. When these are removed, all methods except fpocket experience an increase in recall of 5-10%, yet the method ranking does not change (Figure 4.27 and Figure 4.28). Due to its integrative approach, features, diversity and size – covering >30% of PDB and >20% of BioLiP – LIGYSIS is the most inclusive and representative dataset of protein-ligand interactions.

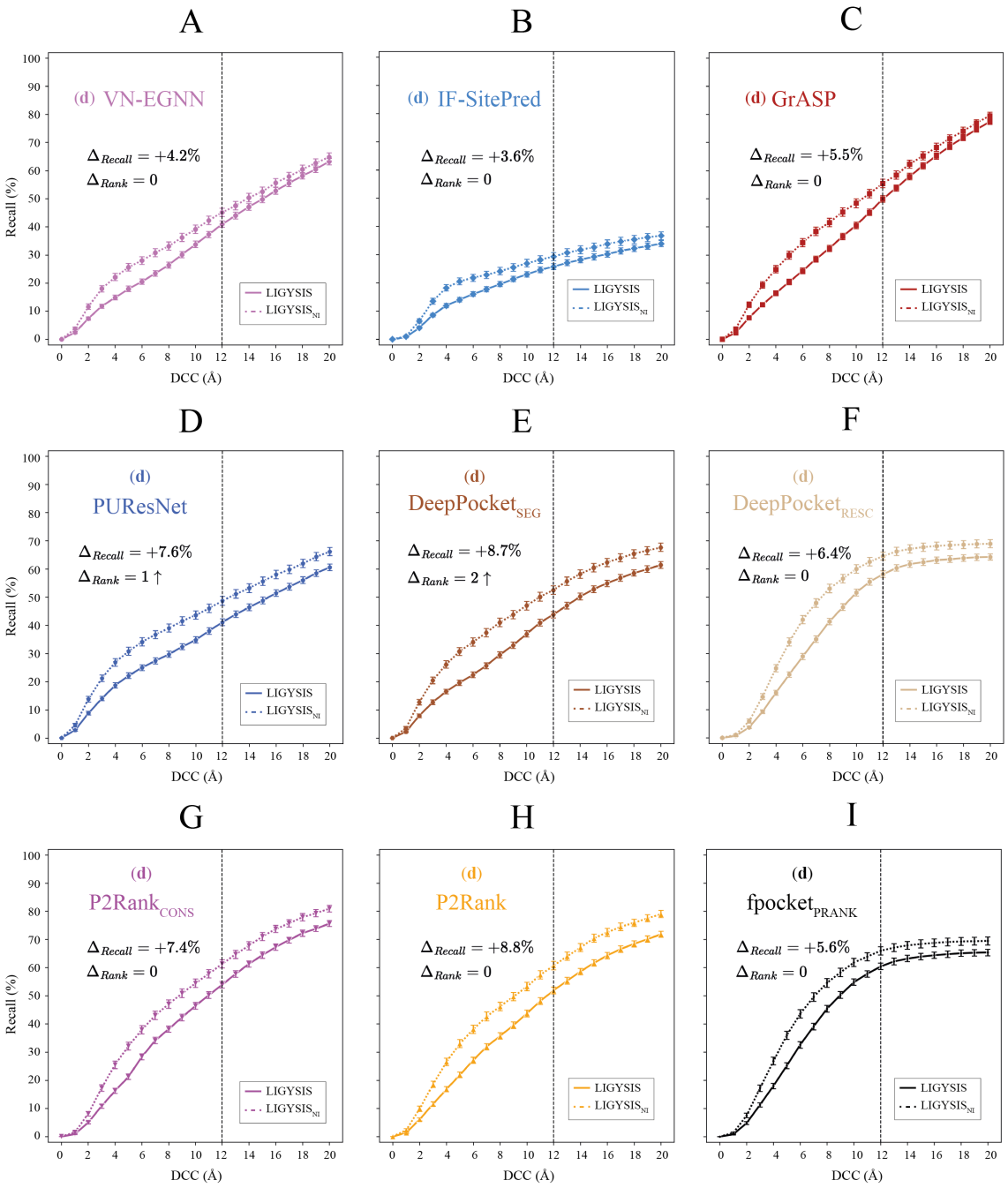

Figure 4.27. Change in top-*N*+2 recall for LIGYSIS *vs* LIGYSIS$_{NI}$ (I). Recall is calculated considering top-*N*+2 pockets at DCC = 12 Å and default methods (**d**). LIGYSIS$_{NI}$ is a subset of LIGYSIS sites containing at least one non-ion ligand – *N* = 4141/6882 (60%). Solid lines indicate recall curve on LIGYSIS and dashed lines for LIGYSIS$_{NI}$. The relative change in recall and rank are indicated by Δ_{Recall} and Δ_{Rank}. These changes are relative to performance on LIGYSIS (including ions). All machine learning-based methods present an increase in recall when removing ion binding sites. This is expected as none of the methods were trained on ion sites. However, ion sites were kept on the main benchmark to challenge and test the limits of the methods. **(A)** VN-EGNN; **(B)** IF-SitePred; **(C)** GrASP; **(D)** PUResNet; **(E)** DeepPocket$_{SEG}$; **(F)** DeepPocket$_{RESC}$; **(G)** P2Rank$_{CONS}$; **(H)** P2Rank; **(I)** fpocket$_{PRANK}$.

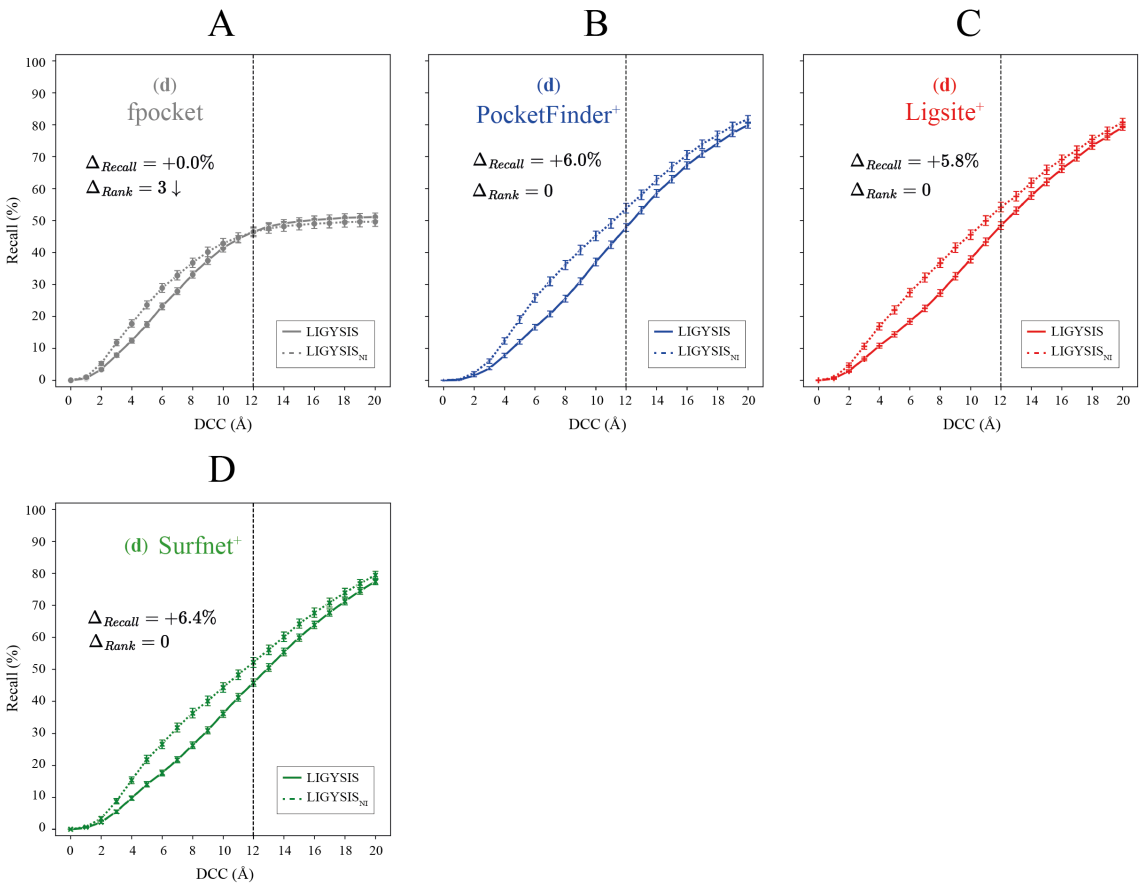

Figure 4.28. Change in top-*N*+2 recall for LIGYSIS *vs* LIGYSIS$_{NI}$ (II). Recall is calculated considering top-*N*+2 pockets at DCC = 12 Å and default methods (**d**). LIGYSIS$_{NI}$ is a subset of LIGYSIS sites containing at least one non-ion ligand – *N* = 4141/6882 (60%). Solid lines indicate recall curve on LIGYSIS and dashed lines for LIGYSIS$_{NI}$. The relative change in recall and rank are indicated by Δ_{Recall} and Δ_{Rank}. These changes are relative to performance on LIGYSIS (including ions). Geometry and energy-based methods also present an increase in recall when removing ion binding sites.

Aggregating protein-ligand interactions across structures of the same protein is likely to be beneficial not only for testing, but also when training these methods. Most available methods train on datasets where a protein is represented by a single structure interacting with a single ligand. For example, in 100% of sc-PDB and 50% of entries for binding MOAD training sets. Methods consider as ligand binding (positives) those residues within a certain distance of the ligand and all other residues negatives. In doing so, residues of the same protein that bind ligands on other structures, but not the one present on these sets, will be incorrectly labelled as "non-ligand-binding" (FN). This mislabelling of residues could lead to a lower prediction performance. This issue is to a certain extent approached by P2Rank and GrASP, which enriched their training datasets by including ligands from

other chains or homologous structures. This noise in the training dataset might be more prevalent for DeepPocket, PUResNet and VN-EGNN, which seem to rely fully on 1:1 protein-ligand interactions. The usage of LIGYSIS, or any other dataset that aggregates ligand interactions across structures, might alleviate this issue and hints at potential room for improvement in the field of ligand binding site prediction.

4.5 Conclusions

The conclusions resulting from the analysis described in this Chapter are as follows:

- LIGYSIS aggregates non-redundant biologically relevant protein-ligand interactions across multiple structures for a protein and sets a new test set standard for the benchmark of ligand binding site prediction tools.
- The use of duplicated protein-ligand interfaces in asymmetric units results in the overestimate of both precision and recall when benchmarking ligand site predictors. Only unique protein-ligand interfaces in biological units should be considered for a more accurate benchmark of the performance of these methods.
- A DCC threshold of 4 Å is too conservative, and to obtain comparable results between DCA and DCC recall, a threshold of DCC of 10-12 Å should be employed.
- Ligand binding site prediction methods differ significantly in the number of predicted sites, their size, proximity and overlap.
- Recall is a more informative measure of the performance of a ligand site prediction tool, rather than precision and so it must be reported. Precision, though a useful metric, should always be contextualised with recall.
- All authors of ligand site prediction tools should use top-N+2 recall as "success rate" for consistency. Benchmarking code should also be shared by the authors for the sake of reproducibility.

- Pocket-level metrics (recall, precision) are a more adequate representation of the ability of ligand site prediction methods than residue-level metrics (F1, MCC).
- Re-scoring of fpocket predictions, like $\text{fpocket}_{\text{PRANK}}$ or $\text{DeepPocket}_{\text{RESC}}$, present the highest (top-N+2) recall (60%) among the methods reviewed in this Chapter.
- Methods that systematically predict a low number of pockets, e.g., VN-EGNN, GrASP or PUResNet, are very precise (>90%). However, their recall is low and might not be as useful in a drug discovery context.
- The IF-SitePred ligandability score introduced in this Chapter correctly recapitulates observed ligand binding sites and suggests IF-SitePred could benefit greatly from using it in its prediction.
- The work presented in this Chapter objectively evaluates the performance of thirteen canonical ligand binding site prediction methods and represents the largest benchmark of ligand site prediction tools to date.

Chapter 5

Improvement on methods for the prediction of protein-ligand binding sites

Preface

This Chapter explores in detail fifteen non-redundant and scoring variants of the thirteen ligand binding site prediction methods evaluated in Chapter 4. The negative effect of feeble pocket scoring schemes and redundancy in ligand site prediction is demonstrated through the performance evaluation of these variants relative to their default modes using seven informative metrics. The work in this Chapter was solely carried out by *me*.

Publications

Utgés, J.S. and Barton, G.J. Comparative evaluation of methods for the prediction of protein-ligand binding sites. *J. Cheminform.* **16**, 126 (2024). https://doi.org/10.1186/s13321-024-00923-z.

5.1 Introduction

In Chapter 4 the human component of the LIGYSIS dataset was employed to carry out the largest critical assessment of ligand binding site prediction tools to date [342]. This evaluation included a set of thirteen methods combining the latest machine learning methods such as VN-EGNN [169], IF-SitePred [168] or GrASP [167], established methods like P2Rank [115, 150], PRANK [149] or fpocket [120, 123] and earlier geometry/energy-based methods such as PocketFinder^{+} [129], Ligsite^{+} [121] and Surfnet^{+} [122]. These methods were thoroughly evaluated at the residue and pocket level using more than ten different metrics.

Beyond ranking the methods by several metrics, Chapter 4 identified VN-EGNN, IF-SitePred and DeepPocket$_{\text{SEG}}$ as having predicted pockets within very high spatial proximity (<5 Å) and significant residue overlap ($I_{rel} > 0.5$). Both of these features hint at redundancy in pocket prediction. Additionally, PUResNet, PocketFinder^{+}, Ligsite^{+} or Surfnet^{+} were highlighted as they do not report scores, nor explicit rank for their predicted pockets. Both of these issues, pocket prediction redundancy and lack of scoring scheme, are likely to have a considerable negative effect on the methods' performance. This Chapter explores in detail both of these aspects and evaluates the performance of fifteen novel scoring and non-redundant variants of the thirteen methods evaluated in Chapter 4.

Pocket prediction redundancy is defined here as the prediction of pockets with centroids very close in space ($D \leq 5$Å) or with overlapping residues ($I_{rel} \geq 0.75$). This indicates multiple predictions of the same potential ligand binding site. Most ligand site prediction tools predict not only the location of the pocket by means of a centroid or pocket residues, but also a pocket confidence, and an associated rank among all the predicted pockets. Ligand site predictors tend to be evaluated by considering the top-N, or top-N+2 ranking pockets, where N is the number of observed sites for a given protein. The redundant prediction of pockets can result in a sub-optimal ranking, thus negatively affecting the performance of the predictors.

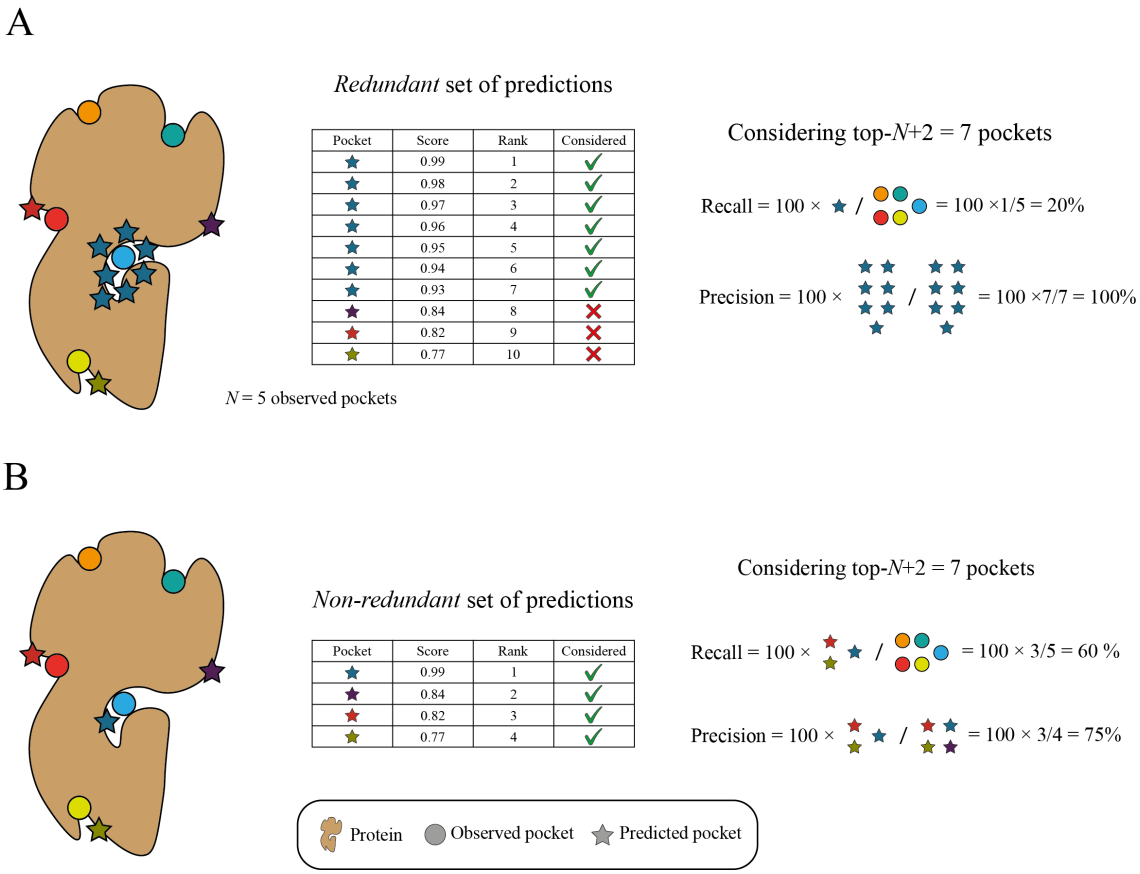

Figure 5.1. The issue of redundancy in ligand binding site prediction. (A) A set of predictions where 6/10 (60%) predictions are redundant, resulting in a low recall of 1/5 (20%) and inflated precision of 7/7 (100%); **(B)** When redundancy is removed, only four predictions remain and recall increases to 3/5 (60%) and precision decreases to 3/4 (75%).

Figure 5.1 shows an example protein with $N = 5$ observed pockets. A ligand site predictor returns 10 predictions, but the top-7 are all within 3 Å of one of the observed pockets, and >12 Å from any of the other four observed pockets. If the top-N+2 (top-7) predictions were considered, this would only recall a single unique pocket, as six of the top-7 predictions are redundant. Top-N+2 recall would then be 20% (1/5). Precision, however, within this top-7 would be 100% (7/7), as the seven predictions are correctly recalling an observed pocket (which happens to be the same). In this case, both the low recall and the high precision are artefacts resulting of the redundancy (Figure 5.1 A). Redundancy in prediction can often result in the overestimate of the precision and the underestimate of the recall. Figure 5.1 B illustrates what happens when redundant predictions are removed, keeping always higher-scoring predictions. When the six redundant predictions

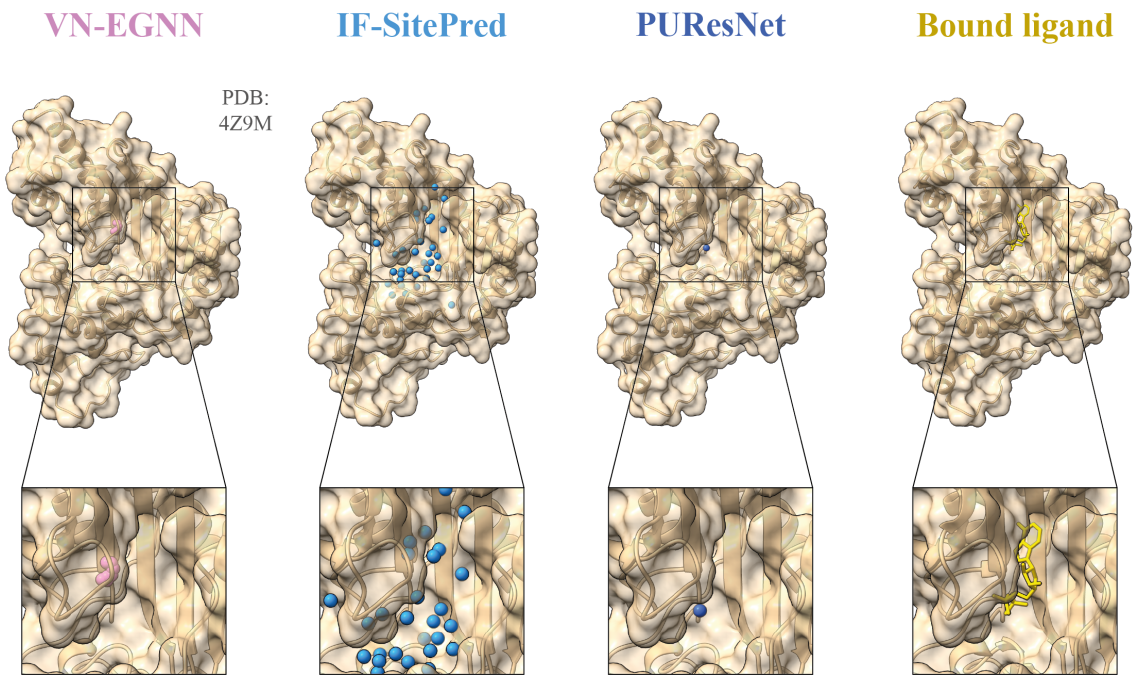

Figure 5.2. Example of redundant predictions. Predictions by VN-EGNN, IF-SitePred and PUResNet, on chain D of PDB: 4Z9M [543] of human creatine kinase (P17540) where ADP binds. For this ADP binding site, VN-EGNN reports 7 predictions, IF-SitePred 33 and PUResNet a single prediction. These three methods correctly predict this site. However, VN-EGNN and IF-SitePred report redundant pocket predictions, which centroids are very close ($\leq$ 5 Å) in space and residues overlap ($\geq$ 0.75).

(blue stars) are removed, the other three predictions, which are of different pockets, are considered as now fall within the top-N+2 predictions. This increases the recall to 60%, as 3/5 observed pockets are now correctly predicted. However, precision decreases, as only three out of the four predictions made overlap with an observed pocket. Pocket rank #2 has a high score but is not observed. This is a *false positive* in this context, however it might be a candidate pocket yet to be resolved and could prove interesting as a drug target.

Figure 5.2 showcases PDB: 4Z9M of human creatine kinase S-type, mitochondrial (P17540) as an example of this phenomenon, where VN-EGNN and IF-SitePred redundantly predict the same pocket 7 and 33 times, whereas PUResNet returns a single prediction. All three methods correctly predict the site, just the difference is in the number of returned predictions. These redundant predictions of the same observed site would count as 7 and 33 TPs when calculating precision, leading to an inflated artificial precision.

5.2 Methods

5.2.1 Generation of non-redundant sets of predictions

Figure 4.22 B shows that prediction redundancy is an issue particularly for VN-EGNN, IF-SitePred, and to a lesser extent, DeepPocket$_{SEG}$. To assess the effect that redundancy had on the performance of these methods, non-redundant subsets of predictions were obtained and labelled with the subscript "NR". A predicted pocket i is considered redundant if there exists a pocket $j \neq i$ so that the distance between their centroids $D_{i,j} \leq 5$ Å or their residue overlap $JI_{i,j} > 0.75$, i.e., they share at least 3/4 (75%) of their residues. Refer to Figure 5.3 for the closest predicted sites for each method. Redundancy filtering was carried out for each method keeping always the higher-scoring pocket. Redundancy (%) was calculated as the proportion of redundant pockets relative to the original total number of pockets. VN-EGNN presents the highest percentage of redundant pockets with 9066/13,582 (67%) redundant pockets, followed by IF-SitePred with 22,232/44,948 (49%) and DeepPocket$_{SEG}$ with 6744/21,718 (31%). For other methods, redundancy was minimal (<1%).

5.2.2 Pocket re-scoring strategies

PUResNet, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+} do not score, nor explicitly rank their pockets, and so pockets were taken in the order given by their ID. This means that when sorting across the dataset, the order of all pockets with the same rank is arbitrary. Multiple strategies were employed to obtain scores for these pockets. Firstly, a score was obtained as the number of pocket amino acids, resulting in variants PUResNet$_{AA}$, PocketFinder$^{+}{}_{AA}$, Ligsite$^{+}{}_{AA}$ and Surfnet$^{+}{}_{AA}$. Secondly, PRANK pocket scoring was employed, resulting in variants PUResNet$_{PRANK}$, PocketFinder$^{+}{}_{PRANK}$, Ligsite$^{+}{}_{PRANK}$ and Surfnet$^{+}{}_{PRANK}$. IF-SitePred uses a simple pocket scoring scheme, which assigns to each centroid the number of clustered cloud points it results from. In this Chapter, newly defined IF-SitePred pocket scores were calculated as the sum of squares (SS) of the ligandability scores (LS_i),

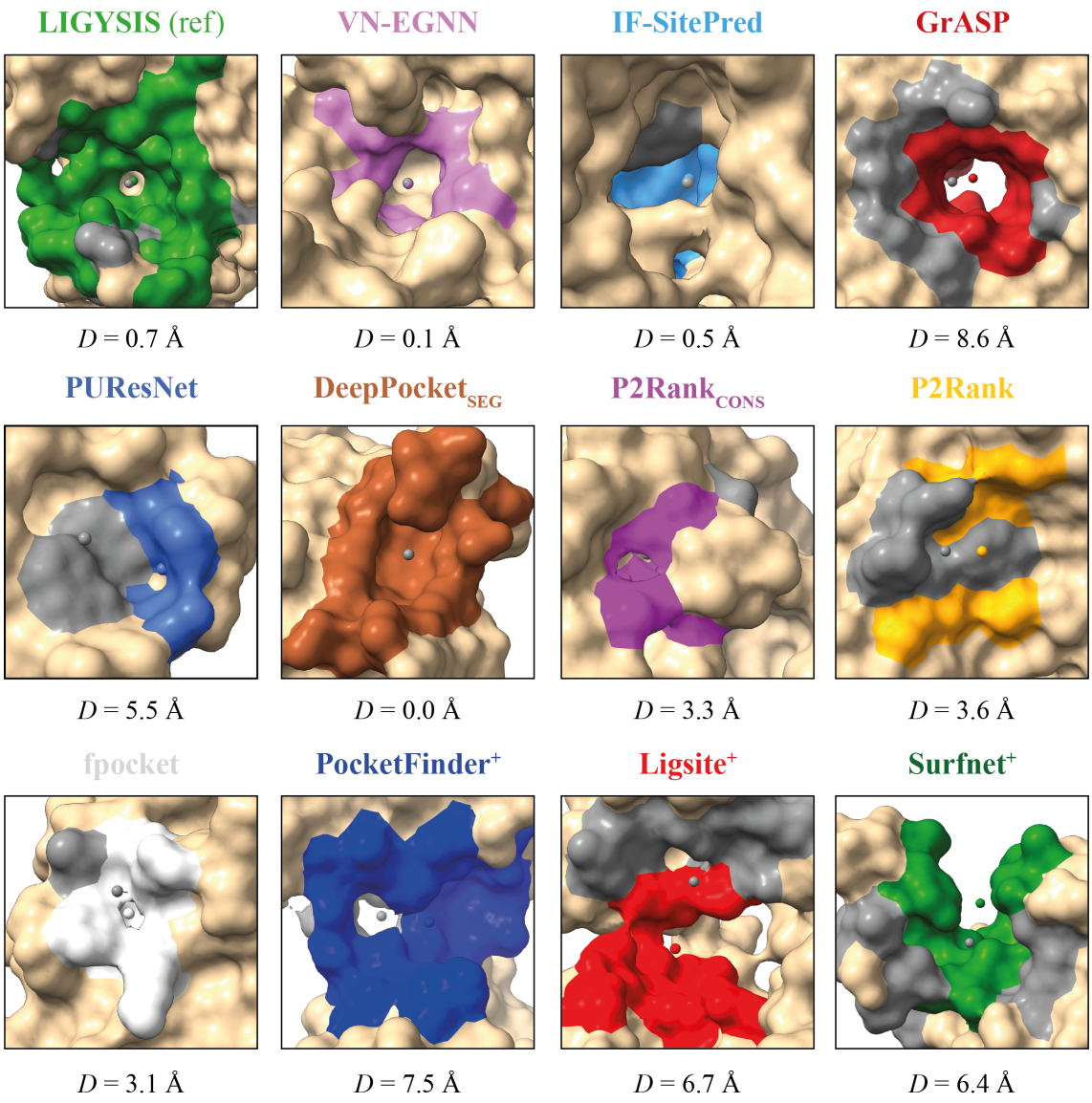

Figure 5.3. Closest predicted pockets for each method. LIGYSIS is a reference dataset, not a prediction method. For each method, the two closest predicted pockets across all protein chains are shown. This is the pair of pockets with the minimum Euclidean distance between their centroids. Protein surface is coloured in tan. The larger pocket (more residues) and centroid is coloured in the method colour and the other pocket's in grey. A distance threshold of $D = 5$ Å was selected to determine whether a pocket prediction was redundant. VN-EGNN, IF-SitePred and DeepPocket$_{SEG}$ clearly differ from other methods presenting distances < 1 Å. Examples for each method are from top to bottom and left to right: P00492 – PDB: 3GEP [544], chain: B; Q96KS0 – PDB: 5V1B [545], chain: A; P31645 – PDB: 5I73 [546], chain: A; Q04724 – PDB: 1GXR [547], chain: A; Q5W0Z9 – PDB: 7KHM [548], chain: B; Q06187 – PDB: 1B55 [549], chain: B; Q9UQG0 – PDB: 7SR6 [550], chain: G; P13866 – PDB: 7YNI [551], chain: A; Q14534 – PDB: 6C6P [455], chain: A; P31321 – PDB: 4DIN [552], chain: B; P78527 – PDB: 7SUD [538], chain: A; Q14416 – PDB: 7EPB [553], chain: A.

calculated with Equation 4.10, of the K residues on a site (Equation 5.1) resulting in IF-SitePred$_{RESC}$. The same was done for PocketFinder$^+$, Ligsite$^+$ and Surfnet$^+$, but instead of

residue scores, grid point scores (GS_i) were used (Equation 5.2). This resulted in further variants PocketFinder$^{+}{}_{SS}$, Ligsite$^{+}{}_{SS}$ and Surfnet$^{+}{}_{SS}$. This is the same approach introduced by Krivák *et al.* [150] and later adopted by Smith *et al.* [167].

$$SS_{\text{IF–SitePred}} = \sum_{i=1}^{K} LS_i^2 \tag{5.1}$$

$$SS_{\text{PocketFinder}^+} = SS_{\text{Ligsite}^+} = SS_{\text{Surfnet}^+} = \sum_{i=1}^{K} GS_i^2 \tag{5.2}$$

5.2.3 Performance evaluation

This Chapter evaluates the performance of fifteen novel non-redundant and scoring variants of the thirteen canonical ligand binding site prediction methods surveyed in Chapter 4. These variants do not affect prediction at the residue level, so performance is only assessed at the pocket level.

Mainly three metrics are discussed in this Chapter. Recall (Equation 4.16) is the percentage of observed binding sites in the reference data that are correctly predicted by a given method for a given DCC, rank or I_{rel} threshold. Precision (Equation 4.15) is the percentage of predicted sites that are correct, i.e., match a pocket in the reference. In this case, Precision$_{1K}$ is reported, where all predictions by a method across the LIGYSIS reference set are sorted by score and precision reported as predictions are considered up to the 1000th highest scoring prediction. In a similar way, ROC$_{100}$ [511] reports cumulative TP *vs* cumulative FP until 100 FP are reached. Finally, relative residue overlap (Equation 4.20) and relative volume overlap (Equation 4.21) represent how well predicted sites match the observed site in terms of residue overlap and shape (%). Refer to Section 4.2.9.2 for more details.

5.2.4 Statistics and reproducibility

ChimeraX v1.7.1 [70] was used for structural visualisation. Performed statistical tests were two-tailed and $\alpha = 0.05$. Sample sizes and measures of significance are reported in

text, figures and legends. For more details on method selection and execution, refer to Section 4.2.10.

5.2.5 Data and code availability

The main results tables and files necessary to replicate the analysis described in this Chapter can be found here: https://doi.org/10.5281/zenodo.13121414 [523]. Software developed to carry out this analysis is found in this GitHub repository: https://github.com/bartongroup/LBS-comparison [524].

5.3 Results

5.3.1 Effect of redundancy and pocket score on ranking

Since VN-EGNN uses $K = 8$ virtual nodes by default, a maximum of eight predicted pockets are possible. However, only seven are observed in this dataset, i.e., in all cases at least one virtual node gets clustered with another, resulting in seven "unique" predictions. Figure 5.4 A illustrates the issue of prediction redundancy and how it affects the scoring and ranking of the pockets. Predictions of the same pocket are reported multiple times as distinct virtual nodes or pocket centroids. These nodes are very close to each other and present similar scores. This is why there is no apparent difference in the distribution of scores across the pocket ranks for VN-EGNN, unlike all other methods. This is no longer the case after removing redundancy and obtaining VN-EGNN$_{\text{NR}}$ (Figure 5.4 B). IF-SitePred predictions are also highly redundant. However, these pockets, despite being close to each other, present different scores (number of points). That is why higher ranks (1, 2, 3...) present higher scores (Figure 5.4 C). Redundancy removal can be observed in Figure 5.4 D as the scatter plot is less crowded and the maximum rank across the dataset is 60 as opposed to 120. Figure 5.4 E shows the non-redundant set of re-scored IF-SitePred predictions, IF-SitePred$_{\text{RESC-NR}}$. This distribution is wider, i.e., scores take values from a larger value distribution, which might yield a better scoring scheme.

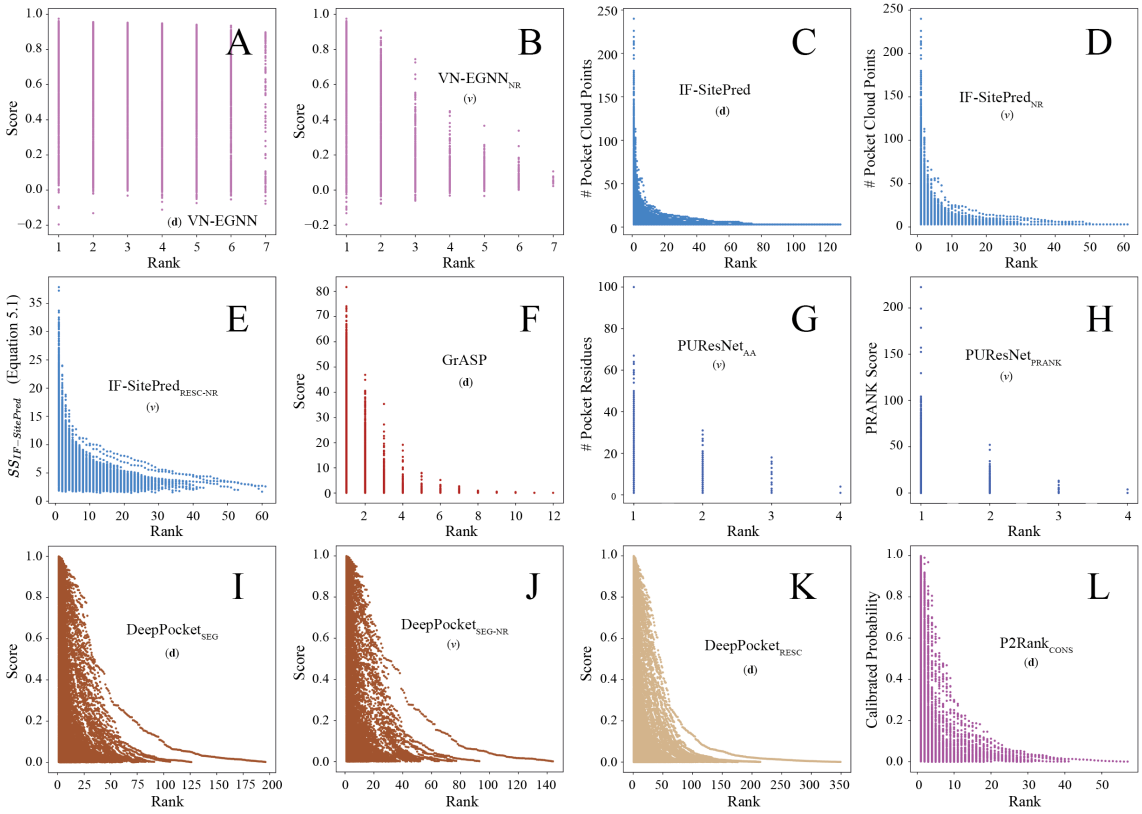

Figure 5.4. Pocket score *vs* pocket ranking (I). (A) VN-EGNN reported pocket scores; **(B)** Non-redundant VN-EGNN predictions (VN-EGNN$_{NR}$); **(C)** Default IF-SitePred predictions are ranked based on the number of pocket cloud points; **(D)** Non-redundant variant of IF-SitePred (IF-SitePred$_{NR}$); **(E)** Re-scored non-redundant IF-SitePred predictions (IF-SitePred$_{RESC-NR}$). Score is calculated as sum of squares of residue ligandability scores (Equation 5.1); **(F)** GrASP; **(G)** PUResNet does not score its pockets. PUResNet$_{AA}$ uses the number of pocket amino acids as a score; **(H)** PRANK-scored PUResNet pockets; **(I)** DeepPocket$_{SEG}$; **(J)** Non-redundant DeepPocket$_{SEG}$ predictions (DeepPocket$_{SEG-NR}$); **(K)** DeepPocket$_{RESC}$; **(L)** P2Rank$_{CONS}$. (**d**) and (*v*) indicate whether methods are default or a variant generated in this work.

There is no clear difference between Figure 5.4 G-H, meaning that using PRANK to score PUResNet predictions does not alter the overall ranking of the predictions made within a protein. This makes sense, as only 10% of proteins present >1 predicted pocket by this method. Nevertheless, this new score could help in the ranking of pockets across the dataset and not just within a protein. The distribution of scores does not change when removing the redundancy from Deep-Pocket$_{SEG}$ predictions (Figure 5.4 I-J), but the maximum rank goes from 200 to 140, indicating the decrease in total predictions. The score distributions of fpocket$_{PRANK}$ (Figure 5.5 B) and fpocket (Figure 5.5 C) are completely different. This makes sense, since the ranking of pockets, recall and precision of these two pocket scoring schemes differ considerably as shown in Chapter 4.

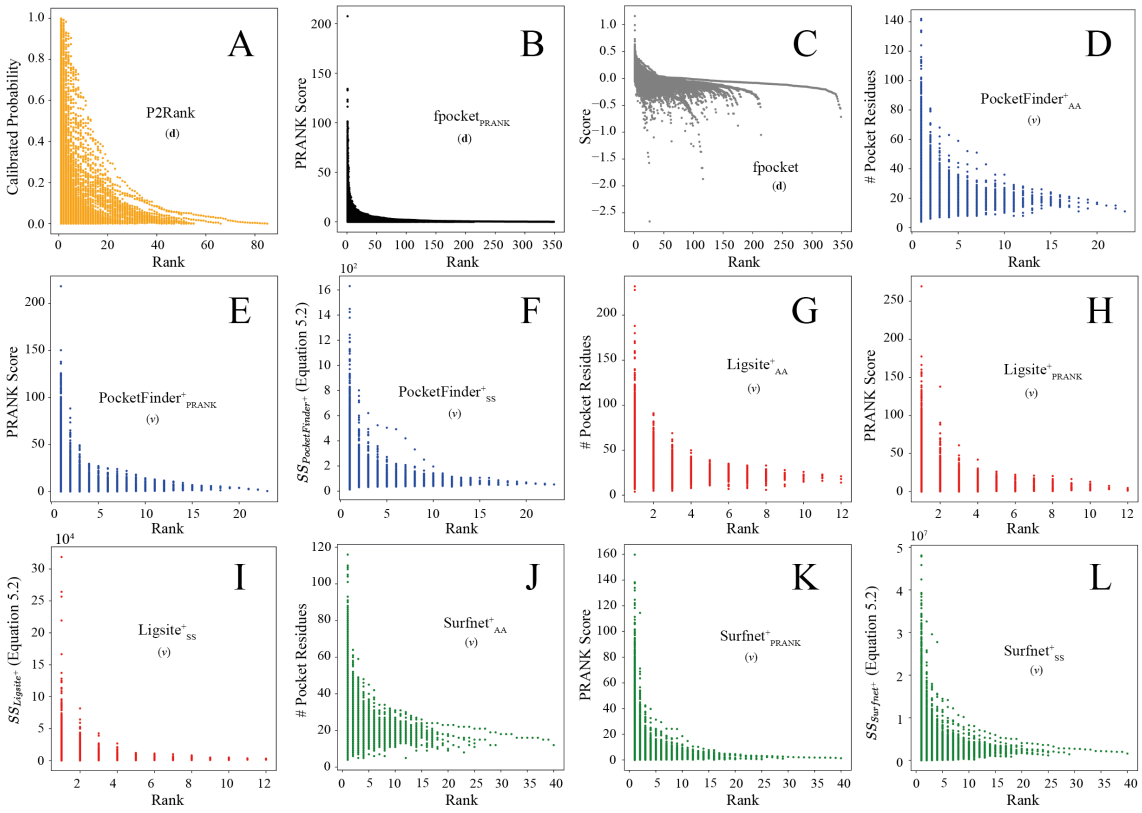

Figure 5.5. Pocket score *vs* pocket ranking (II). (A) P2Rank; **(B)** fpocket$_{PRANK}$; **(C)** fpocket. This distribution differs massively from the re-scored fpocket$_{PRANK}$ one; **(D)** PocketFinder^{+} does not report pocket scores, so the number of pocket residues is displayed for the PocketFinder$^{+}_{AA}$ variant; **(E)** PocketFinder$^{+}_{PRANK}$; **(F)** PocketFinder$^{+}_{SS}$. This variant uses the pocket grid points' scores to calculate a pocket score by summing the squared scores (Equation 5.2); **(G)** Just like PocketFinder^{+}, Ligsite^{+} does not score pockets, Y-axis is number of pocket residues (Ligsite$^{+}_{AA}$); **(H)** Ligsite$^{+}_{PRANK}$; **(I)** Ligsite$^{+}_{SS}$; **(J)** Surfnet$^{+}_{AA}$; **(K)** Surfnet$^{+}_{PRANK}$; **(L)** Surfnet$^{+}_{SS}$. (**d**) and (*v*) indicate whether methods are default or a variant generated in this work.

The score distributions of "$_{AA}$", "$_{SS}$" and "$_{PRANK}$" variants of PocketFinder^{+}, Ligsite^{+} and Surfnet^{+} are similar, suggesting that the number of pocket amino acids might dictate the order in which these pockets are reported and that re-scoring predictions by these methods might not have an effect on their performance (Figure 5.5 D-L).

5.3.2 Effect of redundancy and pocket score on recall

Figure 5.4 and Figure 5.5 demonstrate the drastic effect of redundancy removal in pocket ranking, with VN-EGNN as the clearest example. The following analysis explored the effect of redundancy removal and different pocket scoring schemes in recall for PUResNet, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}, which do not report pocket scores.

Figure 5.4 and Figure 5.5 demonstrate how removing redundancy from predictions can have a drastic effect in the ranking of the predictions, with VN-EGNN being the clearest example. The following analysis explored the effect that redundancy removal and different pocket scoring schemes have on recall for PUResNet, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}, which do not report pocket scores.

Figure 5.6 A-B shows a significant +5.2% increase in recall after removing redundancy for VN-EGNN predictions (Recall = 46.1%). This increase corresponds to 346 extra predictions that fall within the top-N+2 after redundancy removal. An even stronger improvement can be observed for IF-SitePred (Figure 5.6 C-D), where a combination of redundancy removal and pocket re-scoring (Equation 5.1) results in a significant increase of +13.4% (Recall = 39.1%), corresponding to 901 extra predictions within the top-N+2. Most of this change is due to the redundancy removal, as can be seen by the higher recall of IF-SitePred$_{NR}$. Scoring of PUResNet predictions using the number of pocket amino acids (PUResNet$_{AA}$) or PRANK (PUResNet$_{PRANK}$) had no effect on the recall. This was expected as PUResNet predicts a single pocket in 90% of the cases. Consequently, there is no strong need for a score to sort predictions within a protein (Figure 5.6 G-H). Just like VN-EGNN and IF-SitePred, the recall of DeepPocket$_{SEG}$ benefits from redundancy removal, increasing by +5.6% (Figure 5.6 J-K). For PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}, none of the variants had a significant improvement in the recall (Figure 5.7 A-I). This is expected as these predict only a few non-redundant sites per protein, with medians ranging 1-3 pockets per protein. Additionally, these pockets might already be sorted by number of amino acids as suggested by Figure 5.5 D-L.

5.3.3 Effect of redundancy and pocket score on # TP$_{100\ FP}$

There are no negative predictions, either true (TN) or false (FN) in the context of ligand binding site prediction at the pocket level. Accordingly, standard ROC/AUC curves cannot be obtained. Only positives are predicted (pockets). FN can be obtained by examining the observed pockets that are not predicted, but there are not scores for them.

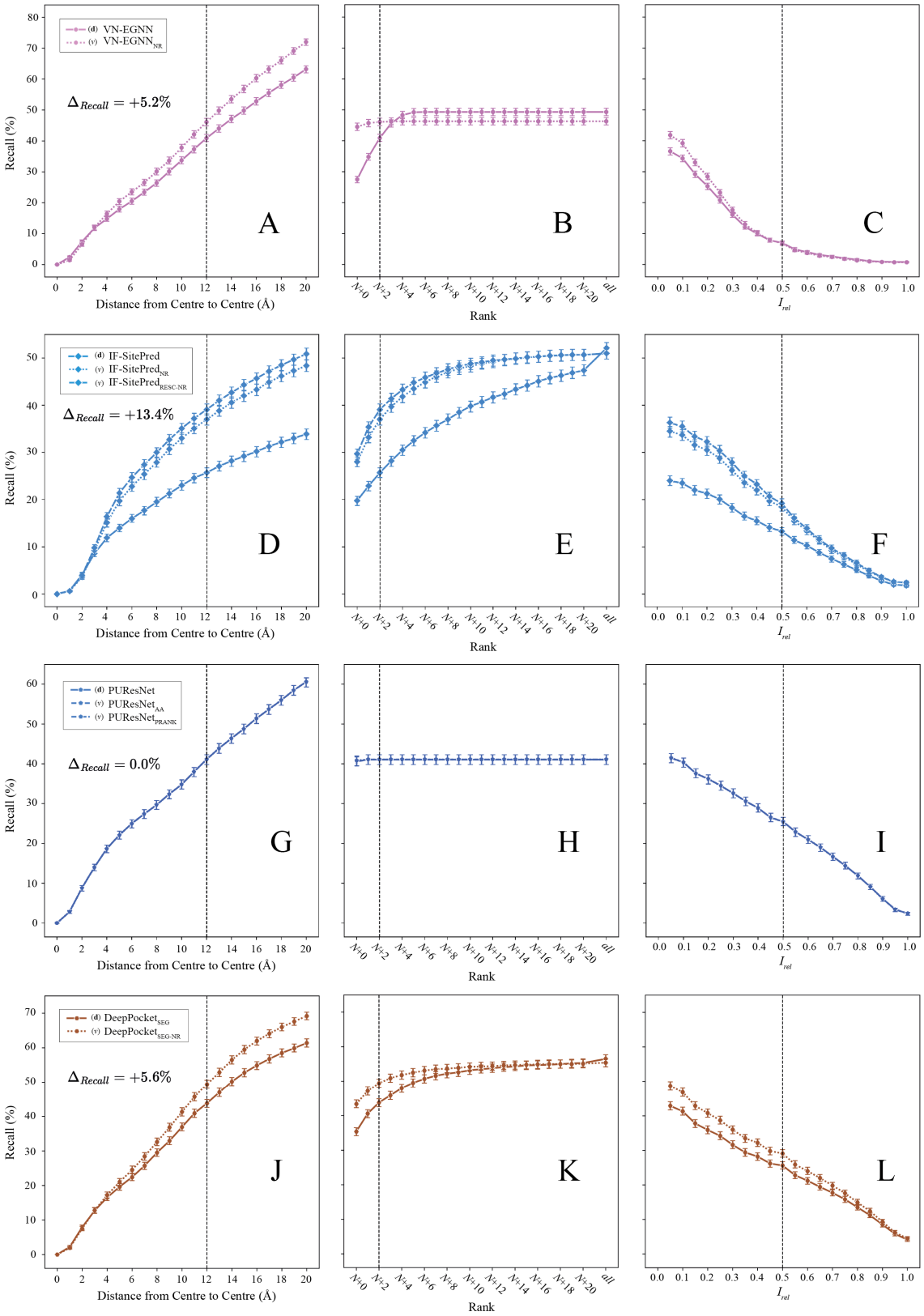

Figure 5.6. Recall curves for method variants (I). Recall curves for different scoring and ranking variants for VN-EGNN **(A-C)**, IF-SitePred **(D-F)**, PUResNet **(G-I)** and DeepPocket$_{SEG}$ **(J-L)**. For each method, panels illustrate how recall changes as DCC, rank and I_{rel} thresholds vary. In this last one, I_{rel} is the criterion used to classify predictions. Dashed lines indicate the thresholds used as reference in this work: DCC = 12 Å, rank = top-N+2, and I_{rel} = 0.5. (**d**) and (*v*) indicate whether methods are default or variants.

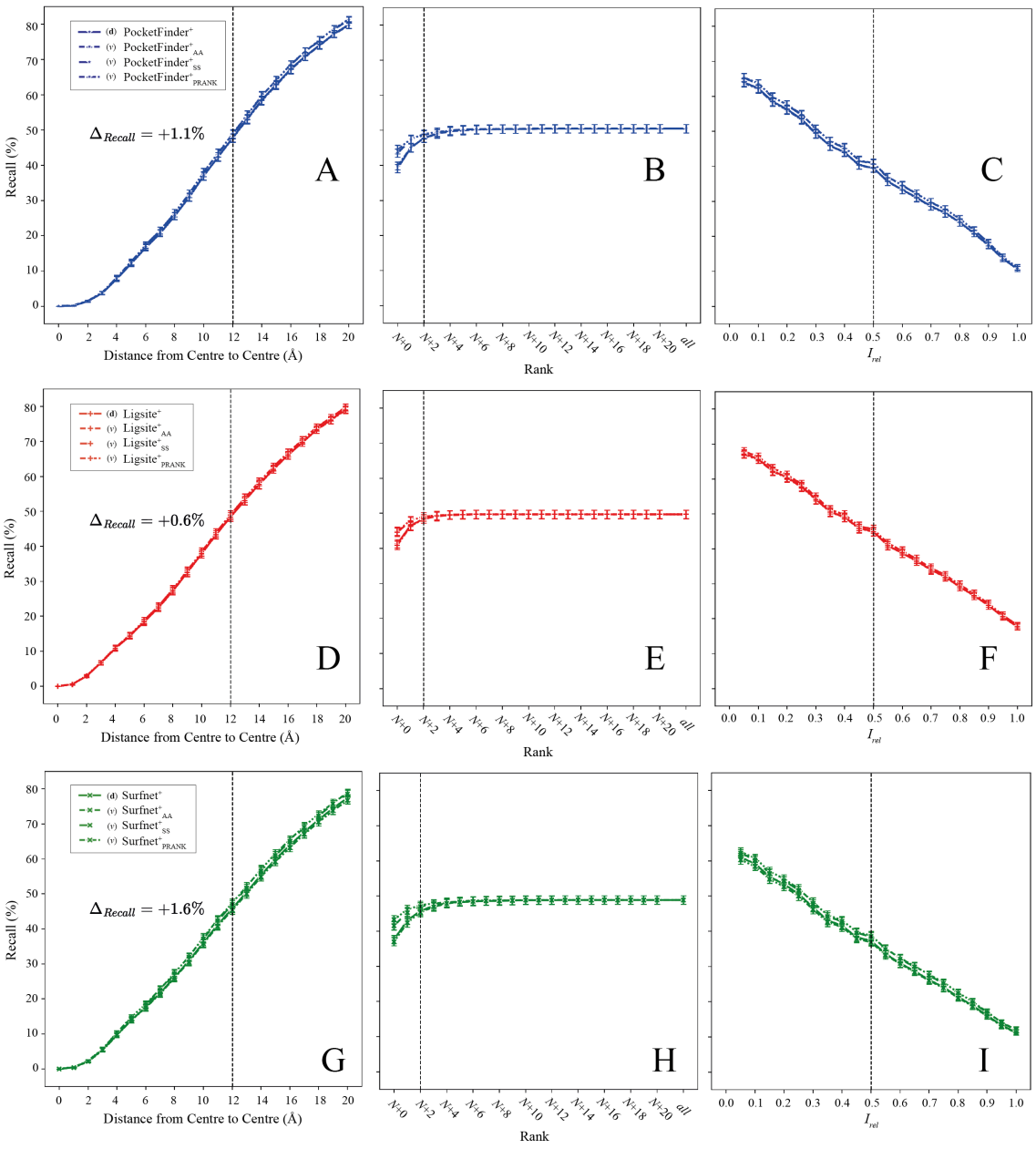

Figure 5.7. Recall curves for method variants (II). Recall curves for different scoring and ranking variants for PocketFinder^{+} **(A-C)**, Ligsite^{+} **(D-F)** and Surfnet^{+} **(G-I)**. For each method, panels illustrate how recall changes as DCC, rank and I_{rel} thresholds vary. In this last one, I_{rel} is the criterion used to classify predictions. Dashed lines indicate the thresholds used as reference in this work: DCC = 12 Å, rank = top-N+2, and I_{rel} = 0.5. (**d**) and (*v*) indicate whether methods are default or variants.

ROC_{100} curves provide an alternative to observe the relationship between true (TP) and false positives (FP). Predictions for each method across the whole reference dataset, LIGYSIS, were sorted based on pocket score, and cumulative TP and FPs were counted until a certain number of FP was reached, in this case, 100. This visualisation provides insight into how well high-scoring predictions match the ground truth. A higher number of

TP at FP = 100 indicates that the high-scoring pockets recapitulate well the ground truth, whereas a low number indicates that pockets scoring high do not match the observed data, given the used threshold of DCC $\leq$ 12 Å. It is important to understand that FPs in this context do not always represent wrong predictions, but could be binding sites that are not considered in the ground truth dataset, composed by biologically relevant protein-ligand interactions [228]. They could also be relevant sites that simply have not been experimentally determined yet. It is also important to contextualise this metric with success rate, or recall, i.e., how many of the observed sites are predicted by each method given the above-mentioned DCC threshold, as well as a rank threshold: top-N+2. A method might present a high number of TP within the first 100 FP, yet have a low recall overall. Figure 5.8 explores how ROC_{100} changes for the non-redundant "$_{NR}$" and re-scored "$_{AA}$", "$_{PRANK}$", "$_{SS}$" and "$_{RESC}$" sets of VN-EGNN, IF-SitePred, PUResNet, $DeepPocket_{SEG}$, $PocketFinder^{+}$, $Ligsite^{+}$ and $Surfnet^{+}$.

Figure 5.8 A illustrates how redundancy can be misleading and overestimate the performance of VN-EGNN. Removing redundancy results in $\Delta_{TP} = -273$ (TP = 1028). This is because redundant predictions by VN-EGNN are very close in space and present very similar scores (Figure 5.4 A). Because of this, in the redundant default set of predictions, multiple TP counts were being added for predictions of the same observed pocket. Even with redundancy removed, VN-EGNN reached 1028 TP for the first 100 FP, indicating that the non-redundant higher-scoring pockets recapitulate well the observed data.

There was no difference between IF-SitePred and $IF\text{-}SitePred_{NR}$ (curves overlap completely), which indicates that despite the redundancy in predictions by this method, its scoring scheme can sort sites in a meaningful manner. Considering multiple proteins with redundant predictions for IF-SitePred: the scoring scheme allows for the top-1 site of each of these proteins to rank above any of the other redundant predictions of the other proteins. The re-scored and non-redundant set of IF-SitePred predictions, $IF\text{-}SitePred_{RESC\text{-}NR}$, results in a $\Delta_{TP} = +285$ (TP = 1246), indicating that IF-SitePred could benefit from a more sophisticated scoring scheme, rather than the number of cloud points per binding site (Figure 5.8 B).

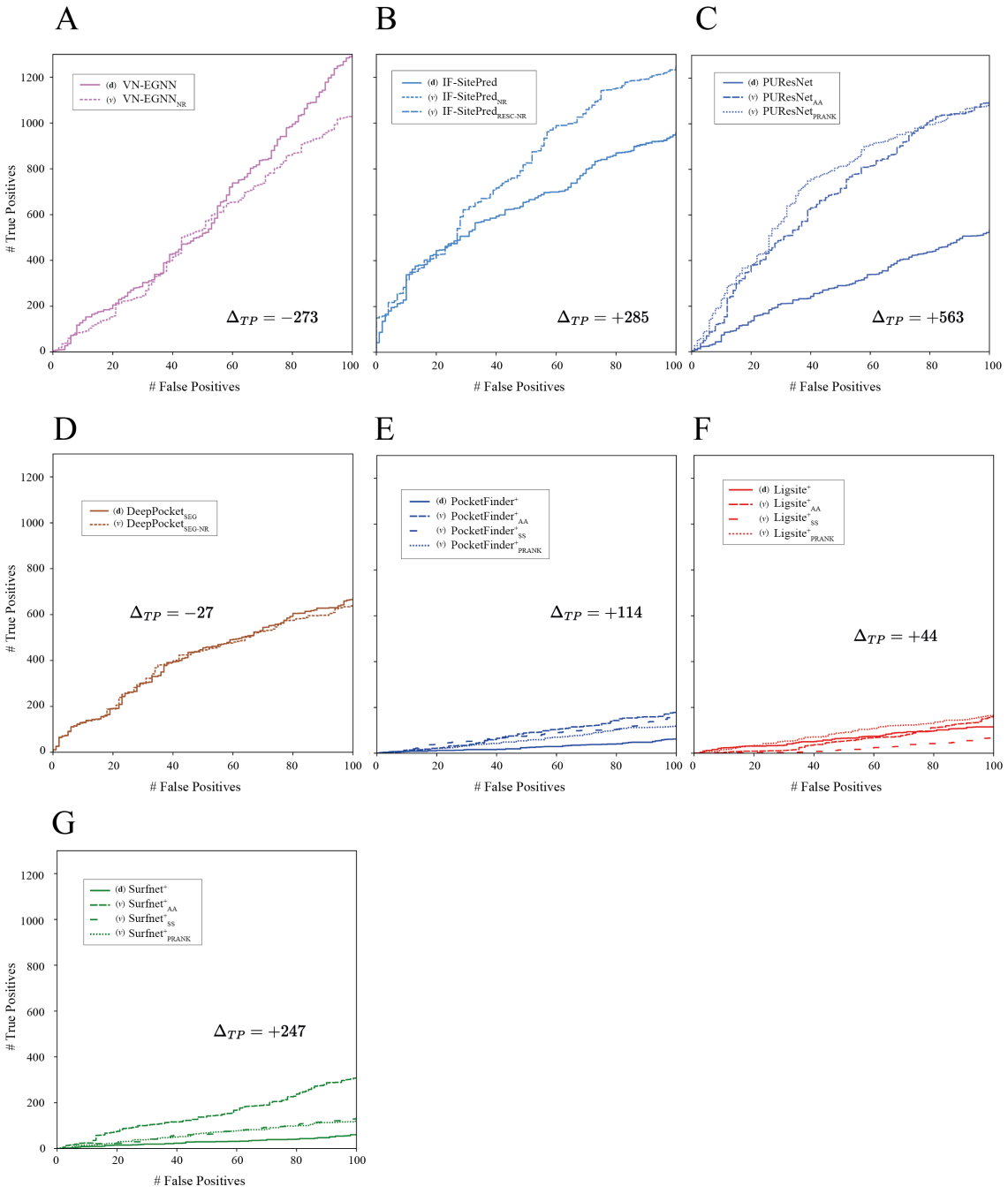

Figure 5.8. ROC_{100} curves for non-redundant and re-scored variants. For each method, predicted pockets across the whole dataset, i.e., all LIGYSIS protein chains, were ranked by their score. This way, pockets with the highest scores were at the top of the list, whereas pockets with the lowest scores located at the bottom. This ranking does not correspond to ranking pockets across proteins by their rank, as a pocket ranked #2, #3 or lower could have a higher score than a pocket #1 on a different protein. Each method has a colour assigned and each variant resulting of redundancy removal or pocket (re-)scoring a different line style. **(A)** VN-EGNN and "$_{NR}$" variant; **(B)** IF-SitePred, non-redundant ("$_{NR}$") and non-redundant re-scored ("$_{RESC\text{-}NR}$") variants; **(C)** PUResNet, "$_{AA}$" and "$_{PRANK}$" variants; **(D)** $DeepPocket_{SEG}$ and "$_{NR}$" variant; **(E)** $PocketFinder^{+}$, "$_{AA}$", "$_{PRANK}$" and "$_{SS}$" variants; **(F)** $Ligsite^{+}$ and variants; **(G)** $Surfnet^{+}$ and variants.

Figure 5.8 C is a perfect example of the importance of scoring pocket predictions. PUResNet does not score its predictions. For this reason, within a protein, pockets were ranked based on the order they are reported, i.e., on their identifier. When sorting across the whole dataset, pockets with the same ID or rank were randomly shuffled. A massive increase in TP was observed when sorting simply by the number of pocket residues. Using PRANK to score this pockets provided an even larger increment in TP of $\Delta_{TP} = +563$ (TP = 1097). An application of this could be running PUResNet on a list of potential drug target proteins. Ranking pocket predictions across different proteins could be a useful criterion to prioritise more druggable targets.

The curve did not change much for DeepPocket$_{SEG}$ with $\Delta_{TP} = -27$ (TP = 643). This indicates that despite the overlap in pockets resulting from DeepPocket's segmentation module, the method's scoring scheme is robust. It is important to consider that this method's pocket scores come from the re-scoring the fpocket candidates, which are not redundant. The redundancy in DeepPocket$_{SEG}$ is therefore unrelated to its scoring scheme, but instead a direct consequence of the shape extraction segmentation module. This suggests that there is a big difference between fpocket candidates (which result in DeepPocket's scores) and the extracted shapes by DeepPocket$_{SEG}$. This difference raises the question of whether it is technically accurate to consider the DeepPocket$_{RESC}$ score for the newly segmented pockets by DeepPocket$_{SEG}$ (Figure 5.8 D).

For the last three methods, earlier and geometry/energy-based PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}, the results agree in that simply using the number of pocket amino acids results in the maximum TP for 100 FP: $\Delta_{TP} = +114$ (TP = 178) (Figure 5.8 F), $\Delta_{TP} = +44$ (TP = 159) (Figure 5.8 G) and $\Delta_{TP} = +247$ (TP = 308) (Figure 5.8 H) for PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}, respectively. This is surprising, as sum of squares "$_{SS}$" and "$_{PRANK}$" scoring schemes worked better for other methods. This result might be related to the fact that pockets predicted by these three methods tend to be larger than those predicted by other methods.

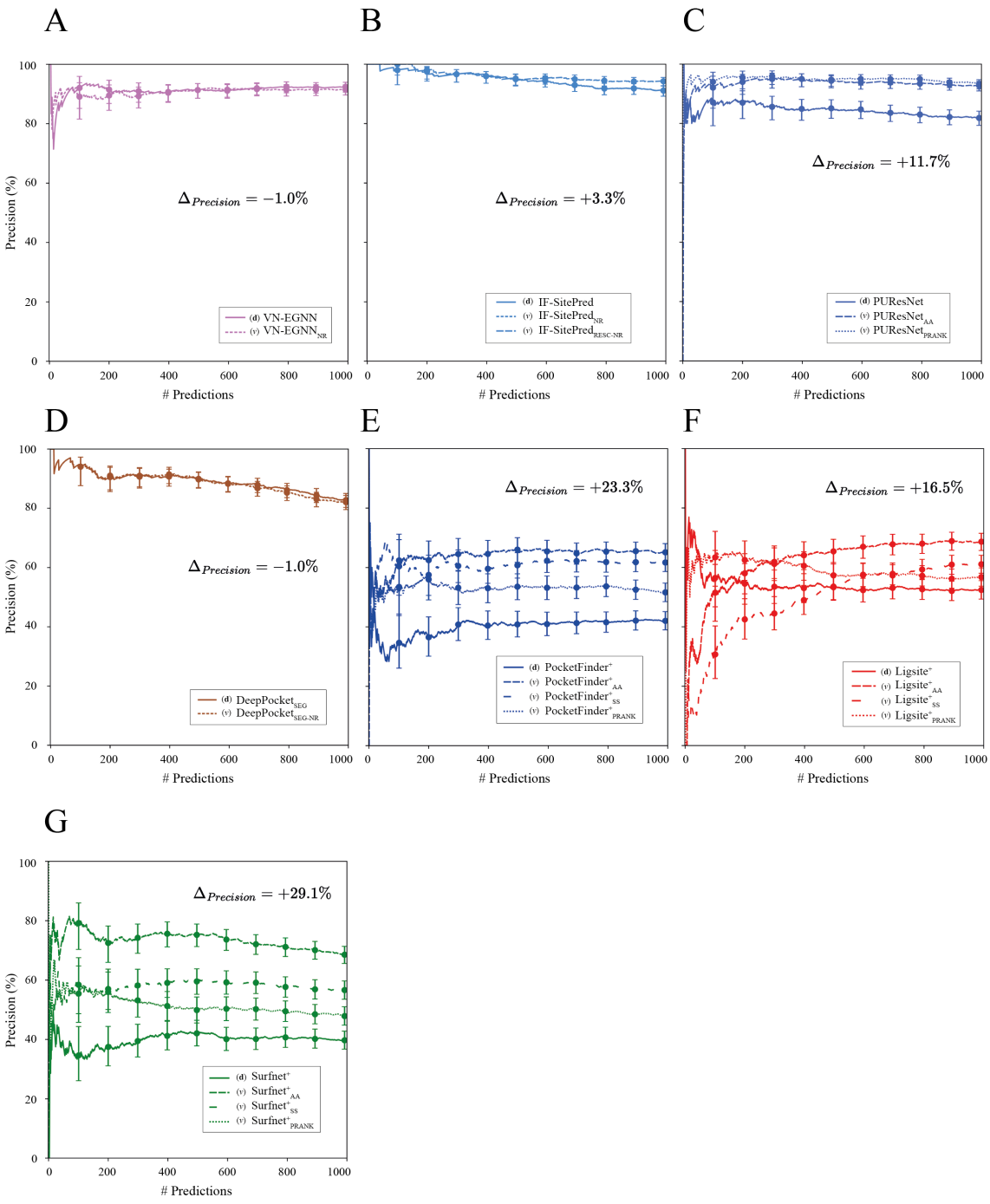

Figure 5.9. Precision$_{1K}$ curves for non-redundant and re-scored variants. Precision$_{1K}$ represents the precision (%) calculated for the top-scoring 1000 predictions. For each method, predicted pockets across the whole LIGYSIS set were ranked by score. This way, pockets with the highest scores were at the top of the list, whereas pockets with the lowest scores at the bottom. Each method has a colour assigned and each scoring variant its own line style. $\Delta_{Precision}$ indicates the difference in precision between the selected method variant and the default one (%). **(A)** VN-EGNN and "$_{NR}$" variant; **(B)** IF-SitePred, "$_{NR}$" and "$_{RESC-NR}$" variants; **(C)** PUResNet, "$_{AA}$" and "$_{PRANK}$" variants; **(D)** DeepPocket$_{SEG}$ and "$_{NR}$" variant; **(E)** PocketFinder^{+}, "$_{AA}$", "$_{PRANK}$" and "$_{SS}$" variants; **(F)** Ligsite^{+} and variants; **(G)** Surfnet^{+} and variants. Error bars indicate 95% CI of the precision (100 $\times$ proportion) and are displayed every 100 predictions.

5.3.4 Effect of redundancy and pocket score on precision

Precision-recall curves cannot be calculated for ligand site prediction at the pocket level for the same reason as ROC/AUC: false negatives (FN) are not predicted, and therefore not scored. Nevertheless, precision can be measured. For this, as it was done for ROC_{100}, all predictions for a method were sorted by pocket score and precision calculated as more predictions with lower scores were considered.

Figure 5.9 portrays the precision curve for the top-1000 predictions for the non-redundant and re-scored variants for VN-EGNN, IF-SitePred, PUResNet, $DeepPocket_{SEG}$, PocketFinder$^+$, Ligsite$^+$ and Surfnet$^+$. There was no significant ($p > 0.05$) change in precision between VN-EGNN and $VN\text{-}EGNN_{NR}$ within the first 1000 predictions, $Precision_{1K}$ = 91.5% (Figure 5.9 A). The same can be said for IF-SitePred with a $Precision_{1K}$ = 94.3% (Figure 5.9 B). Using PRANK to score PUResNet pockets resulted in a significant +11.7% increase with precision = 93.3% (Figure 5.9 C). $DeepPocket_{SEG\text{-}NR}$, as the other redundant methods, did not experience a significant change in precision as redundancy is removed: precision = 81.6% (Figure 5.9 D). Using the number of pocket amino acids as a score ("$_{AA}$") resulted in a precision increase of +23.3% ($Precision_{1K}$ = 65.3%), +16.5% ($Precision_{1K}$ = 68.8%) and 29.1% ($Precision_{1K}$ = 68.8%) for PocketFinder$^+$ (Figure 5.9 E), Ligsite$^+$ (Figure 5.9 F) and Surfnet$^+$ (Figure 5.9 G), respectively.

5.3.5 Evaluation of predictive performance

Figure 5.10 and Table 5.1 compare the performance of the thirteen methods evaluated in this Chapter, which now include six canonical methods (**d**) and seven novel variants (*v*) first introduced in this Chapter. In terms of recall, the ranking of the methods does not change much. The increase obtained by the re-scoring and non-redundant variants, though considerable (+13.4% for IF-SitePred) is not enough to reach the recall achieved by the top-performing original methods (Figure 5.10 A-C). However, a shift in the ranking can be observed in $Precision_{1K}$ and # $TP_{100\ FP}$. PUResNet, PocketFinder$^+$, Ligsite$^+$ and Surfnet$^+$, which originally did not score their predictions, benefit greatly of scoring their predictions

by simply using the number of amino acids of predicted pockets as a score. This results in increases of up to ≈560 TP at 100 FP for PUResNet$_{PRANK}$ and 29.1% in precision for Surfnet$^{+}_{AA}$. The improvement of the best seven novel variants relative to their default counterparts is summarised in Table 5.2.

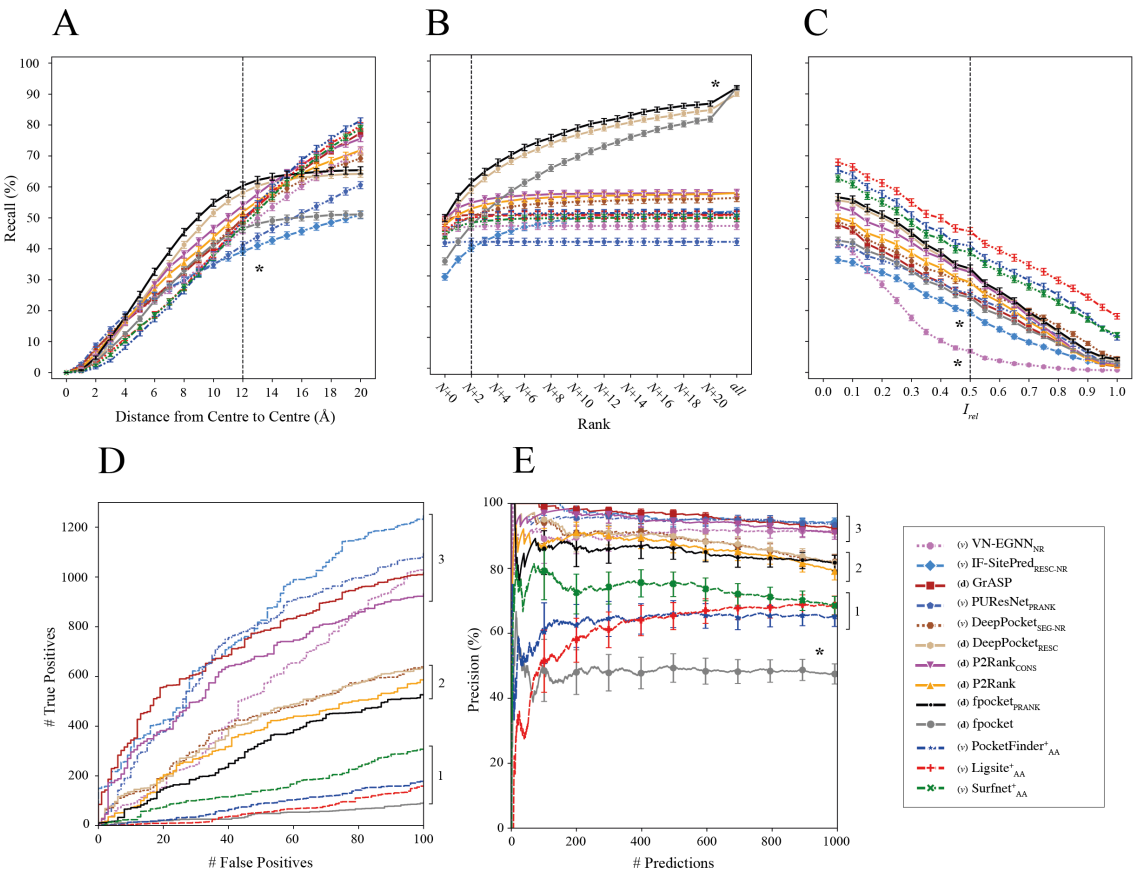

Figure 5.10. Ligand binding site prediction at the pocket level (*best* variants). Only the top-performing, i.e., highest top-*N*+2 recall, variant of each method is drawn on this figure, e.g., IF-SitePred$_{RESC-NR}$ or VN-EGNN$_{NR}$ instead of their default modes. **(A)** Top-*N*+2 recall according to a DCC threshold. Reported recall on Table 5.1 corresponds to DCC = 12 Å; **(B)** Recall using DCC = 12 Å but considering increasing rank thresholds. *all* represents the maximum recall of a method, obtained by considering all predictions, regardless of their rank or score; **(C)** Recall curve using I_{rel} as a criterion; **(D)** ROC$_{100}$ curve (cumulative TP against cumulative FP until 100 FP are reached); **(E)** Precision curve for the top-1000 predictions of each method across the LIGYSIS dataset. Error bars represent 95% CI of the recall **(A-C)** and precision **(E)**, which are 100 × proportion. Numbers at the right of the panels indicate groups or blocks of methods that perform similarly for each metric. Stars "*" indicate outlier methods, or methods that perform very differently than the rest. (**d**) and (*v*) indicate whether methods are default or variant, respectively.

Method	Recall$_{\text{top-}N}$ (%)	Recall$_{\text{top-}N+2}$ (%)	Recall$_{\text{max}}$ (%)	Precision$_{\text{1K}}$ (%)	# TP$_{\text{100 FP}}$	RRO (%)	RVO (%)
(*v*) VN-EGNN$_{\text{NR}}$	44.5 (#7)	46.1 (#11)	46.3 (#11)	91.5 (#4)	1028 (#3)	**31.6 (#11)**	**26.7 (#11)**
(*v*) IF-SitePred$_{\text{RESC-NR}}$	**29.7 (#12)**	**39.1 (#13)**	51.6 (#12)	**94.3 (#1)**	**1246 (#1)**	49.3 (#10)	43.7 (#9)
(**d**) GrASP	48.0 (#2)	49.9 (#5)	50.0 (#8)	92.5 (#3)	1017 (#4)	54.5 (#7)	59.8 (#6)
(*v*) PUResNet$_{\text{PRANK}}$	40.8 (#10)	41.1 (#12)	**41.1 (#12)**	93.3 (#2)	1092 (#2)	61.0 (#4)	63.9 (#4)
(*v*) DeepPocket$_{\text{SEG-NR}}$	43.4 (#8)	49.4 (#6)	55.4 (#5)	81.6 (#7)	643 (#6)	58.4 (#5)	61.3 (#5)
(**d**) DeepPocket$_{\text{RESC}}$	46.6 (#4)	58.1 (#2)	89.3 (#2)	81.7 (#6)	637 (#7)	52.6 (#9)	38.2 (#10)
(**d**) P2Rank$_{\text{CONS}}$	**48.8 (#1)**	53.9 (#3)	57.0 (#3)	90.7 (#5)	932 (#5)	56.4 (#6)	43.8 (#8)
(**d**) P2Rank	46.7 (#3)	51.9 (#4)	57.0 (#4)	79.2 (#8)	586 (#8)	54.4 (#8)	58.2 (#7)
(**d**) fpocket$_{\text{PRANK}}$	**48.8 (#1)**	**60.4 (#1)**	**91.3 (#1)**	81.7 (#6)	526 (#9)	52.6 (#9)	38.2 (#10)
(**d**) fpocket	38.8 (#11)	46.5 (#10)	**91.3 (#1)**	**47.3 (#12)**	**94 (#13)**	52.6 (#9)	38.2 (#10)
(*v*) PocketFinder$^{+}_{\text{AA}}$	44.5 (#6)	48.9 (#8)	50.5 (#7)	65.3 (#11)	178 (#11)	72.3 (#2)	75.9 (#2)
(*v*) Ligsite$^{+}_{\text{AA}}$	44.9 (#5)	49.0 (#7)	49.7 (#9)	68.8 (#9)	159 (#12)	**77.6 (#1)**	**77.0 (#1)**
(*v*) Surfnet$^{+}_{\text{AA}}$	43.3 (#9)	47.4 (#9)	48.9 (#10)	68.6 (#10)	308 (#10)	71.7 (#3)	72.0 (#3)

Table 5.1. Pocket level evaluation (*best* variants). Only the top-performing, i.e., highest top-*N*+2 recall, variant of each method is present on this table. Recall (%) for each method considering top-*N*, *N*+2 and *all* predictions (max), i.e., maximum recall. Precision (%) of the method for the top-1000 scored predictions. Number of TP reached for the first 100 FP (# TP$_{\text{100 FP}}$). Mean relative residue overlap (RRO) for correctly predicted sites and relative volume overlap (RVO) only for sites that have a volume, i.e., are pockets or cavities, and not fully exposed sites, which do not have a volume. RRO and RVO represent the overlap in residues and volume relative to the observed site. Bold font indicates the best (blue) and worst (orange) performing methods for each metric. (**d**) and (*v*) indicate whether methods are default or variant, respectively.

Default method (d)	**Best variant (*v*)**	$\Delta_{\mathbf{Recall_{top\text{-}N}}}$ (%)	$\Delta_{\mathbf{Recall_{top\text{-}N+2}}}$ (%)	$\Delta_{\mathbf{Recall_{max}}}$ (%)	$\Delta_{\mathbf{Precision_{1K}}}$ (%)	$\Delta_{\mathbf{\#\ TP_{100\ FP}}}$
VN-EGNN	$\text{VN-EGNN}_{\text{NR}}$	+17.0	+5.2	−3.0	−1.0	−273
IF-SitePred	$\text{IF-SitePred}_{\text{RESC-NR}}$	+9.9	+13.4	−0.5	+3.3	+285
PUResNet	$\text{PUResNet}_{\text{PRANK}}$	+0.2	0.0	0.0	+11.7	+558
$\text{DeepPocket}_{\text{SEG}}$	$\text{DeepPocket}_{\text{SEG-NR}}$	+8.0	+5.6	−1.1	−1.0	−27
PocketFinder^{+}	$\text{PocketFinder}^{+}{}_{\text{AA}}$	+5.3	+1.1	0.0	+23.3	+114
Ligsite^{+}	$\text{Ligsite}^{+}{}_{\text{AA}}$	+3.6	+0.6	0.0	+16.5	+44
Surfnet^{+}	$\text{Surfnet}^{+}{}_{\text{AA}}$	+6.0	+1.6	0.0	+29.1	+247

Table 5.2. Methods improvement summary. Summary of the performance improvement for the seven methods for which non-redundant or re-scoring variants were generated in this Chapter: VN-EGNN, IF-SitePred, PUResNet, $\text{DeepPocket}_{\text{SEG}}$, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}. Performance increase is calculated for each metric from the *best* variant (*v*), i.e., highest top-*N*+2 recall, relative to the default original method (**d**). Maximum recall is reduced for VN-EGNN, IF-SitePred and $\text{DeepPocket}_{\text{SEG}}$ as their non-redundant variants present fewer pockets. The same happens for $\text{Precision}_{1\text{K}}$ and $\#\ \text{TP}_{100\ \text{FP}}$ for VN-EGNN. Overall, the method variants introduced in this work have a significant positive effect on performance.

5.4 Discussion

This Chapter shows that prediction redundancy underestimates recall and overestimates precision, therefore providing a misleading performance assessment. Redundancy removal and subsequent pocket re-ranking can yield a significant increase in recall. A robust pocket scoring scheme can also have a major impact in performance, both in recall and precision and emphasis should be put into this area. Even if a single site is predicted per protein – as it is the case for PUResNet – a pocket score can be highly useful when ranking predicted pockets across different proteins or conformations, e.g., when having a list of potential drug targets and deciding which protein might be best to target therapeutically. Table 5.2 summarises the performance improvements accomplished in this Chapter by removing redundant predictions (VN-EGNN, IF-SitePred and $\text{DeepPocket}_{\text{SEG}}$) as well as using more sophisticated pocket scoring schemes (IF-SitePred, PUResNet, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}). The magnitude of these improvements is notable: increase in top-N+2 Recall by >15%, $\text{Precision}_{\text{1K}}$ by ≈30% and # $\text{TP}_{\text{100 FP}}$ by >500. Nevertheless, the overall ranking of the methods after including novel variants (Table 5.1) does not change much relative to the benchmark of the default methods (Table 4.6). This is due to the fact that the difference in $\text{Recall}_{\text{top-}N\text{+2}}$ between the top-performing methods (fpocket and re-scored predictions) and the ones for which variants were generated (VN-EGNN, IF-SitePred and $\text{DeepPocket}_{\text{SEG}}$) was larger than the increase in recall resulting from the variants. While removing redundancy post-prediction has a significant improvement in performance ($\text{VN-EGNN}_{\text{NR}}$ and $\text{IF-SitePred}_{\text{NR}}$), approaching this issue before prediction would be better. For VN-EGNN, which predicts a maximum of eight sites, ensuring all of these predictions are non-redundant is more convenient than removing redundant ones ending up with 1/8 predictions. The same applies to IF-SitePred. This method would also benefit from more sophisticated residue and pocket scoring schemes, as well as a different cloud point clustering and site definition algorithm. It is likely that after tweaking their algorithm and approaching the issues highlighted in this and Chapter 4, these methods could perform better and perhaps even out-perform fpocket and P2Rank.

5.5 Conclusions

The conclusions resulting from the work presented in this Chapter are as follows:

- Redundancy in ligand binding site prediction leads to the underestimate of recall and the overestimate of precision. The removal of such redundancy and subsequent re-ranking of the remaining pockets results in a drastic increase in recall.
- A robust pocket scoring scheme is crucial for the correct ranking and prioritisation of predicted sites in downstream analysis, e.g., docking, simulation. Additionally, it has a significant positive effect on both precision and recall.
- IF-SitePred benefits significantly from pocket re-scoring, which suggests that protein inverse folding embeddings, which are not directly dependent on the input structure, represent great promise in the field of ligand site prediction.

Chapter 6

Conclusions

Preface

This Chapter brings together the main findings presented in the four previous Chapters of results of this Thesis and contextualises them with the current state of the art within the field. Additionally, it covers the direction in which this research is heading by describing the next steps that could be taken to take the work presented in this Thesis further.

6.1 Introduction

The work presented in this Thesis sheds light into the nature of ligand binding sites by developing methods to define and characterise them. A pipeline is described to characterise experimentally determined fragment screening and biologically relevant ligand binding sites in the PDBe resulting in the LIGYSIS dataset. A web resource is presented for the exploration of 65,000 ligand sites across 25,000 proteins, as well as for the analysis of user sets of protein-ligand complexes. Finally, the human component of the LIGYSIS set is employed to carry out the largest independent critical comparative assessment of ligand binding site prediction tools to date.

6.2 Fragment screening sites analysis

Chapter 2 describes novel methods for the definition and characterisation of 293 ligand binding sites derived from 37 publicly available fragment screening experiments [353]. In this Chapter, sites are defined by grouping ligands based on their interactions with the protein, therefore avoiding completely the need for structural superposition. Ligand binding sites were later grouped into four robust clusters (C1-C4) by their relative solvent accessibility profile. The four clusters differed in solvent accessibility, evolutionary divergence across homologues, relative enrichment in human missense variation, and most importantly, enrichment in known functional sites. C1 sites were the most buried, conserved across homologues, depleted in missense variation and enriched in functional sites, whereas C4 sites were the complete opposite.

This information could be used in early-stage drug discovery to prioritise between targets and between sites within the same target protein, potentially leading to a reduction in attrition in clinical trials. Cluster 1 sites are ≈28 times more likely to be functional than sites with a C4 label. Consequently, targeting a C1 site is much more likely to have an effect on the target protein function. Considering the evolutionary divergence profile of the site can also provide insight into off-target effects, i.e., targeting a conserved site might also affect other members of the same protein family. In a similar manner, the missense enrichment profile of the site offers insight into potential pharmacogenetic effects. For example, a site with a high enrichment in genetic variation might not be desirable, as individuals with distinct alleles might respond differently to the drug targeting the site. This information could be leveraged in a different way if looking for a site for a molecular glue [554], targeted modification [555–557] or degradation [558, 559] of a protein. In this case, easily accessible sites that are divergent across homologues and do not alter the protein function in any way might be of interest, i.e., C3-C4 sites.

6.3 The LIGYSIS dataset and web resource

Chapter 3 extends the ligand site definition and characterisation methodology introduced in Chapter 2 and applies it to the entire PDBe, resulting in the LIGYSIS dataset. LIGYSIS comprises 65,000 ligand binding sites from biologically relevant protein-ligand complexes across 25,000 proteins. Sites are characterised in terms of conservation, variation and accessibility in the same manner as the fragment screening sites. In addition, a score is defined that is indicative of the likelihood of function for a given ligand binding site.

Chapter 3 also describes the architecture and implementation of LIGYSIS-web, a resource for the analysis of protein-ligand binding sites. LIGYSIS-web is a Python Flask web application to dynamically explore the full LIGYSIS dataset. Furthermore, users can submit their set of protein-ligand complexes for analysis and subsequent visualisation and download of results. The LIGYSIS dataset can be browsed by UniProt accession and provides rich output with hyperlinks to the relevant databases, downloadable tabular data, alignments, images and structural data in PyMOL or ChimeraX format. LIGYSIS-web can be found here: https://www.compbio.dundee.ac.uk/ligysis/.

6.4 Assessing ligand binding site prediction tools

Chapter 4 uses the human component of the LIGYSIS dataset to carry out the largest independent comparative assessment of ligand binding site prediction tools to date [342]. Thirteen canonical methods are evaluated at the pocket and residue level employing fourteen informative metrics. LIGYSIS defines ligand binding sites by aggregating unique biologically relevant protein-ligand interfaces across the biological assemblies of multiple structures available for a given protein. Other datasets, previously used to train and test ligand binding site prediction tools, consider redundant protein-ligand interfaces, single ligand-protein complexes and asymmetric instead of biological units. For these reasons, LIGYSIS presents and advantage over these other sets and is therefore proposed as a new reference set to test on for future methods or benchmarks.

Re-scored fpocket predictions by PRANK (60.4%) and DeepPocket (58.1%) present the highest recall considering a threshold of DCC = 12 Å and predictions within the top-N+2, demonstrating the paramount importance of a robust pocket scoring scheme. VN-EGNN, IF-SitePred and other machine learning-based methods are very precise (>90%). However their recall is low (<40%) due to the small number of predicted pockets (PUResNet), the redundancy of their predictions (VN-EGNN), sub-optimal clustering of points (IF-SitePred) or score thresholding (GrASP). Chapter 4 presents clear evidence to pinpoint the areas in which the ligand site prediction community should focus on in order to improve the quality of the methods as well as the rigorosity with which these are evaluated.

6.5 Improving ligand binding site prediction tools

Chapter 5 carries on the work in Chapter 4 by exploring two factors identified to have an effect on ligand binding site prediction performance: pocket scoring and prediction redundancy [342]. Fifteen non-redundant and scoring variants are analysed for methods that result in redundant predictions: VN-EGNN, IF-SitePred and DeepPocket$_{SEG}$ and methods that do not score their pockets: PUResNet, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}. This Chapter demonstrates the detrimental effect of redundancy in pocket prediction by showing improvements in recall of >5% for VN-EGNN and DeepPocket$_{SEG}$ and up to 13% for IF-SitePred after re-scoring its predictions too. Additionally, pocket scoring can result in improvements of up to 29% in Precision$_{1K}$ (Surfnet^{+}) and >500 # TP$_{100\,FP}$ (PUResNet).

The improvement in performance accomplished in this Chapter is limited by the fact that it takes place post-prediction. It makes sense to think that approaching these issues prior to prediction, in the source code of the methods, would have an even higher positive impact on their performance. For example, making use of the probabilities returned by the 40 IF-SitePred models and improving their clustering implementation, ensuring non-redundant predictions for VN-EGNN, lowering the atom ligandability threshold on GrASP, or returning more than a single prediction for PUResNet. Authors are encouraged to consider these aspects for the benefit of their tools and the community.

6.6 Future steps

Future work on LIGYSIS-web would focus on the analysis of heterometric protein-ligand complexes and the implementation of functionalities to improve the accessibility and usability of the database. Examples of these improvements are: a function to search by ligand name or type, the selection of multiple sites or residues at once, the export of the current structure view – instead of a default one – or the alignment of all structures of a protein on the same coordinate reference. Additionally, an integration of LIGYSIS-web with Jalview would be really enriching, as it would boost access to the LIGYSIS resource and allow for enhanced integrative analysis with the many features, databases and tools that Jalview offers.

Another area of interest would be the ligand site functional score. The current score, reported on LIGYSIS-web, is purely based on the solvent accessibility profile of a given ligand site. Exploring the relationship between the evolutionary divergence, missense enrichment profiles and known functional sites would be of interest and potentially provide further insight into the likelihood of function of a given ligand binding site. For example, a new functional score could be developed that integrates structural, protein sequence and genetic variation data, amongst other features, to predict likelihood of function.

An in-depth characterisation of the experimentally determined *pocketome*, i.e., LIGYSIS, would also be of interest. This would extend the characterisation introduced in Chapter 2 to include features such as pocket surface area, volume, hydrophobicity, charge, orientation relative to the main axis or centroid, interaction type and so on. The thorough evaluation of ligand site prediction methods in Chapter 4 and Chapter 5 strongly suggests fpocket, PRANK and P2Rank as the best predictors. A combination of these methods could be used to predict on all the human proteome making use of 3D structure models as provided by AlphaFold DB. The same characterisation pipeline could then be employed, thus providing additional information into the biological, evolutionary and structural context of the human pocketome, reaching beyond the confines of experimentally determined protein-ligand complexes.

6.7 Concluding remarks

Overall, the work presented in this Thesis aims to advance the understanding of ligand binding sites. This is done by carrying out a binding site characterisation integrating structural, divergence and variation data as well as a comprehensive evaluation of methods for binding site prediction. This characterisation could be integrated with other features or scores, such as the *P* and *FP* scores recently defined by Ibrahim *et al.* [560, 561]. These scores make use of molecular dynamics to estimate the residence time of a ligand in a protein pocket and the interaction energy of the complex. This information is in turn leveraged to identify hotspot residues with a greater contribution to the energetic stability and affinity of the complex. The combination of these two approaches would merge structural and evolutionary data with chemical and energetic information, which could be employed to screen structure databases like the PDBe or AFDB for those pockets, and residues within them, with more favourable features for therapeutic targeting.

The synthesis of these diverse perspectives provides a promising avenue of research for developing a more holistic and complete understanding of ligand binding sites. These insights are pivotal to modern drug discovery, where they could accelerate target identification, lead optimisation and reduce attrition and off-target effects. Furthermore, this research has broader implications, potentially extending beyond drug discovery to applications such as protein function annotation and evolution or enzyme engineering.

Bibliography

[1] Isayama, H. “Attack on Titan”. Vol. 34. Kodansha Comics, 2021. Chap. 137.

[2] Schidlowski, M. *et al.* Carbon isotope geochemistry of the 3.7×10^9-yr-old Isua sediments, West Greenland: implications for the Archaean carbon and oxygen cycles. *Geochim. Cosmochim. Acta* **43**, 189–199 (1979).

[3] Schopf, J. *et al.* Evidence of archean life: stromatolites and microfossils. *Precambrian Res.* **158**, 141–155 (2007).

[4] Betts, H. *et al.* Integrated genomic and fossil evidence illuminates life’s early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556–1562 (2018).

[5] Mora, C. *et al.* How many species are there on earth and in the ocean? *PLOS Biol.* **9**, 1–8 (2011).

[6] Hug, L. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).

[7] Locey, K. and Lennon, J. Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 5970–5975 (2016).

[8] Costello, M. *et al.* Can we name Earth’s species before they go extinct? *Science* **339**, 413–416 (2013).

[9] Koonin, E. *The logic of chance: the nature and origin of biological evolution*. FT press, 2011.

[10] Koonin, E. *et al.* The ancient virus world and evolution of cells. *Biol. Direct* **1**, 29 (2006).

[11] Navarro, B. *et al.* Advances in viroid-host interactions. *Annu. Rev. Virol.* **8**, 305–325 (2021).

[12] Crick, F. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).

[13] Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

[14] Watson, J. and Crick, F. Molecular structure of nucleic acids: a structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).

[15] Levene, P. and Jacobs, W. Über Inosinsäure. *Ber. Dtsch. Chem. Ges.* **42**, 1198–1203 (1909).

[16] Crick, F. *et al.* Codes without commas. *Proc. Natl. Acad. Sci. U.S.A.* **43**, 416–421 (1957).

[17] Gamow, G. Possible relation between deoxyribonucleic acid and protein structures. *Nature* **173**, 318–318 (1954).

[18] Gramatikoff, K. Genetic Code. (2008) https://en.wikipedia.org/wiki/Genetic_code#/media/File:GeneticCode21-version-2.svg.

[19] Sneath, P. Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* **12**, 157–195 (1966).

[20] Taylor, W. The classification of amino acid conservation. *J. Theor. Biol.* **119**, 205–218 (1986).

[21] Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1–12 (1976).

[22] Zvelebil, M.J. *et al.* Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961 (1987).

[23] Jalview Team. Jalview – a multiple sequence alignment editor and analysis workbench (2025) https://www.jalview.org/.

[24] Livingstone, C. and Barton, G. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Bioinformatics* **9**, 745–756 (1993).

[25] Chothia, C. and Lesk, A. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).

[26] Barton, G. Protein multiple sequence alignment and flexible pattern matching. *Methods Enzymol.* **183**, 403–428 (1990).

[27] Dayhoff, M. *et al.* A model of evolutionary change in proteins. *Atlas Protein Seq. Struct.* **5**, 345–352 (1978).

[28] Henikoff, S. and Henikoff, J. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915–10919 (1992).

[29] Barton, G. and Sternberg, M. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J. Mol. Biol.* **198**, 327–337 (1987).

[30] Needleman, S. and Wunsch, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).

[31] Smith, T. and Waterman, M. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).

[32] Livingstone, C. and Barton, G. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**, 745–756 (1993).

[33] Higgins, D. and Sharp, P. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244 (1988).

[34] Higgins, D. *et al.* CLUSTAL V: improved software for multiple sequence alignment. *Bioinformatics* **8**, 189–191 (1992).

[35] Thompson, J. *et al.* CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).

[36] Jeanmougin, F. *et al.* Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **23**, 403–405 (1998).

[37] Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

[38] Katoh, K. *et al.* MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

[39] Katoh, K. and Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).

[40] Katoh, K. and Standley, D. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

[41] Edgar, R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

[42] Edgar, R. Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).

[43] Waterhouse, A. *et al.* Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

[44] Utgés, J. *et al.* Ankyrin repeats in context with human population variation. *PLOS Comput. Biol.* **17**, e1009335 (2021).

[45] Zuckerkandl, E. and Pauling, L. Evolutionary divergence and convergence in poteins. *Evol. Genes Proteins* **1**, 97–166 (1965).

[46] Valdar, W. Scoring residue conservation. *Proteins* **48**, 227–241 (2002).

[47] Wu, T. and Kabat, E. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* **132**, 211–250 (1970).

[48] Jores, R. *et al.* Resolution of hypervariable regions in T-cell receptor beta chains by a modified Wu-Kabat index of amino acid diversity. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 9138–9142 (1990).

[49] Lockless, S. and Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299 (1999).

[50] Sander, C. and Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**, 56–68 (1991).

[51] Shenkin, P. *et al.* Information-theoretical entropy as a measure of sequence variability. *Proteins* **11**, 297–313 (1991).

[52] Gerstein, M. and Altman, R. Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* **251**, 161–175 (1995).

[53] Karlin, S. and Brocchieri, L. Evolutionary conservation of RecA genes in relation to protein structure and function. *J. Bacteriol.* **178**, 1881–1894 (1996).

[54] Thompson, J. *et al.* The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882 (1997).

[55] Landgraf, R. *et al.* Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng. Des. Sel.* **12**, 943–951 (1999).

[56] Pilpel, Y. and Lancet, D. The variable and conserved interfaces of modeled olfactory receptor proteins. *Protein Sci.* **8**, 969–977 (1999).

[57] Armon, A. *et al.* ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447–463 (2001).

[58] Valdar, W. and Thornton, J. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **42**, 108–124 (2001).

[59] Williamson, R. Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. *J. Theor. Biol.* **174**, 179–188 (1995).

[60] Mirny, L. and Shakhnovich, E. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol* **291**, 177–196 (1999).

[61] Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

[62] Rost, B. and Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599 (1993).

[63] Rost, B. and Sander, C. Conservation and prediction of solvent accessibility in protein families. *Proteins* **20**, 216–226 (1994).

[64] Lichtarge, O. *et al.* An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358 (1996).

[65] Glaser, F. *et al.* A method for localizing ligand binding pockets in protein structures. *Proteins* **62**, 479–488 (2006).

[66] Marks, D. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLOS ONE* **6**, 1–20 (2011).

[67] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[68] Pauling, L. *et al.* The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **37**, 205–211 (1951).

[69] Lietzan, A. *et al.* Microbial β-glucuronidases drive human periodontal disease etiology. *Sci. Adv.* **9**, eadg3390 (2023).

[70] Pettersen, E. *et al.* UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

[71] Kendrew, J. *et al.* A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181**, 662–666 (1958).

[72] Bernal, J. and Crowfoot, D. X-ray photographs of crystalline pepsin. *Nature* **133**, 794–795 (1934).

[73] Friedrich, W. *et al.* Interferenzerscheinungen bei Röntgenstrahlen. *Ann. Phys.* **346**, 971–988 (1913).

[74] Bragg, W. and Bragg, W. The structure of some crystals as indicated by their diffraction of X-rays. *Proc. R. Soc. Lond. A* **89**, 248–277 (1913).

[75] Berman, H. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

[76] Williamson, M. *et al.* Solution conformation of proteinase inhibitor IIA from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry. *J. Mol. Biol.* **182**, 295–315 (1985).

[77] Wüthrich, K. *et al.* Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *J. Mol. Biol.* **155**, 311–319 (1982).

[78] Wüthrich, K. *et al.* Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton-proton distances. *J. Mol. Biol.* **180**, 715–740 (1984).

[79] Emwas, A. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods Mol. Biol.* **1277**, 161–193 (2015).

[80] Henderson, R. and Unwin, P. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* **257**, 28–32 (1975).

[81] Henderson, R. *et al.* Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **213**, 899–929 (1990).

[82] Yan, C. *et al.* Structure of a yeast activated spliceosome at 3.5 Å resolution. *Science* **353**, 904–911 (2016).

[83] Kosinski, J. *et al.* Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science* **352**, 363–365 (2016).

[84] Bartesaghi, A. *et al.* Structure of β-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 11709–11714 (2014).

[85] Chiu, W. *et al.* Evolution of standardization and dissemination of cryo-EM structures and data jointly by the community, PDB, and EMDB. *J. Biol. Chem.* **296**, 100560 (2021).

[86] Couty, S. *et al.* The discovery of potent ribosomal S6 kinase inhibitors by high-throughput screening and structure-guided drug design. *Oncotarget* **4**, 1647–1661 (2013).

[87] Debye, P. Interferenz von Röntgenstrahlen und Wärmebewegung. *Ann. Phys.* **348**, 49–92 (1913).

[88] Waller, I. Zur Frage der Einwirkung der Wärmebewegung auf die Interferenz von Röntgenstrahlen. *Z. Phys.* **17**, 398–408 (1923).

[89] Sun, Z. *et al.* Utility of B-Factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chem. Rev.* **119**, 1626–1665 (2019).

[90] Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients. *Eur. J. Med. Chem* **19**, 71–78 (1984).

[91] Laguerre, M. *et al.* MLPP: a program for the calculation of molecular lipophilicity potential in proteins. *Pharm. Pharmacol. Commun.* **3**, 217–222 (1997).

[92] Efremov, R. *et al.* Molecular lipophilicity in protein modeling and drug design. *Curr. Med. Chem.* **14**, 393–415 (2007).

[93] Gaillard, P. *et al.* Molecular lipophilicity potential, a tool in 3D QSAR: method and applications. *J. Comput. Aided. Mol. Des.* **8**, 83–96 (1994).

[94] Coulomb, C. Premier Mémoire sur l'Électricité et le Magnétisme. *Hist. Acad. R. Sci.* **1**, 569–577 (1785).

[95] Zhou, H. and Pang, X. Electrostatic interactions in protein structure, folding, binding, and condensation. *Chem. Rev.* **118**, 1691–1741 (2018).

[96] Gorham, R. *et al.* Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization. *Ann. Biomed. Eng.* **39**, 1252–1263 (2011).

[97] Kukic, P. and Nielsen, J. Electrostatics in proteins and protein-ligand complexes. *Future Med. Chem.* **2 4**, 647–66 (2010).

[98] Vascon, F. *et al.* Protein electrostatics: from computational and structural analysis to discovery of functional fingerprints and biotechnological design. *Comput. Struct. Biotechnol. J.* **18**, 1774–1789 (2020).

[99] Lee, B. and Richards, F. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–IN4 (1971).

[100] Shrake, A. and Rupley, J. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79**, 351–371 (1973).

[101] Van der Waals, J. *Over de Continuïteit van den Gas- en Vloeistoftoestand*. Leiden University, Leiden, Netherlands, (1873).

[102] Richards, F. Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys.* **6**, 151–176 (1977).

[103] Connolly, M. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709–713 (1983).

[104] Connolly, M. *J. Appl. Crystallogr.* **16**, 548–558 (1983).

[105] Daberdaku, S. Surface diagram visualization (2025) https://www.cgl.ucsf.edu/chimerax/docs/user/commands/surface-diagram.png.

[106] Rose, G. *et al.* Hydrophobicity of amino acid residues in globular proteins. *Science* **229**, 834–838 (1985).

[107] Miller, S. *et al.* Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656 (1987).

[108] Tien, M.Z. *et al.* Maximum allowed solvent accessibilites of residues in proteins. *PLOS ONE* **8**, e80635 (2013).

[109] Dill, K. Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155 (1990).

[110] Jones, S. and Thornton, J. Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132 (1997).

[111] Goldman, N. *et al.* Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–458 (1998).

[112] Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–637 (1983).

[113] Rees, D. *et al.* Fragment-based lead discovery. *Nat. Rev. Drug Discov.* **3**, 660–672 (2004).

[114] Volkamer, A. *et al.* Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.* **50**, 2041–2452 (2010).

[115] Krivák, R. and Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminform.* **10**, 39 (2018).

[116] Parikh, S. *et al.* Insights into the genetic variations of human cytochrome P450 2C9: structural analysis, characterization and comparison. *Int. J. Mol. Sci.* **22**, (2021).

[117] Gu, Y. *et al.* Crystal structure of human glutathione S-transferase A3-3 and mechanistic implications for its high steroid isomerase activity, *Biochemistry* **43**, 15673–15679 (2004).

[118] Elkins, J. *et al.* Human DYRK2 bound to curcumin (2019) https://doi.org/10.2210/pdb6hdr/pdb.

[119] Coleman, J. *et al.* X-ray structures and mechanism of the human serotonin transporter. *Nature* **532**, 334–339 (2016).

[120] Le Guilloux, V. *et al.* Fpocket: an open source platform for ligand pocket detection. *BMC Bioinform.* **10**, 168 (2009).

[121] Hendlich, M. *et al.* LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **15**, 359–363, 389 (1997).

[122] Laskowski, R. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* **13**, 323–330, 307–308 (1995).

[123] Schmidtke, P. *et al.* fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* **38**, W582–W589 (2010).

[124] Weisel, M. *et al.* PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **1**, 7 (2007).

[125] Brady Jr, G.P. and Stouten, P.F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided. Mol. Des.* **14**, 383–401 (2000).

[126] Liang, J. *et al.* Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **7**, 1884–1897 (1998).

[127] Kleywegt, G.J. and Jones, T.A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D* **50**, 178–185 (1994).

[128] Levitt, D.G. and Banaszak, L.J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* **10**, 229–234 (1992).

[129] An, J. *et al.* Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteom.* **4**, 752–761 (2005).

[130] Ngan, C.H. *et al.* FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics* **28**, 286–287 (2012).

[131] Ghersi, D. and Sanchez, R. EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics* **25**, 3185–3186 (2009).

[132] Laurie, A. and Jackson, R. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **21**, 1908–1916 (2005).

[133] An, J. *et al.* Comprehensive identification of "druggable" protein ligand binding sites. *Genome Inform.* **15**, 31–41 (2004).

[134] Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28**, 849–857 (1985).

[135] Xie, Z.R. and Hwang, M.J. Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. *Bioinformatics* **28**, 1579–1585 (2012).

[136] Pupko, T. *et al.* Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18 Suppl 1**, S71–S77 (2002).

[137] Wass, M. *et al.* 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.* **38**, W469–W473 (2010).

[138] Yang, J. *et al.* Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **29**, 2588–2595 (2013).

[139] Lee, H.S. and Im, W. Ligand binding site detection by local structure alignment and its performance complementarity. *J. Chem. Inf. Model.* **53**, 2462–2470 (2013).

[140] Brylinski, M. and Feinstein, W.P. *e*FindSite: improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *J. Comput. Aided. Mol. Des.* **27**, 551–567 (2013).

[141] Roy, A. *et al.* COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* **40**, W471–W477 (2012).

[142] Gutteridge, A. *et al.* Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.* **330**, 719–734 (2003).

[143] Huang, B. and Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **6**, 19 (2006).

[144] Halgren, T.A. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **49**, 377–389 (2009).

[145] Capra, J. *et al.* Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLOS Comput. Biol.* **5**, e1000585 (2009).

[146] Huang, B. MetaPocket: a meta approach to improve protein ligand binding site prediction. *OMICS* **13**, 325–330 (2009).

[147] Bray, T. *et al.* SitesIdentify: a protein functional site prediction tool. *BMC Bioinform.* **10**, 379 (2009).

[148] Brylinski, M. and Skolnick, J. FINDSITE^{LHM}: a threading-based approach to ligand homology modeling. *PLOS Comput. Biol.* **5**, e1000405 (2009).

[149] Krivák, R. and Hoksza, D. Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features. *J. Cheminform.* **7**, 12 (2015).

[150] Krivák, R. and Hoksza, D. P2RANK: knowledge-based ligand binding site prediction using aggregated local features. *Algorithms for Computational Biology*, 41–52 (2015).

[151] Jimenez, J. *et al.* DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **33**, 3036–3042 (2017).

[152] Jendele, L. *et al.* PrankWeb: a web server for ligand binding site prediction and visualization. *Nucleic Acids Res.* **47**, W345–W349 (2019).

[153] Santana, C. *et al.* GRaSP: a graph-based residue neighborhood strategy to predict binding sites. *Bioinformatics* **36**, i726–i734 (2020).

[154] Kozlovskii, I. and Popov, P. Spatiotemporal identification of druggable binding sites using deep learning. *Commun. Biol.* **3**, 618 (2020).

[155] Stepniewska-Dziubinska, M. *et al.* Improving detection of protein-ligand binding sites with 3D segmentation. *Sci. Rep.* **10**, 5035 (2020).

[156] Kandel, J. *et al.* PUResNet: prediction of protein-ligand binding sites using deep residual neural network. *J. Cheminform.* **13**, 65 (2021).

[157] Mylonas, S. *et al.* DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics* **37**, 1681–1690 (2021).

[158] Yan, X. *et al.* PointSite: a point cloud segmentation tool for identification of protein ligand binding atoms. *J. Chem. Inf. Model.* **62**, 2835–2845 (2022).

[159] Li, P. *et al.* RecurPocket: recurrent Lmser network with gating mechanism for protein binding site detection. *2022 IEEE Int. Conf. Bioinform. Biomed.*, 334–339 (2022).

[160] Aggarwal, R. *et al.* DeepPocket: ligand binding site detection and segmentation using 3D convolutional neural networks. *J. Chem. Inf. Model.* **62**, 5069–5079 (2022).

[161] Jakubec, D. *et al.* PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures. *Nucleic Acids Res.* **50**, W593–W597 (2022).

[162] Abdollahi, N. *et al.* NodeCoder: a graph-based machine learning platform to predict active sites of modeled protein structures. *arXiv* (2023).

[163] Evteev, S. *et al.* SiteRadar: utilizing graph machine learning for precise mapping of protein-ligand-binding sites. *J. Chem. Inf. Model.* **63**, 1124–1132 (2023).

[164] Li, P. *et al.* GLPocket: a multi-scale representation learning approach for protein binding site prediction. *Proc. 32nd Int. Joint Conf. Artif. Intell.*, 4821–4828 (2023).

[165] Zhang, Y. *et al.* EquiPocket: an E(3)-equivariant geometric graph neural network for ligand binding site prediction. *arXiv* (2024).

[166] Liu, Y. *et al.* RefinePocket: an attention-enhanced and mask-guided deep learning approach for protein binding site prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **20**, 3314–3321 (2023).

[167] Smith, Z. *et al.* Graph attention site prediction (GrASP): identifying druggable binding sites using graph neural networks with attention. *J. Chem. Inf. Model.* **64**, 2637–2644 (2024).

[168] Carbery, A. *et al.* Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures. *J. Cheminform.* **16**, 32 (2024).

[169] Sestak, F. *et al.* VN-EGNN: E(3)-equivariant graph neural networks with virtual nodes enhance protein binding site identification. *arXiv* (2024).

[170] Kandel, J. *et al.* PUResNetV2.0: a deep learning model leveraging sparse representation for improved ligand binding site prediction. *J. Cheminform.* **16**, 66 (2024).

[171] Bairoch, A. *et al.* The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159 (2005).

[172] Bairoch, A. and Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).

[173] Berman, H. *et al.* Announcing the worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* **10**, 980 (2003).

[174] PDBe-KB consortium. PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res.* **48**, D344–D353 (2019).

[175] Coulter-Mackie, M. and Rumsby, G. Genetic heterogeneity in primary hyperoxaluria type 1: impact on diagnosis. *Mol. Genet. Metab.* **83**, 38–46 (2004).

[176] Vihinen, M. Individual genetic heterogeneity. *Genes* **13**, (2022).

[177] Darwin, C. *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. John Murray, London, 1859.

[178] Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).

[179] Gluecksohn-Waelsch, S. Lethal genes and analysis of differentiation. *Science* **142**, 1269–1276 (1963).

[180] Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

[181] McLaren, W. *et al.* The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

[182] Kumar, P. *et al.* Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).

[183] Adzhubei, I. *et al.* Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, 7.20.1–7.20.41 (2013).

[184] Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

[185] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

[186] Kasianowicz, J. *et al.* Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13770–13773 (1996).

[187] Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

[188] Bentley, D. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).

[189] Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).

[190] Rothberg, J. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).

[191] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[192] Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

[193] Gong, S. and Blundell, T. Structural and functional restraints on the occurrence of single amino acid variations in human proteins. *PLOS ONE* **5**, 1–12 (2010).

[194] Beer, T. de *et al.* Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 Genomes Project dataset. *PLOS Comput. Biol.* **9**, 1–15 (2013).

[195] David, A. and Sternberg, M. The contribution of missense mutations in core and rim residues of protein–protein interfaces to human disease. *J. Mol. Biol.* **427**, 2886–2898 (2015).

[196] Sivley, R. *et al.* Comprehensive analysis of constraint on the spatial distribution of missense variants in human protein structures. *Am. J. Hum. Genet.* **102**, 415–426 (2018).

[197] Petrovski, S. *et al.* Genic intolerance to functional variation and the interpretation of personal genomes. *PLOS Genet.* **9**, 1–13 (2013).

[198] Gussow, A. *et al.* The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol.* **17**, 9 (2016).

[199] Li, B. *et al.* The 3D mutational constraint on amino acid sites in the human proteome. *Nat. Commun.* **13**, 3273 (2022).

[200] MacGowan, S. *et al.* Human missense variation is constrained by domain structure and highlights functional and pathogenic residues. *bioRxiv*, 127050 (2017).

[201] Fisher, R. The logic of inductive inference. *J. R. Stat. Soc.* **98**, 39–54 (1935).

[202] Hublin, J. *et al.* New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*. *Nature* **546**, 289–292 (2017).

[203] MacGowan, S.A. *et al.* A unified analysis of evolutionary and population constraint in protein domains highlights structural features and pathogenic sites. *Commun. Biol.* **7**, 447 (2024).

[204] Landrum, M. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2013).

[205] Hughes, J. *et al.* Principles of early drug discovery. *Br. J. Pharmacol.* **162**, 1239–1249 (2011).

[206] Cui, W. *et al.* Discovering anti-cancer drugs *via* computational methods. *Front. Pharmacol.* **11**, 733 (2020).

[207] Schenone, M. *et al.* Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.* **9**, 232–240 (2013).

[208] Paul, S. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203–214 (2010).

[209] Emmerich, C. *et al.* Improving target assessment in biomedical research: the GOT-IT recommendations. *Nat. Rev. Drug Discov.* **20**, 64–81 (2021).

[210] Sinha, S. and Vohora, D. Drug discovery and development: an overview. *Pharm. Med. Transl. Clin. Res.*, 19–32 (2018).

[211] Shou, W. Current status and future directions of high-throughput ADME screening in drug discovery. *J. Pharm. Anal.* **10**, 201–208 (2020).

[212] Shegokar, R. Preclinical testing – understanding the basics first. *Drug Deliv. Aspects*, 19–32 (2020).

[213] Umscheid, C. *et al.* Key concepts of clinical trials: a narrative review. *Postgrad. Med.* **123**, 194–204 (2011).

[214] Suvarna, V. Phase IV of drug development. *Perspect. Clin. Res.* **1**, 57–60 (2010).

[215] Murray, C. and Rees, D. The rise of fragment-based drug discovery. *Nat. Chem.* **1**, 187–192 (2009).

[216] Congreve, M. *et al.* A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today* **8**, 876–877 (2003).

[217] Schiebel, J. *et al.* Six biophysical screening methods miss a large proportion of crystallographically discovered fragment hits: a case study. *ACS Chem. Biol.* **11**, 1693–1701 (2016).

[218] Patel, D. *et al.* Advantages of crystallographic fragment screening: functional and mechanistic insights from a powerful platform for efficient drug discovery. *Prog. Biophys. Mol. Biol.* **116**, 92–100 (2014).

[219] Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

[220] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699–2699 (2018).

[221] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2022).

[222] Varadi, M. *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

[223] Guex, N. *et al.* Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PDBViewer: a historical perspective. *Electrophoresis* **30**, S162–S173 (2009).

[224] Bienert, S. *et al.* The SWISS-MODEL Repository – new features and functionality. *Nucleic Acids Res.* **45**, D313–D319 (2016).

[225] Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).

[226] wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).

[227] Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res.* **39**, D28–D31 (2010).

[228] Yang, J. *et al.* BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **41**, D1096–D1103 (2013).

[229] Lexa, K. and Carlson, H. Full protein flexibility is essential for proper hot-spot mapping. *J. Am. Chem. Soc.* **133**, 200–202 (2011).

[230] Ghanakota, P. *et al.* Large-scale validation of mixed-solvent simulations to assess hotspots at protein-protein interaction interfaces. *J. Chem. Inf. Model.* **58**, 784–793 (2018).

[231] Alvarez-Garcia, D. and Barril, X. Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites. *J. Med. Chem.* **57**, 8530–8539 (2014).

[232] Faller, C. *et al.* Site identification by ligand competitive saturation (SILCS) simulations for fragment-based drug design. *Methods Mol. Biol.* **1289**, 75–87 (2015).

[233] Shin, J. and Cho, D. PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res.* **33**, D238–D241 (2005).

[234] Kozakov, D. *et al.* Optimal clustering for detecting near-native conformations in protein docking. *Biophys. J.* **89**, 867–875 (2005).

[235] McGreig, J. *et al.* 3DLigandSite: structure-based prediction of protein-ligand binding sites. *Nucleic Acids Res.* **5**, W13–W20 (2022).

[236] Pearce, N. *et al.* A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. *Nat. Commun.* **8**, 15123 (2017).

[237] Krissinel, E. and Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).

[238] MacGowan, S.A. *et al.* The Dundee Resource for Sequence Analysis and Structure Prediction. *Protein Sci.* **29**, 277–297 (2020).

[239] Jubb, H. *et al.* Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.* **429**, 365–371 (2017).

[240] Keedy, D.A. *et al.* An expanded allosteric network in PTP1B by multitemperature crystallography, fragment screening, and covalent tethering. *eLife* **7**, (2018).

[241] Barton, G.J. OC – a cluster analysis program. (1993, 2002, 2004) https://www.compbio.dundee.ac.uk/downloads/oc/.

[242] Russell, R.B. and Barton, G.J. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* **14**, 309–323 (1992).

[243] Pettersen, E. *et al.* UCSF Chimera – a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

[244] Boutet, E. *et al.* UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol. Biol.* **1374**, 23–54 (2016).

[245] Eddy, S. Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 114–120 (1995).

[246] Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

[247] Szumilas, M. Explaining odds ratios. *J. Am. Acad. Child Adolesc. Psychiatry* **19**, 227–229 (2010).

[248] Mann, H. and Whitney, D. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).

[249] Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

[250] Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).

[251] Sørensen, T. *A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on Danish commons*. Munksgaard in Komm., 1948.

[252] Sokal, R. and Michener, C. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* **38**, 1409–1438 (1958).

[253] Ward, J. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).

[254] Mead, A. Review of the development of multidimensional scaling methods. *J. R. Stat. Soc. D* **41**, 27–39 (1992).

[255] Rousseeuw, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

[256] Thorndike, R. Who belongs in the family? *Psychometrika* **18**, 267–276 (1953).

[257] Caliński, T. and Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **3**, 1–27 (1974).

[258] Davies, D. and Bouldin, D. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**, 224–227 (1979).

[259] Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

[260] Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* **2**, 303–314 (1989).

[261] Chollet, F. *et al.* Keras. Simple. Flexible. Powerful (2025) https://keras.io/.

[262] Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., B: Stat. Methodol.* **57**, 289–300 (1995).

[263] Cuff, J. and Barton, G. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40**, 502–511 (2000).

[264] Wilson, E. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **22**, 209–212 (1927).

[265] Rand, W. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).

[266] Hubert, L. and Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).

[267] Vinh, N. *et al.* Information theoretic measures for clusterings comparison: is a correction for chance necessary? *Proc. 26th Annu. Int. Conf. Mach. Learn.* **1**, 1073–1080 (2009).

[268] Vinh, N. *et al.* Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11**, 2837–2854 (2010).

[269] Steinley, D. *et al.* The variance of the adjusted Rand index. *Psychol. Methods* **21**, 261–272 (2016).

[270] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

[271] Nightingale, A. *et al.* The Proteins API: accessing key integrated protein and genome information. *Nucleic Acids Res.* **45**, W539–W544 (2017).

[272] Harris, C. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

[273] McKinney, W. Data structures for statistical computing in Python. *9th Python Sci. Conf.* **1**, 56–61 (2010).

[274] The Pandas Development Team. pandas-dev/pandas: Pandas v1.5.0rc0 (2022) https://doi.org/10.5281/zenodo.7018966.

[275] Hunter, J. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

[276] Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

[277] Utgés, J.S. bartongroup/FRAGSYS: second release v2.0.0 (2024) https://doi.org/10.5281/zenodo.10606595.

[278] Wollenhaupt, J. *et al.* F2X-iniversal and F2X-entry: structurally diverse compound libraries for crystallographic fragment screening. *Structure* **28**, 694–706.e5 (2020).

[279] Chang, C. *et al.* Crystal structures of MAP kinase p38 complexed to the docking sites on its nuclear substrate MEF2A and activator MKK3b. *Mol. Cell* **9**, 1241–1249 (2002).

[280] Rodgers, J. and Nicewander, W. Thirteen ways to look at the correlation coefficient. *Am. Stat.* **42**, 59–66 (1988).

[281] Bowley, A. The standard deviation of the correlation coefficient. *J. Am. Stat. Assoc.* **23**, 31–34 (1928).

[282] Dubianok, Y. *et al.* PanDDA analysis group deposition of models with modelled events (e.g. bound ligands) – crystal structure of NUDT5 in complex with Z56983806 (2018) https://doi.org/10.2210/pdb5qjl/pdb.

[283] Nowak, R. *et al.* Crystal structure of the catalytic domain of human JARID1B in complex with 2,5-dichloro-N-(pyridin-3-yl)thiophene-3-carboxamide (N08137b) (ligand modelled based on PanDDA event map, SGC - Diamond I04-1 fragment screening) (2016) https://doi.org/10.2210/pdb5fz0/pdb.

[284] MacKinnon, S. *et al.* PanDDA analysis group deposition of models with modelled events (e.g. bound ligands) – crystal structure of HAO1 in complex with FMOPL000388a (2018) https://doi.org/10.2210/pdb5qib/pdb.

[285] Petrick, J. *et al.* PanDDA analysis group deposition – crystal structure of *T. cruzi* FPPS in complex with FMOPL000500a (2020) https://doi.org/10.2210/pdb5qpm/pdb.

[286] Snee, M. *et al.* PanDDA analysis group deposition – crystal structure of JMJD1B in complex with XS039080d (2020) https://doi.org/10.2210/pdb5ran/pdb.

[287] Pinkas, D. *et al.* PanDDA analysis group deposition of models with modelled events (e.g. bound ligands) – crystal structure of human FAM83B in complex with FMOPL000622a (2018) https://doi.org/10.2210/pdb5qhn/pdb.

[288] Schuller, M. *et al.* Fragment binding to the NSP3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking. *Sci. Adv.* **7**, eabf8711 (2021).

[289] Nelson, E. *et al.* PanDDA analysis group deposition – crystal structure of DCP2 (NUDT20) in complex with Z100435060 (2019) https://doi.org/10.2210/pdb5qp9/pdb.

[290] Krojer, T. *et al.* PanDDA analysis group deposition of models with modelled events (e.g. bound ligands) – crystal structure of NUDT7 in complex with FMOPL000710a (2019) https://doi.org/10.2210/pdb5qp9/pdb.

[291] Nichols, C. *et al.* Mining the PDB for tractable cases where X-ray crystallography combined with fragment screens can beused to systematically design protein-protein inhibitors: two test cases illustrated by IL1β-IL1R and p38α–TAB1 complexes. *J. Med. Chem.* **63**, 7559–7568 (2020).

[292] Newman, J. *et al.* PanDDA analysis group deposition – crystal structure of human Brachyury G177D variant in complex with Z2856434778 (2019) https://doi.org/10.2210/pdb5qsa/pdb.

[293] Bradshaw, W. *et al.* Regulation of inositol 5-phosphatase activity by the C2 domain of SHIP1 and SHIP2. *Structure* **32**, 453–466.e6 (2024).

[294] Diaz-Saez, L. *et al.* PanDDA analysis group deposition – crystal structure of human NUDT22 in complex with N13369a (2020) https://doi.org/10.2210/pdb5r55/pdb.

[295] Grosjean, H. *et al.* PanDDA analysis group deposition – crystal structure of PHIP in complex with Z57261895 (2020) https://doi.org/10.2210/pdb5rju/pdb.

[296] Godoy, A. *et al.* PanDDA analysis group deposition – crystal structure of Zika virus NS3 helicase in complex with Z198194396 (2020) https://doi.org/10.2210/pdb5rhi/pdb.

[297] Newman, J. *et al.* PanDDA analysis group deposition – crystal structure of DCLR-E1A in complex with FMOPL000150a (2018) https://doi.org/10.2210/pdb5q1z/pdb.

[298] Russell, R. and Barton, G. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J. Mol. Biol.* **244**, 332–350 (1994).

[299] Luo, D. *et al.* The flavivirus NS2B-NS3 protease-helicase as a target for antiviral drug development. *Antivir. Res.* **118**, 148–158 (2015).

[300] Tian, H. *et al.* The crystal structure of Zika virus helicase: basis for antiviral drug design. *Protein Cell* **7**, 450–454 (2016).

[301] Mottin, M. *et al.* Molecular dynamics simulations of Zika virus NS3 helicase: insights into RNA binding site activity. *Biochem. Biophys. Res. Commun.* **492**, 643–651 (2017).

[302] Raubenolt, B. *et al.* Molecular dynamics simulations of allosteric motions and competitive inhibition of the Zika virus helicase. *J. Mol. Graph. Model.* **108**, 108001 (2021).

[303] Durgam, L. and Guruprasad, L. Molecular mechanism of ATP and RNA binding to Zika virus NS3 helicase and identification of repurposed drugs using molecular dynamics simulations. *J. Biomol. Struct. Dyn.* **40**, 12642–12659 (2022).

[304] Godoy, A. *et al.* PanDDA analysis group deposition – crystal structure of Zika virus NS3 helicase in complex with Z235341991 (2020) https://doi.org/10.2210/pdb5rhg/pdb.

[305] Naqvi, A. *et al.* Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: structural genomics approach. *Biochim. Biophys. Acta Mol. Basis. Dis.* **1866**, 165878 (2020).

[306] Yue, K. *et al.* The stalk domain of SARS-CoV-2 NSP13 is essential for its helicase activity. *Biochem. Biophys. Res. Commun.* **601**, 129–136 (2022).

[307] Shu, T. *et al.* SARS-Coronavirus-2 NSP13 possesses NTPase and RNA helicase activities that can be inhibited by bismuth salts. *Virol. Sin.* **35**, 321–329 (2020).

[308] Zeng, J. *et al.* Identifying SARS-CoV-2 antiviral compounds by screening for small molecule inhibitors of NSP13 helicase. *Biochem. J.* **478**, 2405–2423 (2021).

[309] Romeo, I. *et al.* Targeting SARS-CoV-2 NSP13 helicase and assessment of druggability pockets: identification of two potent inhibitors by a multi-site *in silico* drug repurposing approach. *Molecules* **27**, (2022).

[310] Ricci, F. *et al. In silico* insights towards the identification of SARS-CoV-2 NSP13 helicase druggable pockets. *Biomolecules* **12**, (2022).

[311] Newman, J. *et al.* Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nat. Commun.* **12**, 4848 (2021).

[312] Yan, L. *et al.* Architecture of a SARS-CoV-2 mini replication and transcription complex. *Nat. Commun.* **11**, 5874 (2020).

[313] Bhattacharyya, S. *et al.* Tenascin-C in fibrosis in multiple organs: translational implications. *Semin. Cell Dev. Biol.* **128**, 130–136 (2022).

[314] Zuliani-Alvarez, L. and Piccinini, A. A virological view of tenascin-C in infection. *Am. J. Physiol. Cell Physiol.* **324**, C1–C9 (2023).

[315] Wang, Y. *et al.* Tenascin-C: a key regulator in angiogenesis during wound healing. *Biomolecules* **12**, (2022).

[316] Khomtchouk, B. *et al.* Targeting the cytoskeleton and extracellular matrix in cardiovascular disease drug discovery. *Expert Opin. Drug Discov.* **17**, 443–460 (2022).

[317] Lepucki, A. *et al.* The role of extracellular matrix proteins in breast cancer. *J. Clin. Med.* **11**, (2022).

[318] Coker, J. *et al.* PanDDA analysis group deposition – crystal structure of fibrinogen-like globe domain of huamn tenascin-C in complex with Z19735067 (2020) https://doi.org/10.2210/pdb5r60/pdb.

[319] Yee, V. *et al.* Crystal structure of a 30 kDa C-terminal fragment from the gamma chain of human fibrinogen. *Structure* **5**, 125–138 (1997).

[320] Akhtar, M. *et al.* Mechanism and stereochemistry of enzymic reactions involved in porphyrin biosynthesis. *Phil. Trans. R. Soc. Lond. B* **273**, 117–136 (1976).

[321] Munakata, H. *et al.* Purification and structure of rat erythroid-specific δ-aminolevulinate synthase. *J. Biochem.* **114**, 103–111 (1993).

[322] Srivastava, G. *et al.* Regulation of 5-aminolevulinate synthase mRNA in different rat tissues. *J. Biol. Chem.* **263**, 5202–5209 (1988).

[323] Bailey, H.J. *et al.* Human aminolevulinate synthase structure reveals a eukaryotic-specific autoinhibitory loop regulating substrate binding and product release. *Nat. Commun.* **11**, 2813 (2020).

[324] Whatley, S. *et al.* C-terminal deletions in the *ALAS2* gene lead to gain of function and cause X-linked dominant protoporphyria without anemia or iron overload. *Am. J. Hum. Genet.* **83**, 408–414 (2008).

[325] Ducamp, S. *et al.* Sideroblastic anemia: molecular analysis of the *ALAS2* gene in a series of 29 probands and functional studies of 10 missense mutations. *Hum. Mutat.* **32**, 590–597 (2011).

[326] Bezerra, G. *et al.* PanDDA analysis group deposition – crystal structure of human ALAS2A in complex with Z730649594 (2019) https://doi.org/10.2210/pdb5qr0/pdb.

[327] Furuyama, K. and Sassa, S. Interaction between succinyl CoA synthetase and the heme-biosynthetic enzyme ALAS-E is disrupted in sideroblastic anemia. *J. Clin. Investig.* **105**, 757–764 (2000).

[328] Douangamath, A. *et al.* Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nat. Commun.* **11**, 5047 (2020).

[329] DasGupta, D. *et al.* Computational identification of possible allosteric sites and modulators of the SARS-CoV-2 main protease. *J. Chem. Inf. Model.* **62**, 618–626 (2022).

[330] Santana, C. *et al.* GRaSP-web: a machine learning strategy to predict binding sites based on residue neighborhood graphs. *Nucleic Acids Res.* **50**, W392–W397 (2022).

[331] Bekes, M. *et al.* PROTAC targeted protein degraders: the past is prologue. *Nat. Rev. Drug Discov.* **21**, 181–200 (2022).

[332] Siriwardena, S. *et al.* Phosphorylation-inducing chimeric small molecules. *J. Am. Chem. Soc.* **142**, 14052–14057 (2020).

[333] Simpson, L. *et al.* An affinity-directed phosphatase, AdPhosphatase, system for targeted protein dephosphorylation. *Cell Chem. Biol.* **30**, 188–202 e6 (2023).

[334] Heitel, P. Emerging TACnology: heterobifunctional small molecule inducers of targeted posttranslational protein modifications. *Molecules* **28**, (2023).

[335] Peng, Y. *et al.* Targeted protein posttranslational modifications by chemically induced proximity for cancer therapy. *J. Biol. Chem.* **299**, 104572 (2023).

[336] Han, B.G. *et al.* Protein 4.1R core domain structure and insights into regulation of cytoskeletal organization. *Nat. Struct. Biol.* **7**, 871–875 (2000).

[337] Munzker, L. *et al.* Fragment-based discovery of non-bisphosphonate binders of *Trypanosoma brucei* farnesyl pyrophosphate synthase. *ChemBioChem* **21**, 3096–3111 (2020).

[338] Gabelli, S.B. *et al.* Structure and mechanism of the farnesyl diphosphate synthase from *Trypanosoma cruzi*: implications for drug design. *Proteins* **62**, 80–88 (2006).

[339] Cui, D.S. *et al.* Leveraging reciprocity to identify and characterize unknown allosteric sites in protein tyrosine phosphatases. *J. Mol. Biol.* **429**, 2360–2372 (2017).

[340] Francis, D.M. *et al.* The differential regulation of p38α by the neuronal kinase interaction motif protein tyrosine phosphatases, a detailed molecular study. *Structure* **21**, 1612–23 (2013).

[341] Pearce, N.M. *et al.* Partial-occupancy binders identified by the Pan-Dataset Density Analysis method offer new chemical opportunities and reveal cryptic binding sites. *Struct. Dyn.* **4**, 032104 (2017).

[342] Utgés, J. and Barton, G. Comparative evaluation of methods for the prediction of protein-ligand binding sites. *J. Cheminform.* **16**, 126 (2024).

[343] Utgés, J.S. and MacGowan, S.A. LIGYSIS-web (2024) https://www.compbio.dundee.ac.uk/ligysis/.

[344] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2020).

[345] Armstrong, D. *et al.* PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res* **48**, D335–D343 (2020).

[346] Ellaway, J. *et al.* Identifying protein conformational states in the Protein Data Bank: toward unlocking the potential of integrative dynamics studies. *Struct. Dyn.* **11**, 034701 (2024).

[347] PDBe Team. PDBe GRAPH API superposition endpoint (2024) https://www.ebi.ac.uk/pdbe/graph-api/uniprot/superposition/.

[348] PDBe-KB Team. PDBe-KB superposition FTP site (2024) https://ftp.ebi.ac.uk/pub/databases/pdbe-kb/superposition/.

[349] PDBe-KB Consortium. PDBe-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res.* **50**, D534–D542 (2022).

[350] Cock, P. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

[351] Velankar, S. *et al.* SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* **41**, D483–D489 (2012).

[352] Dana, J. *et al.* SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* **47**, D482–D489 (2018).

[353] Utgés, J. *et al.* Classification of likely functional class for ligand binding sites identified from fragment screening. *Commun. Biol.* **7**, 320 (2024).

[354] Golebiowski, A. *et al.* Synthesis of quaternary α-amino acid-based arginase inhibitors via the Ugi reaction. *Bioorg. Med. Chem. Lett.* **23**, 4837–4841 (2013).

[355] Grinberg, M. *Flask web development: developing web applications with Python*. O'Reilly Media, Inc., 2018.

[356] Pallets. Jinja: a fast and modern templating engine for Python (2024) https://palletsprojects.com/projects/jinja/.

[357] Bootstrap Team. Build fast, responsive sites with Bootstrap (2025) https://getbootstrap.com/.

[358] Duckett, J. *HTML & CSS: design and build websites*. Wiley Publishing, 2011.

[359] Flanagan, D. *JavaScript: the definitive guide*. 2006.

[360] Rego, N. and Koes, D. 3Dmol.js: molecular visualization with WebGL. *Bioinformatics* **31**, 1322–1324 (2014).

[361] Seshadri, K. *et al.* The 3Dmol.js learning environment: a classroom response system for 3D chemical structures. *J. Chem. Educ.* **97**, 3872–3876 (2020).

[362] Chart.js Team. Chart.js: simple yet flexible JavaScript charting for the modern web (2024) https://www.chartjs.org/.

[363] jQuery Team. jQuery: write less, do more (2025) https://jquery.com/.

[364] Garrett, J. Ajax: a new approach to web applications. *Adaptive Path* **1**, (2005).

[365] Warowny, M. *et al.* Slivka and Slivka-bio: a lightweight, unopinionated framework for executables as web services, and its application to bioinformatics. *F1000Res.* **10**, 707 (poster) (2021).

[366] Warowny, M. Slivka-bio: convenient remote access to bioinformatics tools (2025) https://www.compbio.dundee.ac.uk/slivka/.

[367] Utgés, J.S. and MacGowan, S.A. bartongroup/LIGYSIS-web: pre-release v1.0.0 (2025) https://doi.org/10.5281/zenodo.14178405.

[368] Utgés, J.S. bartongroup/LIGYSIS: first release v1.0.0 (2025) https://doi.org/10.5281/zenodo.14178252.

[369] Utgés, J.S. bartongroup/LIGYSIS-custom: first release v1.0.0 (2025) https://doi.org/10.5281/zenodo.14178216.

[370] Breitenlechner, C. *et al.* Structure-based optimization of novel azepane derivatives as PKB inhibitors. *J. Med. Chem.* **47**, 1375–1390 (2004).

[371] Gaßel, M. *et al.* Mutants of Protein kinase A that nimic the ATP-binding site of Protein kinase B (AKT). *J. Mol. Biol.* **329**, 1021–1034 (2003).

[372] Jmol Development Team. Jmol: an open-source Java viewer for chemical structures in 3D. (2025) https://jmol.sourceforge.net/.

[373] Choudhary, P. *et al.* PDBe tools for an in-depth analysis of small molecules in the Protein Data Bank. *bioRxiv* (2024).

[374] Breitenlechner, C. *et al.* Protein kinase A in complex with Rho-kinase inhibitors Y-27632, fasudil, and H-1152P: structural basis of selectivity. *Structure* **11**, 1595–1607 (2003).

[375] Liu, S. *et al.* G Protein-Coupled Receptors: a century of research and discovery. *Circ. Res.* **135**, 174–197 (2024).

[376] Latorraca, N. *et al.* GPCR dynamics: structures in motion. *Chem. Rev.* **117**, 139–155 (2017).

[377] Hauser, A. *et al.* Trends in GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discov.* **16**, 829–842 (2017).

[378] Insel, P. *et al.* GPCRomics: an approach to discover GPCR drug targets. *Trends Pharmacol. Sci.* **40**, 378–387 (2019).

[379] Zhang, M. *et al.* G protein-coupled receptors (GPCRs): advances in structures, mechanisms and drug discovery. *Signal Transduct. Target. Ther.* **9**, 88 (2024).

[380] Zarzycka, B. *et al.* Harnessing ion-binding sites for GPCR pharmacology. *Pharmacol. Rev.* **71**, 571–595 (2019).

[381] Mannes, M. *et al.* Wandering beyond small molecules: peptides as allosteric protein modulators. *Trends Pharmacol. Sci.* **43**, 406–423 (2022).

[382] Cheng, L. *et al.* Structure, function and drug discovery of GPCR signaling. *Mol. Biomed.* **4**, 46 (2023).

[383] Yeagle, P. *et al.* Studies on the structure of the G-protein-coupled receptor rhodopsin including the putative G-protein binding site in unactivated and activated forms. *Biochemistry* **40**, 11932–11937 (2001).

[384] Liu, S. *et al.* Differential activation mechanisms of lipid GPCRs by lysophosphatidic acid and sphingosine 1-phosphate. *Nat. Commun.* **13**, 731 (2022).

[385] Moukhametzianov, R. *et al.* Two distinct conformations of helix 6 observed in antagonist-bound structures of a β_1-adrenergic receptor. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 8228–8232 (2011).

[386] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, (2021).

[387] Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. *Proc. 39th Int. Conf. Mach. Learn.* **162**, 8946–8970 (2022).

[388] Ke, G. *et al.* LightGBM: a highly efficient gradient boosting decision tree. *31st Conf. Neural Inf. Process. Syst.* **30**, (2017).

[389] Schrödinger, LLC The PyMOL Molecular Graphics System (2025) https://pymol.org/.

[390] Ester, M. *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. Second Int. Conf. Knowl. Discov. Data Min.* **1**, 226–231 (1996).

[391] Capra, J.A. and Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).

[392] Jones, J.E. and Chapman, S. On the determination of molecular fields – from the variation of the viscosity of a gas with temperature. *Proc. R. Soc. Lond. A* **106**, 441–462 (1924).

[393] Schmidtke, P. *et al.* Large-scale comparison of four binding site detection algorithms. *J. Chem. Inf. Model.* **50**, 2191–2200 (2010).

[394] Chen, K. *et al.* A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure* **19**, 613–621 (2011).

[395] Westbrook, J. *et al.* The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics* **31**, 1274–1278 (2015).

[396] Campanacci, V. *et al.* Insight into microtubule nucleation from tubulin-capping proteins. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 9859–9864 (2019).

[397] Paul, N. *et al.* Recovering the true targets of specific ligands by virtual screening of the Protein Data Bank. *Proteins* **54**, 671–680 (2004).

[398] Kellenberger, E. *et al.* sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model.* **46**, 717–727 (2006).

[399] Meslamani, J. *et al.* sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics* **27**, 1324–1326 (2011).

[400] Desaphy, J. *et al.* sc-PDB: a 3D-database of ligandable binding sites – 10 years on. *Nucleic Acids Res.* **43**, D399–D404 (2015).

[401] Hu, L. *et al.* Binding MOAD (Mother Of All Databases). *Proteins* **60**, 333–40 (2005).

[402] Benson, M. *et al.* Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res.* **36**, D674–D678 (2008).

[403] Ahmed, A. *et al.* Recent improvements to binding MOAD: a resource for protein-ligand binding affinities and structures. *Nucleic Acids Res.* **43**, D465–D469 (2015).

[404] Smith, R. *et al.* Updates to binding MOAD (Mother of All Databases): polypharmacology tools and their utility in drug repurposing. *J. Mol. Biol.* **431**, 2423–2433 (2019).

[405] Hubbard, T. *et al.* SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **25**, 236–239 (1997).

[406] Hubbard, T. *et al.* SCOP, structural classification of proteins database: applications to evaluation of the effectiveness of sequence alignment methods and statistics of protein structural data. *Acta Crystallogr. D* **54**, 1147–1154 (1998).

[407] Lo Conte, L. *et al.* SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **28**, 257–259 (2000).

[408] Hartshorn, M. *et al.* Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **50**, 726–741 (2007).

[409] Zhang, Z. *et al.* Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* **27**, 2083–2088 (2011).

[410] Wang, R. *et al.* The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **47**, 2977–2980 (2004).

[411] Wang, R. *et al.* The PDBbind database: methodologies and updates. *J. Med. Chem.* **48**, 4111–4119 (2005).

[412] Cheng, T. *et al.* Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **49**, 1079–1093 (2009).

[413] Li, Y. *et al.* Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *J. Chem. Inf. Model.* **54**, 1700–1716 (2014).

[414] Liu, Z. *et al.* PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).

[415] Liu, Z. *et al.* Forging the basis for developing protein-ligand interaction scoring functions. *Acc. Chem. Res.* **50**, 302–309 (2017).

[416] Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).

[417] Fixman, M. Radius of gyration of polymer chains. *J. Chem. Phys.* **36**, 306–310 (1962).

[418] Nicoludis, J. *et al.* Interaction specificity of clustered protocadherins inferred from sequence covariation and structural analysis. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 17825–17830 (2019).

[419] Hong, S. *et al.* Methyl-dependent and spatial-specific DNA recognition by the orthologous transcription factors human AP-1 and Epstein-Barr virus Zta. *Nucleic Acids Res.* **45**, 2503–2515 (2017).

[420] Nar, H. *et al.* Structure of human FAPalpha in complex with linagliptin (2021) https://doi.org/10.2210/pdb6y0f/pdb.

[421] Lee, S. *et al.* Structural insights into mitochondrial calcium uniporter regulation by divalent cations. *Cell Chem. Biol.* **23**, 1157–1169 (2016).

[422] Chen, C. and Makhatadze, G. ProteinVolume: calculating molecular van der Waals and void volumes in proteins. *BMC Bioinform.* **16**, 101 (2015).

[423] Euclid. *Elements*. Greek text by J.L. Heiberg. Modern English translation by Richard Fitzpatrick. 300 BCE.

[424] Jiang, M. *et al.* Human IFT-A complex structures provide molecular insights into ciliary transport. *Cell Res.* **33**, 288–298 (2023).

[425] Kollman J.M.and Pandi, L. *et al.* Crystal structure of human fibrinogen. *Biochemistry* **48**, 3877–3886 (2009).

[426] Sikora, M. *et al.* Desmosome architecture derived from molecular dynamics simulations and cryo-electron tomography. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 27132–27140 (2020).

[427] Harrison, O. *et al.* Structural basis of adhesive binding by desmocollins and desmogleins. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7160–7165 (2016).

[428] Pulavarti, S. *et al.* Solution NMR Structure of Zinc finger protein Eos from *Homo sapiens*, Northeast Structural Genomics Consortium (NESG) target HR7992A (2013) https://doi.org/10.2210/pdb2ma7/pdb.

[429] Shigdel, U. *et al.* Genomic discovery of an evolutionarily programmed modality for small-molecule targeting of an intractable protein surface. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 17195–17203 (2020).

[430] Takahashi, T. *et al.* Structural insights into two distinct binding modules for Lys63-linked polyubiquitin chains in RNF168. *Nat. Commun.* **9**, 170 (2018).

[431] Kümmel, D. *et al.* Complexin cross-links prefusion SNAREs into a zigzag array. *Nat. Struct. Mol. Biol.* **18**, 927–933 (2011).

[432] Ren, L. *et al.* Structural insight into substrate specificity of human intestinal maltase-glucoamylase. *Protein Cell* **2**, 827–836 (2011).

[433] Dupuy, J. *et al.* Crystal structure of human iron regulatory protein 1 as cytosolic aconitase. *Structure* **14**, 129–139 (2006).

[434] Augustin, P. *et al.* Structure and biochemical properties of recombinant human dimethylglycine dehydrogenase and comparison to the disease-related H109R variant. *FEBS J.* **283**, 3587–3603 (2016).

[435] Yoshida, T. *et al.* Fused bicyclic heteroarylpiperazine-substituted l-prolylthiazolidines as highly potent DPP-4 inhibitors lacking the electrophilic nitrile group. *Bioorg. Med. Chem.* **20**, 5033–5041 (2012).

[436] Ferguson, K. *et al.* Structural basis for discrimination of 3-phosphoinositides by pleckstrin homology domains. *Mol. Cell* **6**, 373–384 (2000).

[437] Kozlov G.and Banville, D. *et al.* Solution structure of the PDZ2 domain from cytosolic human phosphatase hPTP1E complexed with a peptide reveals contribution of the β2–β3 loop to PDZ domain–ligand interactions. *J. Mol. Biol.* **320**, 813–820 (2002).

[438] Enmon, J. *et al.* Solution structure of Eps15's third EH domain reveals coincident Phe−Trp and Asn−Pro−Phe binding sites. *Biochemistry* **39**, 4309–4319 (2000).

[439] Beer, T. de *et al.* Structure and Asn-Pro-Phe binding pocket of the Eps15 homology domain. *Science* **281**, 1357–1360 (1998).

[440] Jubb, H. pdbtools. GitHub repository (2019) https://github.com/harryjubb/pdbtools/tree/master.

[441] Jendele, L. *et al.* PrankWeb: ligand binding site prediction (2017) https://prankweb.cz/.

[442] Comajuncosa-Creus, A. *et al.* Comprehensive detection and characterization of human druggable pockets through binding site descriptors. *Nat. Commun.* **15**, 7917 (2024).

[443] Jaccard, P. Distribution de la florine alpine dans la Bassin de Dranses et dans quelques regiones voisines. *Bull. Soc. Vaud. Sci. Nat.* **37**, 241–272 (1901).

[444] Jaccard, P. The distribution of the flora in the alpine zone. *New Phytol.* **11**, 37–50 (1912).

[445] Durrant, J. *et al.* POVME: an algorithm for measuring binding-pocket volumes. *J. Mol. Graph. Model.* **29**, 773–776 (2011).

[446] Durrant, J. *et al.* POVME 2.0: an enhanced tool for determining pocket shape and volume characteristics. *J. Chem. Theory Comput.* **10**, 5047–5056 (2014).

[447] Wagner, J. *et al.* POVME 3.0: software for mapping binding pocket flexibility. *J. Chem. Theory Comput.* **13**, 4584–4592 (2017).

[448] Jordan, S. and Chmait, S. Human GKRP bound to AMG2882 and sorbitol-6-phosphate (2015) https://doi.org/10.2210/pdb4px2/pdb.

[449] Newman, J. *et al.* Structure of the helicase domain of DNA polymerase Theta reveals a possible role in the microhomology-mediated end-joining pathway. *Structure* **23**, 2319–2330 (2015).

[450] Karpusas, M. *et al.* The crystal structure of human interferon β at 2.2-Å resolution. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 11813–11818 (1997).

[451] Noland, C. *et al.* Structure-guided unlocking of Na_X reveals a non-selective tetrodotoxin-sensitive cation channel. *Nat. Commun.* **13**, 1416 (2022).

[452] Qin, J. *et al.* Molecular mechanism of agonism and inverse agonism in ghrelin receptor. *Nat. Commun.* **13**, 300 (2022).

[453] Jansson, A. *et al.* Crystal structure of peroxisomal trans 2-enoyl CoA reductase (2005) https://doi.org/10.2210/pdb1yxm/pdb.

[454] Chase, D. *et al.* Development of a small molecule downmodulator for the transcription factor Brachyury. *Angew. Chem. Int. Ed.* **63**, e20231–6496 (2024).

[455] Padyana, A. *et al.* Structure and inhibition mechanism of the catalytic domain of human squalene epoxidase. *Nat. Commun.* **10**, 97 (2019).

[456] Xu, P. *et al.* Structural genomics of the human dopamine receptor system. *Cell Res.* **33**, 604–616 (2023).

[457] Kakuda, S. *et al.* Structural basis for acceptor substrate recognition of a human glucuronyltransferase, GlcAT-P, an enzyme critical in the biosynthesis of the carbohydrate epitope HNK-1. *J. Biol. Chem.* **279**, 22693–22703 (2004).

[458] Bjerregaard-Andersen, K. *et al.* Malonate in the nucleotide-binding site traps human AKAP18γ/δ in a novel conformational state. *Acta Crystallogr. F* **72**, 591–597 (2016).

[459] Han, L. *et al.* Structure and mechanism of the SGLT family of glucose transporters. *Nature* **601**, 274–279 (2022).

[460] Che, T. *et al.* Nanobody-enabled monitoring of kappa opioid receptor states. *Nat. Commun* **11**, 1145 (2020).

[461] Kopec, J. *et al.* Crystal structure of human aminoadipate semialdehyde synthase, saccharopine dehydrogenase domain (in NAD+ bound form) (2017) https://doi.org/10.2210/pdb5l78/pdb.

[462] Duan, J. *et al.* Hormone- and antibody-mediated activation of the thyrotropin receptor. *Nature* **609**, 854–859 (2022).

[463] Elkins, J. *et al.* Structure of dystrophia myotonica protein kinase. *Protein Sci.* **18**, 782–791 (2009).

[464] Jurkowska, R. *et al.* H3K14ac is linked to methylation of H3K9 by the triple Tudor domain of SETDB1. *Nat. Commun.* **8**, 2057 (2017).

[465] Tallant, C. *et al.* Crystal structure of human TDRD1 extended Tudor domain in complex with a symmetrically dimethylated E2F peptide (2016) https://doi.org/10.2210/pdb5m9n/pdb.

[466] Zouridakis, M. *et al.* Crystal structures of free and antagonist-bound states of human α9 nicotinic receptor extracellular domain. *Nat. Struct. Mol. Biol.* **21**, 976–980 (2014).

[467] Su, J. *et al.* Galectin-13, a different prototype galectin, does not bind β-galactosides and forms dimers via intermolecular disulfide bridges between Cys-136 and Cys-138. *Sci. Rep.* **8**, 980 (2018).

[468] Qiu, C. *et al.* Mechanism of activation and inhibition of the HER4/ErbB4 kinase. *Structure* **16**, 460–467 (2008).

[469] Canning, P. *et al.* CDKL family kinases have evolved distinct structural features and ciliary function. *Cell Rep.* **22**, 885–894 (2018).

[470] Lee, J. *et al.* Structural basis for the selective inhibition of Cdc2-Like kinases by CX-4945. *BioMed Res. Int.* **2019**, 6125068 (2019).

[471] Goldenzweig, A. *et al.* Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol. Cell* **63**, 337–346 (2016).

[472] Vollmar, M. *et al.* Crystal structure of the second glutaredoxin domain of human TXNL2 (2011) https://doi.org/10.2210/pdb2yan/pdb.

[473] Liu, X. *et al.* Structure of human phagocyte NADPH oxidase in the activated state. *Nature* **627**, 189–195 (2024).

[474] Piserchio, A. *et al.* Association of the cystic fibrosis transmembrane regulator with CAL: structural features and molecular dynamics. *Biochemistry* **44**, 16158–16166 (2005).

[475] Littlejohn, J. *et al.* Structural definition of hSP-D recognition of *Salmonella enterica* LPS inner core oligosaccharides reveals alternative binding modes for the same LPS. *PLOS ONE* **13**, 1–14 (2018).

[476] Yeo, H. *et al.* Phospholipid transfer function of PTPIP51 at mitochondria-associated ER membranes. *EMBO Rep.* **22**, e51323 (2021).

[477] Huber, B. *et al.* Thyroid hormone receptor-β mutations conferring hormone resistance and reduced corepressor release exhibit decreased stability in the N-terminal ligand-binding domain. *Mol. Endocrinol.* **17**, 107–116 (2003).

[478] Rong, Y. *et al.* TMEM120A contains a specific coenzyme A-binding site and might not mediate poking- or stretch-induced channel activities in cells. *eLife* **10**, e71474 (2021).

[479] Radzimanowski, J. *et al.* Crystal structure of the human Fe65-PTB1 domain. *J. Biol. Chem.* **283**, 23113–23120 (2008).

[480] Zhao, F. *et al.* Structural insights into hormone recognition by the human glucose-dependent insulinotropic polypeptide receptor. *eLife* **10**, e68719 (2021).

[481] Yoo, J. *et al.* GlcNAc-1-P-transferase–tunicamycin complex structure reveals basis for inhibition of N-glycosylation. *Nat. Struct. Mol. Biol.* **25**, 217–224 (2018).

[482] Raththagala, M. *et al.* Structural mechanism of laforin function in glycogen dephosphorylation and Lafora Disease. *Mol. Cell* **57**, 261–272 (2015).

[483] Bai, X. *et al.* Sampling the conformational space of the catalytic subunit of human γ-secretase. *eLife* **4**, e11182 (2015).

[484] Liu, S. *et al.* Crystal structure of PRL-1 complex with compound analogy 3 (2016) https://doi.org/10.2210/pdb5bx1/pdb.

[485] Manthei, K. *et al.* A retractable lid in lecithin:cholesterol acyltransferase provides a structural mechanism for activation by apolipoprotein A-I. *J. Biol. Chem* **292**, 20313–20327 (2017).

[486] Wang, X. *et al.* Molecular insights into differentiated ligand recognition of the human parathyroid hormone receptor 2. *Proc. Natl. Acad. Sci. U.S.A* **118**, e2101279118 (2021).

[487] Li, H. *et al.* Structural basis for heparan sulfate co-polymerase action by the EXT1–2 complex. *Nat. Chem. Biol.* **19**, 565–574 (2023).

[488] Mathea, S. *et al.* Human m7GpppN-mRNA hydrolase (DCP2, NUDT20) catalytic domain (2017) https://doi.org/10.2210/pdb5mp0/pdb.

[489] Cho, H. *et al.* A subcomplex crystal structure of human cytosolic aspartyl-tRNA synthetase and heterotetrameric glutathione transferase-homology domains in multi-tRNA synthetase complex (2018) https://doi.org/10.2210/pdb5y6l/pdb.

[490] Jeong, B. *et al.* Cryo-EM structure of the Hippo signaling integrator human STRIPAK. *Nat. Struct. Mol. Biol.* **28**, 290–299 (2021).

[491] Kastrup, J. *et al.* Two mutants of human heparin binding protein (CAP37): Toward the understanding of the nature of lipid A/LPS and BPTI binding. *Proteins* **42**, 442–451 (2001).

[492] Yu, Y. *et al.* Molecular basis of 1-deoxygalactonojirimycin arylthiourea binding to human α-Galactosidase A: pharmacological chaperoning efficacy on Fabry Disease mutants. *ACS Chem. Biol.* **9**, 1460–1469 (2014).

[493] Xu, Y. *et al.* Molecular insights into biogenesis of glycosylphosphatidylinositol anchor proteins. *Nat. Commun.* **13**, 2617 (2022).

[494] Robinson, R. *et al.* Simultaneous binding of Guidance Cues NET1 and RGM blocks extracellular NEO1 signaling. *Cell* **184**, 2103–2120.e31 (2021).

[495] Gao, S. *et al.* Structure of human $Ca_V2.2$ channel blocked by the painkiller ziconotide. *Nature* **596**, 143–147 (2021).

[496] Giordanetto, F. *et al.* Discovery of AZD2716: a novel secreted phospholipase A_2 ($sPLA_2$) inhibitor for the treatment of coronary artery disease. *ACS Med. Chem. Lett.* **7**, 884–889 (2016).

[497] Ramírez, A. *et al.* Cryo–electron microscopy structures of human oligosaccharyltransferase complexes OST-A and OST-B. *Science* **366**, 1372–1375 (2019).

[498] Saito, K. *et al.* Crystal structure of human Tob1 protein (2007) https://doi.org/10.2210/pdb2z15/pdb.

[499] He, L. *et al.* Structure, gating, and pharmacology of human $Ca_V3.3$ channel. *Nat. Commun.* **13**, 2084 (2022).

[500] Nagae, M. *et al.* Crystal structure of α5β1 integrin ectodomain: atomic details of the fibronectin receptor. *J. Cell Biol.* **197**, 131–140 (2012).

[501] Welin, M. *et al.* Structural studies of tri-functional human GART. *Nucleic Acids Res.* **38**, 7308–7319 (2010).

[502] Taherbhoy, A. *et al.* BMI1–RING1B is an autoinhibited RING E3 ubiquitin ligase. *Nat. Commun.* **6**, 7621 (2015).

[503] Kulkarni, M. *et al.* Two independent histidines, one in human prolactin and one in its receptor, are critical for pH-dependent receptor recognition and activation. *J. Biol. Chem.* **285**, 38524–38533 (2010).

[504] Prakash, A. *et al.* Structural basis of nucleic acid recognition by FK506-binding protein 25 (FKBP25), a nuclear immunophilin. *Nucleic Acids Res.* **44**, 2909–2925 (2016).

[505] Evans, C. *et al.* Coordination of di-acetylated histone ligands by the ATAD2 bromodomain. *Int. J. Mol. Sci.* **22**, (2021).

[506] Pantoom, S. *et al.* RAB33B recruits the ATG16L1 complex to the phagophore via a noncanonical RAB binding protein. *Autophagy* **17**, 2290–2304 (2021).

[507] Guo, R. *et al.* Architecture of human mitochondrial respiratory megacomplex $I_2III_2IV_2$. *Cell* **170**, 1247–1257.e12 (2017).

[508] Oefner, C. *et al.* Crystal structure of human dihydrofolate reductase complexed with folate. *Eur. J. Biochem* **174**, 377–385 (1988).

[509] Païdassi, H. *et al.* C1q binds phosphatidylserine and likely acts as a multiligand-bridging molecule in apoptotic cell recognition. *J. Immunol.* **180**, 2329–2338 (2008).

[510] Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).

[511] Webber, C. and Barton, G. Increased coverage obtained by combination of methods for protein sequence database searching. *Bioinformatics* **19**, 1397–1403 (2003).

[512] Scott, M. and Barton, G. Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinform.* **8**, 239 (2007).

[513] Noguchi, J. *et al.* Crystal structure of the covalent intermediate of human cytosolic β-glucosidase. *Biochem. Biophys. Res. Commun.* **374**, 549–552 (2008).

[514] Sestak, F. VN-EGNN. GitHub repository (2024) https://github.com/ml-jku/vnegnn.

[515] Carbery, A. IF-SitePred. GitHub repository (2024) https://github.com/annacarbery/binding-sites.

[516] Smith, Z. *et al.* Graph Attention Site Prediction (GrASP). GitHub repository (2024) https://github.com/tiwarylab/GrASP.

[517] Smith, Z. *et al.* GrASP. Google Colaboratory notebook (2024) https://colab.research.google.com/github/tiwarylab/GrASP/blob/main/GrASP.ipynb.

[518] Kandel, J. PUResNet V2.0 web server (2024) https://nsclbio.jbnu.ac.kr/tools/jmol.

[519] Aggarwal, R. DeepPocket. GitHub repository (2024) https://github.com/devalab/DeepPocket.

[520] Krivák, R. P2Rank. GitHub repository (2024) https://github.com/rdk/p2rank.

[521] Anaconda Team. fpocket v4.2.1. Anaconda Cloud (2024) https://anaconda.org/conda-forge/fpocket.

[522] Capra, T. ConCavity v0.1. Princeton University (2010) https://compbio.cs.princeton.edu/concavity/.

[523] Utgés, J.S. LBS-Comparison results. Zenodo Dataset (2024) https://doi.org/10.5281/zenodo.13121414.

[524] Utgés, J.S. bartongroup/LBS-comparison: submission release v1.1.0 (2024) https://doi.org/10.5281/zenodo.13171101.

[525] Xu, Q. and Dunbrack, R. Principles and characteristics of biological assemblies in experimentally determined protein structures. *Curr. Opin. Struct. Biol.* **55**, 34–49 (2019).

[526] Pickens, J. *et al.* Anchor-based design of improved cholera toxin and *E. coli* heat-labile enterotoxin receptor binding antagonists that display multiple binding modes. *Chem Biol* **9**, 215–224 (2002).

[527] Hofmann, E. *et al.* Structural basis of light harvesting by carotenoids: peridinin-chlorophyll-protein from *Amphidinium carterae*. *Science* **272**, 1788–1791 (1996).

[528] Williams, L. *et al.* Order and disorder: differential structural impacts of myricetin and ethyl caffeate on human amylase, an antidiabetic target. *J. Med. Chem.* **55**, 10177–10186 (2012).

[529] Chen, C. *et al.* Structure of human POFUT2: insights into thrombospondin type 1 repeat fold and O-fucosylation. *EMBO J.* **31**, 3183–3197 (2012).

[530] Liu, R. *et al.* Structural basis for substrate binding and catalytic mechanism of a human RNA:m^5C methyltransferase NSun6. *Nucleic Acids Res.* **45**, 6684–6697 (2017).

[531] Ramirez-Rios, S. *et al.* VASH1-SVBP and VASH2-SVBP generate different detyrosination profiles on microtubules. *J. Cell. Biol.* **222**, e202205096 (2022).

[532] Lim, S. *et al.* The structure and catalytic mechanism of human sphingomyelin phosphodiesterase like 3a – an acid sphingomyelinase homologue with a novel nucleotide hydrolase activity. *FEBS J.* **283**, 1107–1123 (2016).

[533] Finer-Moore, J. *et al.* Crystal structure of the human tRNA m^1A58 methyltransferase-$tRNA_3^{Lys}$ complex: refolding of substrate tRNA alows access to the methylation target. *J. Mol. Biol.* **427**, 3862–3876 (2015).

[534] Lee, H. *et al.* Crystal structure of human pyridoxal 5'-phosphate phosphatase (Chronophin) mutant - C221S (2017) https://doi.org/10.2210/pdb5gyn/pdb.

[535] Blasiak, L. *et al.* Crystal structure of the non-haem iron halogenase SyrB2 in syringomycin biosynthesis. *Nature* **440**, 368–371 (2006).

[536] Mascarenhas, R. *et al.* Architecture of the human G-protein-methylmalonyl-CoA mutase nanoassembly for B_{12} delivery and repair. *Nat. Commun.* **14**, 4332 (2023).

[537] Milani, M. *et al.* FAD-binding site and NADP reactivity in human renalase: a new enzyme involved in blood pressure regulation. *J. Mol. Biol.* **411**, 463–473 (2011).

[538] Liu, L. *et al.* Autophosphorylation transforms DNA-PK from protecting to processing DNA ends. *Mol. Cell* **82**, 177–189 e4 (2022).

[539] Durairaj, J. *et al.* PLINDER: the protein-ligand interactions dataset and evaluation resource. *bioRxiv*, 2024.07.17.603955 (2024).

[540] Stärk, H. *et al.* EquiBind: geometric deep learning for drug binding structure prediction. *arXiv* (2022).

[541] Qiao, Z. *et al.* State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nat. Mach. Intell.* **6**, 195–208 (2024).

[542] Schneuing, A. *et al.* Structure-based drug design with equivariant diffusion models. *Nat. Comput. Sci.* **4**, 899–909 (2024).

[543] Rabeh, W. *et al.* Crystal structure of human sarcomeric mitochondrial creatine kinase (2015) https://doi.org/10.2210/pdb4z9m/pdb.

[544] Keough, D. *et al.* Inhibition of hypoxanthine-guanine phosphoribosyltransferase by acyclic nucleoside phosphonates: a new class of antimalarial therapeutics. *J. Med. Chem.* **52**, 4391–4399 (2009).

[545] Ahmed, S. *et al.* 1,2,4-triazolo-[1,5-a]pyridine HIF prolylhydroxylase domain-1 (PHD-1) inhibitors with a novel monodentate binding interaction. *J. Med. Chem.* **60**, 5663–5672 (2017).

[546] Coleman, J. *et al.* X-ray structures and mechanism of the human serotonin transporter. *Nature* **532**, 334–339 (2016).

[547] Pickles, L. *et al.* Crystal structure of the C-Terminal WD40 repeat domain of the human Groucho/TLE1 transcriptional corepressor. *Structure* **10**, 751–761 (2002).

[548] Lee, C. *et al.* Bivalent recognition of fatty acyl-CoA by a human integral membrane palmitoyltransferase. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2022050119 (2022).

[549] Baraldi, E. *et al.* Structure of the PH domain from Bruton's tyrosine kinase in complex with inositol 1,3,4,5-tetrakisphosphate. *Structure* **7**, 449–460 (1999).

[550] Baldwin, E. *et al.* Human endogenous retrovirus-K (HERV-K) reverse transcriptase (RT) structure and biochemistry reveals remarkable similarities to HIV-1 RT and opportunities for HERV-K–specific inhibition. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2200260119 (2022).

[551] Cui, W. *et al.* Structures of human SGLT in the occluded state reveal conformational changes during sugar transport. *Nat. Commun.* **14**, 2920 (2023).

[552] Ilouz, R. *et al.* Localization and quaternary structure of the PKA RIβ holoenzyme. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 12443–12448 (2012).

[553] Du, J. *et al.* Structures of human mGlu2 and mGlu7 homo- and heterodimers. *Nature* **594**, 589–593 (2021).

[554] Schreiber, S. The rise of molecular glues. *Cell* **184**, 3–9 (2021).

[555] Brewer, A. *et al.* Mapping the substrate landscape of protein phosphatase 2A catalytic subunit PPP2CA. *iScience* **27**, (2024).

[556] Brewer, A. *et al.* Targeted dephosphorylation of SMAD3 as an approach to impede TGF-β signaling. *iScience* **27**, (2024).

[557] Zhao, J. *et al.* Targeted dephosphorylation of TFEB promotes its nuclear translocation. *iScience* **27**, (2024).

[558] Zengerle, M. *et al.* Selective small molecule induced degradation of the BET bromodomain protein BRD4. *ACS Chem. Biol.* **10**, 1770–1777 (2015).

[559] Gadd, M. *et al.* Structural basis of PROTAC cooperative recognition for selective protein degradation. *Nat. Chem. Biol.* **13**, 514–521 (2017).

[560] Ibrahim, P. *et al.* Accurate prediction of dynamic protein–ligand binding using *P-score* ranking. *J. Comput. Chem.* **45**, 1762–1778 (2024).

[561] Ibrahim, P. *et al.* FMOPhore for hotspot identification and efficient fragment-to-lead growth strategies. *ChemRxiv* (2024).

PhD experience graphical summary

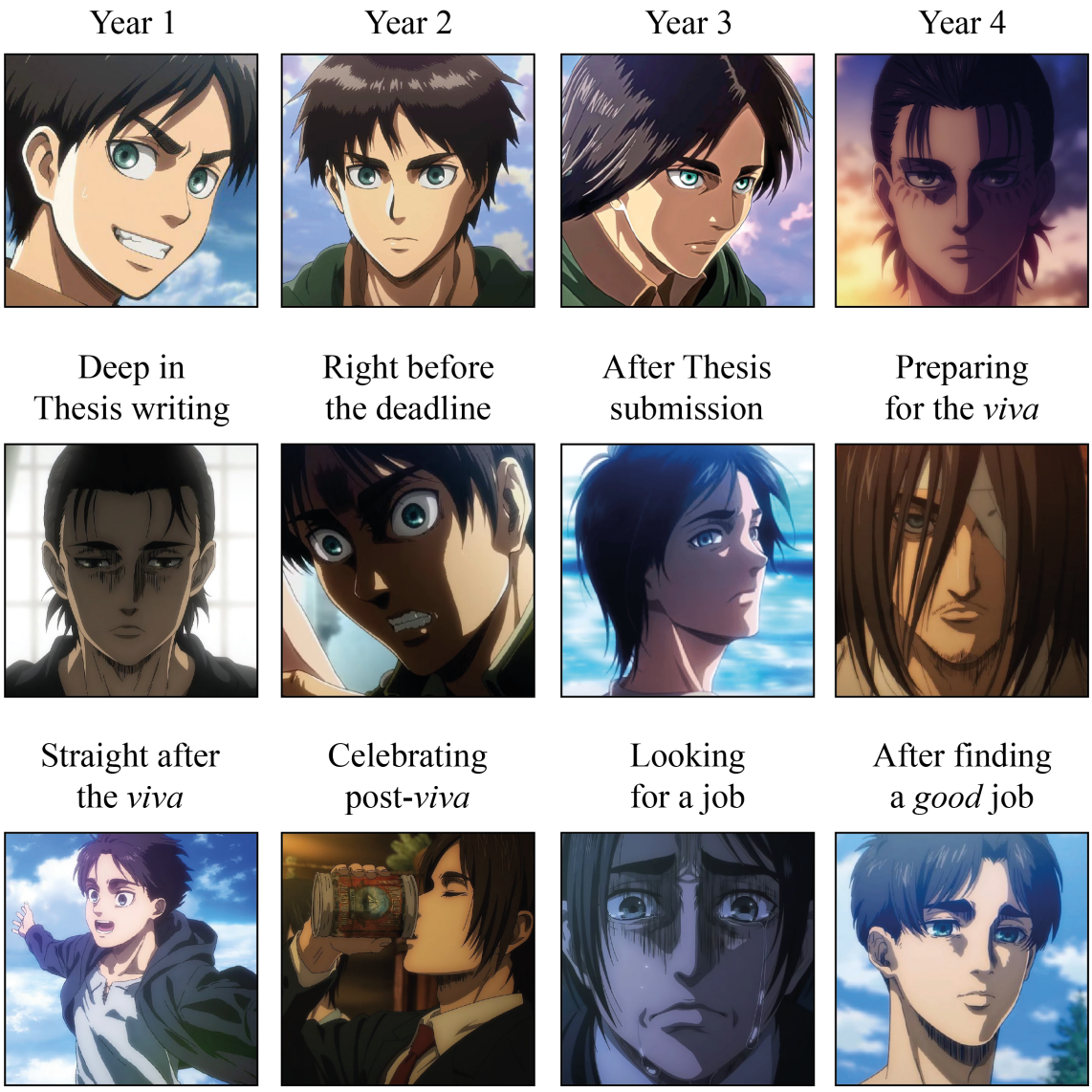

Figure U.1. PhD experience graphical summary. The final figure of this PhD Thesis summarises in a comical and graphical manner the last 4.5 years of *my* life as a PhD student, as well as *my* immediate future after the submission of this Thesis. The character appearing in this figure is *Eren Yeager*, the main character of the Japanese manga *Attack on Titan* by Hajime Isayama [1]. Images are from the anime adaptation by the Wit and MAPPA animation studios. Read from left to right and top to bottom.