Computational Strategies to Explore Glycan Recognition and Protein Design

Reyes Núñez Franco

Doctoral Thesis

2024

Thesis Director:

Prof. Dr. Gonzalo Jiménez-Osés

Universidad del País Vasco
Euskal Herriko Unibertsitatea

CIC bioGUNE
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

A mi familia, por ser apoyo incondicional en cada paso del camino,
por ser siempre el hogar al que volver.

Agradecimientos

Quien me diría que iba a estar escribiendo los agradecimientos de una tesis doctoral. Yo que me negaba en rotundo a hacerla. Cuando terminé el Grado, estaba llena de dudas, pero sí había una cosa que tenía clara era que no quería hacer un doctorado. Afortunadamente, a veces la vida nos hace tragarnos nuestras propias palabras porque hubiese perdido una muy gran oportunidad.

Cuando pensé en escribir los agradecimientos no sabía cómo comenzar y efectivamente no sé cómo hacerlo. Pero me parece una oportunidad maravillosa para hacer un parón de este ritmo frenético y reflexionar. Y es que cuando pienso a quien quiero agradecer, aparece un listado prácticamente interminable en mi mente y eso es una suerte.

Durante este tiempo he tenido la inmensa fortuna de aprender de personas maravillosas, de sorprenderme con la "simple" observación de la naturaleza, de cuestionarme todas las creencias que había establecido en mi mente de forma inconsciente, de hacer grandes amigos y de vivir muy buenos momentos. A ver si soy capaz de desmenuzar todo esto sin que ocupen más los agradecimientos que la propia tesis.

Me gustaría comenzar contigo Gonzalo. Gracias por acoger con los brazos abiertos desde el minuto uno a la pequeña Reyecitas que llegó a Bilbao un 2 de septiembre de 2019. Gracias por guiarme en este camino y ayudarme a crecer no solo como científica sino como persona. Gracias por enseñarme el valor del trabajo meticuloso, del detalle, de profundizar en los conceptos hasta comprender cada pequeña parte de ellos. Gracias por despertar en mi la curiosidad científica, esa que tantos quebraderos de cabeza me (nos) ha dado y que ahora no puedo despegar de mi persona. Gracias por darme alas y libertad para experimentar, probar, equivocarme, aprender y *succeed*. Tengo millones de motivos por los que estarte agradecida, y eso es de las cosas más grandes que me llevo de este periodo.

Gracias también a nuestro lab, al presente y a todas las personitas que han ido pasando y he tenido la suerte de conocer. Gracias a Claudio, o como a mí me gusta llamarle a Caludio. Gracias por ser ejemplo, gracias por estar dispuesto siempre (y cuando digo siempre es siempre) a ayudar, gracias por tu paciencia infinita, por tu calma y por tu guasa (tu baile nupcial nunca será olvidado). Gracias por ser mi postdoc de confianza. Has sido y eres un pilar fundamental. Gracias a ti Francesca, eres digna de admirar. Gracias por tu bondad, por tu apoyo, por tu comprensión, por tu guía. Cada día aprendo algo nuevo contigo. Es un gustazo y una suerte trabajar con personas como tú al lado. Solo deseo que

llegues todo lo lejos que te propongas porque te lo mereces como nadie y porque sería una suerte para todos los demás, que puedas poner en práctica todas esas ideas que tienes en mente. Ojalá verlo de cerquita. Gracias Angelito, por traer la gracia sureña al grupo. Nuestro representante jerezano. Gracias por todos los buenos ratos que hemos compartido y por hacer del norte un lugar un poquito más cálido. Gracias a ti también Sara, por tantos momentos divertidos, las risas contigo sabemos que están aseguradas y eso no tiene precio. Gracias a Elena, Matteo y Riccardo porque habéis sido también personas clave en esta experiencia. Matteo siempre tendremos un huequito para ti en nuestros *coras.* Gracias por venir cargado de buena energía, fue un tiempo muy guay. Mención también para nuestras dos nuevas incorporaciones, Cris y Mila. Os ha tocado llegar en mi época *fully stressed*, *sorry!* Os deseo que saquéis el máximo partido a esta experiencia, pienso que tenéis suerte de comenzar la tesis en el momento científico en que nos encontramos y que si lo aprovecháis podéis hacer de estos años, un viaje muy chulo. ¡Contad conmigo para todo lo que os pueda ayudar para ello!

Dentro de bioGUNE me gustaría hacer mención especial al grupo de Óscar Millet, *Precision Medicine and Metabolism Lab.* Gracias Óscar por abrirnos las puertas de tu casa de par en par con tanta amabilidad. Gracias por darnos acceso tanto a vuestro laboratorio como a tantas personas que han sido claves en nuestra evolución. Habéis sido todo un descubrimiento para mí. En primer lugar, me gustaría agradecer a Gabi, has sido un pilar fundamental en este último tramo del camino. Gracias no solo por todo lo que me has ayudado en el *wetlab,* sino por estar siempre dispuesto a debatir y *discuss* sobre todas mis inquietudes. Me encanta poder ir con mi batiburrillo de ideas, incordiarte un poquito y después salir con todas esas ideas perfectamente ordenadas. Gracias, es una suerte contar contigo. Muchas muchas gracias a los que hacen que el lab sea un lugar tan divertido para mí al que siempre tengo ganas de ir. *Special mention to* Jon, Tania, Marga, Alba, Pablo, Ángela, Sara, Laura. De verdad para mi es una alegría compartir tiempo con vosotros, salgo siempre con las pilas cargadas. Gracias por acogerme como una más. *Thanks a lot also to Andreas, thanks for sharing all your knowledge and thanks for always being so much fun!*

Gracias al grupo de Jesús Jiménez, *Chemical Glycobiology Lab*. Gracias por adentrarme en el mundo de las *galectins*, del RMN y de todas sus peculiaridades. Ha sido un placer poder trabajar y aprender con y de vosotros. Gracias a todas las personas de CIC bioGUNE que han formado parte de esta experiencia y que me han aportado tantas y diferentes cosas. Muy brevemente no me gustaría dejar pasar la oportunidad de agradecer a Mikel y Felix, de la Plataforma de Proteómica, gracias por vuestro trabajazo y simpatía, es un gusto siempre estar con vosotros. Gracias también a Alfonso Martínez, por transmitir su entusiasmo

por la investigación y por valorar siempre tanto nuestro trabajo. Antonio Franconetti (también conocido como Dr. Antoine), gracias por tu apoyo incondicional desde tiempos inmemoriales. Gracias por adentrarme en el mundo científico cuando era toda una niñita y por seguir apoyándome en cada paso del camino.

I would also like to express my gratitude to David Baker. Thank you for allowing me to join your laboratory and delve into the fascinating world of Protein Design. Thanks for helping me immersing in this field from which I don't want to move away. Thanks for being always present, for listening to my concerns, and provide unwavering support. My time at the IPD marked an inflexion point in both my scientific and personal life. Big thank you to Basile, Lukas and Indrek for your mentorship during this period; your generous investment of time and guidance has been invaluable. Thanks to Susana, Marc, Alfredo, Begoña, Miguel, Fernando, Kiera, Anna, Mohamad, Sam, Jeremiah, Sidney, Aditya, Phil, Robert, David, Linna, Shingo, and many others for welcoming me from day one. Thanks to the cotilleo *crew. I will never forget that period of my life and all the special moments we shared.*

Por supuesto gracias a mi familia. Pilar fundamental e imprescindible en mi vida. Hay un texto que le encanta a mi padre:

Enseñarás a volar, pero no volarán tu vuelo.
Enseñarás a soñar, pero no soñarán tu sueño.
Enseñarás a vivir, pero no vivirán tu vida.
Sin embargo, en cada vuelo, en cada vida, en cada sueño, perdurará siempre la huella del camino enseñado.

No le falta ni una pizca de razón. Gracias a mis padres por enseñarme a volar, por impulsar mi vuelo, pero sobre todo por ser siempre hogar para volver. Gracias por apoyarme en cada uno de mis pasos, y sobre todo gracias por la familia fuerte y unida que nos habéis dado. Gracias a mis hermanos, Marcelo, Juan y Bea. Gracias por formar un grupo de cuatro hermanos de lo más variopinto que os podáis imaginar. Gracias por todo lo que me aportáis cada uno de vosotros. Gracias por conseguir estar presente en nuestro día a día a pesar de que cada uno haya volado a sus anchas. Gracias por los regalos más bonitos que me habéis podido hacer: mis sobrinos. La pandillita. Juanito, Alma, Alexis, Hugo, gracias por llenarme el corazón en cada abrazo y recargarme las pilas con cada sonrisa.

Como no podría ser de otra manera gracias a mis niñas. Mi piña. No sabéis lo inmensamente afortunada que me siento de poder conservar a mis amigas del cole, mis amigas de toda la vida, las que me han visto crecer en todas las etapas

de mi vida. En las buenas y en las no tan buenas. La gente siempre se sorprende cuando digo todas las que somos y les cuesta creer la relación que tenemos. Y es que no es habitual que un grupo tan grande y diferente pueda mantenerse tan unido en el tiempo. Por eso siempre pienso que es algo mágico y que debemos cuidarlo con toda nuestra energía. Ojalá estar a la altura. Gracias a ellas por estar siempre, por seguir achuchándome con la misma fuerza por más aventuras que decida recorrer. Cuando vuelvo, ellas siempre están y el valor que tiene eso es incalculable. Gracias de todo corazón, por ser vosotras.

La vida me ha permitido conocer y disfrutar de uno de los sitios más bonitos que la naturaleza ha decidido crear. No voy a decir nombre porque si no alguna de las que incluyo por aquí me daría un tirón de orejas. Tengo la suerte de contar con muchas personas especiales en mi vida, y ellas sin duda lo son. Gracias al Norte por regalarme la oportunidad de conocerlas. Andre, Sil, Clau, Cris, Inés, María, gracias por vuestra amistad sincera, por vuestro apoyo incondicional y por ser diversión asegurada.

Moverse tiene muchas ventajas e inconvenientes, pero uno de sus puntos fuertes es que te permite conocer gente que después piensas, ¿pero que habría sido de mi sin ellas? Y cuando yo vine a Bilbao hecha pedacitos, la vida en su curiosa manera de compensar, me puso al lado a mi Raqpe. Vaya golpe de suerte. ¡Cuántas cosas hemos vivido desde entonces! No puedo ni imaginar cómo habrían sido todos estos años sin ti. Gracias por ser casa. Gracias por hacer de nuestro hogar un lugar de lo más divertido. (Decir que nos lo pasamos bien en el confinamiento tiene guasa, pero es que así fue, ¡no me extraña que tengas premio!). Somos el claro ejemplo del *yin y el yang*. Deseando ver de cerquita todo lo que la vida nos sigue regalando.

No me olvido ni mucho menos de mi pandillita bilbaína. Paloma (Mita), gracias por convertirte en una pieza clave en Bilbao, gracias por comprenderme y por agarrarme la mano fuerte siempre que lo he necesitado. Das paz amiga. Diegui, gracias por los lotes de risas bien empacaditos que nos regalas a todos. Verte es tener asegurado pasarlo bien. Ainhoa, gracias por aportar el punto autóctono al grupo. Gracias por tu energía insaciable y contagiar tu espíritu aventurero. Bea, mi enfermera favorita, te debemos muchas cerves por tirar de nosotros cuando estamos un poquito dispersos. Y Pablis la esencia salamantina, gracias por tu predisposición y tu generosidad. Eres de las personitas con el corazón más puro que conozco.

¡Ah, Gonzalo! Se me había olvidado una cosa, gracias por elegir tan bien los temas musicales para las colaboraciones. Gracias por impulsar la unión entre tierras. Todavía recuerdo como en mis primeros días en bioGUNE, Claudio se

paró en las escaleras a charlar con un chiquito que acababa de volver de su estancia en el IPD – UW. En ese mismo momento le escribí un mensaje a Raquel que no puedo poner aquí, pero madre mía todo lo que ha llovido (literalmente) desde ese día. Al final Gonzalo llevaba razón con sus intuiciones. Gracias Dani por todo lo que me has aportado desde entonces. Gracias por quererme de la manera más sana que se pueda imaginar. Gracias por ayudarme a remover todos mis cimientos y reconstruirlos con las piezas que realmente me encajan. Gracias por ayudarme cada día a intentar ser la mejor versión de mí misma. Me siento inmensamente afortunada de que elijamos cada día disfrutar de este camino juntos. Deseando seguir descubriendo juntos todo lo que la vida tiene por ofrecernos.

En el tiempo en que escribo estos agradecimientos, me encuentro leyendo la obra 'La insoportable levedad del ser', que un Doctor muy inteligente me recomendó. En esta obra, Milan Kundera destaca como ciertos eventos y encuentros casuales pueden tener un impacto significativo en la vida de las personas. El autor sugiere que, dado al carácter efímero de la existencia, las casualidades y las decisiones individuales adquieren una importancia única en la conformación de nuestras vidas. (Gracias Charlie Garcia Pérez Torres). Cuando leí esto a mi cabeza vino instantáneamente el experimento mental del gato de Schrödinger. Al igual que el gato de Schrödinger existe en un estado de posibilidades hasta que es observado, así la vida está llena de posibilidades, donde las decisiones y eventos casuales pueden tener consecuencias impredecibles. Yo entre todas las infinitas posibilidades después de un buen *tute* de llorar, decidí estudiar química (gracias a mi profesora Ángela, primera gran científica influyente en mi vida). Decidí irme un año al extranjero para aprender e intentar descifrar cuales eran los siguientes pasos que quería dar. Decidí después de esto, irme a Barcelona, a estudiar un Máster en Bioinformática, ¿quién me mandaría?! ¡Yo no sabía ni que era una terminal! Estando allí un buen amigo, Antonio, me mandó una oferta para realizar una tesis doctoral. No era la primera vez que esto ocurría, y que yo directamente sin abrir descartaba el mensaje. Pero esta vez, en este sin fin de posibilidades la Reyes del pasado decidió abrirlo y algo dentro de mi hizo *click*. ¡Y menos mal que lo hice! Esa decisión me ha dado unos años maravilloso, en los que he aprendido como nunca y en los que siento que he crecido a pasos de gigante. Por supuesto gracias a todas estas personas que he tenido a mi lado durante el camino, vuestra influencia ha sido clave en todo lo que ha ocurrido desde entonces.

Infinitamente agradecida.

La verdadera ciencia enseña, por encima de todo, a dudar y a ser ignorante.

Miguel de Unamuno

Fortune favors the prepared mind.

Louis Pasteur

Contents

Abbreviations

$\Delta\Delta G^{calc}$	Predicted Unfolding Free Energy Change
ΔC_p	Heat Capacity of Unfolding
ΔG_b	Binding Free energy
ΔG_s	Maximum Thermodynamic Stability
ΔG_{sim}	Free Energy of Unfolding
ΔH_b	Binding Enthalpy
ΔH_m	Enthalpy of Unfolding at T_m
ΔS_b	Binding Entropy
ACP	Acyl Carrier Protein
AI	Artificial Intelligence
AMBER	Assisted Model Building with Energy Refinement
CD	Circular Dichroism
CORCEMA-ST	Complete Relaxation and Conformational Exchange Matrix-Saturation Transfer
CRD	Carbohydrate Recognition Domain
CSP	Chemical Shift Perturbation
CT	C-terminal
CTL	C-type lectins
DC-SIGN	Dendritic cell-specific ICAM-3-grabbing non-integrin
DFT	Density Functional Theory
EGFP	Enhanced Green Fluorescent Protein
FDA	Food and Drug Administration
FF	Force Field
FF_{sim}	Fraction of protein in the native (Folded) state
FXN	Frataxin
GAFF2	General AMBER Force Field
Gal	Galectin
GIST	Grid Inhomogeneous Solvation Theory Method
HIV	Human Immunodeficiency Virus
HSQC	Heteronuclear Single Quantum Correlation
ICAM-3	Intercellular adhesion molecule 3
IEF-PCM	Integral Equation Formalism Polarizable Continuum Model
ISCU	Iron-sulfur cluster assembly enzyme
ISD11	LYR motif-containing protein 4
ITC	Isothermal Titration Calorimetry

k_{cat}	Catalytic constant
K_M	Michaelis-Menten constant
K_{sim}	Equilibrium Constant of Folding
LMCS	Low-Mode Conformational Search
MBP	Maltose-Binding Protein
MD	Molecular Dynamics
MGL	Macrophage Galactose Lectin
MM	Molecular Mechanics
MSA	Multiple-Sequence Alignment
NFS1	Cysteine Desulfurase
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NPL	Natural Language Processing
NT	N-terminal
PDB	Protein Data Bank
pI	Isoelectric Point
QM	Quantum Mechanics
RFU	Relative Fluorescent Units
RMSD	Root-Mean-Square-Deviation
RMSF	Root-Mean-Square-Fluctuation
SASA	Solvent Accessible Surface Area
SBP	Structural Binding Pattern
SE	Shannon Entropy
SEC	Size Exclusion Chromatography
STD	Saturation Transfer Difference
T	Temperature
TEV	Tobacco Etch Virus
T_m	Melting temperature
TS	Transition State
Ts	Temperature of maximum thermodynamic stability
VdW	Van der Waals
WISP	Weighted Implementation of Suboptimal Paths

Resumen

El proceso de reconocimiento proteína-ligando es fundamental en el ámbito de la biología molecular, rigiendo funciones celulares esenciales y mediando en diversos procesos biológicos. En el núcleo de estas interacciones reside la notable capacidad de las proteínas para unirse de forma selectiva a pequeñas moléculas, conocidas como ligandos. Desde las cascadas de señalización que regulan las respuestas celulares hasta las reacciones enzimáticas que impulsan el metabolismo, el reconocimiento proteína-ligando desempeña un papel central en la química de la vida.

En el proceso de asociación entre una proteína y un ligando la influencia de la solvatación, particularmente el comportamiento de las moléculas de agua alrededor de la zona de unión es de gran importancia. El agua, componente integral en la interfaz biomolecular, desempeña un papel crucial al participar en enlaces ubicuos y específicos. Aunque ocupa menos espacio que las cadenas laterales polares de las proteínas, las moléculas de agua pueden participar en múltiples enlaces de hidrógeno gracias a su adaptabilidad. Esta versatilidad le permite interactuar con diversos ligandos y contribuye a la energía de unión mediante procesos de desolvatación y reorganización durante la formación del complejo. Comprender la influencia del agua en las interacciones proteína-ligando es fundamental para optimizar el diseño de ligandos y comprender los componentes energéticos involucrados en estas interacciones. Además, el carácter dinámico de las proteínas les permite responder estructuralmente a la unión de ligandos, lo que a su vez puede desencadenar efectos alostéricos en regiones distantes al sitio de reconocimiento.

Entre estos procesos de reconocimiento, cabe destacar el reconocimiento de carbohidratos por proteínas en los procesos biológicos. A diferencia de las pequeñas moléculas orgánicas, los carbohidratos añaden una dimensión única a estas interacciones. Las proteínas que se unen a carbohidratos, como las lectinas, desempeñan papeles clave en la adhesión celular, señalización, interacciones huésped-patógeno y más. Los patógenos exhiben carbohidratos en su superficie, que sirven para mediar la interacción con las células huésped. Estas interacciones tienen un impacto notable en la respuesta inmunitaria y ofrecen posibles objetivos terapéuticos para modularlas.

En los últimos años, los estudios experimentales empleando técnicas como la cristalografía de rayos X y la espectroscopia de RNM han proporcionado detalles precisos sobre las interacciones entre lectinas y carbohidratos. Además, el uso de diferentes métodos experimentales ha permitido determinar parámetros termodinámicos y cinéticos asociados con el proceso de unión. En paralelo, se han desarrollado métodos computacionales en el campo de la Modelización

Molecular y la Química Computacional. Estos métodos incorporan diversas técnicas a diferentes escalas espaciotemporales, lo que resulta en modelos altamente precisos para caracterizar las interacciones entre lectinas y carbohidratos, y en predicciones de alta calidad.

En la primera parte de la presente tesis doctoral se aborda el estudio del reconocimiento de carbohidratos por lectinas, en concreto galectinas y DC-SIGN. En este estudio se emplean diversas técnicas computacionales, incluyendo cálculos de mecánica cuántica y dinámicas moleculares. Todos estos cálculos son apoyados por experimentos de Resonancia Magnética Nuclear.

Las galectinas son una familia de proteínas que destacan por su capacidad para unir específicamente β-galactósidos. Estas proteínas están distribuidas ampliamente en el cuerpo humano y desempeñan un papel crucial en la regulación de numerosos procesos biológicos, como la adhesión celular, la señalización intracelular y la respuesta inmunológica. Su capacidad para interactuar con carbohidratos presentes en la superficie de células las convierte en componentes clave de la adhesión celular y en la formación de complejos moleculares. Además, las galectinas están implicadas en la progresión de enfermedades como el cáncer, la inflamación y las enfermedades autoinmunes, lo que las convierte en objetivos potenciales para el desarrollo de terapias y diagnósticos.

En el *Capítulo 3*, se lleva a cabo un análisis computacional exhaustivo de las interacciones entre las diferentes formas existentes de galectinas humanas y carbohidratos. Además, se estudia la posible comunicación alostérica presente en estos sistemas y las contribuciones termodinámicas de la solvatación utilizando simulaciones de dinámica molecular. Este estudio proporciona información detallada sobre los aspectos estructurales, dinámicos y energéticos de estas interacciones, destacando la importancia de entender las redes de comunicación alostérica y la naturaleza entálpica o entrópica de las interacciones clave. Así, se ha caracterizado cómo las diferencias en la composición de los sitios de unión de las galectinas son en última instancia responsables de sus especificidades, determinando la estructura y dinámica de la microsolvatación local y la termodinámica general de la hidratación. Estos hallazgos contribuyen al aumento del conocimiento del papel de las zonas de agua conservadas y de los efectos alostéricos, ofreciendo información crucial para el diseño de fármacos o intervenciones terapéuticas.

DC-SIGN es una proteína expresada principalmente en las células dendríticas, que son componentes clave del sistema inmunológico. DC-SIGN juega un papel esencial en la interacción entre las células dendríticas y los patógenos, como virus y bacterias, a través de la unión a carbohidratos específicos presentes en la superficie de estos microorganismos. Esta interacción desencadena una

respuesta inmunológica adaptativa que es fundamental para combatir las infecciones. La capacidad de DC-SIGN para reconocer y unir patógenos lo convierte en un objetivo importante en la investigación de vacunas y terapias antivirales. Además, su papel en la modulación de la respuesta inmunológica también lo hace relevante en la comprensión de enfermedades autoinmunes y procesos inflamatorios. En conjunto, tanto las galectinas como DC-SIGN ejemplifican la importancia crítica de las proteínas que reconocen y se unen a carbohidratos en una variedad de procesos biológicos y médicos. DC-SIGN muestra una notable promiscuidad en la unión a carbohidratos. Esta amplia capacidad de unión de diferentes carbohidratos plantea desafíos en la comprensión de las especificidades y las implicaciones funcionales de estas interacciones. Investigar la base molecular de la promiscuidad de DC-SIGN y su relevancia en las interacciones huésped-patógeno es un área de investigación activa con posibles aplicaciones en el desarrollo de terapias y vacunas dirigidas a respuestas inmunológicas mediadas por DC-SIGN.

En el caso de lectinas de tipo C, como DC-SIGN, el proceso de reconocimiento de carbohidratos implica la coordinación de un catión calcio por dos grupos hidroxilo situados en posiciones vecinales en el monosacárido, lo que resulta en una interacción relativamente débil. Sin embargo, se pueden lograr afinidades más altas mediante contactos adicionales que involucran residuos situados fuera del sitio activo. En el *Capítulo 4* se estudia el reconocimiento de monosacáridos por DC-SIGN, utilizando diversas técnicas computacionales incluyendo mecánica cuántica y dinámica molecular y validando los resultados con experimentos de RMN y cálculos de CORCEMA-ST, con el propósito de comprender los mecanismos moleculares involucrados en el proceso de reconocimiento. En este estudio se presenta el patrón estructural óptimo de menor tamaño que debe presentar un ligando cíclico de seis miembros con múltiples grupos hidroxilo para ser reconocido por la lectina humana DC-SIGN. Basándonos en estos resultados, se ha explorado por primera vez la interacción entre diversos ligandos no nativos que presentan este motivo estructural y DC-SIGN. El conocimiento obtenido sobre las preferencias de unión de los distintos ligandos y las interacciones moleculares que las gobiernan podrían tener aplicación en el campo del reconocimiento molecular y el diseño de medicamentos con capacidad de inhibir esta lectina y dianas similares.

Además de abordar las interacciones proteína-ligando desde una perspectiva fundamental, esta tesis también se adentra en aplicaciones prácticas del diseño de proteínas. El diseño de proteínas es un campo de estudio apasionante que se encuentra intrínsecamente vinculado al desafío del plegamiento proteico. Este es un desafío central en la biología molecular que ha intrigado a la comunidad científica durante mucho tiempo. Implica comprender cómo una proteína, que se sintetiza inicialmente en el ribosoma como una cadena lineal de aminoácidos,

logra plegarse en una estructura tridimensional única y crucial para su función biológica. La información necesaria para identificar la estructura plegada de una proteína, también conocida como su estado nativo, está completamente contenida en la secuencia lineal de aminoácidos de la proteína. La teoría termodinámica de Anfinsen sostiene que esta información está codificada en el perfil energético de la cadena polipeptídica , poseyendo el estado nativo la energía libre más baja. Durante muchos años, este principio sentó las bases para un enfoque general en la predicción de la estructura de proteínas, que implica muestrear diversas conformaciones, evaluarlas en función de la energía y finalmente identificar la conformación con el estado de energía más bajo. El proceso de plegamiento se basa en la necesidad de confinar residuos hidrofóbicos en el núcleo de la proteína, lejos del solvente acuoso, optimizando interacciones de van der Waals y enlaces de hidrógeno intra-proteína, lo que resulta en una estructura tridimensional altamente específica y funcional.

La diversidad de funciones de las proteínas, que abarcan desde catalizar reacciones bioquímicas hasta proporcionar soporte estructural y regular procesos celulares, está intrínsecamente ligada a sus estructuras tridimensionales únicas, determinadas por su secuencia lineal de aminoácidos. Durante décadas, los biólogos estructurales han seguido el paradigma secuencia-estructura-función, que asume que secuencias de proteínas similares generan estructuras y funciones similares. Sin embargo, el crecimiento exponencial de secuencias de proteínas disponibles ha superado las capacidades de los métodos experimentales tradicionales para determinar estructuras proteicas; por ello, en los últimos años se han producido grandes avances en la predicción de estructuras proteicas mediante algoritmos de *deep learning*, ejemplificados por programas como AlphaFold y RoseTTAFold. Estos desarrollos han revelado que secuencias diferentes pueden plegarse en estructuras similares, destacando la importancia de la detección de homología lejana y los métodos de reconocimiento de plegamiento. Esto implica reconocer similitudes en las secuencias de aminoácidos entre proteínas aparentemente no relacionadas, indicando que podrían compartir una estructura tridimensional similar. Este enfoque destaca la necesidad de utilizar métodos específicos de reconocimiento de plegamiento diseñados para predecir cómo se pliegan las proteínas en su estructura tridimensional con base en la secuencia de aminoácidos.

El diseño de proteínas, a menudo descrito como el problema inverso a la predicción de estructuras de proteínas, se centra en inferir una secuencia de aminoácidos que estabilice la conformación deseada de una proteína en lugar de encontrar la conformación más estable para una secuencia de proteína específica. Este campo se divide en dos grupos principales: la ingeniería de proteínas y el diseño *de novo*. La ingeniería de proteínas se centra en modificar proteínas

existentes para mejorar sus capacidades y explorar nuevas funciones, buscando resolver desafíos como la insuficiente estabilidad térmica, la deficiente estabilidad en condiciones adversas, y la baja expresión, actividad y especificidad, y que suelen presentar las proteínas naturales. Por otro lado, el diseño *de novo* comienza con requisitos preestablecidos, como una forma o función deseada, sin depender de proteínas existentes. Dicho diseño se logra mediante algoritmos computacionales y principios estructurales con el fin de crear proteínas desde cero, abriendo posibilidades para diseñar proteínas con funciones a la carta, que no se conocen en la naturaleza.

En la segunda parte de la presente tesis doctoral, dos proyectos ilustran la versatilidad del diseño de proteínas en la modificación y optimización de propiedades con aplicaciones diversas. En un caso, se ha abordado la mejora de la termoestabilidad de la proteína frataxina humana (*Capítulo 5*), mientras el otro (*Capítulo 6*) se centra en el diseño de variantes de la enzima TEV con solubilidad, expresión, estabilidad y actividad mejoradas.

Las terapias basadas en proteínas presentan ventajas significativas en comparación con los fármacos basados en moléculas pequeñas, ya que ofrecen una mayor potencia y una gama más diversa de funciones, como biocatálisis, señalización y transporte. Además, dado que han evolucionado para desempeñar roles altamente especializados, suelen inducir menos efectos secundarios. Sin embargo, el éxito de las terapias basadas en proteínas a menudo se ve limitado por la estabilidad de las proteínas terapéuticas. La disminución de la estabilidad termodinámica *in vitro* a menudo se correlaciona con la degradación temprana de la proteína terapéutica en el organismo debido a su despliegue prematuro y posterior degradación en el proteosoma u otras vías de eliminación proteica. Este desafío puede superarse mediante el diseño de variantes de proteínas que mantengan conformaciones plegadas estables a temperaturas fisiológicas. Para lograrlo, es fundamental que las variantes diseñadas sean semejantes a sus contrapartes naturales con el fin de conservar la función nativa y evitar respuestas inmunitarias. En particular, la estabilidad termodinámica y la solubilidad se pueden mejorar mediante ingeniería de la secuencia de aminoácidos, asegurando la estabilidad y solubilidad de la proteína. Estas estrategias son fundamentales para mejorar la eficacia terapéutica y avanzar en el campo de la terapia génica y proteica.

La ataxia de Friedreich es una enfermedad autosómica recesiva que afecta a pacientes jóvenes y se caracteriza por la disminución de la frataxina, una proteína crucial para la función mitocondrial. Los pacientes con esta enfermedad muestran una deficiencia en la frataxina debido a mutaciones genéticas. En el *Capítulo 5* se describe cómo el uso de herramientas de inteligencia artificial, como AlphaFold y ProteinMPNN, ha permitido predecir con éxito cambios en la

termoestabilidad de las proteínas mediante la ingeniería de su secuencia. En este estudio, se han diseñado variantes de frataxina que han demostrado una mayor termoestabilidad y resistencia a la degradación, lo que las convierte en prometedores candidatos para terapia de reemplazo de proteínas en la ataxia de Friedreich. Este enfoque racional podría tener aplicaciones más amplias en el diseño de proteínas con propiedades mejoradas para usos terapéuticos y biotecnológicos.

Por otro lado, TEV es una proteasa derivada del virus del tabaco ampliamente utilizada en aplicaciones biotecnológicas debido a su capacidad para realizar un corte específico entre glutamina y serina dentro de su secuencia de reconocimiento (ENLYFQ/S). Este proceso de corte se utiliza frecuentemente para eliminar etiquetas de purificación de proteínas producidas por recombinación. Sin embargo, a pesar de su interés, el uso de esta enzima a menudo presenta desafíos significativos en términos de solubilidad, estabilidad térmica y eficiencia catalítica subóptima, lo que puede llevar a tiempos de incubación prolongados e incompletos procesos de proteólisis.

En el *Capítulo 6*, se describe el uso de técnicas de diseño de proteínas basadas en el *deep learning*, en particular ProteinMPNN, para optimizar la solubilidad de la proteasa TEV y abordar sus limitaciones inherentes. Tomando como punto de partida una variante de la proteasa TEV conocida como TEVd que ha demostrado resistencia a la autólisis, se han diseñado variantes que muestran una mejora significativa tanto en la solubilidad, la estabilidad y la eficiencia catalítica de la proteasa TEV, allanando el camino para aplicaciones biotecnológicas más eficientes y efectivas.

Así, el uso combinado de información evolutiva y modelos de d*eep learning* para el diseño de secuencias (ProteinMPNN) y predicción de estructura (AlphaFold) de proteínas ha demostrado ser extremadamente exitoso para la mejora de propiedades en proteínas muy diferentes, como son el caso de frataxina y la proteasa TEV. La selección racional de posiciones a mutar basada en información estructural y funcional, y su posterior muestreo con ProteinMPNN ha sido garantía de éxito en el diseño, tal y como se demuestra por los altos valores de predictibilidad de los modelos generados mediante AlphaFold, y especialmente por los resultados experimentales obtenidos.

Con esta metodología se han logrado mejoras en el rendimiento de expresión de proteínas y en su estabilidad (tanto termodinámica como proteolítica), que son cruciales para su potencial uso como herramientas biotecnológicas o agentes terapéuticos. Esta estrategia racional basada en *deep learning* constituye un enfoque prometedor para el diseño y la ingeniería de proteínas, pudiendo servir como complemento a técnicas de uso común como la mutagénesis aleatoria o la evolución dirigida.

Abstract

Protein-ligand recognition is a central to molecular biology, governing essential cellular functions and mediating various biological processes. At its core lies the remarkable ability of proteins to selectively bind small molecules, known as ligands. From regulating cellular responses to driving metabolic reactions, protein-ligand recognition plays a central role in life's chemistry. Solvation, particularly water behavior, is vital in protein-ligand interactions. Water, which is crucial at the biomolecular interface, participates in promiscuous and specific hydrogen bonds. It significantly contributes to binding energy through desolvation, reorganization, and displacement during complex formation.

In its first part, this doctoral thesis focuses on carbohydrate recognition by lectins, specifically galectins and DC-SIGN. The study uses a diverse range of computational techniques, including quantum mechanics calculations, molecular dynamics, conformational analysis, and solvation analysis, supported by nuclear magnetic resonance experiments.

Galectins, a protein family known for their specific binding to β-galactosides, play a crucial role in various biological processes such as cell adhesion, intracellular signaling, and immune response. Their ability to interact with carbohydrates on cell surfaces makes them key components in cellular adhesion and molecular complex formation. Galectins have also been implicated in diseases like cancer, inflammation, and autoimmune disorders, making them potential targets for therapies and diagnostics.

In *Chapter 3*, a comprehensive computational analysis is conducted on the interactions between different forms of human galectins and carbohydrates. The study explores potential allosteric communication and solvation thermodynamics using molecular dynamics simulations. Detailed insights into the structural, dynamic, and thermodynamic aspects of these interactions are provided, offering a deeper understanding of the role of conserved water regions and long-distance movements.

Differences in the binding sites of galectins ultimately dictate their specificities, influencing local microsolvation dynamics and overall hydration thermodynamics. These findings may prove valuable for future drug design or therapeutic interventions.

DC-SIGN is a protein primarily expressed in dendritic cells, crucial components of the immune system. It plays an essential role in the interaction between dendritic cells and pathogens, such as viruses and bacteria, by binding to specific carbohydrates on the surface of these microorganisms. This interaction triggers an adaptive immune response vital for combating infections. DC-SIGN's ability

to recognize and bind pathogens makes it a significant target in vaccine research and antiviral therapies. Additionally, its role in modulating the immune response is relevant to understanding autoimmune diseases and inflammatory processes.

Both galectins and DC-SIGN exemplify the critical importance of proteins recognizing and binding to carbohydrates in various biological and medical processes. DC-SIGN exhibits notable promiscuity in carbohydrate binding, posing challenges in understanding the specificities and functional implications of these interactions. Investigating the molecular basis of DC-SIGN's promiscuity and its relevance in host-pathogen interactions is an active research area with potential applications in developing therapies and vaccines targeting DC-SIGN-mediated immune responses.

In the case of C-type lectins like DC-SIGN, the carbohydrate recognition process involves calcium coordination by two neighboring hydroxyl groups in the monosaccharide, resulting in a relatively weak interaction. However, higher affinities can be achieved through additional contacts involving residues outside those conserved in secondary sites. *Chapter 4* focus on monosaccharide recognition by DC-SIGN, using different computational techniques such as quantum mechanics and molecular dynamics, and validating predictions through NMR experiments and CORCEMA ST calculations. The goal is to understand the molecular mechanisms involved in carbohydrate recognition by DC-SIGN and its potential applications in future research and applications.

The study introduces the optimal structural pattern that a six-membered cyclic ligand with multiple hydroxyl groups must have to be recognized by the human lectin DC-SIGN. Building upon these findings, the interaction between diverse ligands and DC-SIGN is explored for the first time. Insights into binding preferences and molecular interactions could have broad implications in the field of molecular recognition and drug design, potentially guiding future research.

In addition to addressing protein-ligand interactions from a fundamental perspective, the second part of this doctoral thesis delves into practical applications of protein design. Protein design is an exciting field intricately linked to the challenge of protein folding, a central issue in molecular biology that has intrigued the scientific community for a long time. It involves understanding how a protein, initially synthesized as a linear chain of amino acids, folds into a unique three-dimensional structure crucial for its biological function.

The diverse protein functions, from catalyzing biochemical reactions to providing structural support and regulating cellular processes, are inherently

linked to their unique three-dimensional structures, determined by their linear amino acid sequences. Structural biologists have traditionally followed the sequence-structure-function paradigm, assuming that similar protein sequences generate similar structures and functions. However, the exponential growth of available protein sequences has surpassed the capabilities of traditional experimental methods for determining protein structures, leading to significant advances in protein structure prediction through deep learning algorithms, exemplified by programs like AlphaFold and RoseTTAFold. These developments have revealed that different sequences can fold into similar structures, highlighting the importance of detecting distant homology and folding recognition methods.

Protein design, often described as the inverse problem of protein structure prediction, focuses on inferring an amino acid sequence that stabilizes the desired conformation of a protein rather than finding the most stable conformation for a specific protein sequence. This field is divided into two main groups: protein engineering and *de novo* design. Protein engineering aims to modify existing proteins to enhance their capabilities and explore new functions. It seeks to address challenges such as marginal thermal stability, low expression, low activity and specificity, and deficient stability under adverse conditions of natural proteins. On the other hand, *de novo* design starts with pre-established requirements, such as a desired shape or function, without relying on existing proteins. This is achieved through computational algorithms and structural principles to create proteins from scratch, opening possibilities for designing proteins with custom functionalities that may not exist in nature.

In this thesis, two projects illustrate the versatility of protein design in modifying and optimizing properties with diverse applications. In one case, the challenge of improving the thermostability of the Frataxin protein is addressed (*Chapter 5*), while the other (*Chapter 6*) focuses on designing variants of the TEV enzyme with enhanced solubility, expression, stability, and activity.

Protein-based therapies offer significant advantages over small molecule drugs, providing greater potency and a more diverse range of functions, such as catalysis, signaling, and transport. Moreover, as they have evolved to play highly specialized roles, they often induce fewer side effects. However, the success of protein-based therapies is often limited by the stability of therapeutic proteins. The decrease in *in vitro* thermodynamic stability often correlates with early degradation of therapeutic proteins in the body due to premature unfolding and subsequent degradation in the proteasome or other protein elimination pathways. This challenge can be overcome by designing protein variants that maintain stable folded conformations at physiological temperatures. To achieve this, it is crucial for the designed variants to closely

mimic their natural counterparts to preserve native function and avoid immune responses. Thermodynamic stability and solubility can be improved through amino acid sequence engineering, ensuring the stability and solubility of the protein. These strategies are essential for improving therapeutic efficacy and advancing the fields of gene and protein therapy.

Friedreich's ataxia is an autosomal recessive condition affecting young patients and characterized by decreased Frataxin, a protein crucial for mitochondrial function. The use of artificial intelligence tools, such as AlphaFold, has successfully predicted changes in protein thermostability through sequence engineering. In *Chapter 5*, Frataxin variants have been designed to exhibit increased thermostability and resistance to degradation, making them promising candidates for protein replacement therapy in Friedreich's ataxia. This rational approach could have broader applications in designing proteins with enhanced properties for therapeutic and biotechnological uses.

On the other hand, TEV is a widely used protease in biotechnological applications due to its ability to perform specific cleavage between glutamine and serine within its recognition sequence (ENLYFQ/S). This cleavage process is used to remove purification tags from recombinantly produced proteins. However, despite its interest, the use of this enzyme often presents significant challenges in terms of solubility, thermal stability, and suboptimal catalytic efficiency, leading to prolonged incubation times and incomplete cleavage processes.

In *Chapter 6*, protein design techniques based on deep learning, particularly ProteinMPNN, are applied to optimize the solubility of the TEV protease and address its inherent limitations. Starting with a variant of the TEV protease known as TEVd, which has shown resistance to autolysis, variants have been designed that exhibit significant improvement in the solubility, stability, and catalytic efficiency of the TEV protease, paving the way for more efficient and effective biotechnological applications.

The combined use of evolutionary information and deep learning models for sequence design (ProteinMPNN) and protein structure prediction (AlphaFold) has proven to be extremely successful for property enhancement in very different proteins, as exemplified by Frataxin and the TEV protease. Rational selection of positions to mutate based on structural and functional information, followed by sampling with ProteinMPNN, has been a guarantee of success, as demonstrated by the high predictability values in AlphaFold models and especially by the results of the presented experimental assays.

This methodology has achieved improvements in protein expression performance and stability (both thermodynamic and proteolytic), crucial for their potential use as biotechnological tools or therapeutic agents. The presented rational deep learning-based approach constitutes a promising strategy for protein design and engineering, serving as a complement to powerful techniques such as random mutagenesis or directed evolution.

Chapter 1

Introduction

Insights into Protein-Carbohydrate Recognition Processes

1. Insights into Protein-Carbohydrate Recognition Processes

1.1 Protein-ligand recognition

Protein-ligand recognition is a crucial phenomenon in the realm of molecular biology, governing essential cellular functions and mediating different biological processes. At the core of this interaction lies the remarkable ability of proteins to selectively bind small molecules known as ligands. From the intricate signaling cascades that regulate cellular responses to the finely tuned enzymatic reactions that drive metabolism, protein-ligand recognition plays a central role in life's chemistry.

At low concentrations, a protein and a ligand can rotate and translate independently without running into one another. The association between the two molecules only takes place after a collision if their mutual attraction is stronger than their mobility. This implies that the energy gained from binding should exceed the energy lost due to restricted molecular movement. With an increase in temperature, the kinetic energy and the frequency of collisions rise, requiring the collision energy to surpass the kinetic energy for successful binding. The hydrophobic effect, which refers to the tendency of non-polar parts of the protein, especially certain amino acid chains, to group together, further enhances complex formation through strong interactions among solvent molecules in solution [1].

Upon association, the protein and the ligand experience constrained movement, losing their independent rotational and translational freedom. Additionally, vibrational motion emerges between the molecules. These increased constraints lead to a redistribution of kinetic energy. However, the loss of rotational and translational freedom is partly compensated by the residual motion of the complex, allowing for some remaining movement. As the strength of the interaction between the molecules increases and the vibrational motion become less flexible, the residual vibration diminishes, ultimately leading to covalent bonds in some extreme cases [2].

a. Enthalpy-entropy compensation effect

The binding free energy (ΔG_b) for a ligand-target complex at a given temperature is determined by Equation 1:

$$\Delta G_b = \Delta H_b - T\Delta S_b \qquad \text{(Equation 1)}$$

In the context of protein-ligand recognition, this equation reflects an intrinsic and temperature-dependent enthalpy-entropy compensation. Specifically, as the complex formation becomes tighter, there is a notable loss of mobility, as indicated by the opposing entropic ($T\Delta S_b$) and enthalpic (ΔH_b) terms in the equation.

This phenomenon covers the correlation and offsetting nature of changes in enthalpy (ΔH_b) and entropy ($T\Delta S_b$), thereby maintaining a relatively constant free energy change (ΔG_b). Enthalpy-entropy compensation is a prevalent occurrence observed in almost all weak intermolecular associations, with hydrogen bonding in aqueous solutions being a common example [3]. The extent of compensation is limited by two factors:

i) The entropy loss caused by the complete immobilization of molecules. To mitigate this entropic loss, a common strategy is to design a ligand that is intrinsically more rigid. By minimizing the entropic loss, it becomes easier for the complex to overcome this limitation.

ii) The enthalpic gain from new noncovalent interactions formed in the complex, which is the most common strategy used in drug development by maximizing the number and strength of polar contacts, hydrophobic interactions, salt bridges, etc.

In a thermodynamically favorable protein-ligand binding process, the combined contributions of entropic and enthalpic changes must outweigh any interactions between the solvent and the solute.

In protein-ligand recognition processes, enthalpy-entropy compensation can occur in numerous scenarios. Some examples are:

- *Ligand-induced conformational changes*
  Binding of a ligand leads to structural rearrangements in the protein, balancing enthalpic gains from new interactions with entropic loss due to increased rigidity [4].

- *Solvent reorganization*
  Ligand binding causes solvent molecules to rearrange around the complex, resulting in enthalpic gain or loss and potential changes in the system's entropy [5].

- *Hydrophobic interactions*
  Hydrophobic regions of the ligand and the protein come together upon binding, minimizing contact with bulk water molecules, resulting in

favorable enthalpic gain but potentially entropic loss due to ordered water molecules at the interface [6].

- *H-bonding interactions*
  The formation or breaking of hydrogen bonds between the ligand and the protein contributes to enthalpic changes, while the rearrangement of water molecules involved in these interactions can affect entropy [7–9].

- *Ligand flexibility*
  Flexible ligands can adapt their conformation to fit the binding site, gaining entropy, but the restriction of ligand flexibility upon binding may lead to entropic losses [4].

b. Solvation effects

Considering the protein-ligand recognition process, the influence of solvation, particularly the behavior of water molecules, holds immense significance alongside the roles of the protein and the ligand themselves [10]. Water, an integral component at the biomolecular interface, possesses remarkable versatility that profoundly impacts ligands binding. Acting as both hydrogen bond donor and acceptor, water plays a dual role in facilitating bond formation. Despite occupying less space compared to the polar side chains of a protein, water molecules participate in multiple hydrogen bonds. This adaptability confers promiscuous binding capabilities to surfaces, since water molecules possess a degree of flexibility that allows them to form interactions with different types of ligands. However, water also possesses the ability to confer exquisite specificity and increased affinity to interactions, enabling precise and high-affinity binding [11].

Water, the universal solvent of life, plays a crucial role in ligand binding, as evidenced by the significant influence of the hydrophobic effect and the frequent occurrence of water-mediated interactions between proteins and ligands within binding sites [1,12]. The processes of ligand association with its binding partner necessitates, at least, the partial desolvation of the ligand, the displacement of water molecules from the binding site, and the reorganization of water in the surrounding vicinity (Fig. 1). On a molecular scale, the hydration status of the two binding partners undergoes specific changes during the binding process, affecting only a small number of water molecules directly. However, despite their limited presence, water molecules play a significantly role in binding energetics, making substantial contributions to the overall stability and affinity of the complex [13].

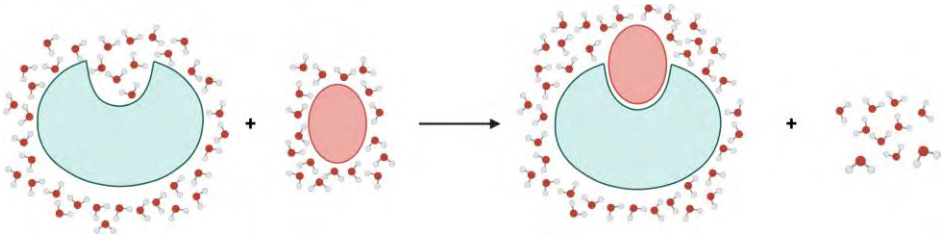


Figure 1. Schematic representation of the solvation/desolvation mechanism in protein-ligand recognition processes. The figure represents how water molecules play a crucial role in hydrating the binding site of the protein (green). Upon ligand (red) binding, water molecules are displaced, leading to the desolvation of the binding site and creating a favorable environment for protein-ligand interactions.

When considering the displacement of water molecules from an active site in favor of a newly introduced non-covalent interaction, several factors come into play to determine its favorability. This displacement is only favorable if the benefits outweigh the cost of breaking the water molecule's hydrogen bonds with the active site. This involves considering the gain in ligand-protein enthalpy and the gain in solvent enthalpy and entropy. This phenomenon represents a complex form of enthalpy-entropy compensation and is a common challenge in ligand design to achieve strong binding [14].

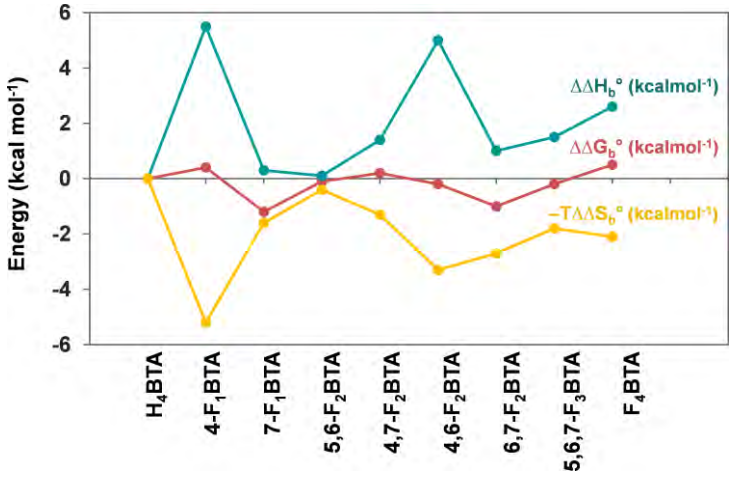


Figure 2. Relative binding enthalpy ($\Delta\Delta H_b°$, in green dots), entropy ($-T\Delta\Delta S_b°$, in yellow), and free energy ($\Delta\Delta G_b°$, in red) measured experimentally by isothermal titration calorimetry (ITC) for a series of fluorinated ligands to human carbonic anhydrase. The observed H/S compensation effect is strongly influenced by the structural and thermodynamic properties of water surrounding the bound ligands. Figure adapted from reference [5].

Hence, understanding the effect of solvation and the role of water in protein-ligand recognition processes is vital for comprehending the intricate energetics and optimizing ligand design strategies.

c. Allostery

Proteins exhibit inherent dynamics that are essential for their functional roles, allowing them to undergo significant conformational transitions upon encountering external stimuli, such as ligand binding. The process of ligand binding can induce alterations in the protein's structure, and due to their densely packed nature, these changes can trigger allosteric effects at distant sites from the binding site. This allosteric communication within a protein is crucial for facilitating essential biochemical processes. Investigating the protein's allosteric behavior resulting from ligand binding holds significance in understanding the thermodynamics of binding, as these remote conformational flexibilities account for a contribution to entropy that has often been overlooked in the study of protein-ligand interactions [15].

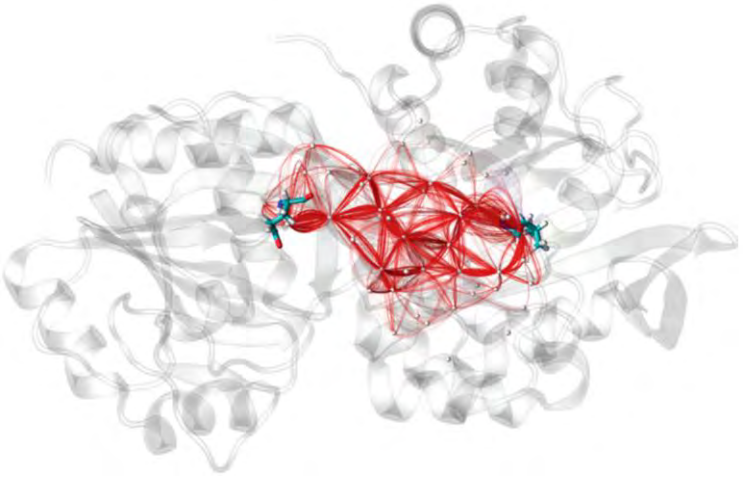


Figure 3. Allosteric pathways between residue Leu50 of HisF and Glu180 of imidazole glycerol phosphate synthase subunit HisH calculated with WISP software. Figure adapted from reference [16]

1.2 Importance of carbohydrate recognition by proteins

The recognition of carbohydrates by proteins holds significant importance in the field of protein-ligand recognition. While conventional protein-ligand interactions predominantly focus on small organic molecules, particularly in the field of drug discovery, the involvement of carbohydrates introduces a distinct dimension to these recognition events.

Carbohydrate-binding proteins, such as lectins, play a crucial role in cell adhesion [17], signaling events [18], host-pathogen interactions [19], cancer development [20], and many more. In numerous cases, the interaction between proteins and carbohydrates is not limited to a singular event but instead acts as the initial step within a broader sequence of interconnected events and interactions [17, 18]. These subsequent processes can give rise to intricate signaling cascades, underscoring the multi-faceted nature of protein-carbohydrate interactions.

Pathogens, including viruses, bacteria, parasites, and fungi, present carbohydrates on their surfaces. These glycans serve as the initial point of interaction with host cells, making them potential targets for interventions aimed at preventing infection [23]. Carbohydrate-protein interactions also play a pivotal role in the immune response and pathogen recognition. Lectins recognize specific carbohydrate patterns on pathogens, facilitating their clearance and initiating immune response through binding carbohydrate-containing antigens [21]. Thus, carbohydrate-protein interactions contribute to diverse immune processes, including both infection responses and defense mechanisms.

The implications of carbohydrate recognition in protein-ligand recognition extend beyond fundamental molecular understanding, offering potential targets for therapeutic interventions. Understanding the intricate recognition and binding events between carbohydrates and target proteins provides opportunities to design therapeutic agents that can effectively modulate such interactions.

1.3 Importance of lectins

Lectins are sugar-binding proteins that lack catalytic function (they are not enzymes) and do not directly trigger immune responses (they are not antibodies). However, the specificity observed in the binding between lectins and specific carbohydrates can be comparable to that of antigen-antibody or substrate-enzyme interactions [24]. The recognition of carbohydrates by lectins is widely acknowledged for its crucial role in diverse biological processes, underscoring its importance. These proteins possess an extraordinary ability to specifically bind carbohydrates present on cell surfaces. This recognition process facilitates crucial interactions involved in cell adhesion, immune response modulation, and pathogen recognition.

In the realm of plants, lectins can be classified into 12 distinct families, while in animals, there are a minimum of 14 families of lectins [25]. Among the various

lectin families, C-type, I-type and S-type lectins are particularly relevant in humans. Of these, C-type lectins, which rely on calcium (Ca^{2+}) for their function, are found in both transmembrane and soluble protein forms. Certain lectins within this family, such as DC-SIGN, langerin and MGL, play critical roles in pathogen recognition and have emerged as targets in the field of drug discovery [26]. Currently, I-type lectins, particularly sialic acid binding immunoglobulin-type lectins (Siglecs), have become a subject of significant interest due to their critical roles in immune regulation [27]. Galectins, earlier referred to as S-type lectins, are a family of proteins that specifically binds to β-galactosides. This family consists of 16 members and is widely distributed throughout the human body [28].

Lectins can bind to free sugars as well as sugar residues present in polysaccharides, glycoproteins, or glycolipids, which can exist either in free form or as part of cell membranes. Nevertheless, it is important to note that lectin-monosaccharide interactions typically exhibit relatively weak affinities, with dissociation constant frequently falling within the micromolar to millimolar range [29–31]. This is in part due to the shallow binding pockets on the surface of lectins, as they are exposed to competitive interactions with the solvent. In biological systems, multivalency comes into play to overcome the challenge of low-affinity binding by facilitating simultaneous synergistic interactions between the receptor and the ligand [32–34]. Multivalent presentation provides a range of mechanisms to enhance affinity, such as chelation, subsite binding, steric stabilization, statistical rebinding and clustering effects, among others [35–39].

Lectin bind glycans at the CRD (carbohydrate recognition domain) in a noncovalent manner. Carbohydrates engage in interactions with lectins through hydrogen bonds, metal coordination, van der Waals, and hydrophobic interactions [40]. The abundance of hydroxyl groups on carbohydrates makes them well-suited for participating in complex networks of hydrogen bonds. These hydrogen bonds are typically cooperative, with the hydroxyl group serving as both a donor and an acceptor. Amino acid side chains like aspartic acid (Asp), asparagine (Asn) and histidine (His), as well as main-chain amide hydrogens and carbonyl oxygen, commonly contribute to these hydrogen bonds, while other amino acids are less frequently involved. The hydrogen bonds between proteins and carbohydrates can occur directly or can be water-mediated [40–42].

Divalent cations such as Ca^{2+} and Mn^{2+} play important roles in carbohydrate recognition. They can indirectly shape the binding site, as seen in legume lectins [40], or directly bind to the carbohydrate as in the C-type lectins.

Despite the high polarity of carbohydrates, the arrangement of hydroxyl groups creates hydrophobic patches on their surfaces that can interact with hydrophobic regions in protein residues. Notably, aromatic amino acids such as phenylalanine (Phe), tyrosine (Tyr) or tryptophan (Trp) can engage in stacking interactions with monosaccharides [43].

Due to their mainly uncharged nature, saccharides typically do not form complexes with proteins through ionic interactions. However, there are exception such as the heparin-antithrombin III complex [44].

1.4 Computational characterization of lectin-carbohydrate recognition processes

In the recent years, experimental studies using techniques such as X-ray crystallography and NMR spectroscopy have provided detailed insights into lectin-carbohydrate interactions. Moreover, the use of different experimental methods has enabled the determination of thermodynamic and kinetic parameters associated with the binding process. Alongside experimental data, computational methods have been developed in the field of Molecular Modeling and Computational Chemistry. These methods incorporate diverse multi-scale techniques, leading to highly accurate models for lectin-carbohydrate interactions and resulting in high-quality predictions [45,46].

As explained previously, the binding affinity depends on the balance between enthalpic and entropic contributions, with interactions between the ligand and binding site residues stabilizing the complex. However, unfavorable contributions arise from desolvation of binding partners, entropic loss during complex formation, and conformational changes within the interacting partners. Given the highly flexible and polar nature of carbohydrates, this is particularly relevant in protein-carbohydrates interactions. Consequently, the binding affinity results from a small difference between the enthalpy and entropy values.

According to the equation $\Delta G = -RT\ln(K)$, the energetic difference between a millimolar and a nanomolar ligand (i.e. six orders of magnitude) is less than 10 kcal mol^{-1}, underscoring the necessity for theoretical models to precisely characterize the carbohydrate recognition process by lectins [47]. Such models should consider a myriad of factors, including intermolecular interactions, desolvation energies, and entropy changes, among others.

In the field of Computational Chemistry, two prominent methods, Molecular Mechanics (MM) and Quantum Mechanics (QM), offer distinct approaches to investigate protein-carbohydrates interactions (Fig. 4).

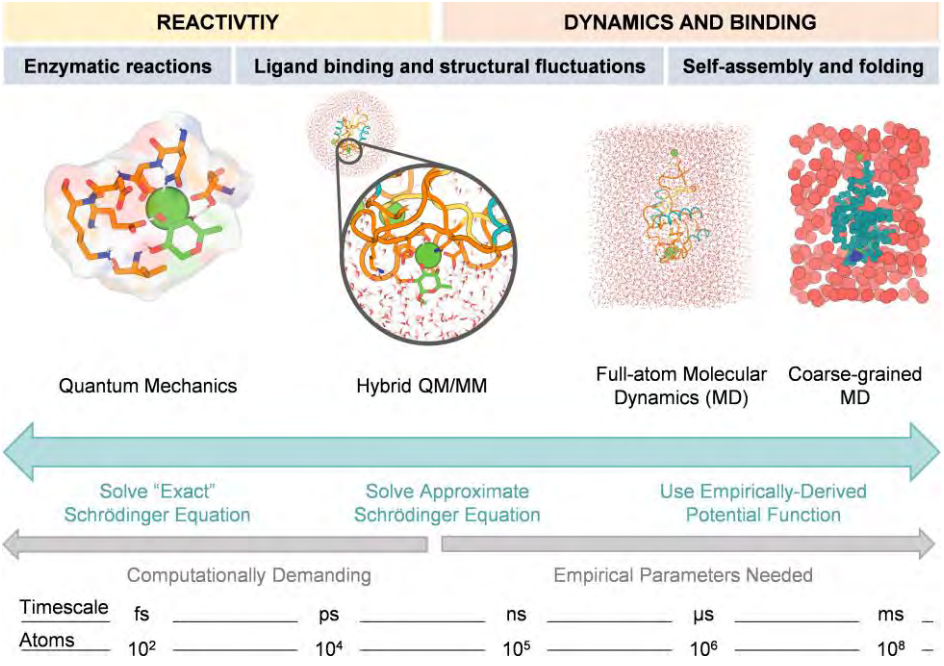


Figure 4. Computational Chemistry methods overview: from highly accurate QM methods used for studying chemical reactions to efficient MM methods suited for large-scale simulations.

a. Molecular Mechanics Methods

Remarkably, predicting *relative* affinity within a given ligand series can be sufficient for their comparative evaluation. Given the necessity of sampling different conformations, a computationally efficient treatment of the protein-ligand model becomes crucial [48]. In this regard, classical force fields and molecular mechanics (MM) are commonly employed. In simulations, ensuring that the so-called 'ensemble average' converges is crucial. This convergence refers to the stabilization of the average value of a property after exploring a broad spectrum of the system's molecular configurations. When this average stabilizes numerically, it indicates that statistically robust results can be obtained, offering a comprehensive understanding of the system. Further simulations are unlikely to substantially alter the results [49,50]. Achieving this convergence, which is necessary for several commonly used binding free energy methods, demands considerable sampling. Given their low computational cost, MM are indispensable in enabling efficient and extensive conformational sampling. This positions them as invaluable tools for studying protein-ligand interactions and obtaining accurate binding free energy predictions.

MM methods, especially time-resolved Molecular Dynamics (MD) simulations, are exceptionally well-suited for elucidating the structure and dynamics of protein-carbohydrate complexes, which is a considerable challenge since these systems are intrinsically more dynamic than many other protein-ligand complexes [51]. MM methods describe the system as point charges connected by "springs" to account for the bonds between atoms, while angles and dihedrals within the system are addressed using simple mathematical functions (i.e., potentials). MD simulations apply Newton's equations of motion to compute the trajectories of atoms as they interact with each other under the influence of the forces described by the MM potential energy function.

In MD simulations, a molecule composed of multiple atoms is represented by a set of atomic positions (r_N), forming a molecular geometry (R) at a given time (t). Successive system configurations are generated by integrating Newton's laws of motion for each atom. Key points to consider are:

- *Representation of molecular geometry*
  A molecular with N atoms is expressed as $R = (r_1, r_2, \ldots, r_N)$, where each atom has its own set of atomic coordinates

- *Newton's Second Law*
  For each atom, Newton's second law ($F = ma$) is applied, where F represents the force acting on the atom. The acceleration (a) is the second derivative of position (r) with respect to time (t). Hence, $F_i = m_i \frac{\partial^2 r_i}{\partial t^2}$ for the *ith* atom.

- *Force-position relationship*
  The force of an atom is also related to the first derivative of the potential energy.

$$F_i = -\frac{\partial V}{\partial r_i} \qquad \text{(Equation 2)}$$

- *Numerical integration*
  Analytical solutions are impractical for systems with more than two atoms. Instead, a numerical approach, such as the finite difference method, is used. Time is divided into small time steps (δt), typically on the order of femtoseconds, to solve the equations of motion at each step.

- *Predicting positions and velocity*
  At each step, forces are calculated, and from these, accelerations are determined.

$$a_i = \frac{F_i}{m_i} = -\frac{1}{m_i}\frac{\partial V}{\partial r_i} \qquad \text{(Equation 3)}$$

Using the acceleration, the positions and velocities of atoms are predicted at the next time step ($t + \delta t$).

- *Trajectory generation*
  The process is repeated for many steps, generating a trajectory that shows how particle positions and velocities change over time.

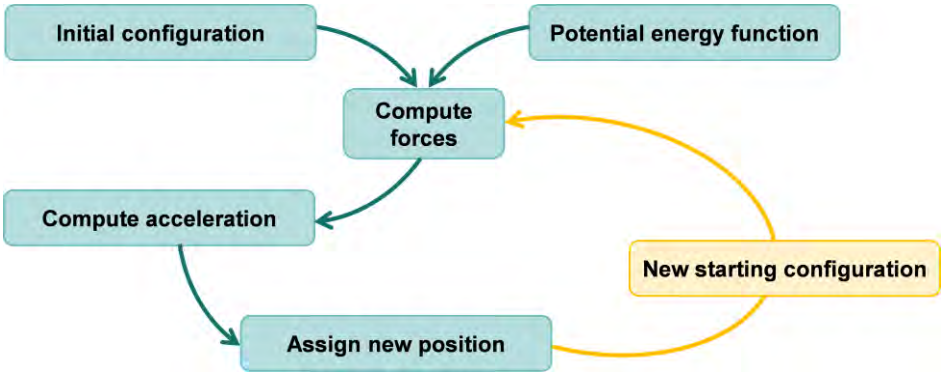


Figure 5. Simplified representation showing a Molecular Dynamics simulation protocol.

The starting point for MD simulations in the context of studying protein-ligand interactions can vary depending on the available information. When the X-ray crystallographic structure of the complex is available, starting from this often offers a more accurate representation of the actual binding conformation. However, when this is not available or when one wants to explore different conformations from the minimum energy X-ray pose, starting from docking poses provide a computationally predicted starting binding pose. Molecular docking is a highly parametrized MM technique that computationally predicts and calculates the most favorable interaction position between a ligand and a target (typically a protein), although with quite limited accuracy.

Force fields (FF) define the potential energy functions and parameters that dictate interactions between system particles. They are constructed based on the Born-Oppenheimer approximation, excluding electronic motions and focusing solely on nuclear positions. This classical mechanics approach in MD allows for the study of large systems, for which would be infeasible for quantum mechanical methods that incorporate electrons. A typical force field includes various energy components: bond lengths and angles are represented with harmonic potentials, dihedral angles utilize periodic functions, while

electrostatic and van der Waals interactions are described by Coulomb's law and Lennard-Jones potentials, respectively. A general force field function, such as the one used in AMBER, is shown below:

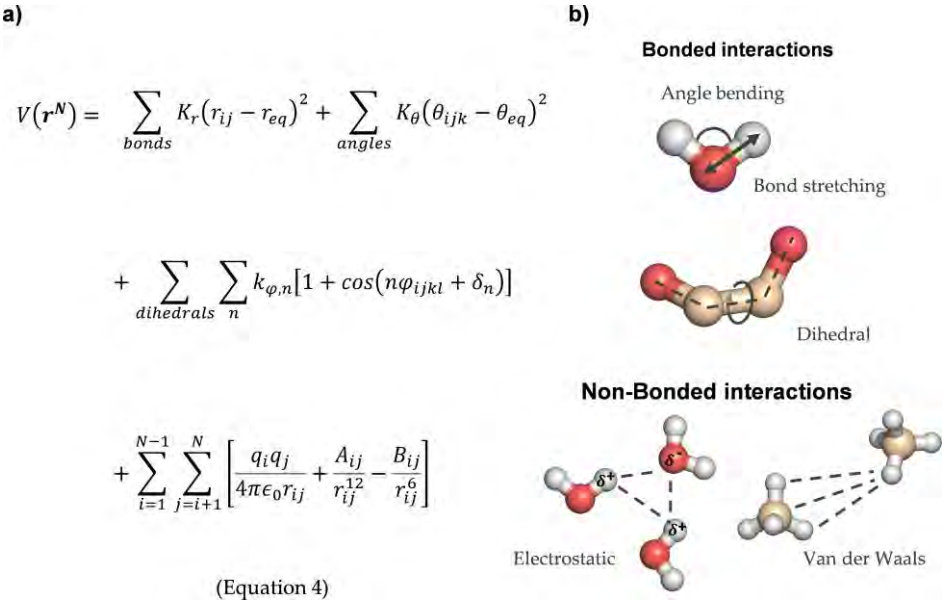


Figure 6. a) Energy function used to derive atomic forces. b) Representation of the 4 basic terms typically found in molecular force fields.

where r^N denotes the positions of all system particles. The first summation term expands to all covalent interactions between pairs of linked atoms *ij*. The second sum is related to the bond angle among groups of consecutive atoms *ijk*, while the third represents the cumulative torsional angles determined by groups of *ijkl* atoms. The fourth term, which requires the most computational effort to determine, relates to both intra and inter-molecular non-bonded interactions.

The derivation of parameters for each contribution can be achieved through both experimental and *in-silico* methods. The primary objective is to establish parameters that reflect experimental results. Determining parameters for every atom in a molecule represents a significant computational challenge and is often impractical. To address this challenge, existing general-purpose parameters might be utilized to characterize various systems. One of the key characteristics of MM is that, within a particular force field, parameters are considered transferable. This is based on the understanding that the same functional group might appear in different molecules and its response is believed to be largely consistent across diverse settings.

Force fields in computational chemistry are not universally applicable or inherently 'correct'. Their effectiveness is measured by their performance in specific scenarios. While some may excel in particular systems, they might fail in others. Selection often depends on the balance between accuracy and computational feasibility. Presently, there are force fields tailored for specific molecular classes, such as those designed exclusively for proteins, optimizing the balance between specificity and computational efficiency.

Numerous FFs are available for modeling various molecular systems under different contexts, as outlined in Table 1. Within this dissertation, all MD simulations were conducted using the AMBER suite dedicated to biomolecular simulations. The GAFF2, ff14SB, and GLYCAM_06j force fields were used to parametrize organic molecules (excluding non-modified sugars), proteins, and carbohydrates, respectively.

Table 1. Some examples of force fields employed in computational simulations. The force fields shown here are periodically updated and extended.

Force field	Application
ff14SB	Proteins
GAFF	Small organic molecules
OPLS	Small organic molecules
GLYCAM_06	Carbohydrates
LIPID17	Lipids
CHARMM36	Proteins, DNA, RNA, lipids
AMOEBA	Polarizable FF
MARTINI	Coarse-grained proteins, lipids, carbohydrates, nanoparticles
REF15	Proteins, DNA, RNA, small molecules

Apart from studying the dynamic behavior of the complex, MD simulations are also employed to study the role of solvation and investigate the dynamic effects of solvent molecules in the protein-carbohydrate recognition process. In their free states, both the carbohydrate and the protein are hydrated, and crystallographic [52] and MD simulations [53] have revealed that these water molecules often occupy the carbohydrate-binding site. These water molecules are typically found in positions similar to those of the hydroxyl groups of the bound ligand in the complex. Therefore, there is limited enthalpy gain expected upon replacing these waters with ligand hydroxyl groups.

However, significant changes in entropy could arise from displacing bound waters from each interacting surface [54,55], and entropic contributions have been identified as a major factor in determining carbohydrate-binding affinity

such as in the binding of xylooligosaccharides to non-primary subsites of *Cs*CBM6-1 protein (xylan-specific type B carbohydrate-binding module, CBM, from *Clostridium stercorarium*) [56]. Different methods based on MD simulations (e.g.: MM-GB/SA, MM-PB/SA) offer a unique tool to analyze lectin-carbohydrate interactions and their associated energies [57–60], which can greatly aid in the development of glycomimetics as potential therapeutic agents.

As stated before, due to their computational efficiency and their ability to handle large systems containing thousands of atoms, MD simulations are a powerful technique not only to estimate the relative binding affinities but also to analyze other important features of the molecular recognition event. The examination of possible conformational changes, the role of solvation, the presence of allosteric networks, the hydrogen bond networks taking place between protein and carbohydrate residues, among others, are important features that can be explored by using MD.

Despite all these advantages, MM force fields have known limitations in describing electronic effects, such as polarization, as electrons are not explicitly modeled. To achieve accurate binding free energies, accounting for ligand polarization within the protein environment is increasingly recognized as crucial [48]. Polarization has been demonstrated to substantially influence the electrostatic interactions occurring between protein and ligands [61].

Traditional force field parametrization uses fixed partial charge models that lack important physical details significantly contributing to binding free energies in some cases. Nevertheless, despite this limitation, MM force fields remain a cornerstone in the field due to their computational efficiency and broad applicability. Polarizable MM force field models such as AMOEBA have been developed to successfully address polarization effects by considering induced dipoles or induced charges. An alternative approach involves using quantum mechanical (QM) methods, which provide an explicit representation of electrons, allowing the calculation of properties reliant on the electronic distribution. However, QM calculations are computationally demanding and typically limited to studying tens of atoms, making them less feasible for larger systems and extended dispersion interactions.

b. Quantum Mechanics and QM/MM Methods

Quantum Mechanics methods offer representation of chemical bond formation and cleavage, enabling the exploration of reaction mechanisms. Specially, in the case of enzymes, QM methods are widely used to provide a precise description of the electronic interactions and changes in electronic distribution that occur during enzyme-catalyzed reactions. In these methods, typically within the

Density Functional Theory (DFT) framework, the electronic structure of a system is described by approximately solving the Schrödinger equation.

While QM methods provide a more direct representation of polarization, their much higher computational cost makes them challenging for larger systems such as proteins. Consequently, when using QM methods to model enzyme active sites, a small yet crucially selected region of the enzyme needs to be chosen for QM treatment representing the active site. This method, known as the quantum chemical cluster approach [62] has yielded valuable insights into enzymatic reactions. To account for the enzyme's local environment, peripheral atoms in the model are kept fixed based on available X-ray structures, while the model is immersed inside a dielectric cavity to incorporate polarization effects exerted by the protein. As a consequence, the quantum chemical cluster approach inevitably overlooks the detailed influence of the enzyme's environment. Furthermore, a slow convergence has been observed for the cluster approach in relation to the cluster size, with calculated energies potentially varying by more than 10 kcal mol^{-1} for models containing up to 230 atoms [63]. Evidently, a more realistic depiction of a reaction mechanism necessitates incorporating the full enzyme environment into the computations [64]. Here, the combined QM/MM approach offers a solution to the intrinsic limitations of both MM and QM. The hybrid QM/MM approach is a molecular simulation method that combines the strengths of *ab initio* QM calculations (accuracy) and MM (speed) approaches to model molecular systems (Fig. 4). The key considerations in QM/MM treatments are defining the boundaries between the QM and MM regions and determining the proper size of the QM region.

Overall, the choice between MM, QM or QM/MM methods depends on the system's size and the level of accuracy required for calculations.

Protein Design

2. Protein Design

2.1 The protein folding problem

The protein folding problem has long stood as a central challenge in molecular biology. It involves understanding how a protein, which is initially synthesized in the ribosome as a linear chain of amino acids, folds into a unique, three-dimensional structure that is crucial for its biological function. The information needed to identify a protein's folded structure (its native state), according to refolding experiments, is entirely included in the protein's linear amino acid sequence [65–67]. Anfinsen's thermodynamic theory states that this information is encoded in the polypeptide's energy landscape, where the native state has the lowest free energy [68,69]. For many years this premise laid the groundwork for a general approach in protein structure prediction. This approach integrates the sampling of alternative conformations, scoring them based on energy, and ultimately identifying the conformation with the lowest energy state [70,71].

The primary obstacle to the success of the energy-guided approach for protein folding on the biological timescale is the large array of alternative conformations, which grows exponentially with chain length rapidly reaching astronomical numbers, despite each amino acid only having only a limited and distinct set of potential backbone states [72]. The answer to this challenge can be found in the understanding that finding the native state does not require exploring the entire conformational space. The energy landscape does not resemble a flat 'golf course' with a single target 'hole'; instead, it adopts a funnel shape with directional indications, guiding the sampling process towards near-native conformations (Fig. 7) [71,73].

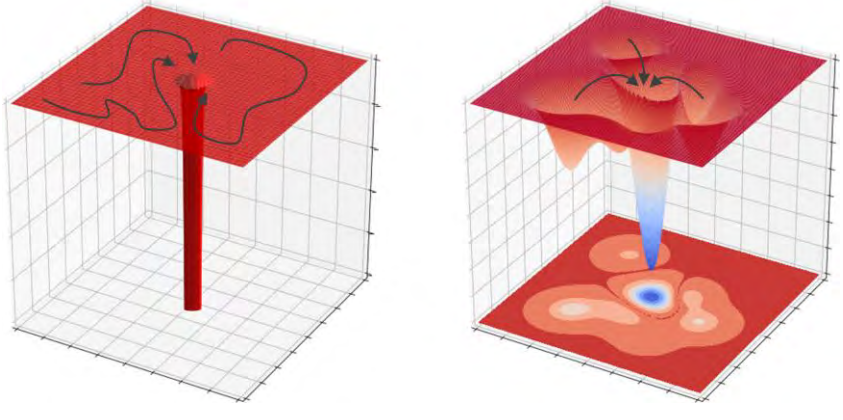


Figure 7. a) Simplified energy landscape in the shape of a 'golf course', where finding the native energy minimum requires extensive exploration of the surface. b) Simplified energy landscape in the shape of a 'funnel', where a simple downhill search from most starting points effectively leads to the native state.

The primary driving force behind protein folding lies in burying hydrophobic residues within the protein's core, away from the solvent [74]. Achieving a minimized cavity size in water and maximizing van der Waals forces necessitate the close packing of side chains in the core, while avoiding energetically unfavorable atomic overlaps. Polar groups that interact with the solvent in the unfolded state and become buried upon folding must form intra-protein hydrogen bonds to compensate the lack of solvation; otherwise, the substantial energy cost of hiding these groups to the solvent will impede the folding process [75]. The range of observed amino acid torsion angle distributions is effectively constrained by this packing, which also considers the strong torsional preferences of both the backbone and side chains. As a result, sidechain flexibility is diminished to a limited number of rotamers at each position [76].

2.2 Sequence-Structure-Function Relationship

Proteins exhibit a remarkable diversity of functions, ranging from catalyzing biochemical reactions to providing structural support and regulating cellular processes. Their ability to perform these functions is intimately linked to their unique three-dimensional structures [68,77], which are ultimately determined by their lineal amino acid sequence.

For decades, structural biologists have been guided by the sequence-structure-function paradigm, which posits that similar protein sequences give rise to similar structures and functions. As the number of available protein sequences has grown exponentially, traditional experimental methods for determining protein structures have struggled to keep pace. Nevertheless, the field has undergone a profound transformation due to recent advancements in protein structure prediction and a renewed focus on machine learning techniques, exemplified by groundbreaking methods such as AlphaFold [78] and RoseTTAFold [79]. Despite existing challenges in dealing with disordered sequences, large complexes, multiple chains, and protein-protein interactions, these developments are quickly bridging the gap between sequence and structure.

In a recent study [80], approximately 200000 protein structures for non-redundant microbial sequences were predicted providing intriguing examples that questioned the traditional sequence-structure-function relationship. Surprisingly, it was observed that dissimilar sequences could fold into similar structures, revealing that sequence diversity surpasses structural diversity. This remarkable discovery emphasizes the significance of distant homology detection and fold recognition methods in predicting structures for diverse sequences.

In contrast, traditional approaches to protein structure prediction often relied heavily on sequence homology, assuming that similar sequences would lead to similar structures and functions. However, this limited the scope of structural predictions to those with close sequence similarities, leaving out a vast portion of the protein universe that exhibited dissimilar sequences but potentially similar functions and structures.

This shift in perspective has opened up exciting possibilities for understanding the intricate world of protein structure-function relationships.

2.3 Protein structure prediction

Predicting protein structure from its amino acid sequence has long been a complex challenge in computational biophysics, due to both its inherent scientific appeal and wide-ranging applications, including genome interpretation and protein function prediction. Over the past few years, methods for predicting and designing protein structures have seen substantial progress. The combination of enhanced computing power and rapid expansion in protein sequence and structure databases has driven the development of sophisticated, data-intensive strategies for structure prediction.

Although significantly different, protein structure prediction methods can be broadly grouped into three main classes: template-based modelling, template-free modelling, and deep-learning-based models.

a. Template-based modelling

Also known as homology modeling, it is based on the hypothesis that the 3D structure of proteins exhibits higher conservation than their amino acid sequences. This implies that if two protein sequences share significant similarity (typically over 30%), they are likely to adopt similar 3D structures [81]. In the process of template-based modelling, the key steps involve choosing an appropriate structural template, aligning the target sequence to the template structure, and modeling mutations, insertions, and deletions. Closely related templates can be detected through single-sequence search methods to scan the Protein Data Bank (PDB) sequences, like BLAST [82], while more distantly related templates require target sequence profiles [83,84].

b. Template-free modelling

This method does not depend on global similarity to known structures, but on ‘the first principles’ of protein folding [85,86], enabling its application to proteins with unique, non-characterized folds. Template-free modelling approaches are suitable for proteins that lack global structural similarity to any known protein in the PDB. It

involves a conformational sampling strategy to generate models and an energetic ranking criterion to select native-like conformations. The process begins with a multiple-sequence alignment (MSA) of the target protein and related sequences. Predicted structural features guide the construction of 3D models, which are then refined, ranked, and compared to select the best predictions.

Historically, these methods have been distinct, with template-based approaches focusing on identifying and aligning with related known structures, while template-free methods involve extensive conformational sampling and the use of physics-based energy functions. However, recent advances have blurred the boundaries between these approaches [76]. Template-based methods now incorporate energy-guided model refinement, while template-free methods harness machine learning and fragment-based sampling to extract insights from the structural database.

c. Machine learning-based modelling

Up until very recently, computational structural biology relied mostly on physically based models using force fields and energy functions for biomolecular prediction and design. However, these models faced challenges in dealing with the vast protein conformational space and accuracy of force fields. Recent advancements in machine learning, particularly deep learning methods such as RoseTTAFold [79] and AlphaFold [78], have revolutionized the field. These methods, with millions of non-physical parameters and no assumptions about atomic interactions, achieve remarkable accuracy by learning from vast sets of experimentally determined protein structures and sequences. While these methods are trained on evolutionary data from alignments of homologous sequences, they have demonstrated the ability to predict protein structures accurately from single amino acid sequences. This indicates that they possess a rich understanding of sequence-structure relationships, making evolutionary data unnecessary for simpler systems.

Deep learning methods using multiple sequence alignments (MSAs) have demonstrated high accuracy in protein structure prediction. However, the requirement for high-quality MSAs presents challenges. There are instances of orphan sequences or “lineage-specific genes” with limited or no homologs in current databases. Predictions of these sequences using MSA-based methods often fall short in accuracy. Additionally, methods based solely on single sequences can be computationally more efficient, given that MSA-based approaches involve resource-intensive database searches, potentially slowing down predictions for large protein databases. Recent research has begun to explore prediction methods solely based on

single-sequence inputs employing natural language processing (NPL) methods, such as ESMFold[87] and OmegaFold [88].

2.4 Protein design

Protein design is often described as the reverse problem of protein structure prediction. Rather than trying to find the most stable conformation for a specific protein sequence, the aim is to infer an amino acid sequence that will stabilize the desired protein conformation [76]. It is possible to roughly classify protein design efforts into two groups: protein engineering and *de novo* design.

a. Protein engineering

Protein engineering focuses on modifying existing proteins to enhance their capabilities and explore new protein functions. Generally speaking, natural proteins often fall short in meeting real-world application requirements due to factors such as marginal thermal stability, low recombinant expression levels, limited activity, specificity, and poor stability in harsh conditions. The limited stability of natural proteins is attributed to their energy landscape (Fig. 8), which includes energetically closely misfolded states that hinder proper function. These misfolded states can restrict expression in different hosts and reduce the protein's longevity *in vitro*. The aim of stability design is to enlarge the energy gap between the correctly folded state and misfolded or unfolded states. This can be achieved by reducing the energy of the native state and eliminating as many misfolded states as possible, while increasing the energy of the remaining ones. This approach not only enhances thermal stability but also improves expression levels by inducing unfrustrated folding, thereby enlarging the gap between the native and misfolded states and positively impacting both thermal stability and expressibility [89].

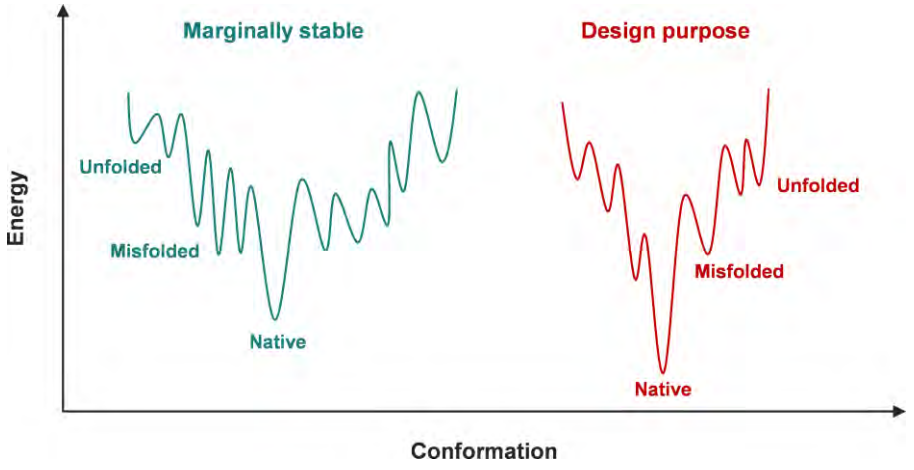


Figure 8. Left) Schematic illustration of the folding landscape of a marginally stable protein compared to the design procedure's purpose. The hardly stable protein has several competing misfolded states that are only slightly more energetic than the original form, thus frustrating folding. Right) In the folding landscape of a designed protein certain misfolded states are removed, increasing the energy difference between native and unfolded states as well as the misfolded states. Figure adapted from reference [89]

Protein engineering involves techniques such as rational design and directed evolution. Rational design aims at strategically modifying specific amino acids in a protein's sequence to achieve desired changes in its function or stability. This approach relies on prior knowledge of the protein's structure-function relationships to guide the design process. Directed evolution, on the other hand, emulates natural selection to optimize proteins for specific tasks. It involves generating diverse protein variants through random mutagenesis or recombination, and then screening and selecting those with improved properties. The application of these methods is sometimes hindered by limitations when screening a large number of sequence variants, since traditional laboratory-based screening processes can be time-consuming and costly. The introduction of *in silico* techniques revolutionized the protein engineering process by enabling virtual screening of mutations before conducting expensive and time-consuming experiments [90]. Computational approaches have facilitated the identification of promising mutations and have reduced the number of variants needed to be experimentally tested for detecting hits.

Computational protein engineering is useful in a diverse range of applications, such as:

- *Stabilizing proteins*

  The use of computer-based methods to enhance proteins thermodynamic stability has gained significant interest over the years [89,91,92]. Increased stability is associated with benefits such as improved recombinant protein expression and reduced risk of aggregation. It has been observed that naturally occurring proteins are not always optimized for stability, prompting the exploration of global redesigns where all residues are allowed to mutate. These simulations have shown that thermal unfolding temperatures can be significantly increased, sometimes by over 30 °C [91,93]. However, for practical applications, a more selective approach is preferred to raise stability. One effective strategy involves combining computational design simulations with information from multiple sequence alignments (MSA) [94]. By identifying homologues of the target protein in the NCBI non-redundant protein database, one can replace poorly represented amino acids with the most evolutionary selected ones at specific positions, resulting in increased protein stability, [95,96] such as in the PROSS approach [97].

- *Improving ligand affinity*

  Improving protein binding affinities through engineering has been a significant challenge for computational protein design [98]. Strategies include N-terminal or C-terminal extensions to increase protein-peptide interactions [99]. These extensions interact beyond traditional peptide binding sites, strengthening affinity. Similarly, protein engineering has been effectively used to increase the binding affinity of an antibody fragment to the I-domain of the integrin VLA1 [100], a critical component involved in cellular adhesion and signal transduction. Additionally, redesigning the Fc-Fcγ receptor interaction [101] led to over a hundred-fold enhancement of *in vitro* effector function.

- *Tailoring substrate specificity*

  Given that natural enzymes often have restricted affinity for chemically and/or structurally diverse substrates and might not exhibit strong catalytic activity towards a wide range of compounds, a key objective for industry is to amplify enzyme substrate versatility, allowing for a wider substrate specificity [102]. Computational techniques such as structure-based methods, enable strategic

mutational design to broaden substrate scope. Engineered lipases, such as those from *Pseudomonas aeruginosa*, exemplify the potential of these methods for expanding substrate tolerance [103]. Glycosyl transferases have been engineered to create O-, N-, or S-glycosidic bonds, expanding their substrate recognition capabilities [104]. Additional examples include the redesign of the specificity of an endonuclease [105] and the alteration of calmodulin specificity [106]. Notably, OptZyme [107] represents a pioneering computational strategy aimed at amplifying enzyme activity for novel substrates. The main idea is to use transition state (TS) analogue compounds, which are well-established for numerous reactions, as proxies for the often unknown TS structures of interest. This approach identifies mutations minimizing the interaction energy between the enzyme and its corresponding TS analogue, as opposed to the substrate, diminishing the energy barrier associated with the formation of the TS and thus increasing reaction rate towards the unnatural substrates.

- *Improving enzymatic activity*

Enhancing enzyme activity frequently requires the introduction of multiple mutations near the active site [108]. An illustrative example of this is the redesign of an ω-transaminase using a mechanism-guided computational enzyme design strategy [109]. Employing quantum mechanics and with the aid of RosettaDesign [110,111], different variants were designed. The most active among them, featuring five mutations, exhibited an astounding 1660-fold rise in catalytic efficiency as measured by k_{cat}/K_M. A successful example of this design goal is seen in the development of FuncLib [112,113]. This innovative approach centers on the creation of multipoint mutations within enzyme active sites, employing a combination of phylogenetic analysis and RosettaDesign calculations.

- *Customizing stereoselectivity*

Computational protein design is not only able to refine enzyme activity and specificity, but also to alter stereoselectivity. The potential of protein engineering in customizing stereoselectivity is exemplified by *Bacillus* sp. YM555-1 aspartase [114], which was converted into a library of hydroamination biocatalysts through structure-based computational enzyme design, enabling the production of enantiopure β-amino acids. Another illustrative example is the case of CYP105AS1, a cytochrome P450 enzyme from *Amycolatopsis orientalis* [115]. The natural stereoselectivity of this enzyme

must be inverted for producing the cholesterol-lowering agent pravastatin. Using the Rosetta CoupledMoves protocol [116], a virtual library of mutants was created to bind the substrate in the desired orientation. Rational analysis led to the identification of eight promising variants resulting in >99% stereoselective hydroxylation of the substrate compactin to pravastatin. Enzymatic stereodivergent synthesis to access multiple stereoisomers is very challenging, but by employing a strategy called "focused rational iterative site-specific mutagenesis" (FRISM), *Candida antarctica* lipase B was engineered into four highly complementary variants, each producing specific stereoisomers, achieving over 90% selectivity for all possible stereoisomers in a transesterification reaction [117]. Furthermore, an alcohol dehydrogenase from *Thermus thermophilus* [118] had its native enantioselectivity transformed through a modified CASCO (catalytic selectivity by computational design) workflow, integrating Rosetta and molecular dynamics simulations to yield four variants with reversed enantioselectivity.

b. *De novo* protein design

Contrary to protein engineering, *de novo* protein design starts with a predetermined requirement, such as a desired shape or function without relying on existing protein templates. *De novo* design involves creating proteins from scratch by using computational algorithms and structural principles. It offers the exciting potential to design novel proteins with tailor-made functionalities that may not exist in Nature. This approach expands the boundaries of protein design by exploring uncharted territory and enabling the development of highly stable and specialized proteins for specific applications in medicine, materials science, and synthetic biology.

De novo protein design has evolved significantly over time, starting from parametric approaches to more sophisticated machine learning techniques. In the early stages, parametric design involved the development of mathematical models to predict protein structures and functions based on known principles and physical properties [119–121]. However, this approach was limited by its inability to explore the vast sequence and conformational spaces of proteins.

A breakthrough in *de novo* protein design was achieved by the development of fragment-based methods and the Rosetta software suite [110]. These methods combine experimental data with computational methods to generate accurate protein structures and design novel sequences that fold into desired conformations. This marked a significant advancement in protein design capabilities and paved the way for designing new proteins with specific functions.

For many years, *de novo* computational protein design workflows primarily followed four sequential steps [122] (Fig. 7). Determining the protein target topology was the first stage of the protocol. The best backbone generation technique was then selected to create thousands of models compatible with the target topology based on the size and fold type. After that, these models were subsequently filtered before performing full-sequence design calculations. Total energy, side chain packing, amino acid composition, total charge, hydrogen bond networks or secondary structure prediction are some of the most common filters used. Sequence-structure compatibility was typically used to rank the top-ranked backbone-sequence pairs and those exhibiting funnel-shaped energy landscapes were selected for experimental validation.

Backbone building methods in classical *de novo* methods can be classified based on the degree of experimental knowledge they incorporate. These methods can be divided into two subcategories:

a) *ab initio* methods. These methods use minimal or no preexisting information. They start from scratch, without relying on existing protein structures, and build backbone purely based on computational algorithms and physical principles [123–125].
b) Knowledge-based methods. On the other extreme, there are methods which rely heavily on prior knowledge. SEWING, for instance, recombines pieces of known secondary structure elements and their connections to generate a new-to-Nature backbone [126].

Another way to differentiate these backbone generation methods is whether they produce continuous or discontinuous backbones:

a) Fragment-assembly methods. Some methods, like fragment-assembly methods, build secondary structure elements along with their loop connections, generating completely connected backbones. These methods assemble fragments of known protein structures to construct a continuous backbone.
b) Parametric generation methods. Other approaches such as the parametric generation of backbones, start by arranging secondary structure elements, such as helix bundles, and then require a loop closure step to connect all the elements into a continuous amino acid chain.

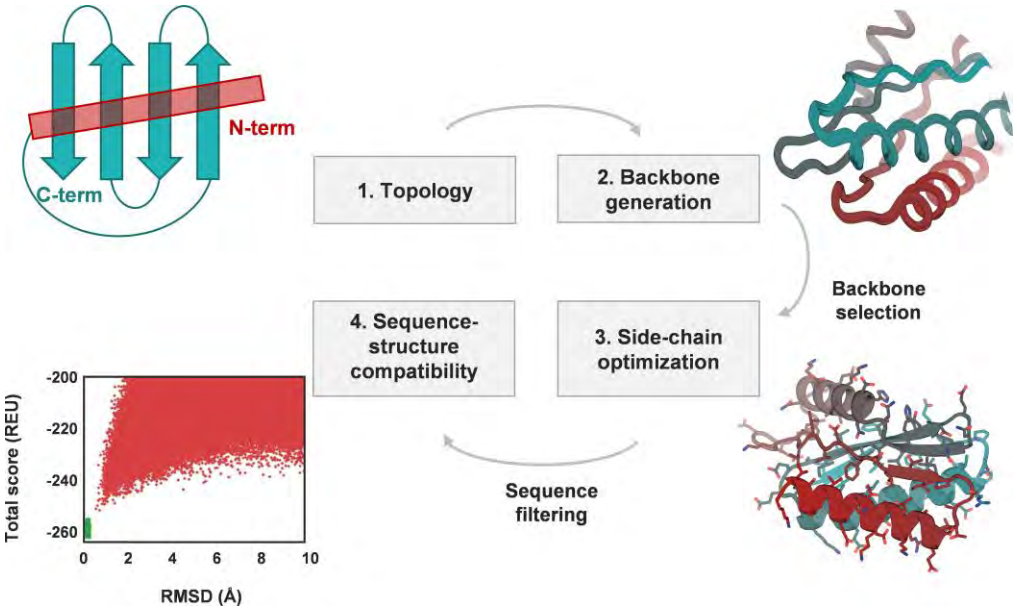


Figure 7. Schematic workflow of a classical *de novo* protein design protocol. It starts by defining the desired protein topology, including various length combinations. Based on the protein's size and folding pattern, an appropriate method for generating backbones is selected. This generates multiple models consistent with the target topology, which are then refined through filtering before undergoing complete sequence design calculations. The most promising backbone-sequence combinations are assessed for compatibility between their sequences and structures, with a focus on identifying those displaying favorable energy landscapes resembling funnels. Ultimately, these selected designs are subjected to experimental validation. The protocol involves several crucial decisions at each stage of design. Figure adapted from reference [122].

After generating the protein backbone model or set of models, the next step in the design process involves identifying an amino acid sequence that stabilizes the desired conformation or binding event. Sequence optimization programs incorporate two main components: an energy function that assesses the favorability of a specific sequence, and a protocol to search for more favorable sequences. In both, high-resolution protein structure prediction and protein design, similar energy functions are commonly used since they are aimed at optimizing the same physical properties, such as side-chain packing, hydrogen bonding, hydrophobic burial, and backbone and side-chain strain.

De novo protein design revolutionized bioengineering and molecular design, achieving remarkable milestones through computational tools and structural principles. This transformative field has generated novel protein structures and functions, including enzymes with customized catalytic activities, protein nanoparticles for drug delivery, engineered protein materials with unique mechanical traits, and foldable mini-

proteins. *De novo* techniques have enabled the creation of enzymes that catalyze entirely new chemical reactions, such as the Diels-Alder [127], Kemp elimination [128] or Retro-Aldol reactions [129].

More recently, machine learning algorithms have revolutionized *de novo* protein design. This revolution has been strongly influenced by the developments in protein structure prediction as described previously. Machine learning techniques, and in particular deep learning methods, leverage vast databases of protein sequences and structures to train models that can predict protein stability, folding, and function. Machine learning approaches have shown remarkable success in designing proteins with novel functions and properties, making *de novo* protein design more accessible and efficient.

Deep learning methods have enabled simultaneous design of sequence and structure. One of the pioneering methods in this field, named protein *hallucination* [130] allows creating novel protein structures with matching sequences, bypassing the need for a protein backbone. Essentially, *hallucination* generates a stabilized structure using a random input sequence. This method has been utilized to scaffold functional sites without the necessity to predefine the fold or secondary structure of the scaffold [131]. This was achieved using a protocol known as "*constrained hallucination*" that optimizes sequences to ensure that their predicted structures will contain the targeted functional site.

There is another approach for facing the same challenge also developed by the Baker Lab that is termed *inpainting* [131]. This approach initiates from a functional site and fills in additional sequence and structure information to provide a viable protein scaffold. This is achieved in a single forward pass through a specifically trained RosettaFold network.

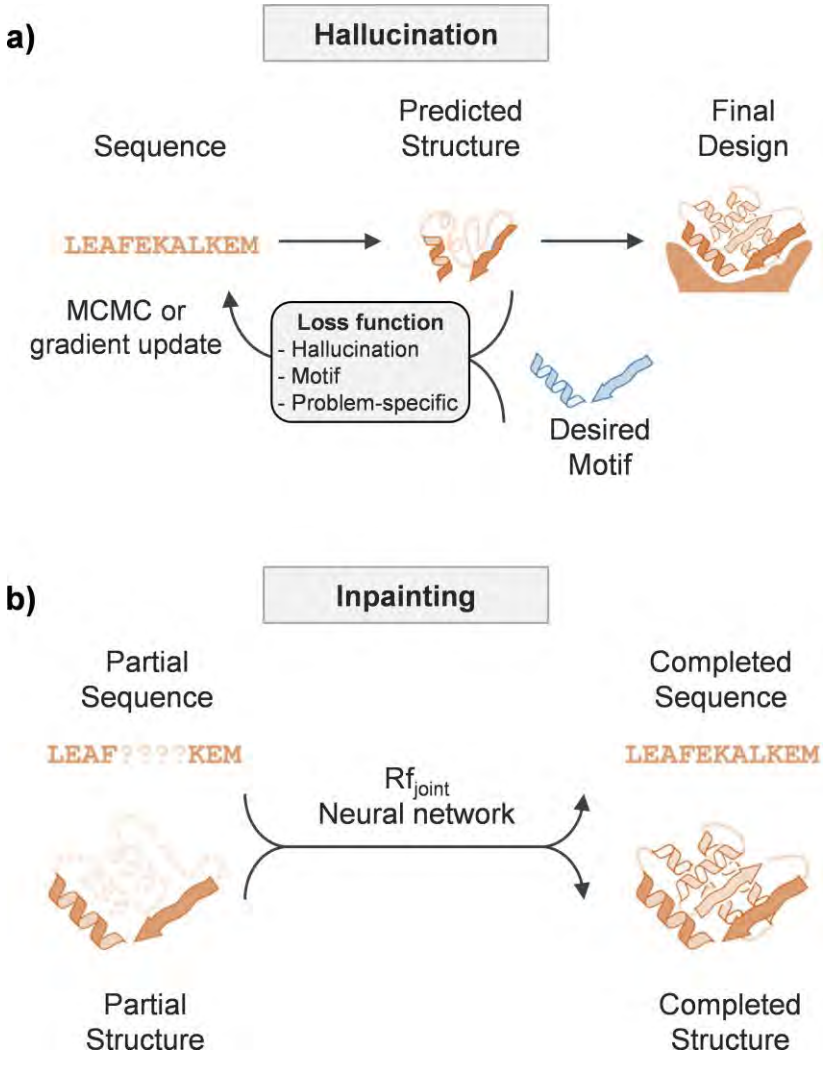


Figure 8. a) *Constrained hallucination* process: each iteration involves feeding a sequence into either the trRosetta or RoseTTAFold neural network. This network then predicts 3D coordinates and inter-residue distances and orientations. A specialized loss function evaluates these predictions, promoting both the structural confidence and the motif representation, alongside other task-specific parameters. MCMC stands for Markov chain Monte Carlo. b) Missing information recovery (*inpainting*) approach: inputting partial sequence and/or structural information into a modified RoseTTAFold network (called RFjoint) produces a complete sequence and structure outcome. Figure adapted from reference [131].

While both *hallucination* and *inpainting* successfully generate protein sequences and backbones for both monomeric and oligomeric novel structures, experimental validations frequently reveal solubility issues. In response, just one year after the release of the *hallucination* methodology, a deep learning–based protein sequence

design method named ProteinMPNN [132] was developed. ProteinMPNN generates highly stable sequences for a designed backbone. For native backbones it produces sequences that are predicted to fold into the intended structures more confidently than their native counterparts. It has been proved that ProteinMPNN generates sequences encoding protein structures with enhanced solubility and stability and with a much higher propensity to crystallize.

One of the latest advances in this area is the development of a guided diffusion model for generating *de novo* proteins called RFdiffusion [133]. Across a diverse set of challenges such as topology-constrained protein monomer design, protein binder design, symmetric oligomer design, enzyme active site scaffolding, and symmetric motif scaffolding for therapeutic and metal-binding protein design, RFdiffusion surpasses current protein design techniques.

The combination of these methods for structure and sequence design has made the field of protein design today more advanced and efficient, paving the way for unprecedented innovations and applications in molecular biology, medicine, biotechnology, material science, to name a few. Beyond the individual strengths of structure and sequence design methods, there are now advanced tools such as ProteinGenerator [134] that integrate the best of both worlds. By leveraging the capacity of RoseTTAFold to model protein sequences and structure concurrently, ProteinGenerator can design not only protein backbones but also sequences. This allows for designing proteins with any desired combination of sequence and structural attributes, marking a significant evolution in protein design capabilities.

Today, *de novo* protein design continues to advance, with an increasing emphasis on using artificial intelligence to accelerate the design process and enhance the accuracy of predictions. These advancements hold tremendous potential for creating customized proteins with tailored functionalities, making *de novo* protein design a powerful tool in various scientific and practical applications.

Chapter 2

Objectives

The main goals for this Doctoral Thesis are the following:

1. To provide full-atom models of the structure and dynamics of carbohydrate-lectin complexes, mainly using Molecular Dynamics (MD) simulations. Such studies involve integrating experimental information derived from crystallographic structures, Nuclear Magnetic Resonance (NMR) spectroscopy, and Isothermal Titration Calorimetry (ITC), with computer models generated in the framework of molecular mechanics/dynamics calculations using the Amber and GLYCAM force fields. The lectins object of study are several human galectins DC-SIGN, and the glycans involve blood group antigens and simple monosaccharides (galactose, mannose, fructose, etc.). Besides the structure and dynamics of the complex in the nano- to microsecond scale, relevant aspects for biomolecular recognition such as allosteric communication and hydration, will be analyzed in selected cases.

2. To combine different computational techniques such as quantum mechanics and MD simulations with Saturation Transfer Difference (STD) NMR spectroscopy to decipher the minimum ligand binding epitope to calcium-dependent lectin DC-SIGN. The main purpose of this study is to determine the minimum number of hydroxyl groups in a carbohydrate-like ligand, and their geometric arrangement, to most effectively bind to both the calcium atom and the surrounding amino acids constituting DC-SIGN's binding site. A special emphasis will be placed on augmenting the capabilities of the CORCEMA method, which uses experimental information mainly from NMR spectroscopy, with the structural and dynamical information of the protein-complex ligand provided by MD simulations.

3. To engineer human frataxin, a protein involved in the progressive neurodegenerative disease known as Friedreich ataxia, to increase the thermodynamic stability of both the wild-type and pathological single mutants compromising its folding. Computational techniques based on bioinformatics (analysis of evolutionary data) and deep-learning (ProteinMPNN and AlphaFold) will be used to generate variants, and their thermodynamic and biological properties will be assayed experimentally, with special emphasis on proteolytic resistance and the ability to bind metal ions and other proteins forming the iron-sulfur cluster.

4. To improve the expression levels and, if possible, the catalytic properties of Tobacco Etch virus (TEV) protease, which is a widely used enzyme in biotechnological applications. Evolutionary conservation, biological function and deep-learning models (ProteinMPNN and AlphaFold) will be used to address these goals. The expression levels, stability and catalytic performance of the designed variants will be tested experimentally, and MD simulations will be used

to characterize their structural and dynamic properties. This work is done as part of the research developed during a research secondment at the Baker Lab, Institute for Protein Design, University of Washington, USA.

5. To propose principles and concepts relevant to both carbohydrate binding and protein design, derived from the results obtained from the projects described above.

Chapter 3

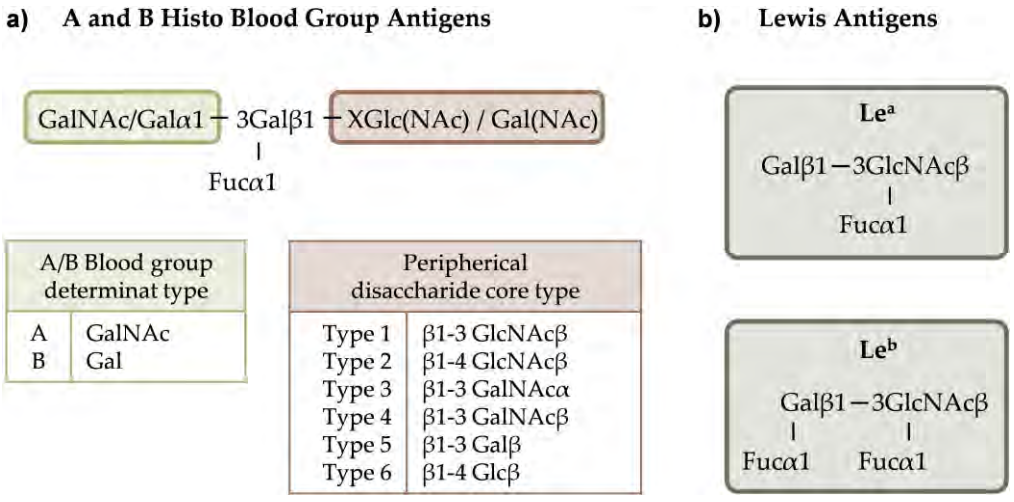
Deciphering Carbohydrate Recognition by Galectins: Interactions, Solvation and Allostery

1. Introduction

Lectins are a diverse group of proteins known for their capacity to selectively bind to specific carbohydrate residues present on the surface of cells or molecules. Among them, galectins, also known as S-type lectins, are probably the most ancient glycan-binding proteins. Through their carbohydrate recognition domain (CRD), galectins can recognize and interact with β-galactoside carbohydrates with high affinity and specificity [28].

Human galectins are involved in a variety of physiological functions including cell adhesion, the regulation of immune and inflammatory responses, cell migration, autophagy and signaling [28]. Galectins have significantly gained importance in recent decades due to their association with diseases such as cancer, fibrosis or diabetes. Understanding the complex carbohydrate recognition processes mediated by galectins is crucial for unraveling their roles in disease development and progression.

Blood type antigens, ubiquitous on the surface of red blood cells, epithelial cells, and various tissues, serve as essential ligands for galectins. The ABO system, consisting of A, B, AB, and O blood types, and the Lewis system with its related antigens, play pivotal roles in immune responses, cell-cell interactions, and disease susceptibility. The ABO antigens are categorized into six groups (Scheme 1a), depending on the peripheral core disaccharide structures, resulting in a diverse family of epitopes presented in different manners. This distinct presentation is recognized to influence their antigenicity [135]. The AB blood group carries both GalNAc and Gal (characteristic of the A and B types, respectively) as terminal carbohydrates, whereas the O blood group lacks such terminal carbohydrates. A total of six antigens belong to the Lewis blood group system, however they are often categorized into two main groups: Lewis A (Le^a) and Lewis B (Le^b) (Scheme 1b).



Scheme 1. a) The A and B histo-blood group antigens and their possible peripheral disaccharide core structures. b) The Lewis A and B antigen

In this chapter, the interaction of different carbohydrates with galectins of various types has been extensively studied in collaboration with the Chemical Glycobiology Lab at CIC bioGUNE led by Prof. Jesús Jiménez-Barbero. By combining NMR experiments, isothermal titration calorimetry (ITC), and molecular dynamics (MD) simulations, the structural determinants of carbohydrate-galectin recognition have been scrutinized.

Galectin structures can be classified in three main groups: dimeric, tandem and chimera (Fig. 1). Dimeric galectins consist of homodimers compromising two identical carbohydrate recognition domains. Chimera galectins can exist either as monomers or in multivalent forms, while tandem galectins possess two distinct CRDs linked by a peptide linker.

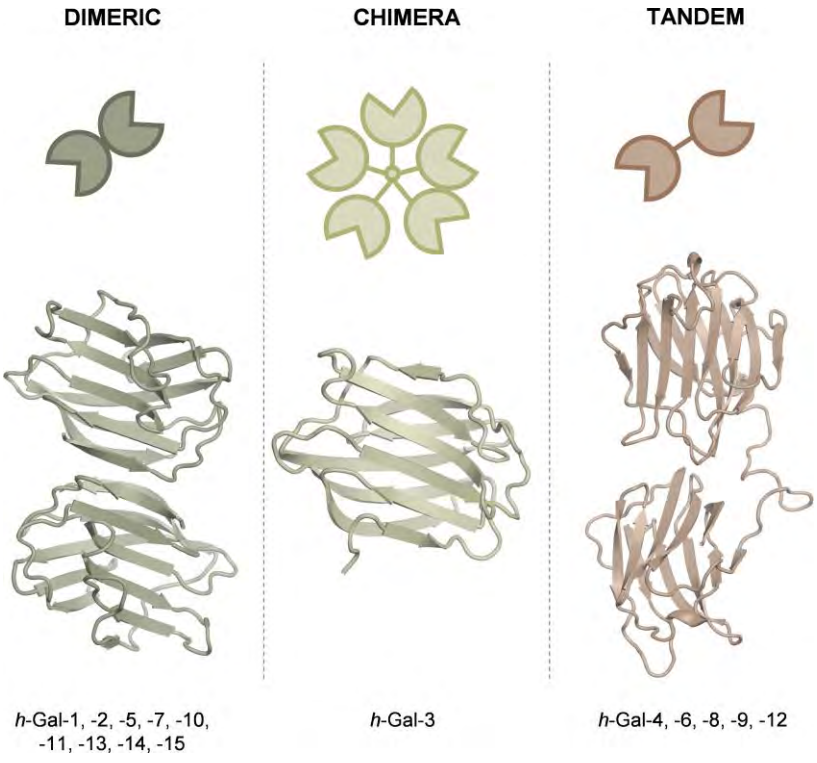


Figure 1. Examples of different galectin structures: *h*-Gal-1 homodimeric galectin, *h*-Gal-3 chimeric galectin and model for *h*-Gal-8 tandem galectin.

Representative examples of human galectins were selected to analyze the three different galectin forms. Homodimeric galectins, specifically *h*-Gal-1 and *h*-Gal-7, were examined. The main difference between these two galectins is the distinct relative disposition of their domains, with the protein-protein interface being localized in different regions for each of them (Fig. 2).

Chimera galectins were investigated through *h*-Gal-3 in its monomeric form. Additionally, *h*-Gal-4 and *h*-Gal-8 were studied as examples of tandem galectins. It is important to note that no crystal structures have been fully solved for full-length tandem galectins, particularly regarding the flexible peptide linker connecting the two CRDs. Therefore, atomistic models were created to represent the complete structure of these systems.

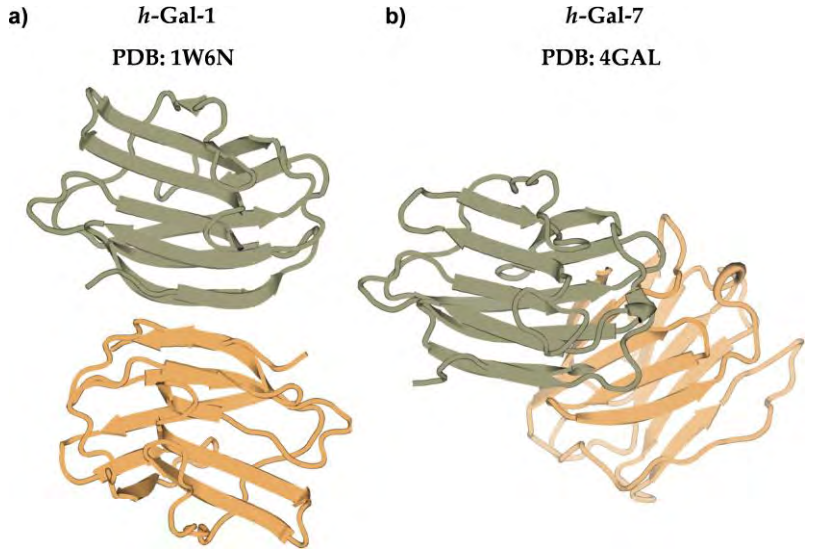


Figure 2. X-ray structures of human galectin-1 (*h*-Gal-1, PDB ID: 1W6N) and human galecitn-7 (*h*-Gal-7, PDB ID: 4GAL). N-terminal and C-terminal domains are colored in green and yellow, respectively.

As mentioned in the Introduction, protein-ligand recognition in buffered aqueous solution can be understood as the sum of numerous interrelated processes. Some of the most relevant ones include:

- Formation of protein-ligand contacts.
- Rearrangement of water initially solvating the protein.
- Rearrangement of water initially solvating the ligand.
- Hydration of the protein-ligand complex.
- Conformational changes in the protein upon binding.
- Conformational changes in the ligand upon binding.
- Alterations in the dynamics of the protein.
- Changes in the dynamics of the ligand upon binding.
- Modifications in the organization and interactions involving buffer ions.

Frequently, even minor differences in the structures of the binding partners or the experimental conditions can lead to compensating changes in enthalpies and entropies of binding, resulting in no net change in affinities [3]. As discussed above, the phenomenon of enthalpy-entropy compensation presents challenges when predicting interactions between biomolecules. In this chapter, various aspects behind this phenomenon are qualitatively explored.

2. Results and discussions

a. Modeled systems

The B type II blood antigen was selected for studying its binding to *h*-Gal-1, -3 and -7 for comparative purposes, given the availability of experimental data obtained from ITC experiments [136,137]. For the *h*-Gal-4 and -8 tandem galectins, given the different binding profiles of their N- and C-termini, the carbohydrate with the highest affinity in each case was selected [135,138] (Table 1 and Fig. 3).

Table 1. Galectin-ligand complexes studied. NT and CT stand for N-terminal and C-terminal domains respectively.

Galectin	Ligand	K_D (μM)	ΔG (kcal mol^{-1})	ΔH (kcal mol^{-1})	−T ·ΔS kcal mol^{-1})
Dimeric					
h-Gal-1	B type II	379	-4.7	-4.3	-0.4
h-Gal-7	B type II	288	-4.8	-9.5	4.7
Chimera					
h-Gal-3	B type II	4.4	-7.4	-8.9	1.6
Tandem					
h-Gal-4 NT	B type VI	51	-5.9	-12.3	6.4
h-Gal-4 CT	A type VI	26	-6.3	-8.3	2.0
h-Gal-8 NT	sialyl-T antigen	2.2	-7.7	-13.0	5.0
h-Gal-8 CT	A type II	13.5	-6.7	-11.2	4.6

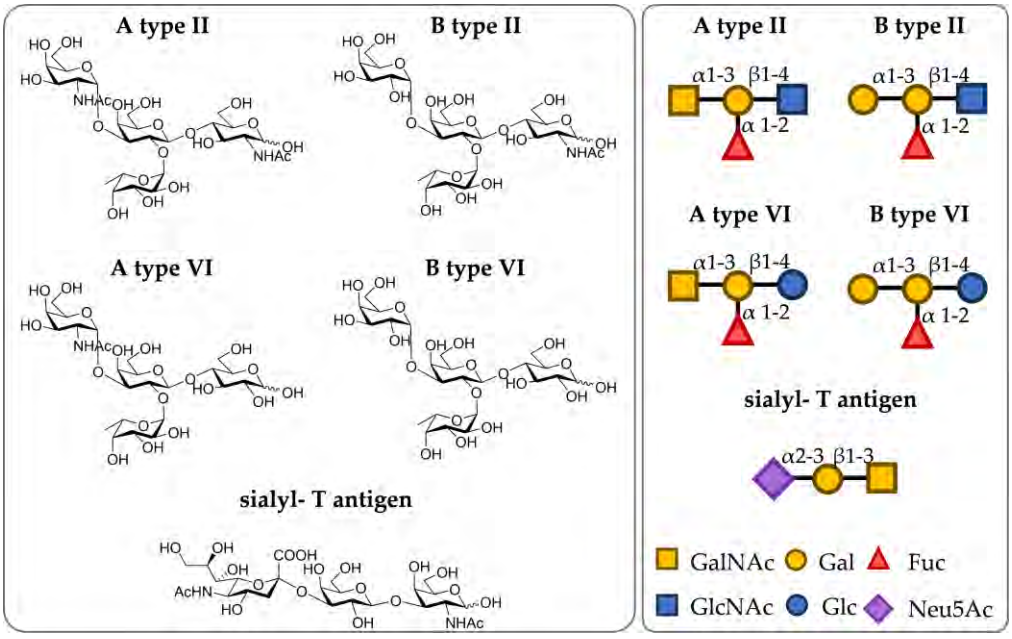


Figure 3. Chemical structure of studied ligands.

b. Galectins binding site description – common interactions

In the context of molecular recognition between carbohydrates and galectins, pivotal interactions such as hydrogen bonds, van der Waals forces, and salt bridges play a central role. Galectin carbohydrate recognition domains (CRD) contain around 130-150 amino acid residues forming two 5- to 6-stranded antiparallel β-sheets arranged in a β-sandwich [139]. The concave surface on the S side creates a shallow groove where the carbohydrate is accommodated. Certain residues of galectin binding sites are conserved among the seven different CRDs studied (Fig. 4). Those are expected to play a major role in the carbohydrate binding interactions. Among them, three different types of interactions have been selected to be analyzed: a CH···π, hydrogen-bond and an ionic interaction (Table 2 and Fig. 5). To characterize these interactions, the distances between the atoms involved in them were monitored during the different MD simulations of the complexes.

Figure 4. Multiple sequence alignment of (*h*-Gal-1, *h*-Gal-3, *h*-Gal-7, *h*-Gal-4 N- and C- terminal domains and *h*-Gal-8 N- and C-terminal domains). Residues conserved in all the systems are colored in dark green.

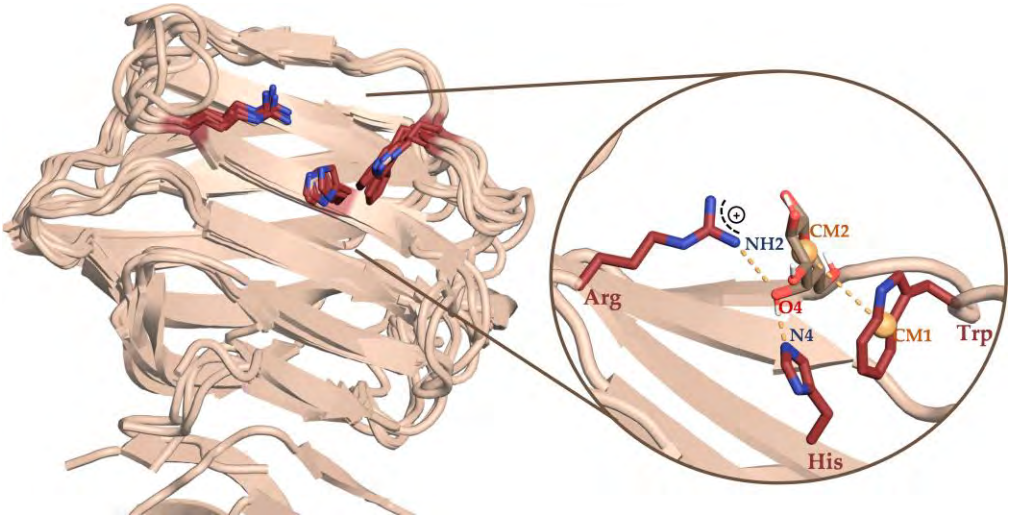


Figure 5. Superposition of galectins' CRDs. Binding site residues conserved in all the studied galectins are shown as red sticks. Represented galectins are *h*-Gal-1, *h*-Gal-3, *h*-Gal-4 CT, *h*-Gal-4 NT, *h*-Gal-7, *h*-Gal-8 NT and *h*-Gal-8 CT. NT and CT stand for N-terminal and C-terminal domains, respectively. The zoom panel focuses on key interactions involved in the ligand recognition process, including CH···π, hydrogen bonds, and ionic interactions. These interactions were analyzed through MD simulations. To simplify, a β OH-galactose residue has been used in the picture (all studied ligands share a galactose core).

Table 2. Description of the analyzed interactions.

Interaction	Distance measured
CH···π	Center of mass of the indole ring of the tryptophan residue of galectins' CRD (CM1) and the center of mass of the galactose ring conserved in all the studied ligands (CM2).
H-bond	N4 atom of histidine residue and O4 atom of the central galactose conserved residue
Ionic	O4 atom of the galactose residue and the NH2 atom of the CRD arginine residue

c. Molecular Dynamics and binding characterization

To investigate the binding process of the selected galectins with the different carbohydrates, microsecond molecular dynamics simulations (μs-MDs) were run for the different protein-ligand complexes and the *apo* form of galectins. Ten independent replicas of 2 μs each were run for each system, achieving a total sampling time of 20 μs.

In order to minimize the influence of other affecting factors, and focus on the very same type of interactions but with slightly different proteins and ligands, very similar

protein-ligand complexes were selected to carried out these type of analysis. To gain insight into the nature of these interactions, the distances obtained from the 20 μs accumulated simulated time for each complex were examined. The histogram distribution of these distances revealed distinct profiles among different systems (Fig. 6). Instead of assuming a Gaussian distribution, which might not accurately represent the asymmetric nature of the data, a more appropriate approach was followed. The distances were fitted to the extreme function, a type of Gumbel distribution. This choice was made due to the observed right skewness (positively skewed) in the distances distributions, as indicated by a longer tail on the right side. The extreme function, characterized by parameters such as amplitude (A), equilibrium distance (x_c), and width (w), provided a better fit for the non-Gaussian data. Through these parameters derived from the distributions, the strength and tightness of specific interactions can be characterized, and their dominant enthalpic or entropic character inferred.

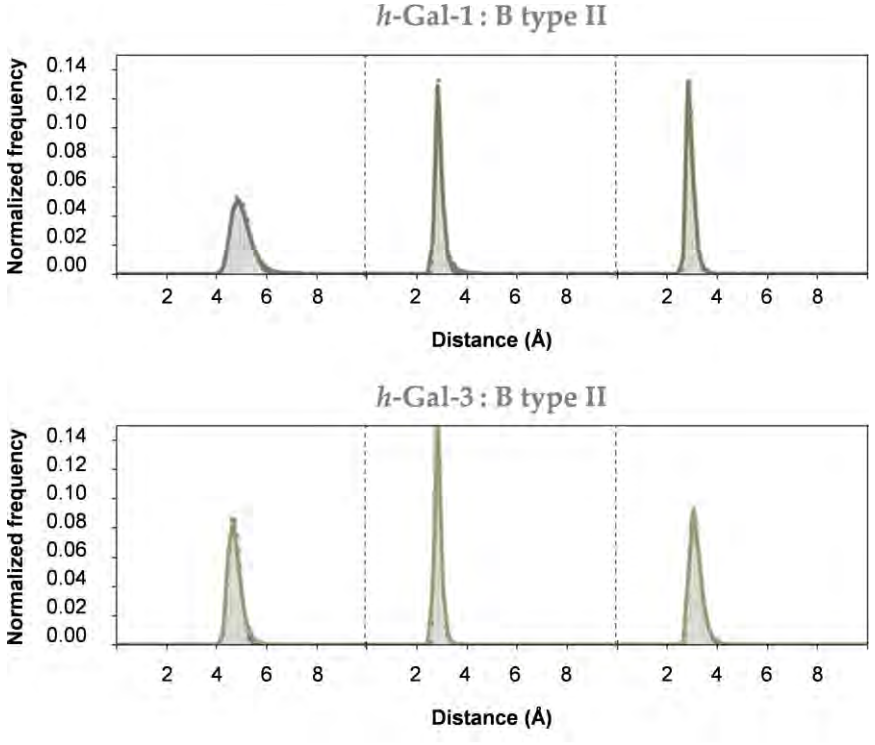


Figure 6. Histograms showing the distributions of distances characterizing the studied protein-ligand interactions in all galectins:glycan complexes.

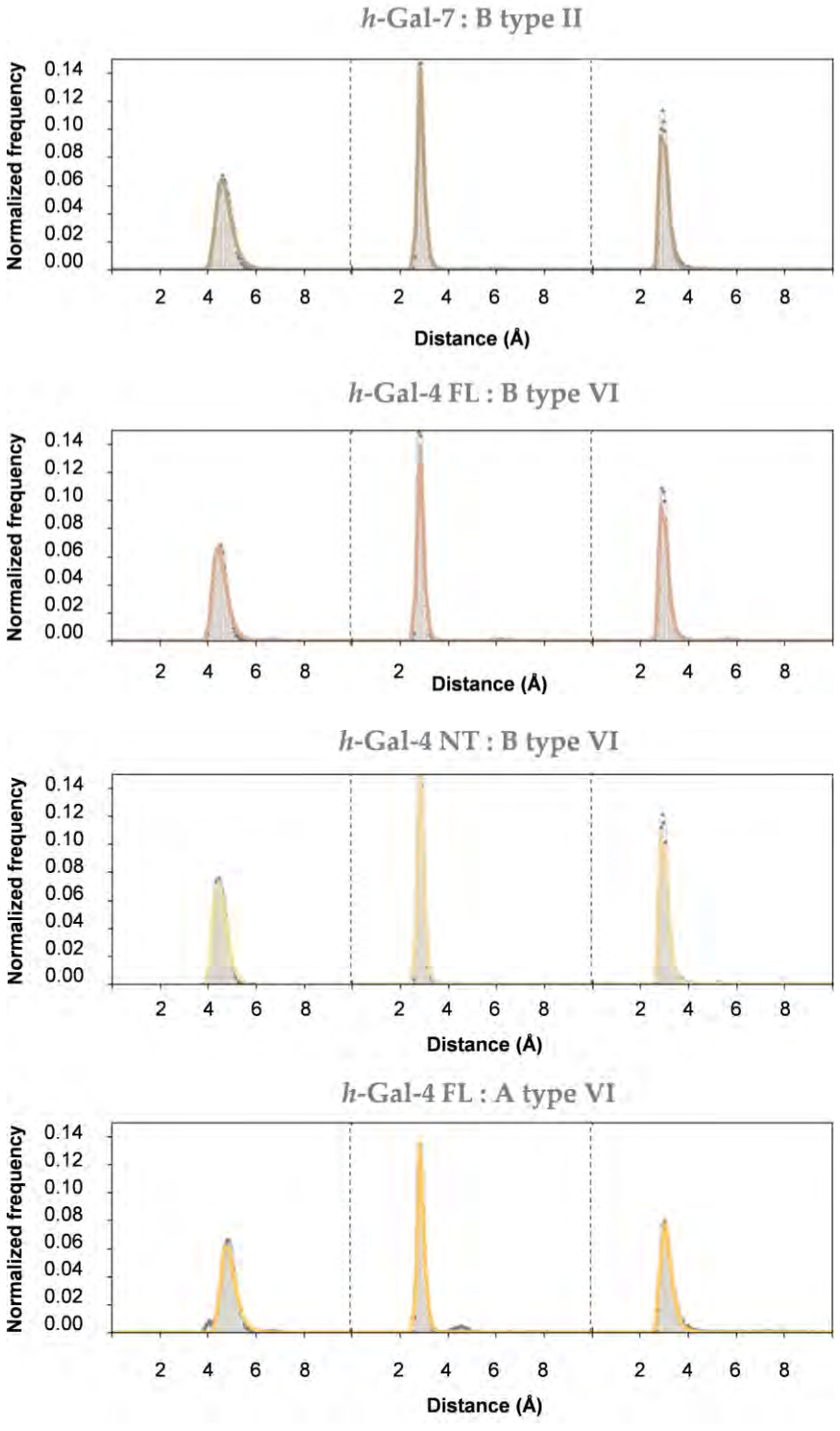


Figure 6 (cont.). Histograms showing the distributions of distances characterizing the studied protein-ligand interactions in all galectins:glycan complexes.

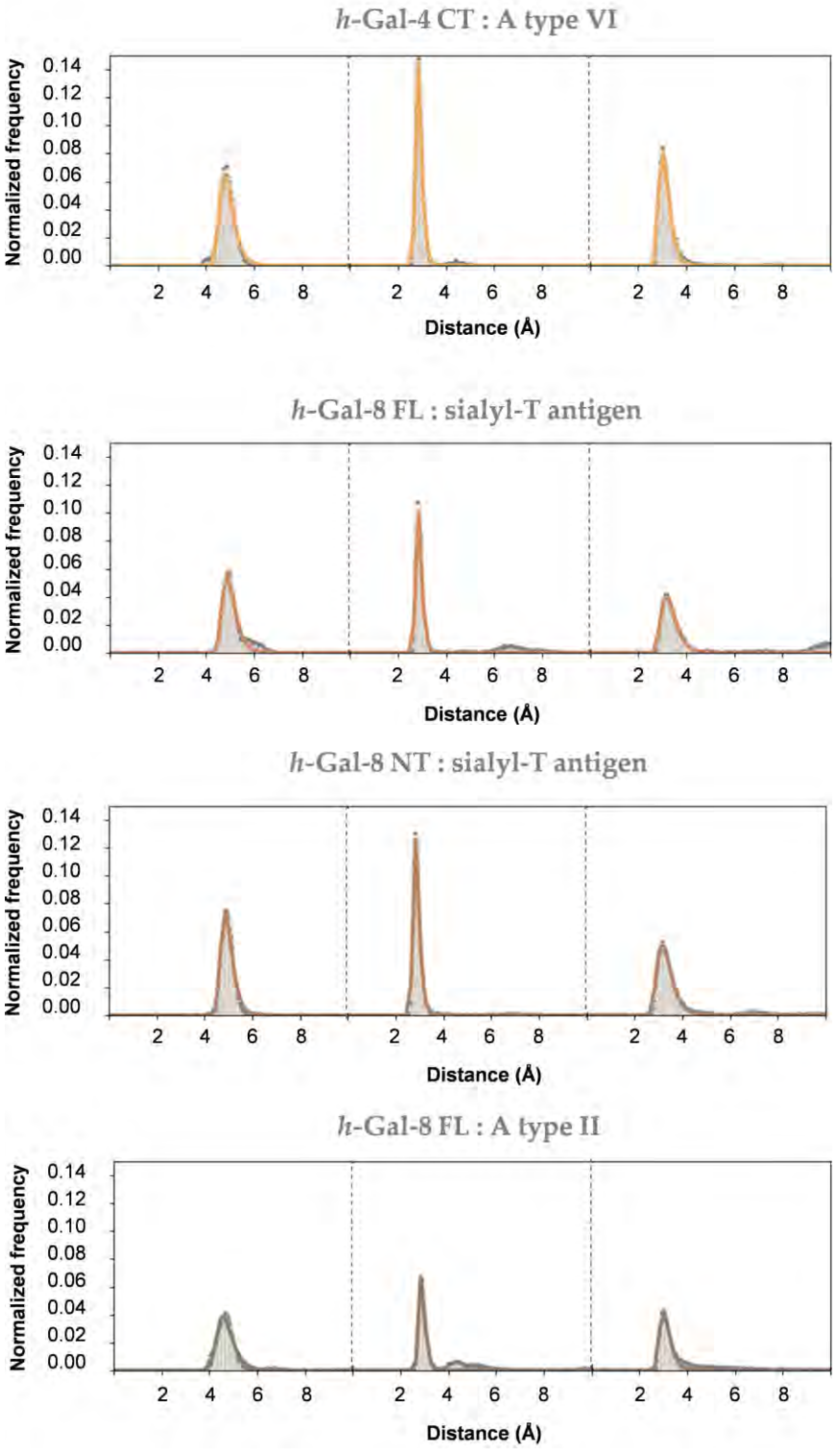


Figure 6 (cont.). Histograms showing the distributions of distances characterizing the studied protein-ligand interactions in all galectins:glycan complexes.

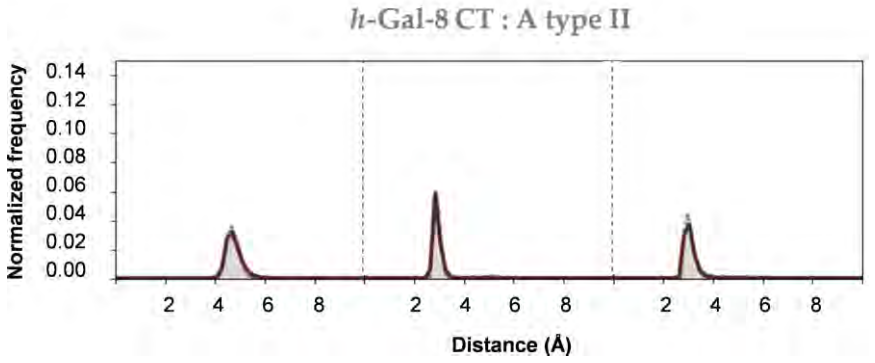


Figure 6 (cont.). Histograms showing the distributions of distances characterizing the studied protein-ligand interactions in all galectins:glycan complexes.

The width parameter of the fitting curve is an indication of the tightness of the interaction (i.e. how much the atoms involved in the contact fluctuate around the energy minimum). As shown in Fig. 7, the tightest interaction among the three studied ones is that occurring between the galactose unit of the ligand and the histidine residue (a hydrogen bond). The galactose-tryptophan CH···π interaction and galactose-arginine salt bridge exhibit contrasting behavior, with the former generally being loose and the latter being tighter. This trend is consistent across all the studied systems, except for *h*-Gal-8:sialyl-T antigen, where the galactose-tryptophane interaction is tighter and the galactose-arginine interaction is looser than the average.

The width distribution is inversely related to the amplitude, with higher amplitudes corresponding to narrower distributions (Fig. 7). Additionally, the equilibrium distances, which are characteristic of each interaction, are highly conserved around the optimal values for all the systems (Fig. 7).

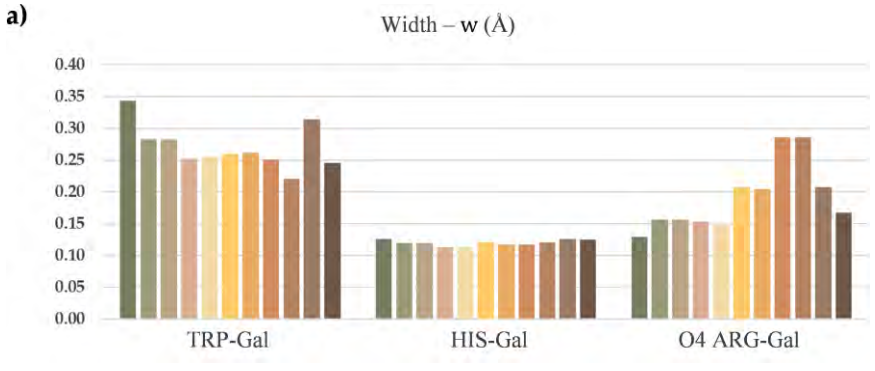


Figure 7. a) Width (w), b) amplitude (A) and c) equilibrium distance (x_c) parameters obtained from the fitting to the extreme function the distributions of Trp-Gal, His-Gal and Arg-Gal distances in the different studied systems. Gal stands for the central galactose unit conserved in all the studied ligands and FL stands for full-length systems.

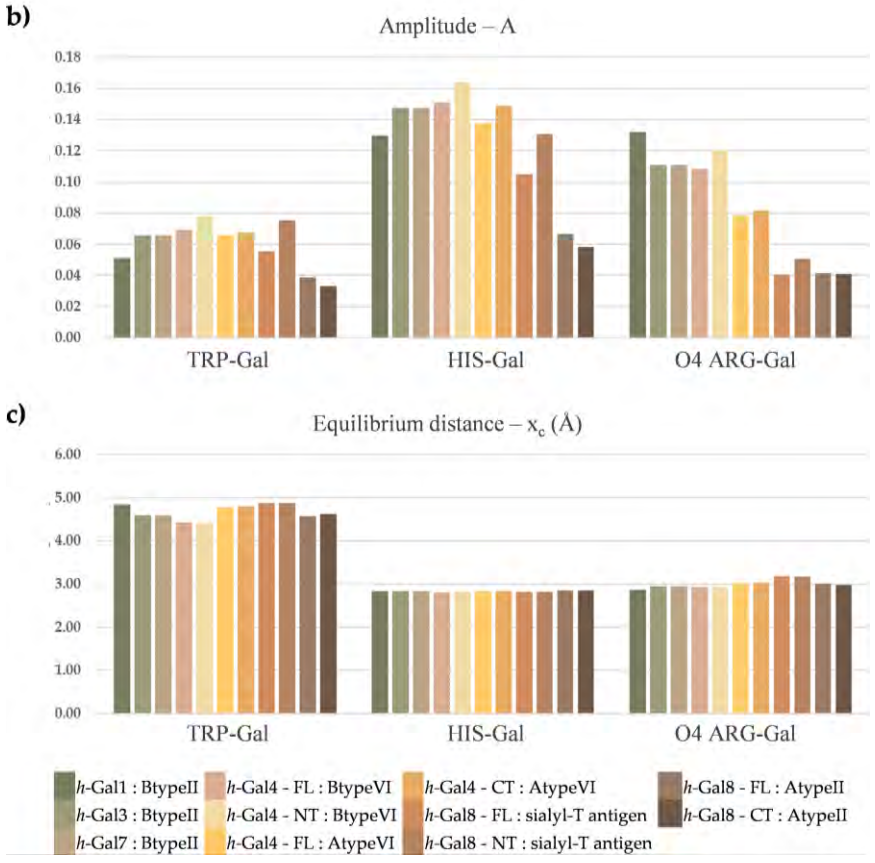


Figure 7 (cont.). a) Width (w), b) amplitude (A) and c) equilibrium distance (x_c) parameters obtained from the fitting to the extreme function the distributions of Trp-Gal, His-Gal and Arg-Gal distances in the different studied systems. Gal stands for the central galactose unit conserved in all the studied ligands and FL stands for full-length systems.

This type of analysis allows characterizing the associated states with an emphasis on describing the intensity and rigidity of molecular contacts. Stronger interactions are often more enthalpically intense but tend to exhibit greater rigidity, resulting in entropic penalties. Regarding the flexibility of the bound states sampled in our simulations (i.e. the unbound states were not characterized), the dispersion-driven CH···π interaction between the tryptophan residue and the galactose ring of B type II exhibits a loose, entropy-driven character in the three galectins mentioned above, particularly in *h*-Gal-1. On the other hand, both the ionic and the hydrogen bond interactions are dominated by electrostatics, tighter, and thus primarily driven by enthalpy.

d. Allosteric communication

Proteins can be treated as a four-dimensional network consisting of discrete nodes (amino acids) connected through space and time. The result of this crosstalk between the constituting amino acids, which can be largely affected by environmental factors (solvent, ionic strength, temperature, substrate binding, etc.), defines the structure, dynamics and ultimately the properties of a given protein. Although usually the most effective communication between these nodes takes place through direct covalent or noncovalent interactions, identification of allosteric pathways between pairs of residues that are not directly linked provides insight into alternative modes of signal transmission and communication within the protein structure. The investigation of allosteric pathways between these pairs of residues has been investigated using dynamical network and correlated residues analysis [140,141].

In this study, the allosteric communication between the active sites and distant regions of the N- and C- terminal domains of all the examined galectins was analyzed through molecular dynamics simulations. Through these simulations, optimal and suboptimal pathways for dynamic correlation between binding residues and any other amino acid in the protein were traced (Fig. 8) using the Weighted Implementation of Suboptimal Paths (WISP) algorithm [16].

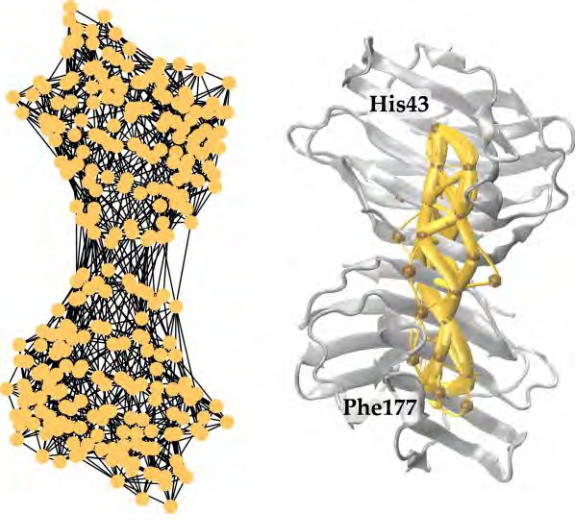


Figure 8. Left) Representation of a dynamic correlation network calculated from MD simulations for *h*-Gal-1; nodes (yellow circles) represent the side chain residues centers of mass and edges (black lines) are defined by a metric quantifying the interdependence among nodes. Right) Example of 100 optimal and suboptimal pathways based on correlated motions for *h*-Gal-1 between His-43 and Phe-177. The width of these pathways is proportionally related to their length, the wider paths correspond to the shorter trajectories, and conversely, narrower paths represent longer trajectories.

Results showed that the correlation motion starting from the binding site histidine propagates through the β-sheets of the CRD core, extending even to the homodimer interface in the case of *h*-Gal-1 – and to a lesser extent in *h*-Gal-7 – when the canonical ligand *N*-acetyllactosamine (LacNAc) is bound (Fig. 9). This observation supports the notion of allosteric communication (i.e. long-range ligand binding effects) operating in homodimeric galectins. In contrast, the correlated motions in *h*-Gal-3 dissipate in the vicinity of the active site. This remarkable feature revealed by molecular dynamics was further supported by relaxation dispersion NMR experiments conducted on *h*-Gal-1 and *h*-Gal-3 [137]. In *h*-Gal-1, binding to LacNAc induces μs-ms dynamics in up to 34 residues (Fig. 10c) [137], whereas the limited motions observed in *h*-Gal-3 are attributed to residual thermal motions.

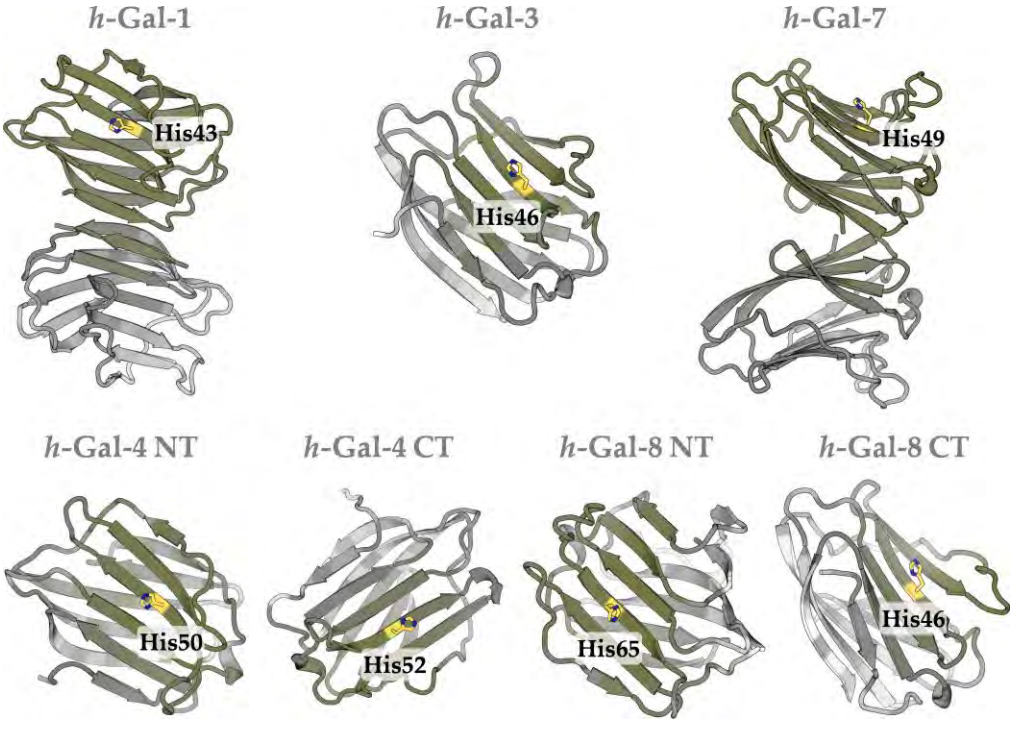


Figure 9. Color-coded distribution of whole-protein allosteric pathways determined from selected residues (shown as yellow-blue sticks) in galectin binding sites through 100 ns MD simulations; green and gray colors indicate residues involved in shorter/efficient and longer/inefficient pathways, respectively.

The presence of such long-range correlated motions in the homodimer might be at the origin of the unusual energy profile of carbohydrate binding to *h*-Gal-1, where entropy does not impose a penalty to ligand recognition [137]. In the case of the binding of LacNAc to *h*-Gal-1, the residues showing concerted dynamics in the micro-to-millisecond timescale upon binding determined by transversal relaxation dispersion

NMR experiments match those appearing at the highest frequency in the calculated pathways (Fig. 6b and c) [137].

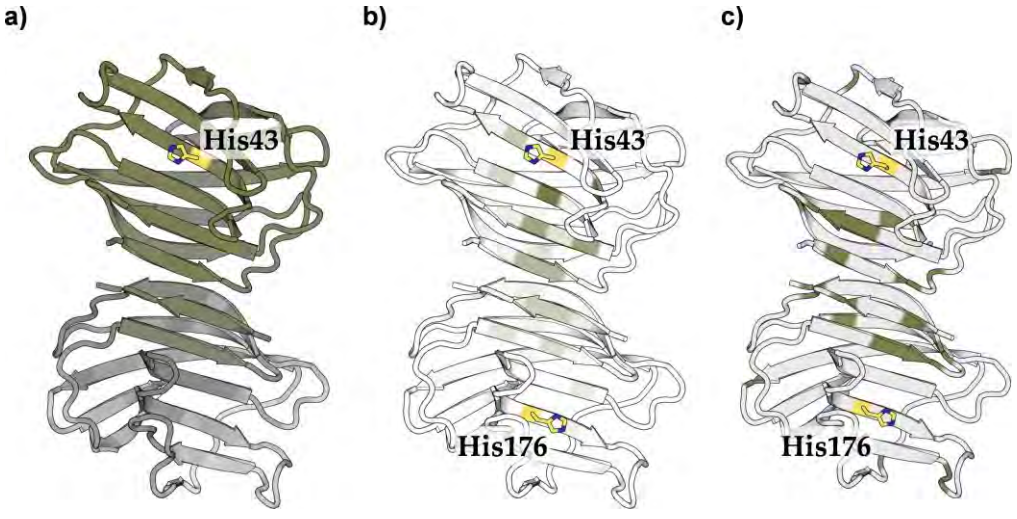


Figure 10. Long-range concerted dynamics observed in *h*-Gal-1. a) Color-coded distribution of whole-protein allosteric pathways determined from a binding site histidine (represented as yellow-blue sticks) in *apo h*-Gal-1 through μs-MD simulations. Shorter/efficient pathways are shown in green, while inefficient/longer pathways are depicted in gray. b) Residues most frequently involved in the optimal and suboptimal pathways (color-coded in a green gradient) calculated from histidine residues (represented as yellow-blue sticks sticks) in both binding sites of *apo h*-Gal-1. c) Residues (in green) exhibiting concerted dynamics at 380 s^{-1} as determined by transversal relaxation dispersion (RD) NMR experiments.

The identification of residues that play a crucial role in allosteric communication provides valuable insights into this phenomenon (Fig. 11a). These residues are determined based on their high frequency of occurrence in the calculated pathways, indicating their significance in mediating allosteric effects. To our delight, these predictions were fully validated by the analysis of NMR chemical shift perturbations (CSP) recorded upon ligand binding. In the case of the heterodimeric galectins *h*-Gal-4 and *h*-Gal-8, CSP analysis revealed perturbed residues located far from the binding site and concentrated again in the internal β-strands region, suggesting the presence of conserved allosteric networks whose effects that are dynamically transmitted throughout the core of protein structures. These findings are summarized in Fig. 11b, which highlights the regions of each galectin CRD that are most affected by ligand binding as determined by both MD simulations and NMR experiments.

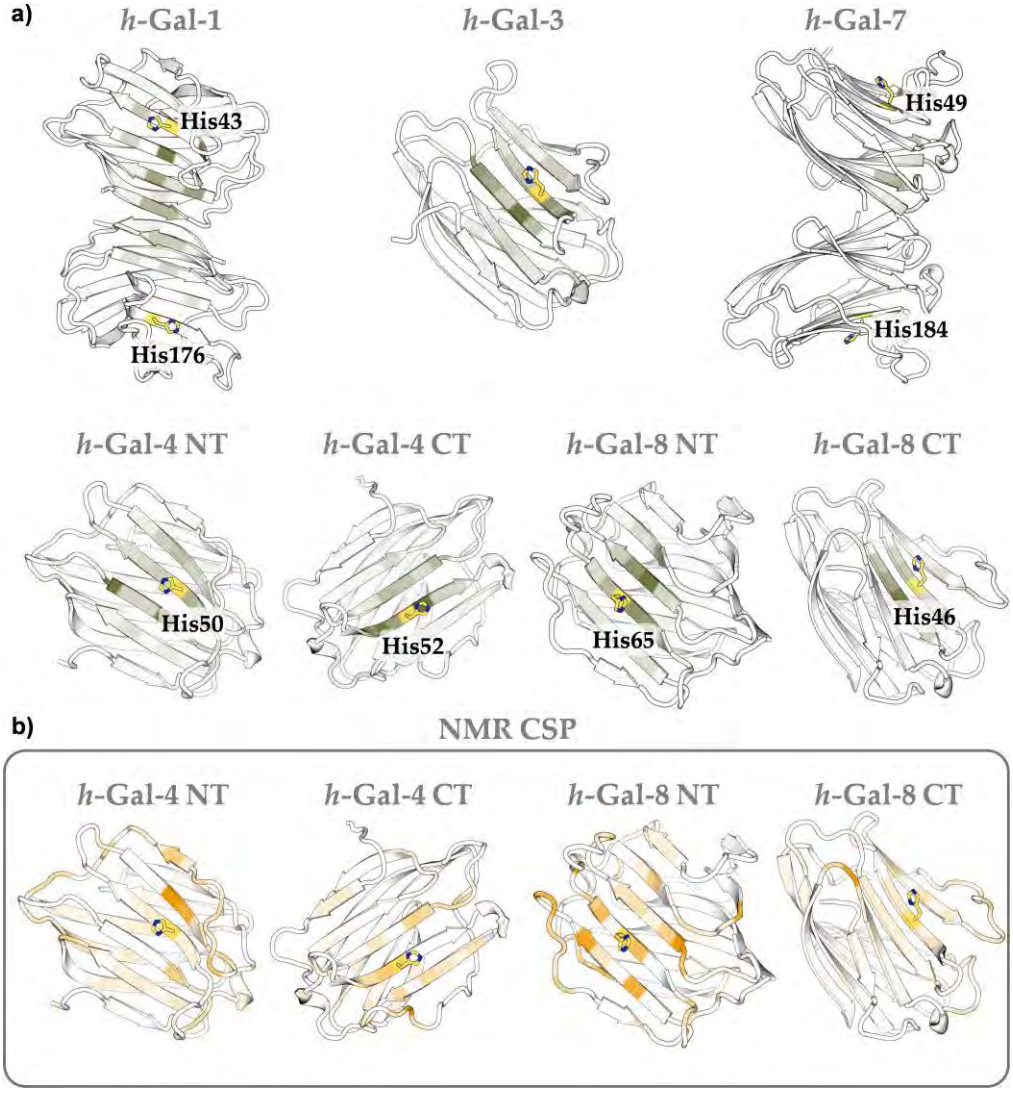


Figure 11. a) Residues most frequently involved in the optimal and suboptimal pathways calculated from selected residues (shown as yellow sticks) in galectin binding sites through 100 ns MD simulations; residues with the highest frequency are highlighted in green. b) Most perturbated residues in NMR CSP analysis for *h*-Gal-4 NT:B type VI, *h*-Gal-4 CT:A type VI complexes, *h*-Gal-8 NT:sialyl-T antigen and *h*-Gal-8 CT:A type II complexes.

e. Hydration profiles

Water, acting as both a hydrogen bond donor and acceptor, exhibits remarkable versatility at the biomolecular interface, significantly impacting ligand binding. Analysis of protein and ligand solvation, using the grid inhomogeneous solvation theory (GIST) [142] method, allows for an in-depth examination of the structural, dynamic, and thermodynamic aspects of water molecules surrounding the bound

carbohydrate ligands. Notably, water-mediated interactions and the displacement of conserved water molecules are critical factors influencing the stability and affinity of protein-ligand complexes [52,143,144].

i. Crystallographic water

The identification of water sites in different high-resolution crystallographic structures of *apo* and ligand-bound crystal structures of the studied galectins, provides valuable insights on the role of water in such molecular recognition events. Hence, overlaying the X-ray structures of these galectins reveals highly ordered water molecules surrounding the binding site (Fig. 12). This observation suggests a relevant role of water molecules in the context of galectin-carbohydrate interactions.

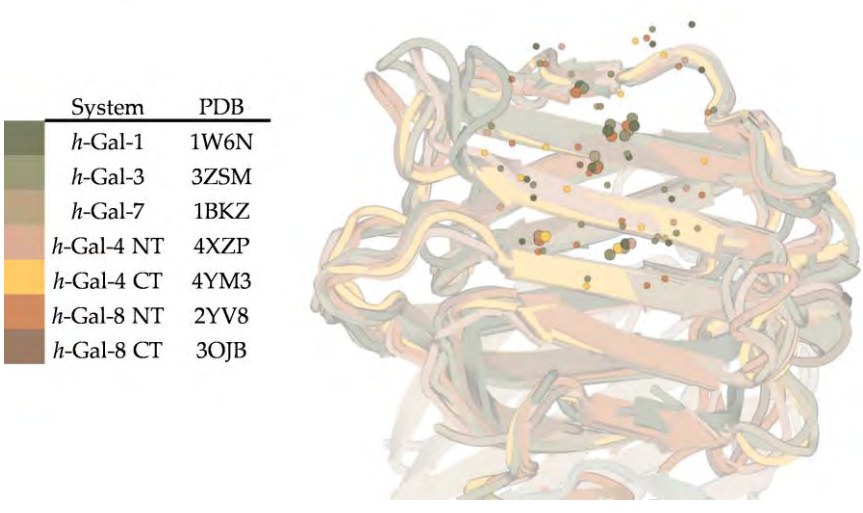


Figure 12. Superposition of crystallographic structures of *h*-Gal-1, *h*-Gal-3, *h*-Gal-4, *h*-Gal-4 CT, *h*-Gal-7, *h*-Gal8 NT, *h*-Gal-8 CT showing as spheres water molecules located 5 Å around the central Gal unit. Wider spheres represent those conserved in at least three crystallographic structures.

Water displacement from the binding site to the bulk solution upon ligand recognition significantly influences the free energy changes associated with the recognition process. This displacement involves the movement of water molecules away from the binding site, leading to the rearrangement of the surrounding water network. Understanding the dynamics and energetics of water displacement is essential for comprehending the mechanism of ligand binding [145].

In the bound state, water sites identified in *apo* galectins are typically occupied by ligands hydroxyl groups (Fig. 13). This occupancy by the ligand hydroxyl groups

serves to maintain the hydrogen bonds that were observed in the *apo* structures. This phenomenon is not only an established aspect of ligand binding but also carries entropic benefits [146]. The hydrophobic effect associated with water displacement reduces the entropic penalty of binding, while enthalpy contributions remain relatively similar because water and ligand hydroxyl groups form highly conserved interactions. As a consequence, the overall favorable energetics of ligand binding are enhanced.

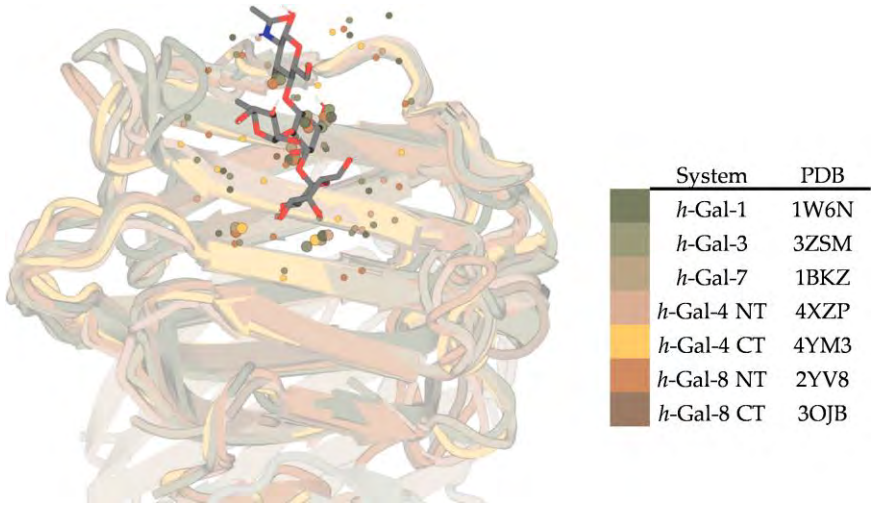


Figure 13. Superposition of crystallographic structures of *h*-Gal-1, *h*-Gal-3, *h*-Gal-4 NT, *h*-Gal-4 CT, *h*-Gal-7, *h*-Gal-8 NT, *h*-Gal-8 CT in complex with B type II antigen. Blue spheres represent highly conserved crystallographic water molecules found in the *apo* crystal structures.

ii. Bridge water molecules in complexes

While previous analyses were based on X-ray structures, a more comprehensive exploration of water pockets was conducted through MD simulations. This type of analysis enables a deeper investigation into the dynamic behaviour of water molecules surrounding the binding site. Unlike X-ray methods, MD provides a temporal dimension, allowing capturing intricate fluctuations and interactions within solvent patches. Simulations, such as MD in aqueous solutions, offer a unique opportunity for a thorough investigation of the interactions between water molecules and solutes, with significant implications for the recognition process, particularly in carbohydrates (Fig. 14) [147–149].

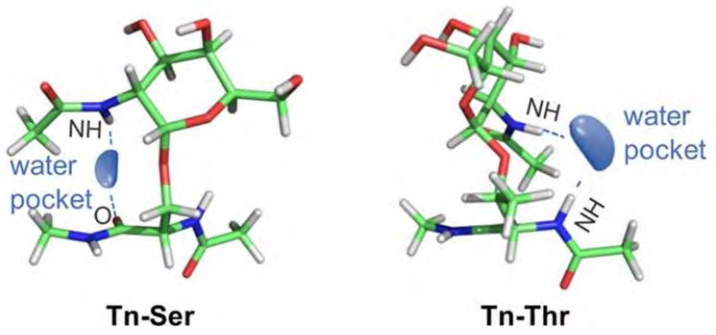


Figure 14. Example of different conformational preferences in two Tn antigen variants: Tn-Ser and Tn-Thr guided by the bridging water density around them. Figure adapted from reference [148].

- **Visualization of water structure**

The global solvation of the studied galectins MD trajectories was assessed through GIST analysis. By using a 25 Å x 25 Å x 25 Å grid centered on the shared lactose unit, the top 5 most populated voxels were selected. High density water pockets were detected in several of the studied systems by superimposing them (Fig 15, water sites A to D). These structural waters act as bridges between the ligand and the binding site. This approach is related to, but qualitatively different from that based on measuring the 2D radial distribution functions to locate high density water patches between specific atom pairs [147–149].

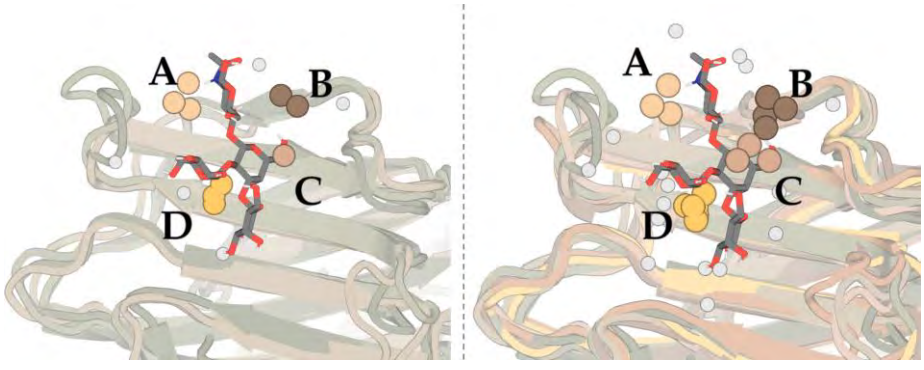


Figure 15. Left) High-dentisity water pockets commonly found in *h*-Gal-1, *h*-Gal-3 and *h*-Gal-7. Right) high- density water pockets commonly found in *h*-Gal-1, *h*-Gal-3, *h*-Gal-4 CT, *h*-Gal-4 NT, *h*-Gal-7, *h*-Gal-8 NT, *h*-Gal-8 CT.

In order to gain a better understanding of the role of water in carbohydrate recognition by galectins, we used the *grid inhomogeneous solvation theory* (GIST)

approach to estimate the solvation thermodynamics of water sites A and D found in *h*-Gal-1, *h*-Gal-3 and *h*-Gal-7, as well as water site B found in *h*-Gal-1 and *h*-Gal-7.

- **Quantitative analysis of local contributions to solvation**

The total referenced entropies, enthalpies and free energies of solvation calculated with GIST from MD simulations are summarized in Fig. 16.

	h-Gal1:B type II			*h*-Gal3:B type II			*h*-Gal7:B type II		
	Water site A	**Water site B**	**Water site D**	**Water site A**	**Water site B**	**Water site D**	**Water site A**	**Water site B**	**Water site D**
ΔHs-w	-5.87	-5.69	-3.16	-0.80		-2.48	-3.94	-2.60	-8.27
ΔHw-w	-0.56	-3.14	-0.98	-0.26		0.46	-0.82	-0.98	-0.47
ΔHsolv	-6.43	-8.82	-4.14	-1.06		-2.02	-4.76	-3.58	-8.74
-T·ΔSorient	0.05	0.15	0.02	-0.04		0.02	-0.10	-0.09	0.29
-T·ΔStrans	0.34	0.63	0.30	-0.04		0.10	0.19	0.21	0.46
-T·ΔSsolv	0.39	0.78	0.32	-0.08		0.12	0.10	0.12	0.76
ΔGsolv	-6.05	-8.04	-3.82	-1.14		-1.91	-4.66	-3.46	-7.99

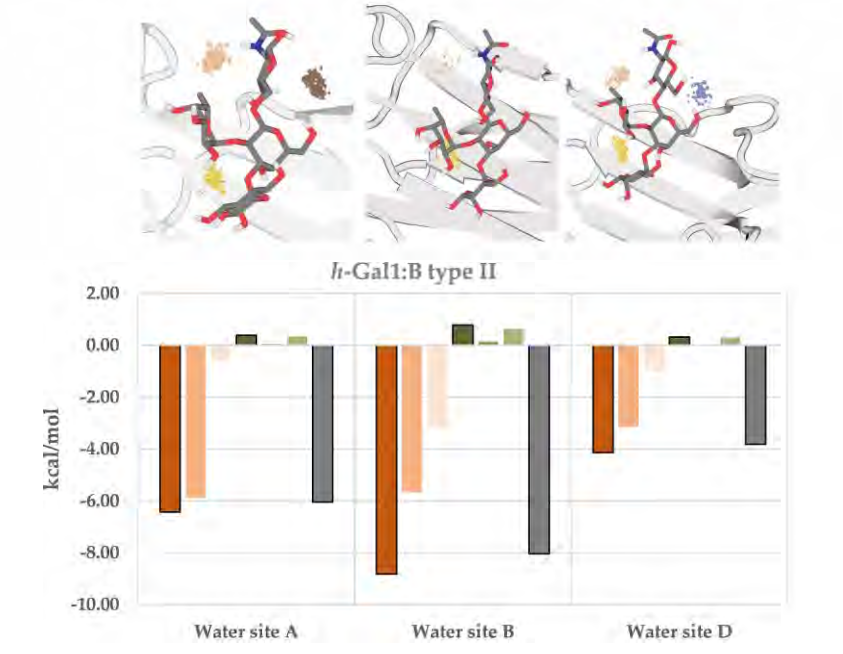


Figure 16. Solvation thermodynamics derived from GIST calculations for *h*-Gal-1, *h*-Gal-3 and *h*-Gal-7 in complex with B type II antigen. Dashed regions colored in wheat, brown and yellow represent water sites A, B and D respectively. Water-water (w-w) and solute-water (s-w) interaction energy, as well as the orientational and translational entropy are shown relative to bulk water (TIP3P). All quantities are given in kcal mol^{-1}. $\Delta S_{solv} = \Delta S_{orient} + \Delta S_{trans}$; $\Delta H_{solv} = \Delta H_{s-w} + \Delta H_{w-w}$; $\Delta G_{solv} = \Delta H_{solv} - T\Delta S_{solv}$. Solvation thermodynamics in water site B could not be calculated for *h*-Gal-3 due to the very low occupancy of water molecules in this region.

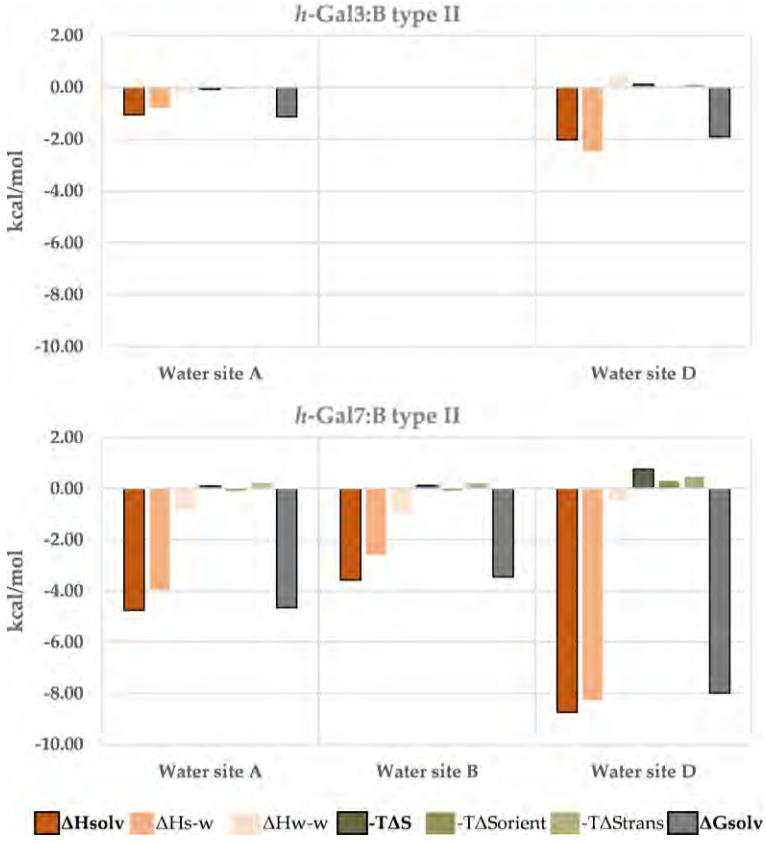


Figure 16 (cont.). Solvation thermodynamics derived from GIST calculations for *h*-Gal-1, *h*-Gal-3 and *h*-Gal-7 in complex with B type II antigen. Dashed regions colored in wheat, brown and yellow represent water sites A, B and D respectively. Water-water (w-w) and solute-water (s-w) interaction energy, as well as the orientational and translational entropy are shown relative to bulk water (TIP3P). All quantities are given in kcal mol^{-1}. $\Delta S_{solv} = \Delta S_{orient} + \Delta S_{trans}$; $\Delta H_{solv} = \Delta H_{s-w} + \Delta H_{w-w}$; $\Delta G_{solv} = \Delta H_{solv} - T\Delta S_{solv}$. Solvation thermodynamics in water site B could not be calculated for *h*-Gal-3 due to the very low occupancy of water molecules in this region.

The analysis results revealed two key findings: first, the free solvation energy of these water sites (ΔG_{solv}) is predominantly governed by the enthalpic component (ΔH_{solv}). Secondly, it was observed that the primary contribution to this solvation energy stems from water-host interactions, conclusively demonstrating the substantial importance of these water pockets in stabilizing the bound state. From the data depicted in Fig. 16, it becomes apparent that water is less structured around the ligand in *h*-Gal-3, constituting a clear difference with respect to the other studied galectins.

Moreover, water sites B and D exhibited a consistent qualitative trend characterized by slightly unfavorable entropic contributions, indicating a higher degree of water ordering compared to the bulk solvent. These water sites also display favorable water-host (either protein or ligand) interactions, thus minimizing water-water interactions.

A similar trend is observed in water site A for *h*-Gal-1 and *h*-Gal-7, where the entropic term opposes the enthalpic one. However, in the case of *h*-Gal-3 water site A shows a very small but favorable entropic term. This result aligns with the experimental observation that recognition of branched tetrasaccharides (B/A type II antigens) by *h*-Gal-3 minimizes the entropy penalty for ligand binding, as described by Jiménez-Barbero and co-workers [136].

Despite their high conservation, there are differences in the composition of galectin binding sites residues involved in the water pockets described above (Fig. 17). In *h*-Gal-1, water site A involves interactions between the *N*-acetamido group of the terminal GlcNAc unit of the ligand and residues Arg47, Asp53, Arg72; additionally, His51 can contribute to stabilization of the solvent molecules in water site A, or be oriented towards water site D. In the case of *h*-Gal-3, water site A involves also two arginine residues (Arg50, Arg74), while the aspartic acid residue in *h*-Gal-1 is replaced by a glutamic acid. In *h*-Gal-7, water site A shares the same interacting residues as *h*-Gal-3 (Arg188, Arg209 and Glu193) but it features an additional residue, Thr191, which contributes to the stabilization of this water site. The presence of four polar residues, as in *h*-Gal-1 and -7, seems to enhance the stabilization of the solvent molecules.

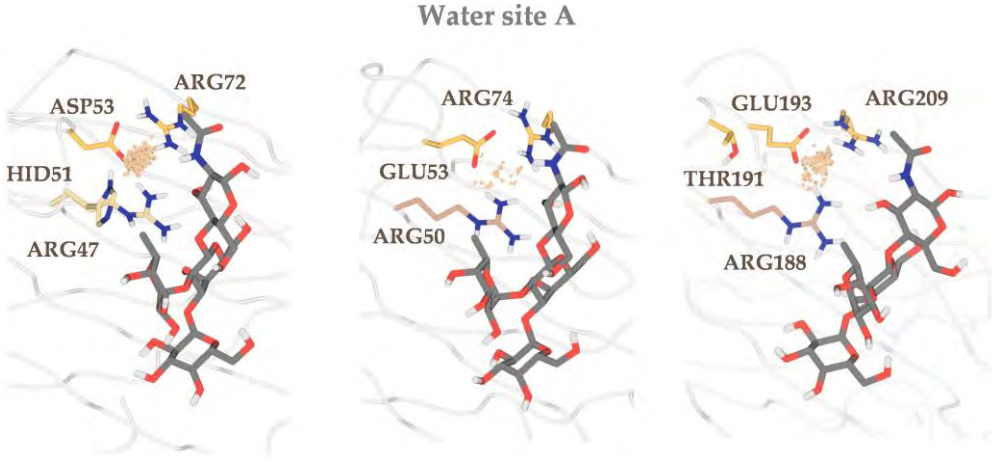


Figure 17. Representation of the residues bridged by the structured water sites between the proteins and the ligands in *h*-Gal-1:B type II, *h*-Gal-3:B type II and *h*-Gal-7:B type II complexes.

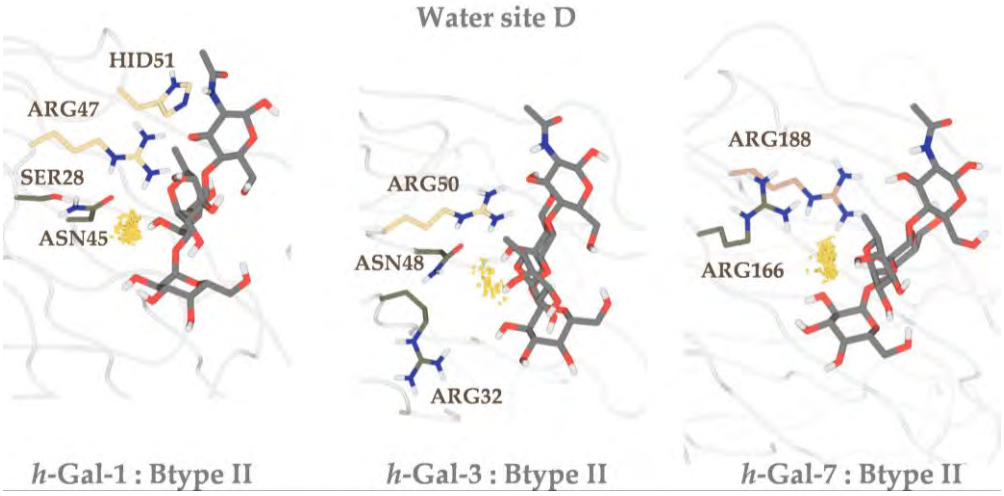


Figure 17 (cont.). Representation of the residues bridged by the structured water sites between the proteins and the ligands in *h*-Gal-1:B type II, *h*-Gal-3:B type II and *h*-Gal-7:B type II complexes.

Water site D exhibits an interesting pattern where one of the arginine residues involved in the interactions with water site A also contributes to water organization in water site D across all three galectins (Arg47 in *h*-Gal-1, Arg50 in *h*-Gal-3 and Arg188 in *h*-Gal-7). In *h*-Gal-1, an additional residue Asn45, assists in structuring water in site D. In the case of *h*-Gal-3, it shares these two residues (Arg32 and Asn48), but an arginine residue in that region occasionally occupies water site D, leading to its desolvation. Similarly, *h*-Gal-7 also possesses an additional arginine residue, Arg166, in that region. However, in this system, Arg166 exhibits reduced mobility and does not occupy the binding site; instead, it plays a role in structuring water molecules within water site D.

These subtle variations in the binding site composition of galectins contribute to differences in local microsolvation among *h*-Gal-1, *h*-Gal-3 and *h*-Gal-7. The specific residues and their dynamic behavior influence the position and dynamics of water molecules within the binding site, thereby affecting the overall hydration thermodynamics.

3. Conclusions

A comprehensive analysis of galectin-carbohydrate interactions, allosteric communication and solvation thermodynamics using molecular dynamics simulations, has been performed. The study provides insights into the structural, dynamic, and thermodynamic aspects of these interactions, shedding light on the role of conserved water patches and long-range motions. The differences in the composition of galectins binding sites are ultimately responsible of their specificities and determine local microsolvation dynamics and overall hydration thermodynamics. These findings might be useful for developing future applications in drug design or therapeutic interventions.

4. Methods

a. System preparation

The initial 3D models for *apo h*-Gal-1, *h*-Gal-3 and *h*-Gal-7 were built based on the X-ray crystallographic structures of *h*-Gal-1:LacNAc (PDB ID: 1W6N), *h*-Gal-3:LacNAc (PDB ID: 1KJL) and *h*-Gal-7:lactose (PDB ID: 4GAL) complexes.

To construct full-length models for *h*-Gal-4 and *h*-Gal-8, the crystallographic coordinates of their N- and C- terminal CRDs were used (Table 3). The peptide linkers connecting the two domains, which were unresolved in the crystal structures, were built in an extended conformation using the *tleap* module of Amber 20 [150]. The three fragments were manually assembled using PyMol [151] and the resulting full-length structure was refined through MD simulations (further datils provided below).

Table 3. PDB codes of the crystallographic structures used for constructing full-length models of *h*-Gal-4 and *h*-Gal-8 .

System	PDB
h-Gal-4 N-term	4XZP
h-Gal-4 C-term	5CBL
h-Gal-8 N-term	4BMB
h-Gal-8 C-term	3OJB

b. Molecular Dynamics Simulations

Molecular dynamics (MD) simulations were run with Amber 20 suite [150] using the *ff14SB* [152] and *GLYCAM 06j-1* [153] force fields for the proteins and carbohydrate ligands, respectively. Binding histidine residues (H43, H51, H176 and H184 for *h*-Gal-

1, H46 for *h*-Gal-3, H49 for *h*-Gal-7, H50 for *h*-Gal-4 N-T, H52 for *h*-Gal-4 C-T, H65 for *h*-Gal-8 N-T and H46 for *h*-Gal-8 C-T) were modelled in their Nδ1-H tautomeric state (residue name HID in Amber). Initial structures were neutralized with either Na^+ or Cl^- counterions and set at the center of a cubic TIP3P water [154] box with a buffering distance between solute and box of 10 Å. A two-stage geometry optimization approach was performed. The first stage minimizes only the positions of solvent molecules and ions, and the second stage is an unrestrained minimization of all the atoms in the simulation cell. The systems were then heated by incrementing the temperature from 0 to 300 K under a constant pressure of 1 atm and periodic boundary conditions. Harmonic restraints of 10 $kcalmol^{-1}$ were applied to the solute, under the Andersen temperature coupling scheme [155]. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Water molecules were treated with the SHAKE algorithm [156] such that the angle between the hydrogen atoms is kept fixed through the simulations. Long-range electrostatic effects were modelled using the particle mesh Ewald method. [157] An 8 Å cut-off was applied to Lennard-Jones interactions. Each system was equilibrated for 2 ns with a 2 fs timestep at a constant volume and temperature of 300 K. Ten independent production trajectories were run for additional 2.0 μs under the same simulation conditions, leading to accumulated simulation times of 20 μs for each system (*h*-Gal-1 *apo*, *h*-Gal-1:B type II, *h*-Gal-3 *apo*, *h*-Gal-3:B type II, *h*-Gal-7 *apo*, *h*-Gal-7:B type II, *h*-Gal-4 N-T *apo*, *h*-Gal-4 N-T:B type VI, *h*-Gal-4 C-T *apo*, *h*-Gal-4 C-T:A type VI, *h*-Gal-8 N-T *apo*, *h*-Gal-8 N-T:sialyl T antigen, *h*-Gal-8 C-T *apo*, *h*-Gal-8 C-T:A type II).

c. Cluster search

To obtain the most representative structure for full-length *h*-Gal-4 and *h*-Gal-8, initially modelled as fully extended conformations, the production trajectories for the *apo* and bound systems were combined, resulting in an accumulated simulation time of 20 μs for each system. Conformations were sampled every 40 ns and clustered based on the root-mean-square deviation (RMSD) of the N-terminal domain residues. The DBSCAN clustering algorithm [158] implemented in *cpptraj* [159] module of Amber 20 [150], was used for clustering analysis. A distance cut-off of 2 Å was applied for forming a cluster, with a minimum requirement of 50 points to form a cluster.

d. Protein-ligand interactions

To gain insight into the nature of these interactions, the aforementioned distances were extracted from the 20 μs accumulated trajectories for each complex. Histogram distributions were then fitted to the extreme function (Eq. 1, Fig. 18) for each distance in each system. The three parameters that define these distributions were extracted from the fitting (Fig. 18b): the amplitude (A), the equilibrium distance (x_c) and the

width (w). Tight interactions are characterized by high amplitude values and low width values (Fig. 18c), while loose interactions exhibit the opposite pattern (Fig. 18c).

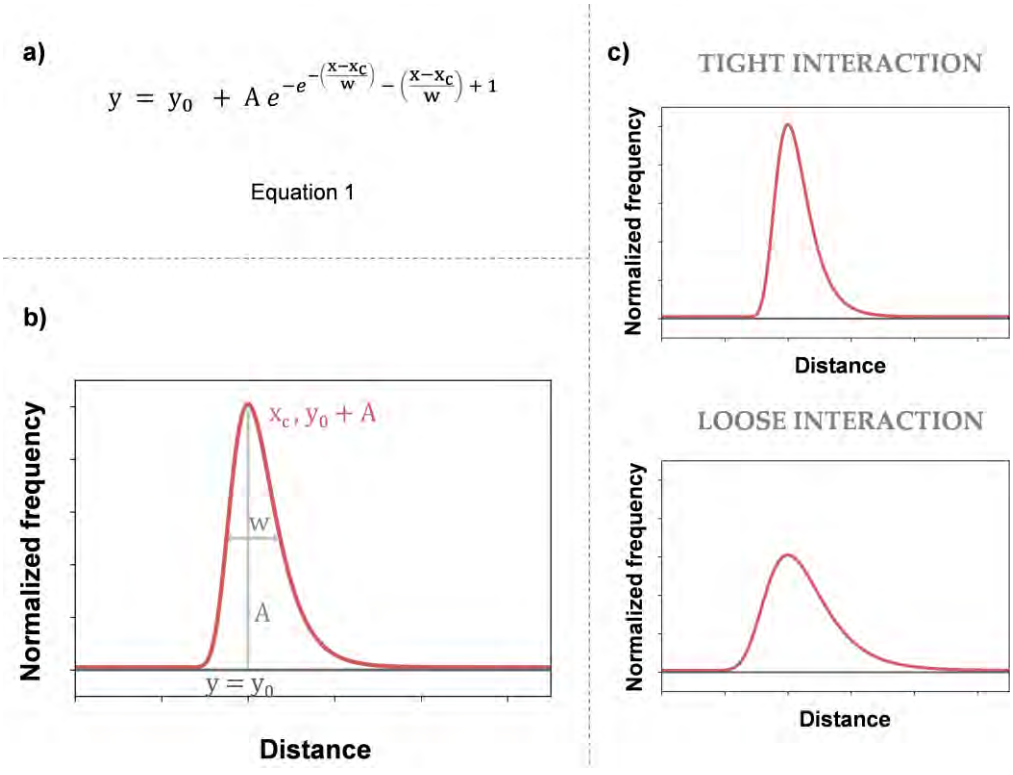


Figure 18. a) Definition of the extreme distribution. b) Example of a representation of an extreme distribution function. "A" stands for amplitude, "w" for width and "x_c" for the equilibrium distance. c) Representative examples of the distances distributions shapes for: a tight vs a loose interaction.

e. Allosteric pathways calculation

First, a correlation matrix (C_{ij}) is constructed using 1000 snapshots extracted every 100 ps from a window 100 ns within a converged MD trajectory. The correlation motion among pairs of nodes is calculated using Eq. 2. In this model, nodes are defined by the center of mass of side-chain residues, and nodes are considered to be in contact if the mean distance between them along the MD simulation is 6 Å or less. The length of the edges connecting these nodes quantifies the degree of dynamic communication between pairs of connected nodes, as defined in Eq. 3. The pathway length is inversely proportional to the correlation motion between nodes, indicating that shorter w_{ij} values denote tightly correlated or anticorrelated nodes, while larger values indicate less correlated nodes.

$$C_{ij} = \frac{\langle \Delta\vec{r}_i(t) \cdot \Delta\vec{r}_j(t) \rangle}{\sqrt{\langle \Delta\vec{r}_i(t)^2 \rangle \langle \Delta\vec{r}_j(t)^2 \rangle}} \quad \text{(Equation 2)}$$

$$w_{ij} = -\log(|C_{ij}|) \quad \text{(Equation 3)}$$

Then, Dijkstra's algorithm is used to generate all force-node paths, finding the shortest (i.e. optimal) path. WISP employs a bidirectional search approach to identify not only the optimal but also suboptimal pathways. Suboptimal pathways are defined as those closest in length to the optimal one, but not including it. The available code rapidly calculates both optimal and suboptimal communication pathways between two user-specified residues of a protein.

In this study, 100 pathways were computed between the conserved histidine residue of the binding sites and all the other residues in each of the studied galectins.

The so-called length of these calculated pathways can be mapped and color-coded onto the protein structure, providing a visual representation of the internal correlated motions. This allows for an intuitive visualization of the extent of dynamic communication within the protein. Additionally, the frequency with which a given amino acid is involved in the calculated allosteric pathways can be represented. This representation facilitates the identification of residues that operate as critical nodes to relay dynamic information inside the protein (Fig. 10 and 11)

f. Hydration analysis

i. Analysis of crystallographic water molecules in *apo* galectins.

Ordered waters in the *apo* form of galectins were identified from high resolution crystal structures. To this aim, X-ray structures from the highest resolution *apo* (when available) crystal structures of the studied galectins (Table 4) were superimposed in PyMol.

Table 4. PDB codes of human galectins used in this work.

System	PDB	Resolution (Å)
h-Gal-1	1W6N	1.65
h-Gal-3	3ZSM	1.25
h-Gal-7	1BKZ	1.90
h-Gal-4-NT	4XZP	1.48
h-Gal-4-CT	4YM3	1.89 [a]
h-Gal-8-NT	2YV8	1.92
h-Gal-8-CT	3OJB	3.01[b]

[a] *h*-Gal-4-CT:lactose.

[b] *h*-Gal-8-CT unique *apo* structure available.

ii. Prediction of conserved water molecules within protein-ligand complexes via MD simulations.

The structure, dynamics and thermodynamic properties of water molecules surrounding the bound ligands was explored using the grid inhomogeneous solvation theory (GIST) method [142] implemented in *cpptraj* [159] module of Amber 20 [150].

Carbohydrate binding to galectins is an inherently dynamic process, as reflected by the generally low affinities measured experimentally. Thus, MD simulations of such complexes commonly involve highly flexible scenarios both at the ligand and protein binding site level. Bearing this in mind, and to ensure reliable results, the GIST analysis required snapshots from MD simulations in which the solute was restrained to essentially one conformation. Thus, we investigated the conformational variability of them in solution. A clustering analysis was performed on the first 10 ns of trajectories of each system, which were later used for the GIST analysis. Conformations were sampled every 20 ps and clustered attending to the root-mean-square deviation (RMSD). The DBSCAN clustering algorithm [160] was used as implemented in *cpptraj* [159] module of Amber 20. The distance cutoff between points for forming a cluster was set to 3.5 Å. At least 50 points were required to form a cluster. All the studied complexes exhibited a single unique cluster (Table 5), and the entire analyzed trajectories belonged to these clusters (frac = 1). The maximum average distance between points in the cluster was found to be 1.91 Å, with a maximum standard deviation of points in the clusters of 0.28 Å.

Table 5. Cluster analysis of MD trajectories. *Frac* refers to the size of the cluster as a fraction of the total trajectory. *AvgDist* is the average distance between points in the cluster and *Stdev* is the standard deviation.

Complex	Frac	AvgDist	Stdev
h-Gal-1:B type II	1	1.70	0.18
h-Gal-3:B type II	1	1.62	0.15
h-Gal-7:B type II	1	1.91	0.28
h-Gal-4 NT:B type VI	1	1.52	0.13
h-Gal-4 CT:A type VI	1	1.68	0.14
h-Gal-8 NT:sialyl-T antigen	1	1.89	0.20
h-Gal-8 CT:A type II	1	1.68	0.16

Given that the selected trajectory windows maintained a consistent ligand conformation, they were selected as input trajectories for the GIST analysis.

- **Visualization of water structure and dynamics**

  First, the global hydration of galectin binding sites was analyzed by visual inspection of the results of a GIST analysis. For this initial analysis, GIST was performed on a box grid of dimensions 25 Å x 25 Å x 25 Å and a grid spacing of 0.5 Å, centered in the central lactose unit common to all the ligands studied (Fig. 19).

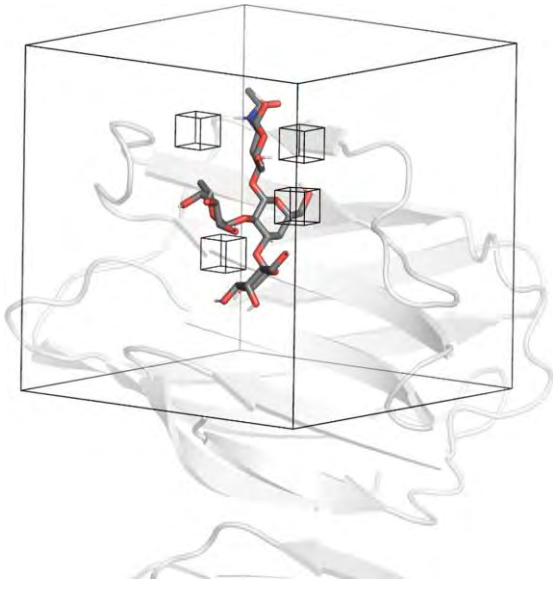


Figure 19. Grids used for the two-stage GIST analysis (*h*-Gal-1:B type II complex shown as an example). The larger box illustrates the grid used for the global hydration analysis whereas the smaller boxes are the ones used for the analysis of the local contributions to solvation (i.e. water sites).

For all the studied complexes, the top 5 most populated voxels were selected (shown as large spheres in Fig. 20). These voxels also met two additional criteria:

- The fraction of oxygen atoms found in the voxel with respect to the bulk density is higher than 3.5.
- The voxel is placed within 10 Å of the center of the box (i.e. the central galactose unit).

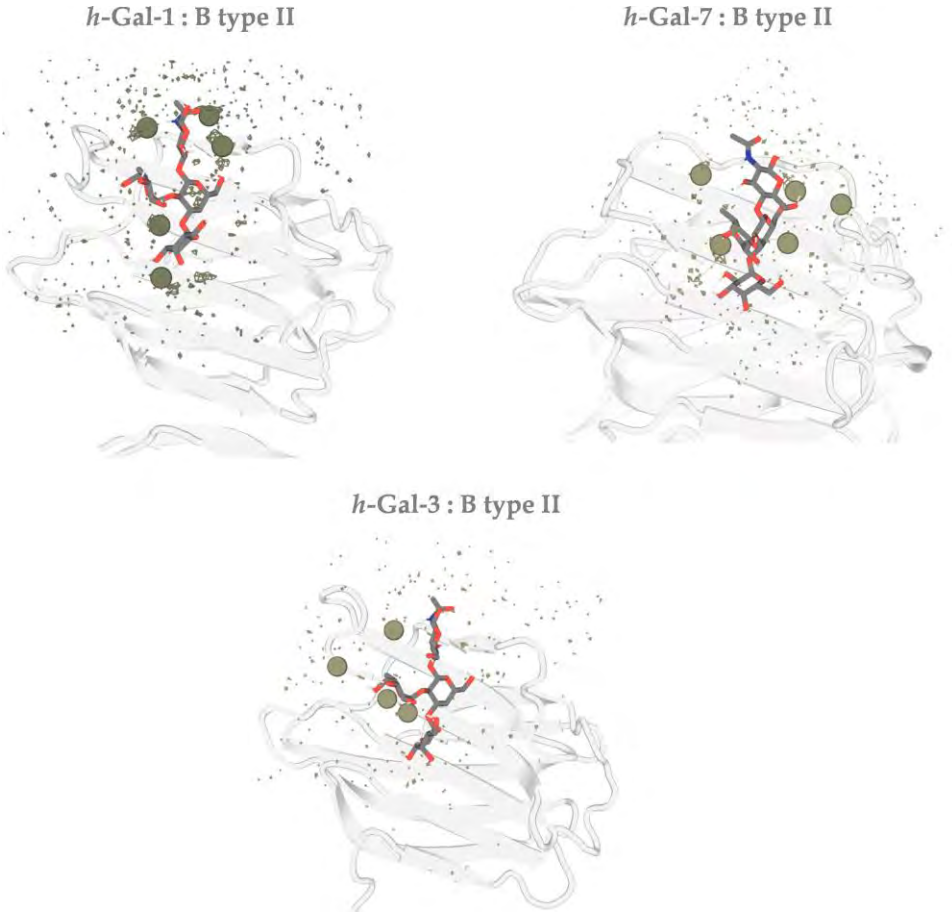


Figure 20. Local solvation patches around galectins binding sites calculated with GIST from MD simulations and represented as volumetric meshes. Large spheres represent the top 5 most populated voxels that are within 10 Å of the galactose rings. Meshed volumes represent the hydration regions found with the help of the GIST analyses.

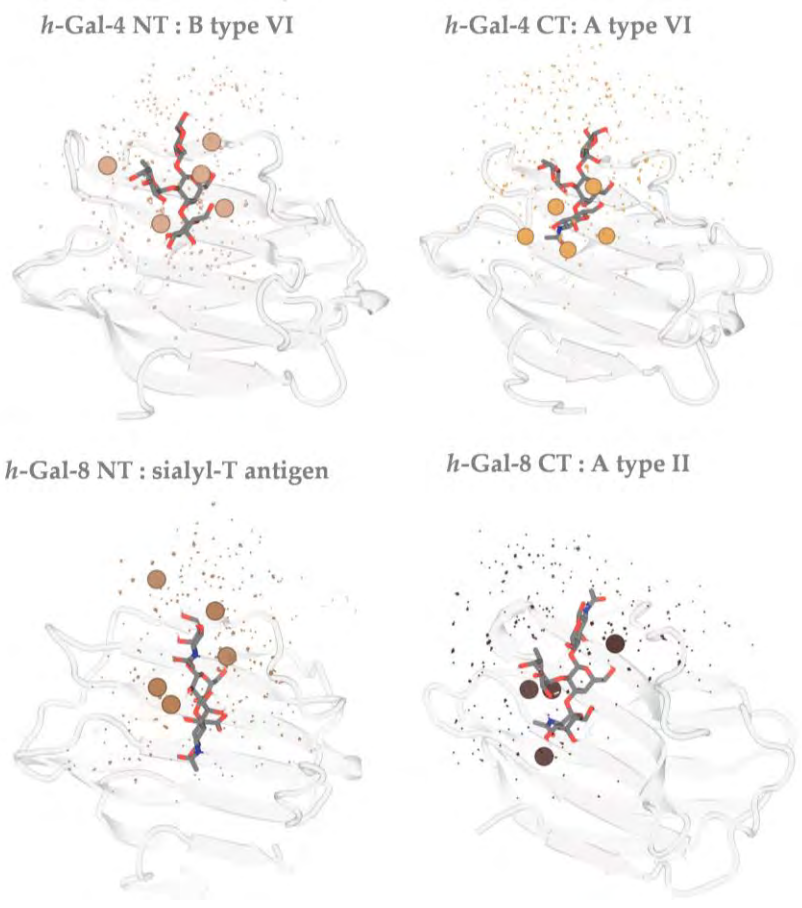


Figure 20 (cont.). Local solvation patches around galectins binding sites calculated with GIST from MD simulations and represented as volumetric meshes. Large spheres represent the top 5 most populated voxels that are within 10 Å of the galactose rings. Meshed volumes represent the hydration regions found with the help of the GIST analyses.

- **Quantitative analysis of local contributions to solvation**

To analyze the local contributions to the hydration thermodynamics of galectins, three subregions were defined. Each subregion corresponds to smaller box grids with dimensions of 2.5 Å x 2.5 Å x 2.5 Å, comprising a total of 10 x 10 x 10 voxels (1000 voxels in total). The grid spacing was set at 0.25 Å, ensuring detailed resolution within the subregion. These boxes were built centered on highly structured waters within subregions A, B, and D (Fig. 18). The enthalpic and entropic contributions are calculated as indicated in the original GIST manuscript [142] (Equations 4-8). The thermodynamic properties of each water site (ΔS_{orient}, ΔS_{trans}, ΔH_{s-w}, ΔH_{w-w}) are internally calculated by the program as the sum of these

values at each voxel comprising the water site grid, multiplied by the voxel volume (0.016 Å^3)

$$\Delta G_{solv} = \Delta H_{solv} - T\Delta S_{solv} \quad \text{(Equation 4)}$$

$$\Delta S_{solv} = \Delta S_{s-w} + \Delta S_{w-w} \quad \text{(Equation 5)}$$

$$\Delta S_{solv} \approx \Delta S_{s-w} \quad \text{(Equation 6)}$$

$$\Delta S_{s-w} = \Delta S_{s-w}^{trans} + \Delta S_{s-w}^{orient} \quad \text{(Equation 7)}$$

$$\Delta H_{solv} = \Delta H_{s-w} + \Delta H_{w-w} \quad \text{(Equation 8)}$$

This work has been published in the following articles:

A Computational Perspective on Molecular Recognition by Galectins. R. Núñez-Franco, F. Peccati, G. Jiménez-Osés. *Curr. Med. Chem.* **2021**, *28*, 1–13.

The two domains of human galectin-8 bind sialyl- and fucose-containing oligosaccharides in an independent manner. A 3D view by using NMR. M. Gómez-Redondo, S. Delgado, R. Núñez-Franco, G. Jiménez-Osés, A. Ardá, J. Jiménez-Barbero, A. Gimeno. *RSC Chem. Biol.* **2021**, *2*, 932–941.

Galectin-4 N-Terminal Domain: Binding Preferences Toward A and B Antigens With Different Peripheral Core Presentations. J. I. Quintana, S. Delgado, R. Núñez-Franco, F. Javier Cañada, G. Jiménez-Osés, J. Jiménez-Barbero, A. Ardá. *Front. Chem.* **2021**, *9*, 664097.

Unravelling the Time Scale of Conformational Plasticity and Allostery in Glycan Recognition by Human Galectin-1. S. Bertuzzi, A. Gimeno, R. Núñez-Franco, G. Bernardo-Seisdedos, S. Delgado, G. Jiménez-Osés, O. Millet, J. Jiménez-Barbero, A. Ardá. *Chem. Eur. J.* **2020**, *26*, 15643–15653.

Chapter 4

Investigating the structural basis of sugar recognition by DC-SIGN: uncovering a minimum binding epitope.

1. Introduction

C-type lectins (CTL) are Ca^{2+}-dependent carbohydrate-binding proteins. They feature a carbohydrate-recognition domain (CRD) that exhibits structural similarities across different C-type lectins. Oligomerization is common among CTLs, increasing their avidity for multivalent ligands and enhancing their recognition via pattern recognition receptors [161]. Some well-known proteins belonging to the C-type lectin family include DC-SIGN, mannose receptor, dectin-1 and selectins.

The dendritic cell-specific ICAM-3-grabbing non-integrin (DC-SIGN or CD209) is one of the most studied C-type lectins over the last twenty years as it is a key piece in the infection by Human Immunodeficiency Virus (HIV) [162]. DC-SIGN is a transmembrane protein primarily expressed on the surface on immature dendritic cells and macrophages [163]. Dendritic cells play a crucial role in the immune response, acting as antigen-presenting cells and triggering adaptative immune responses. DC-SIGN is characterized by its ability to bind a wide range of carbohydrates present on pathogens, such as bacteria, viruses, and fungi (Fig. 1). This interaction has been described as a critical step when capturing and processing pathogens by dendritic cells, initiating in that way specific immune responses.

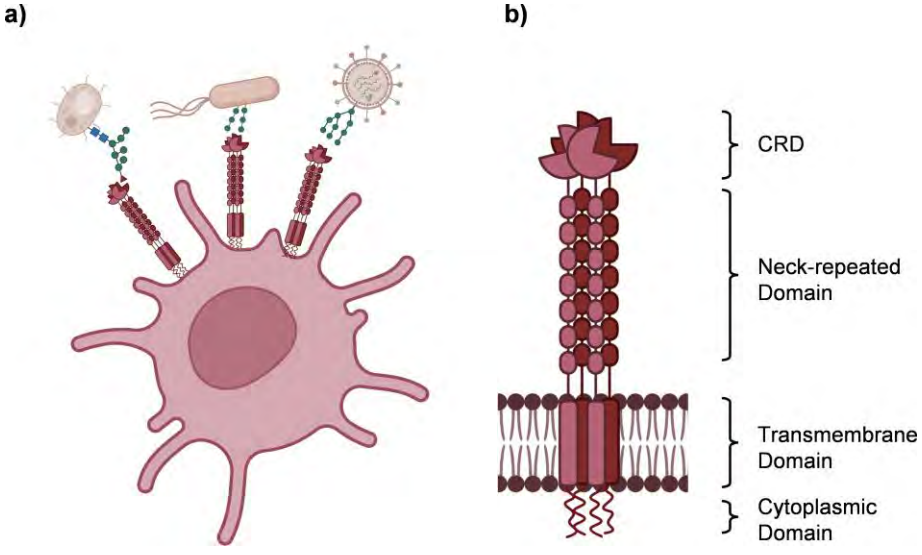


Figure 1. a) Schematic representation of a dendritic cell with DC-SIGN receptors on its membrane. The recognition processes of bacteria, fungi, and virus are exemplified. b) Tetramer DC-SIGN schematic structure formed by: Carbohydrate Recognition Domain (CRD), a neck region composed for seven and one incomplete repeats of a 23 amino acid sequence, and a transmembrane domain followed by a cytoplasmic tail.

DC-SIGN exhibits a remarkable level of promiscuity in its ability to bind carbohydrates, making it a versatile receptor for pathogen recognition, immune cell adhesion, and modulation of immune responses [26]. However, the broad range of carbohydrate ligands also presents challenges in understanding the specificities and functional consequences of these interactions. Elucidating the molecular basis of DC-SIGN's promiscuity and its implications in host-pathogen interactions is an active area of research that can contribute to the development of therapeutics and vaccines targeting DC-SIGN-mediated immune responses.

In most of C-type lectins, the carbohydrate recognition process takes place through the coordination of calcium by two vicinal hydroxyl groups in the monosaccharide (Fig. 2a). This interaction is relatively weak, with affinity in the millimolar range [26]. However, higher affinities can be achieved through additional contacts involving residues outside the conserved ones at secondary sites, and from the multivalent architecture presentation of the lectin allowing the generation of sugar-lectin clusters.

In this chapter, an in-depth investigation is conducted into the recognition of monosaccharides by DC-SIGN, using a wide range of computational techniques. The computational findings are further validated and complemented by NMR experiments and CORCEMA calculations carried out in collaboration with the Chemical Glycobiology Lab at CIC bioGUNE led by Prof. Jesús Jiménez-Barbero. The main objective is to gain a comprehensive understanding of the molecular mechanisms involved in sugar recognition by DC-SIGN, with the potential to guide future applications.

2. Results and discussion

a. Structural characterization of the minimum binding epitope and binding patterns to DC-SIGN

First, a systematic search was conducted on all available crystallographic structures of protein-ligand complexes involving L-Fuc and D-Man oligosaccharides in association with DC-SIGN, including synthetic glycomimetics. This investigation aimed at identifying common structural features irrespective of the specific sugar type, leading to the discovery of a conserved arrangement of atoms referred as the minimum ligand binding epitope (Fig. 2a). Notably, the positions of the vicinal diol (COH-COH) motif, which coordinates the Ca^{2+} ions, remained constant, while the atomic coordinates of the pyranose ring varied (Fig. 2a). The presence of this minimum epitope in all crystallized protein-ligand complexes highlights its importance for ligands to effectively bind to DC-SIGN. To precisely characterize the spatial distribution of the

COH-COH motif, a plane was defined by bisecting the C-C bond of this minimum binding epitope and passing through the three Ca^{2+} ions of the lectin (Fig. 2b). The solvent-exposed side of the plane was arbitrarily assigned as positive (+), with the labeled ^{+}O representing the oxygen atom in that region. Conversely, the side of the plane facing the protein was designated as negative (–), with the labeled ^{-}O representing the corresponding oxygen atom.

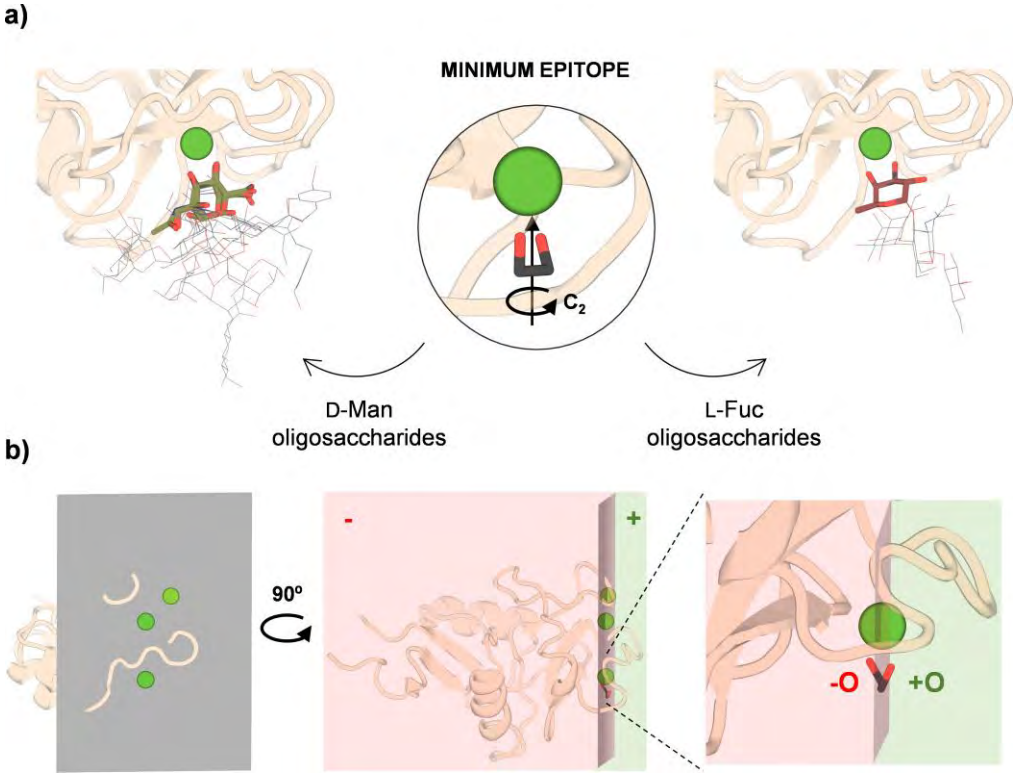


Figure 2. a) Overlay of DC-SIGN's crystallographic structures with D-Man (left) and L-Fuc (right) oligosaccharides. The calcium-interacting monosaccharide in each structure is represented with thick sticks (D-Man is shown in green and L-Fuc in red). Middle: the conserved minimum pattern of atomic types and positions found in all structures (minimum binding epitope). b) The reference plane passing through the three Ca^{2+} ions of the lectin (in gray) bisects the minimum binding epitope through the C-C bond. Each oxygen atom of the minimum binding epitope lies in a different side (– or +) of the plane, and thus are labelled as ^{-}O and ^{+}O respectively. Most of the protein's structure is located at the – side of the plane, whereas the + side is solvent exposed.

In line with the hypothesis of a minimal motif being largely responsible of carbohydrate binding to DC-SIGN, a comprehensive analysis was conducted on the orientations of α-OMe-L-Fuc and α-OMe-D-Man at the binding site by superimposing all possible permutations of their vicinal CO-CO pairs onto the geometry of the minimum binding epitope. Only structures showing a nearly perfect alignment with

the minimum binding epitope (RMSD < 0.15 Å) were considered as potential binding modes to DC-SIGN (Scheme 1, see Methods)

Scheme 1. Protocol followed to analyze the proposed binding modes (BM = binding mode)

Considering the C2 symmetry axis exhibited by the minimum epitope (Fig. 2a), each sugar containing a Ca^{2+}-coordinating COH-COH pair can adopt at least two potential binding orientations. For α-OMe-L-Fuc and α-OMe-D-Man, four binding permutations fulfill the requirements of the minimum binding epitope outlined in Scheme 1 (see Methods). To characterize these binding modes, a two-digit nomenclature was used, indicating the atomic positions of $^{+}$O and $^{-}$O oxygen atoms in the pyranose ring. Thus, the resulting binding poses for α-OMe-L-Fuc and α-OMe-D-Man fulfilling the minimum epitope are labeled as 2-3, 3-2, 3-4, and 4-3, with the 1-2 and 2-1 arrangements being hindered by the presence of the methoxy group (OMe) at the anomeric center.

Considering the six-membered cyclic nature of the ligand, this analysis revealed three distinct Structural Binding Patterns (SBPs) for both α-OMe-L-Fuc and α-OMe-D-Man in the crystallographic DC-SIGN complexes, which remains consistent regardless of the sugar type. These SBPs are denoted as A, B1, and B2 (Fig. 3) and are characterized as follows:

- SBP A: both the $^{+}$O and $^{-}$O oxygen atoms occupy equatorial positions in the pyranose ring.
- SBP B1: the $^{+}$O and $^{-}$O oxygen atoms occupy axial and equatorial position in the pyranose ring, respectively.
- SBP B2: the $^{+}$O and $^{-}$O oxygen atoms occupy equatorial and axial positions in the pyranose ring, respectively.

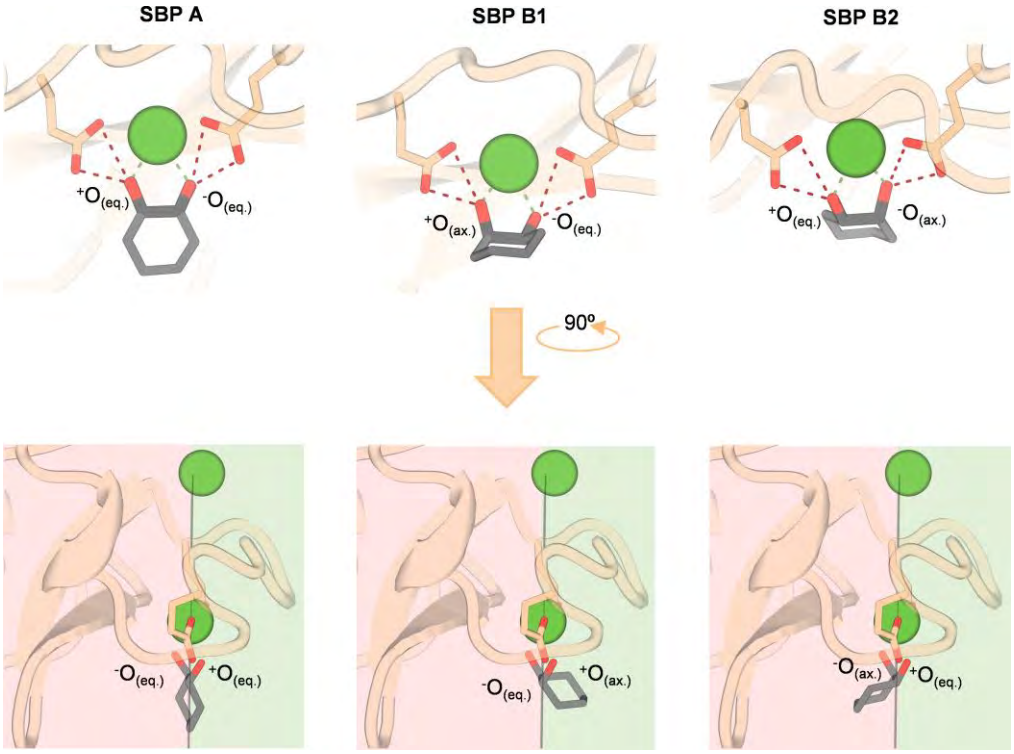


Figure 3. Structural binding patterns (SBP) A, B1, and B2 found in carbohydrate ligands bound to DC-SIGN. Sugar pyranoses have been simplified to a substituted cyclohexane ring (shown in sticks), and only the hydroxyl groups interacting with the Ca^{2+}-binding motif are shown.

Based on the defined structural patterns, the potential binding modes of α-OMe-L-Fuc and α-OMe-D-Man to DC-SIGN can be classified as presented in Table 1.

Table 1. Classification of the binding modes of α-OMe-L-Fuc and α-OMe-D-Man to DC-SIGN based on Structural Binding Patterns (SBPs). eq = equatorial; ax = axial.

Sugar	Binding pose	SBP	$^{+}O/^{-}O$ configuration
α-OMe-L-Fuc	*2-3*	A	eq/eq
	3-2	A	eq/eq
	3-4	B2	eq/ax
	4-3	B1	ax/eq
α-OMe-D-Man	*2-3*	B1	ax/eq
	3-2	B2	eq/ax
	3-4	A	eq/eq
	4-3	A	eq/eq

b. Energetic analysis of Structural Binding Patterns towards DC-SIGN

In order to assess the relative stability of the three Structural Binding Patterns (A, B1, and B2) at the lectin binding site, a quantum mechanical cluster model [62] was employed. This method focuses on modeling the protein's binding site and the ligand, with their positions being fixed either completely or partially [164]. By combining quantum mechanics with a continuum solvent model [165], an accurate description of the system's energetics, specifically ligand binding, was obtained (see Methods for details).

To determine the stability and energetics, the minimum energy conformers of α-OMe-L-Fuc and α-OMe-D-Man bound to the protein cluster model were considered for the four binding poses previously described. Using an analogous di-aquo complex as a representation of the protein's *apo* state, the binding enthalpy (ΔH_{bind}) of the conformers was calculated through an isodesmic reaction (Fig. 4). Furthermore, α-OMe-L-Gal, which is similar to α-OMe-L-Fuc but has a hydroxymethyl group instead of a methyl group at C5, was also included in this analysis. Additionally, the impact of incorporating the sidechain of Val351, known to play a significant role in the recognition of fucosylated sugars by DC-SIGN [166,167], was taken into account (see Methods).

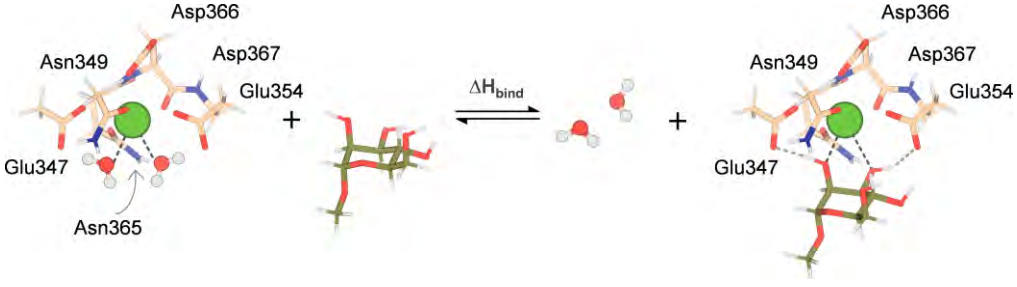


Figure 4. Example of an isodesmic reaction used to calculate the relative binding enthalpy of α-OMe-D-Man to DC-SIGN through a cluster model. Calcium atoms are represented as green spheres. Water molecules are represented as balls and sticks. Carbohydrates are represented as sticks.

The results of the energetic calculations are depicted in Fig. 5. It is observed that the B1 structural binding patterns ($^{+}O_{ax}/^{-}O_{eq}$, Table 2), which corresponds to the 2-3 binding mode for α-OMe-D-Man and the 4-3 binding mode for α-OMe-L-Fuc and αOMe-L-Gal, exhibits higher affinities to DC-SIGN by at least 2 kcal mol^{-1} (>95% population at 25 °C).

These findings suggest an important contribution of the outer-sphere hydrogen bond formed between a third pyranose hydroxyl group adjacent to those coordinated to Ca^{2+}, and the residue Glu354, which is a characteristic feature of this binding mode (Fig. 5c).

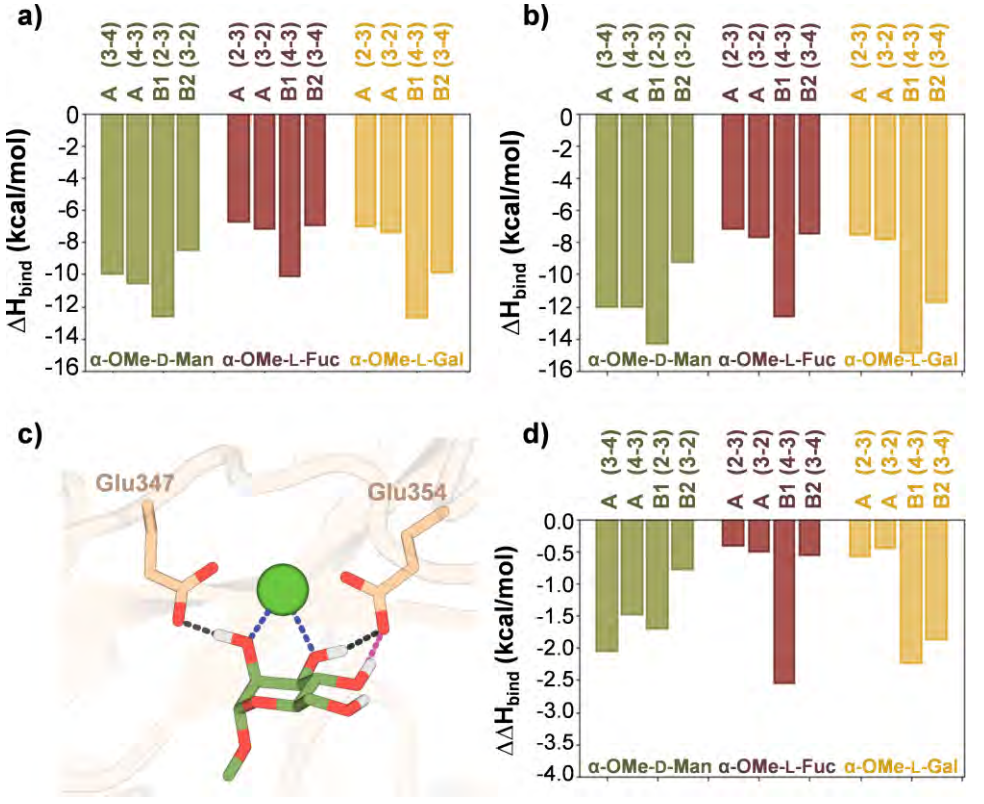


Figure 5. Binding enthalpies (ΔH_{bind}) calculated for each binding mode of α-OMe-D-Man (green), α-OMe-L-Fuc (red), and α-OMe-L-Gal (yellow) a) in the absence and b) in the presence of Val351 sidechain. c) α-OMe-D-Man bound to DC-SIGN in the B1 structural binding pattern. Ca^{2+}-oxygen interactions and protein-ligand hydrogen bonds common to all binding modes are depicted as blue and black dashed lines respectively. The extra hydrogen bond taking place in B1 binding modes is depicted with pink dashed lines. d) Relative binding enthalpies ($\Delta\Delta H_{bind}$) calculated for each binding mode of α-OMe-D-Man (green), α-OMe-L-Fuc (red), and α-OMe-L-Gal (yellow) in the presence vs. absence of Val351 sidechain.

Furthermore, the presence of the Val351 sidechain consistently enhances the binding affinity of all sugars by approximately 2 kcal mol^{-1} through dispersion interactions [41,168,169]. This effect maintains or further strengthens the overall preference for the B1 pattern (Fig. 5d). Notably, this effect is more pronounced for α-OMe-L-Fuc and α-OMe-L-Gal.

c. Stability of Structural Binding Patterns in complex with DC-SIGN in solution

The quantum mechanical cluster model analysis provides insight into the inherent strength of the carbohydrate-lectin binding interactions with DC-SIGN based on existing crystallographic structures. However, it is important to acknowledge that such interactions might be different in solution. Therefore, to account for the influence of surrounding water molecules, ions, and thermal effects, microsecond molecular dynamics simulations were performed on the carbohydrate-lectin CRD complexes, starting from all the different structural binding patterns (see Methods).

The three studied carbohydrates (α-OMe-D-Man, α-OMe-L-Fuc, and α-OMe-L-Gal) exhibited similar ligand residence time profiles throughout the simulations, as depicted in Fig. 6. Consistently with the higher affinity calculated using quantum mechanics, ligands in the B1 binding poses remained bound the entire simulation. In contrast, the other binding modes showed moderate persistence in solution, with the exception of one of the A binding modes that displayed a notably unstable behavior (ligand residence frequency < 20%).

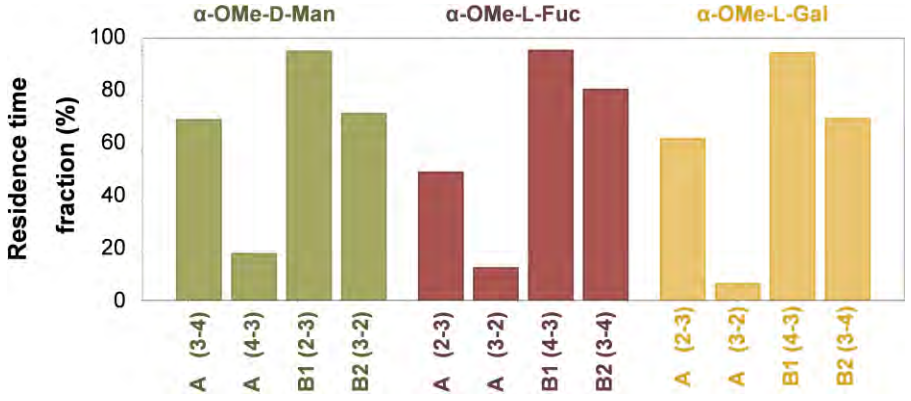


Figure 6. Ligand residence frequencies derived through MD simulations for α-OMe-D-Man (in green), α-OMe-L-Fuc (in red), and α-OMe-L-Gal (in yellow) bound to DC-SIGN in each different pose. Residence frequencies are calculated as the fraction of bound complex over the entire simulation time (1 μs).

The hydrogen bonds between the carbohydrate hydroxyl groups coordinating the Ca^{2+} ion and glutamic acid residues Glu347 and Glu354 were conserved across all structural binding patterns (Fig. 7). Other hydrogen bonds involving the same coordinating OH groups, such as those with Ca^{2+}-binding Asn365, Asp366 and Asn349, were found to be stronger in the B1 motif compared to the other structural patterns. Of note, the characteristic hydrogen bond in motif B1 between the contiguous equatorial hydroxyl group not directly involved in Ca^{2+} binding and residue Glu354 (Fig. 5), was consistently conserved throughout most of the trajectory for the three ligands. These

hydrogen bonds were previously observed in the quantum mechanical study, and MD simulations demonstrated their remarkable persistence even in an aqueous solution.

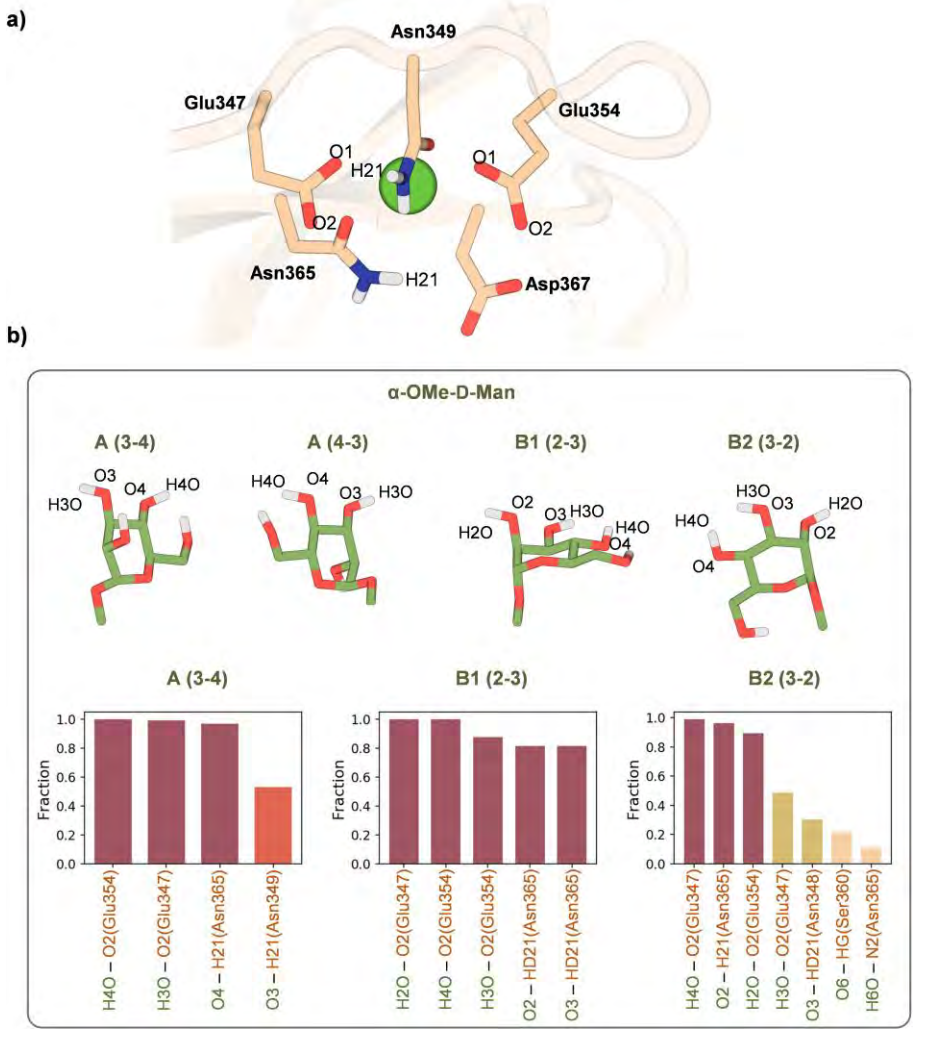


Figure 7. a) Atoms involved in the main hydrogen bond interactions between the carbohydrate ligands and DC-sign binding site. b) Hydrogen bonds measured for each ligand-protein complex in different binding modes during the section of the MD simulation in which the ligand is bound to DC-SIGN. The fraction of the ligand-bound snapshots in which the hydrogen bonds occur is shown in the bar plots.

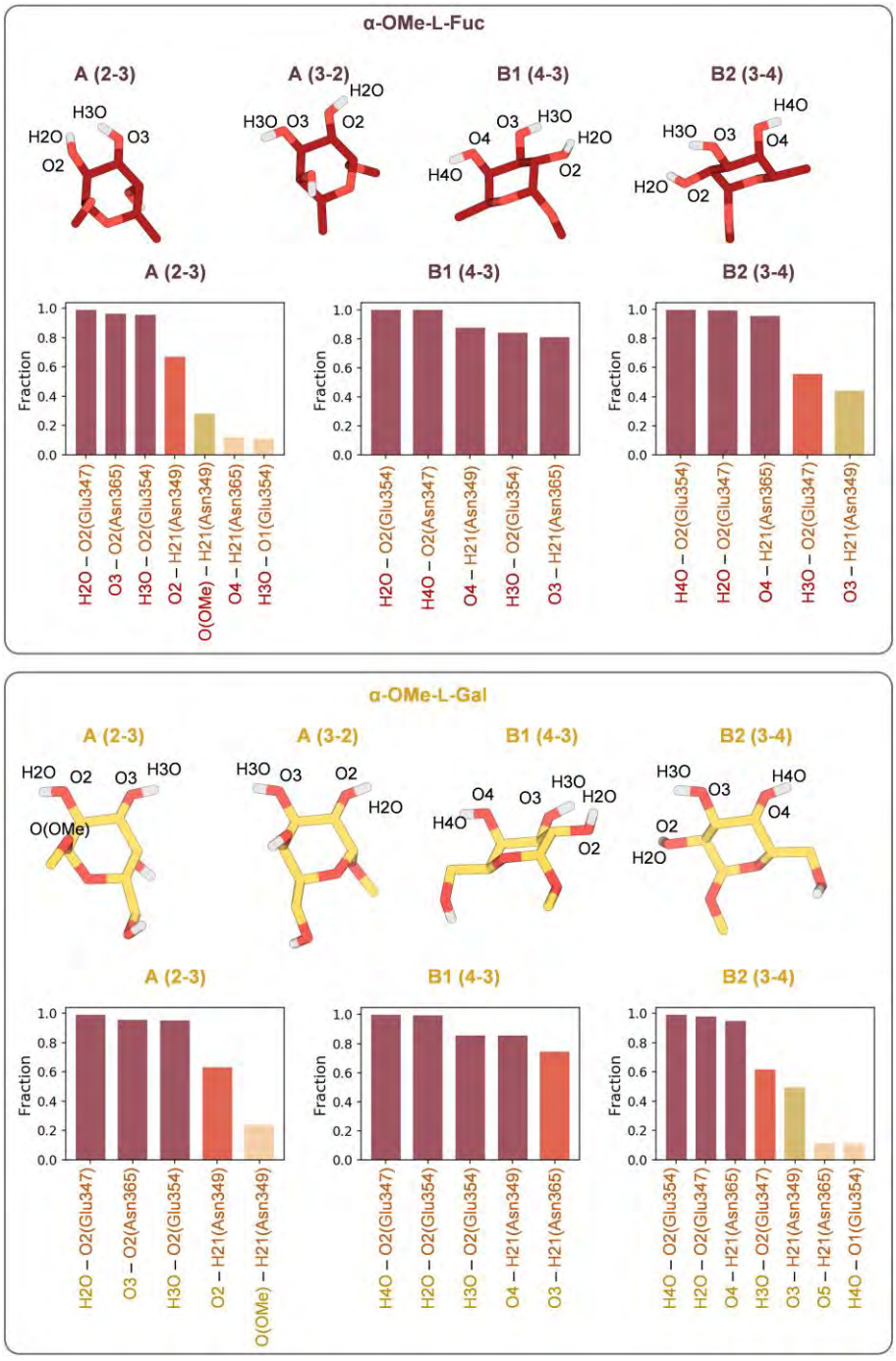


Figure 7 (cont.). a) Atoms involved in the main hydrogen bond interactions between the carbohydrate ligands and DC-sign binding site. B) Hydrogen bonds measured for each ligand-protein complex in different binding modes during the section of the MD simulation in which the ligand is bound to DC-SIGN. The fraction of the ligand-bound snapshots in which the hydrogen bonds occur is shown in the bar plots.

Additionally, specific hydrophobic contacts between the carbohydrate and the protein were exclusively detected in the B1 arrangement. These contacts involve the sidechain of Val351 and the C1H and C2H (with α-OMe-L-Fuc and α-OMe-L-Gal) or the C4H and C6H2 groups (with α-OMe-D-Man), as shown in Fig. 8 and Fig. 9. Such van der Waals interactions have been reported previously in both experimental and computational studies [41,168,169] and may contribute significantly to the enhanced stability of the B1 structural pattern compared to the other alternatives, as computed quantum mechanically in the previous section.

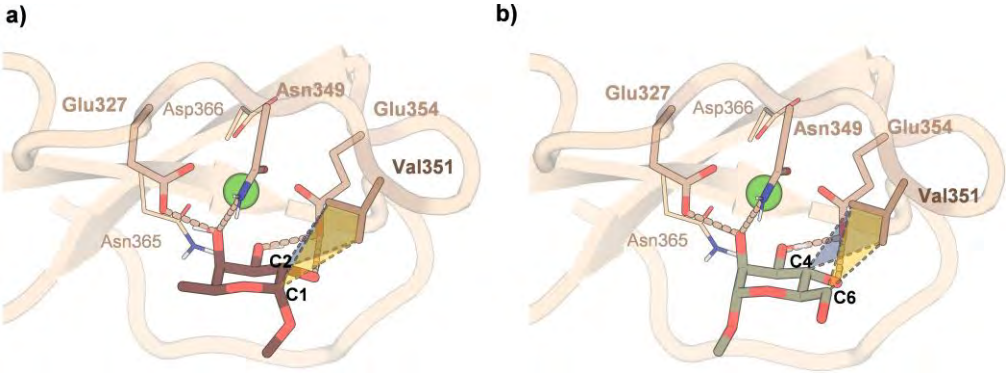


Figure 8. Representative snapshots from MD simulations of a) α-OMe-L-Fuc (in dark red) and b) α-OMe-D-Man (in green) bound to DC-SIGN (in light brown) in the B1 binding structural pattern. Protein residues involved in Ca^{2+} binding are shown as lines, and residues involved in direct contacts with the carbohydrates are shown as sticks. Conserved hydrogen bonds are shown as light brown dashed lines whereas van de Waals contacts with Val351 are represented as yellow and blue planes. Calcium atoms are represented as green spheres.

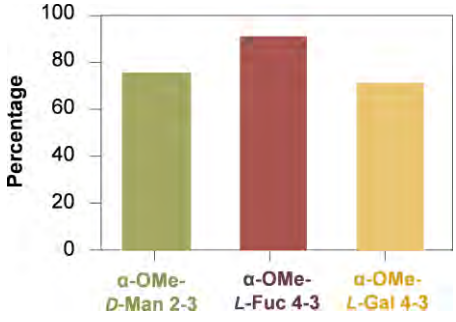


Figure 9. Occurrence of Van der Waals interactions between any non-polar hydrogen of the sugar (CH) and Val351 found on the MD trajectories, expressed as % of the total simulation time in which the ligand is bound to DC-SIGN. Van der Waals (VdW) interactions are considered to take place when the distance between the pair of atoms involved is lower than 1.2 times the sum of their VdW radii. Only the binding modes in which these interactions are found in more than 10% of the simulated time are represented in this graph (all of them correspond to B1 structural patterns).

d. STD-NMR experiments and CORCEMA analysis

To experimentally confirm the computationally predicted prevalence of the B1 structural pattern in the binding of single monosaccharides to the lectin, the interactions between α-OMe-D-Man and α-OMe-L-Fuc with the extracellular domain of DC-SIGN were investigated by the Chemical Glycobiology Lab at CIC bioGUNE using ^{1}H-STD-NMR experiments. Saturation-Transfer Difference (STD) NMR is a useful technique for characterizing the ligand's position at the binding site, thereby providing insight into the existence of distinct binding geometries.

The experiments were conducted with a large excess of the monosaccharides to minimize any potential re-binding effects. The results (Fig. 10) showed that α-OMe-D-Man had the strongest response at proton H4, with moderate effects observed for H3 and H6. For α-OMe-L-Fuc, the highest response was observed for protons H1 and H2, which matched the saturation patterns found in previous studies with DC-SIGN and fucose-containing oligosaccharides [168,169].

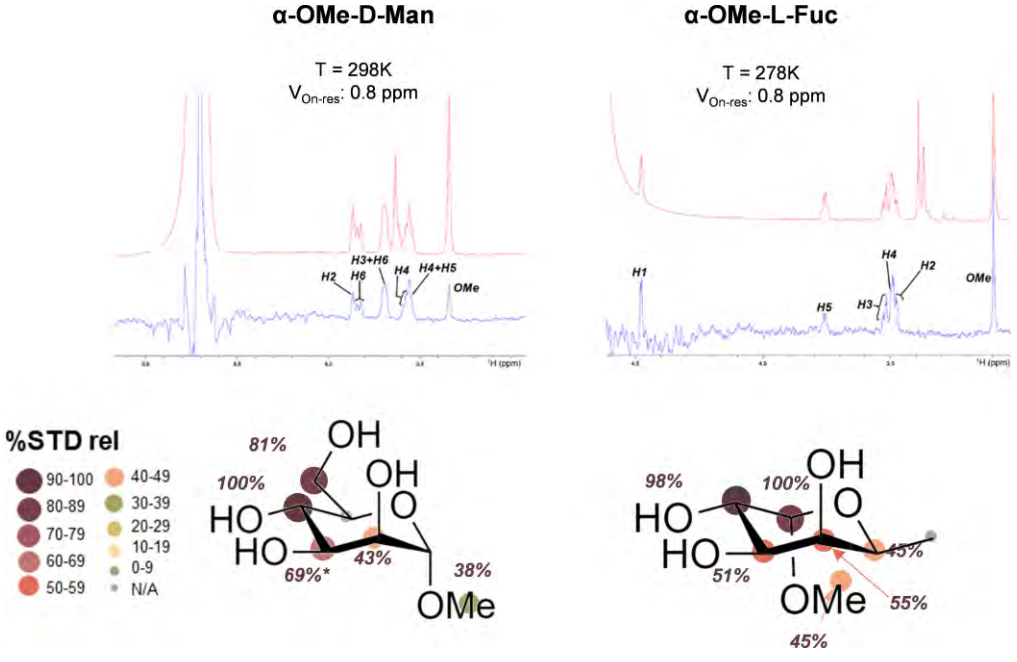


Figure 10. ^{1}H-STD spectra of a α-OMe-D-Man and α-OMe-L-Fuc in the presence of DC-SIGN ECD at t_{sat} = 2 s (blue spectra). The off -resonance spectra are shown in red. Relative STD intensities using saturation grown rates at t_{sat} = 0 s shown below the spectra in each case.

The experimental STDs were quantitatively analyzed using full matrix relaxation calculations (CORCEMA-ST) [170] and the MD trajectories of each complex with the ligand in the different binding poses (see Methods). Unlike using a single averaged

structure to predict theoretical STDs, subsets of structural ensembles derived from the MD simulations were used. This approach accounts better for the dynamic behavior of the molecules in solution. Finally, the average STD from all individual MD snapshots was calculated and compared to the experimental data using a NOE R-factor function (see Methods). In short, the binding mode whose MD ensemble yields the lowest R-factor is the one best reproducing the experimental STD data. This methodology allowed for a more comprehensive evaluation of the binding interactions between the monosaccharides and DC-SIGN.

The B1 structural pattern exhibited the best correlation with the experimental STD intensities in both the α-OMe-D-Man (NOE R-factor 0.47) and α-OMe-L-Fuc (NOE R-factor 0.37) complexes with DC-SIGN (Fig. 11a). The comparison between the averaged STD intensity of the ensemble for the highest-ranked binding pose (resulting in the lowest NOE R-factor) and the experimental STD showed good agreement (Fig. 11b). Interestingly, the other binding poses ranked significantly worse, with the relative deviations from the best pose (B1) ranging from 80% to 100% for both ligands.

Besides introducing the intrinsic dynamics of each binding mode trough an MD ensemble when comparing predicted and experimental STD, it is also worth considering the possibility of multiple binding modes occurring simultaneously in solution when interacting with the receptor, as previously observed for DC-SIGN with mannose oligosaccharides [171]. To investigate this, the BM-Mixer [171] procedure was applied. Results consistently showed that including structures belonging only the B1 binding modes yields the best agreement with the experimental data (Fig. 11c), indicating a single-mode interaction with the receptor. This finding aligns with previous studies that identified the B1 binding mode in D-Man [172]. Interestingly, the crystal structures of D-Man oligosaccharides revealed a *different* binding mode (A), which was not observed with natural monosaccharides. This discrepancy in binding poses between short and long glycans is likely due to the additional interactions of non-Ca^{2+} binding sugars with other amino acids in an extended binding site.

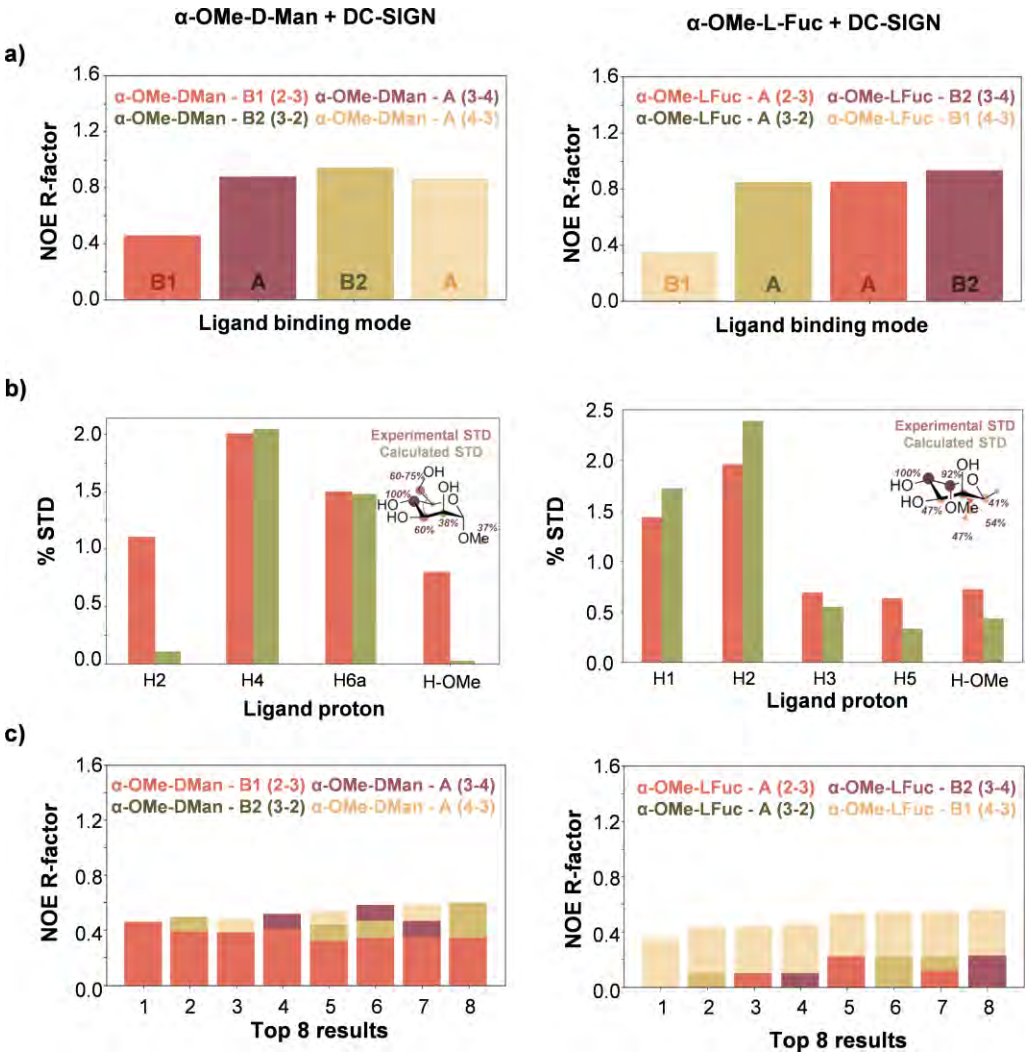


Figure 11. a) NOE R-Factors of the average STD intensities of each ensemble of binding modes calculated by CORCEMA-ST. Each ensemble is represented with a different color. b) Experimental (in orange) vs. CORCEMA-ST predicted (in green) STD of the best scoring ensemble for α-OMe-D-Man (left) and α-OMe-L-Fuc (right) in the presence of DC-SIGN at t_{sat} = 0.5 s. 200 Snapshots from each MD trajectory were used in each case. The corresponding normalized STD values are represented in the sketched carbohydrates. c) Top scoring mixes of binding modes found by BM-Mixer. Note that the best result (i.e. the lowest NOE R-factor) involves structures from only the B1 motif (2-3 pose for α-OMe-D-Man and 4-3 pose for α-OMe-L-Fuc).

e. Exploring novel DC-SIGN binders

Based on the identified optimal binding motif (B1), new potential binders of DC-SIGN were surveyed and experimentally tested. L-Gal, which is structurally similar to L-Fuc

but with an hydroxymethyl group at C5, was selected. Similarly, the natural occurring D-Rha, which is similar to D-Man but with a methyl group at C5, showed potentially similar binding modes. Unlike previously described D-Man and L-Fuc derivatives, these molecules are not *O*-protected at the anomeric position, thus allowing for both the α and β hydroxyl groups to engage in even more potential binding modes that the α-1-methoxy derivatives. Additionally, myo-inositol, a common metabolite precursor in eukaryotes cells, was tested.

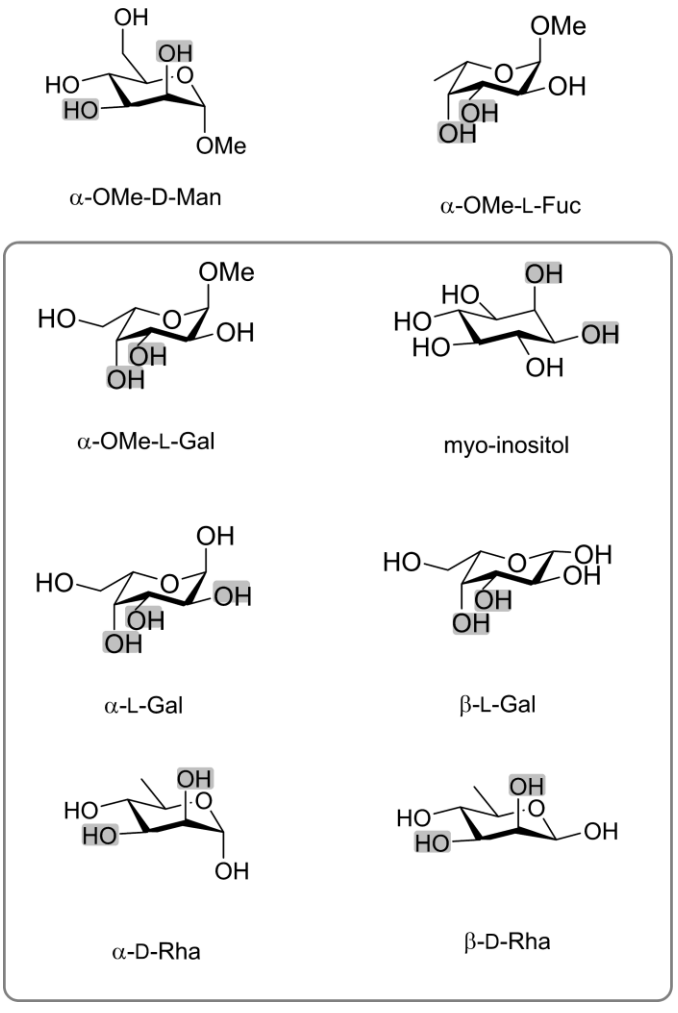


Figure 12. Potential DC-SIGN binders (box) explored based on the B1 binding motif together with previously analyzed DC-SIGN binders α-OMe-D-Man and α-OMe-L-Fuc. The OH groups capable of binding in the B1 mode are highlighted in gray.

The potential binding modes for α/β-L-Gal, α/β-D-Rha, and myo-inositol to DC-SIGN can be classified as presented in Table 3.

Table 3. Classification of the binding modes of α-L-Gal, β-L-Gal, α-D-Rha, β-D-Rha, and myo-inositol to DC-SIGN based on their Structural Binding Patterns (SBPs). eq = equatorial; ax = axial.

Sugar	Binding pose	SBP	$^{+}$O/$^{-}$O configuration
α-L-Gal	*1-2*	B1	ax/eq
	2-1	B2	eq/ax
	2-3	A	eq/eq
	3-2	A	eq/eq
	3-4	B2	eq/ax
	4-3	B1	ax/eq
β-L-Gal	*2-3*	A	eq/eq
	3-2	A	eq/eq
	3-4	B2	eq/ax
	4-3	B1	ax/eq
α-D-Rha β-D-Rha	*2-3*	B1	ax/eq
	3-2	B2	eq/ax
	3-4	A	eq/eq
	4-3	A	eq/eq
myo-inositol	*1-2*	B2	eq/ax
	2-1	B1	ax/eq
	4-5	A	eq/eq
	5-4	A	eq/eq
	6-1	A	eq/eq
	1-6	A	eq/eq

Following the protocol described above, MD simulations were performed to analyze the binding poses of α/β-L-Gal, α/β-D-Rha, and myo-inositol that met the minimum binding epitope requirements. All possible binding modes were considered, including 1-2 and 2-1 poses for α-L-Gal. The residence time profiles of these systems were comparable to those observed for α-OMe-D-Man, α-OMe-L-Fuc, and α-OMe-L-Gal, proving their similar binding properties (Fig. 13). The analysis of protein-ligand interactions also revealed characteristic contacts similar to those in SBP A, B1, and B2 of the α-methoxy derivatives (Figs. 14 and 15).

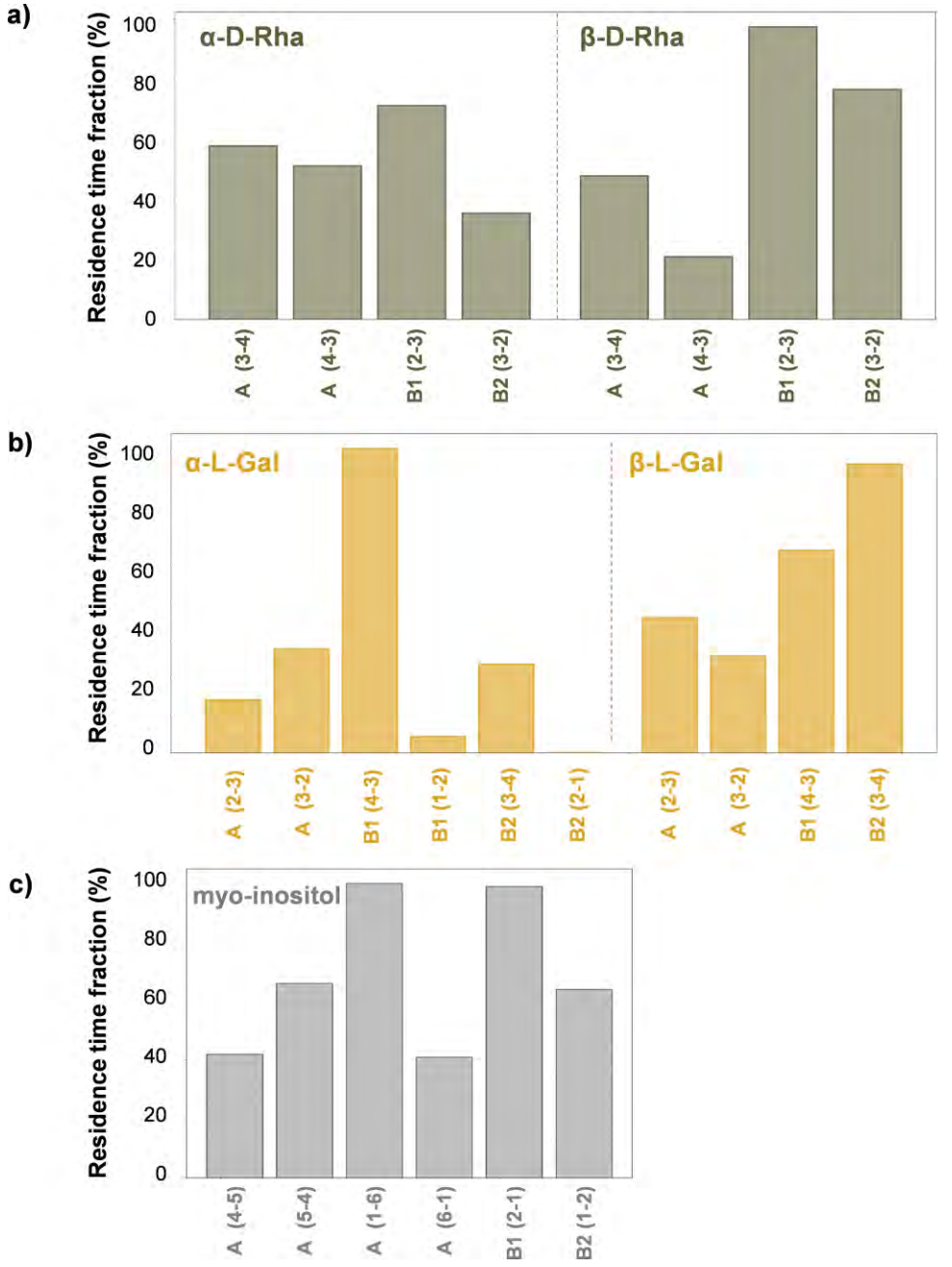


Figure 13. Ligand residence frequencies derived through MD simulations for a) α/β-D-Rha, b) α/β-L-Gal, and myo-inositol at the lectin binding site in each of the different poses.

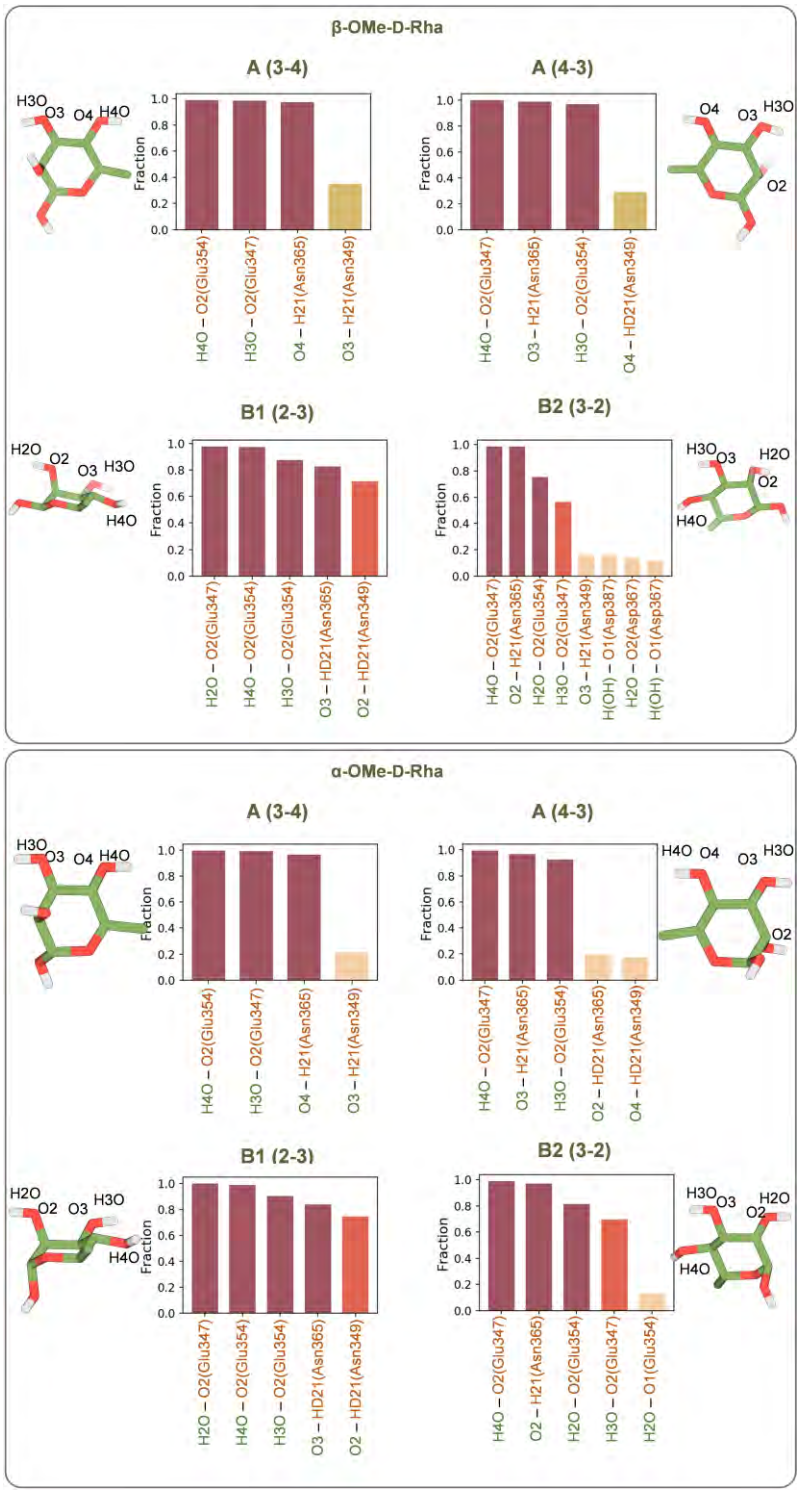


Figure 14. Hydrogen bond interactions measured for each ligand-protein complex during the section of the MD simulation in which the ligand is bound to DC-SIGN. The fraction of the ligand-bound snapshots in which the hydrogen bonds occur is shown in the bar plots.

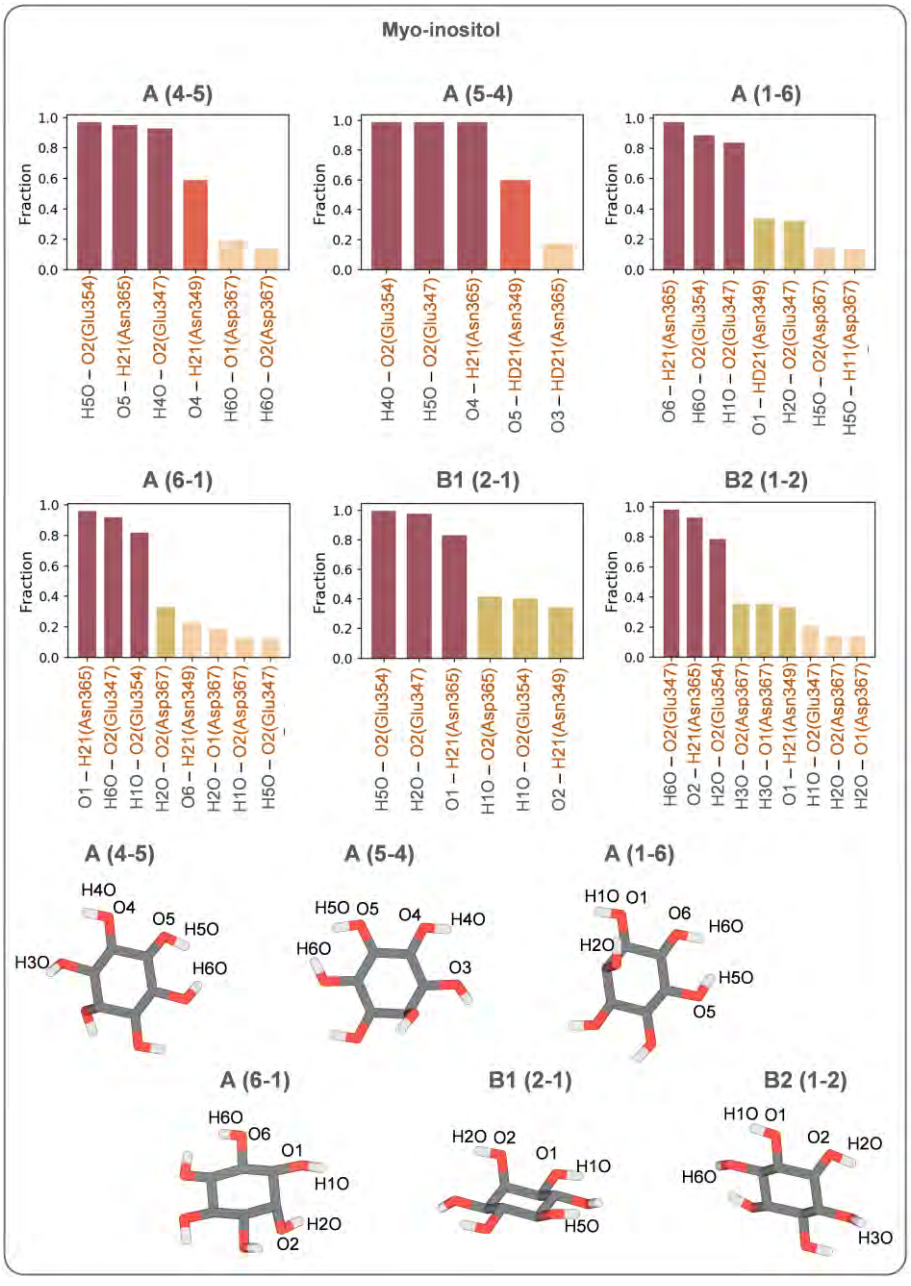


Figure 14 (cont.). Hydrogen bond interactions measured for each ligand-protein complex during the section of the MD simulation in which the ligand is bound to DC-SIGN. The fraction of the ligand-bound snapshots in which the hydrogen bonds occur is shown in the bar plots.

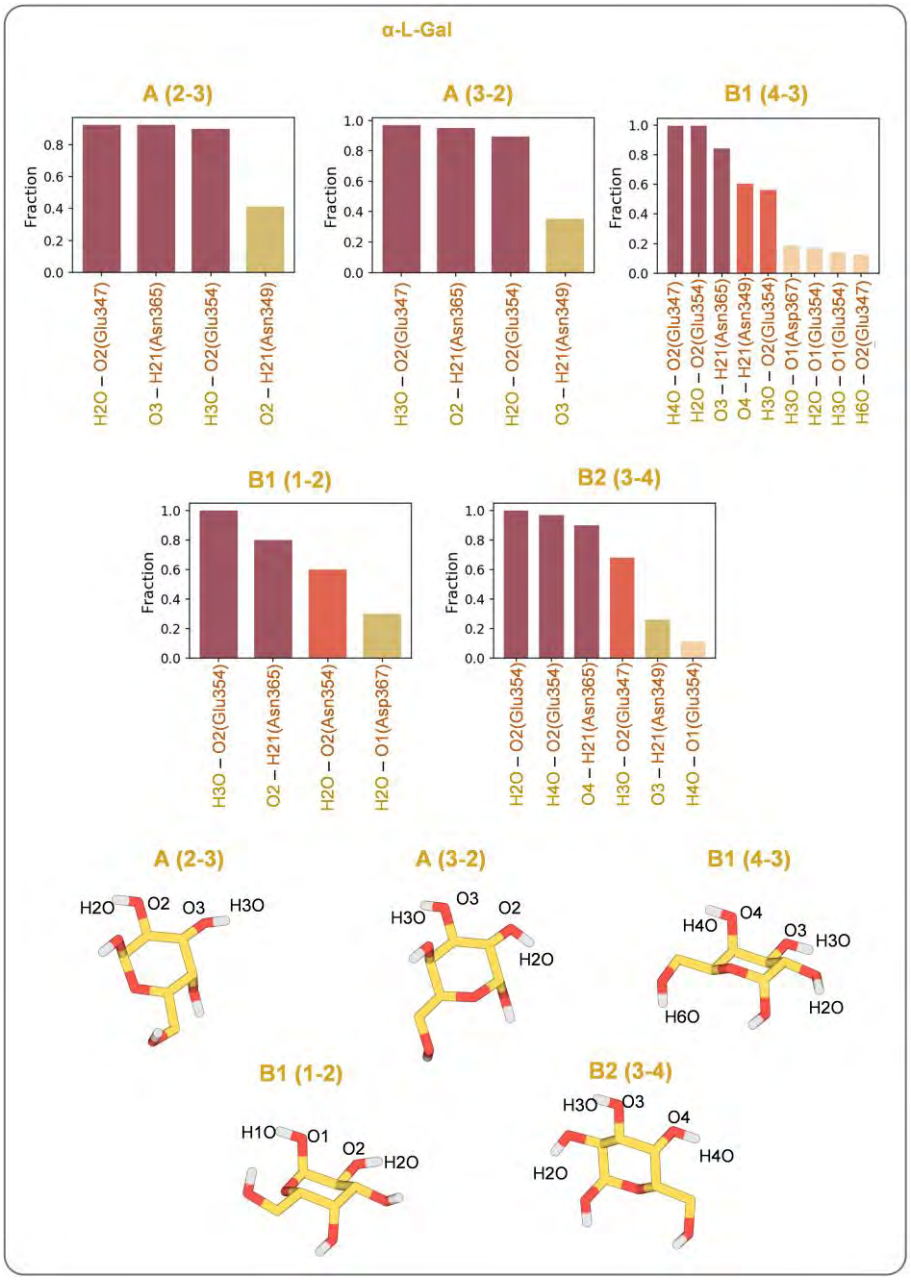


Figure 14 (cont.). Hydrogen bond interactions measured for each ligand-protein complex during the section of the MD simulation in which the ligand is bound to DC-SIGN. The fraction of the ligand-bound snapshots in which the hydrogen bonds occur is shown in the bar plots.

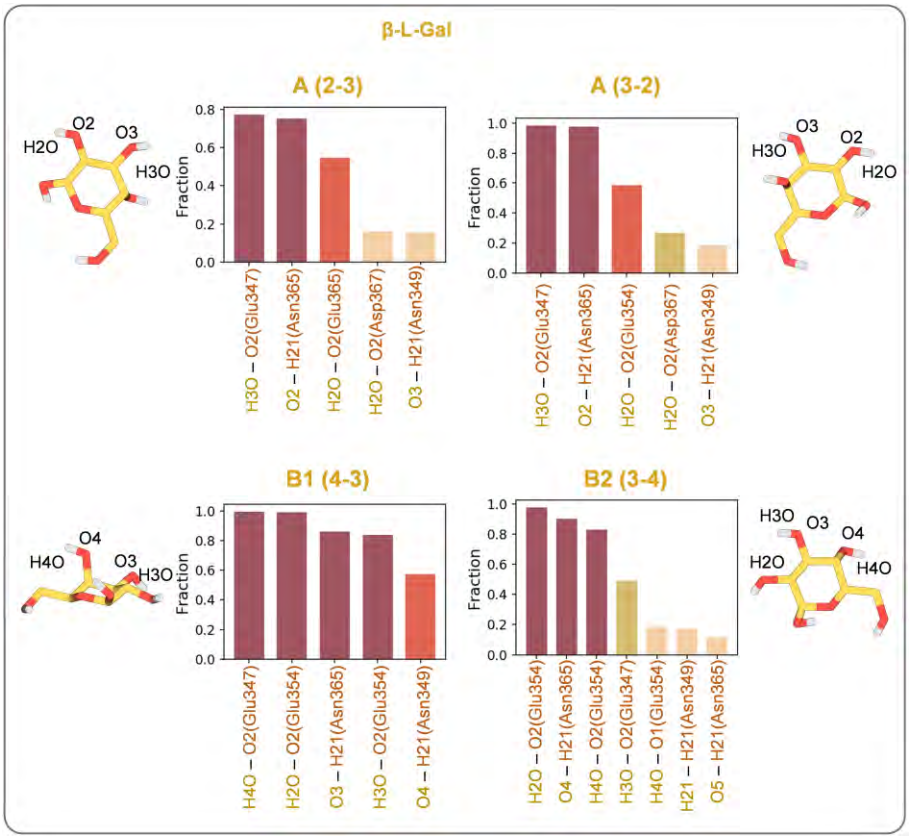


Figure 14 (cont.). Hydrogen bond interactions measured for each ligand-protein complex during the section of the MD simulation in which the ligand is bound to DC-SIGN. The fraction of the ligand-bound snapshots in which the hydrogen bonds occur is shown in the bar plots.

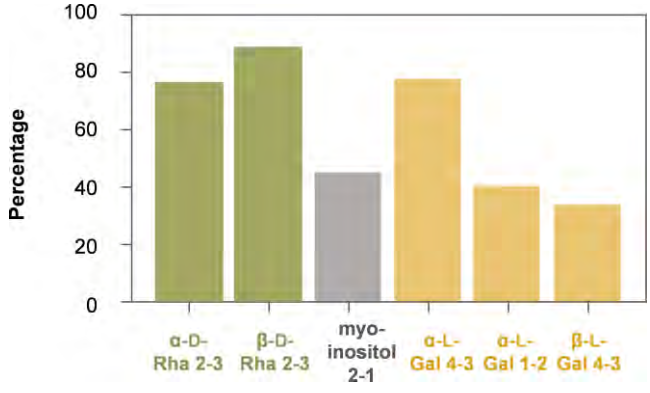
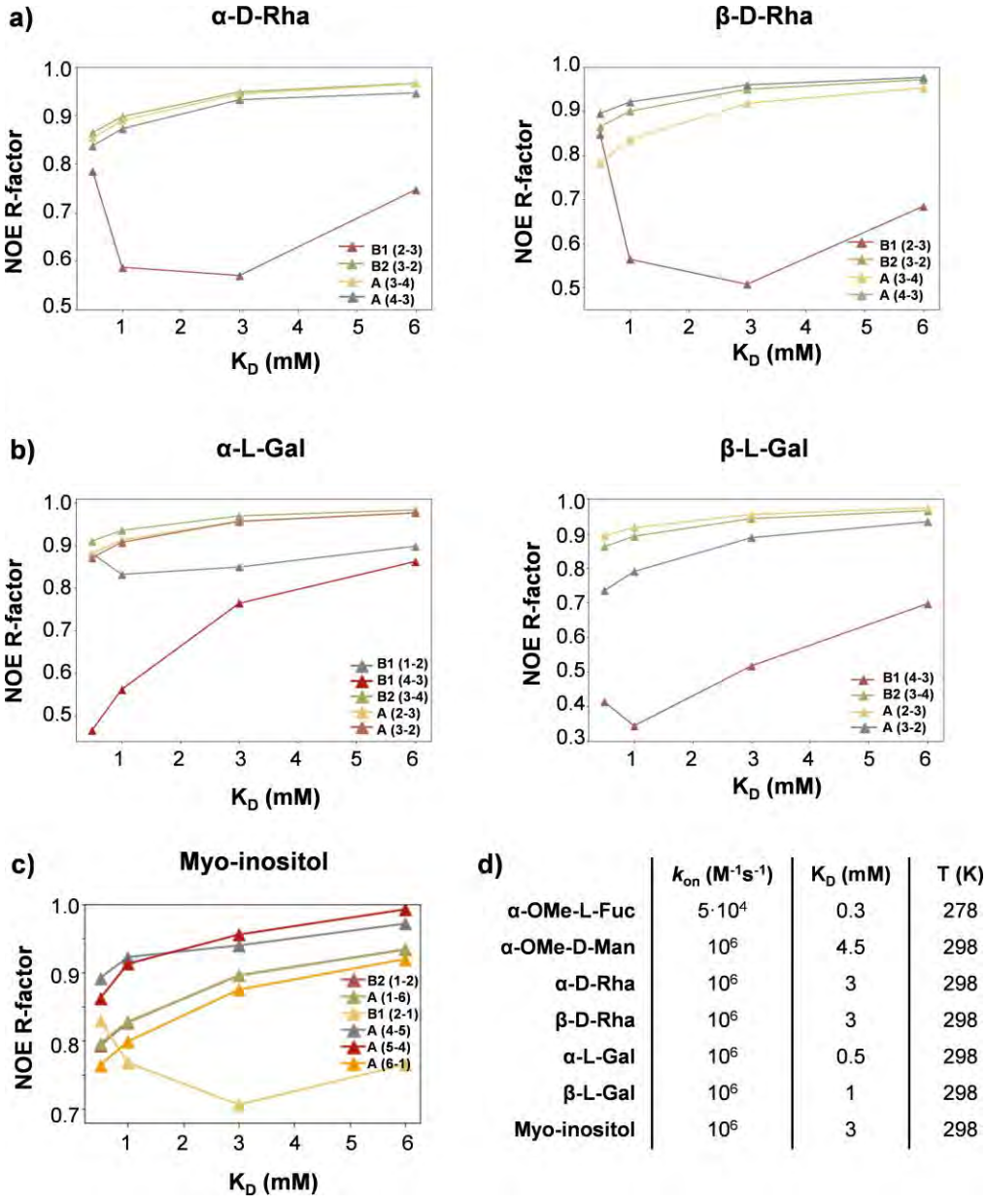


Figure 15. Occurrence of Van der Waals interactions between any non-polar hydrogen of the sugar (CH) and Val351 found on the MD trajectories, expressed as % of the total simulation time in which the ligand is bound to DC-SIGN. Van der Waals (VdW) interactions are considered to take place when the distance between the pair of atoms involved is lower than 1.2 times the sum of their VdW radii. Only the binding modes in which these interactions are found in more than 10% of the simulated time are represented in this graph (all of them correspond to B1 structural patterns).

The B1 pattern was again predicted to show a more persistent binding in the MD simulations, except for β-L-Gal, where the B2 pattern of the 3-4 binding pose was more stable (Fig. 14). This was attributed to the lack of hydrophobic interactions between Val351 and CH1 (now in axial position) in the 4-3 binding mode (B1 pattern) of β-L-Gal (Fig.15). In the case of myo-inositol, the A pattern and the B1 pattern exhibited similar ligand residence times. Interestingly, the 1-2 and 2-1 binding modes in α-L-Gal involving the anomeric hydroxyl group were highly unstable, likely due to the weaker Ca^{2+}-coordinating properties of the hemiacetal group.

STD-NMR experiments were conducted for α/β-L-Gal, α/β-D-Rha, and myo-inositol. All compounds showed measurable STD effects, indicating their interaction with the lectin. The relative STD profiles of the ligands were similar, suggesting a similar binding epitope. However, there were significant differences in the absolute STD values, suggesting variations in the binding affinities.

Applying a similar methodology as in the case of α-OMe-D-Man and α-OMe-L-Fuc, CORCEMA-ST calculations were performed on the MD ensembles of α/β-L-Gal, α/β-D-Rha, and myo-inositol bound to DC-SIGN. Unlike the 1-methoxy derivatives for which the K_D and k_{on} values needed for the calculation of theoretical STDs were simultaneously optimized, different values of K_D were explicitly tested at a fixed value of k_{on} for the potential binders to find the one better reproducing the experimental data (see Methods). The results showed that the binding modes belonging to the B1 pattern consistently fit better to the experimental data in a range of K_D values, yielding minimum NOE R-factors (Fig. 16, data plotted in red for α/β-L-Gal, α/β-D-Rha, and yellow for myo-inositol). In contrast, the other MD ensembles exhibited much worse NOE R-factors.



d)

	k_{on} ($M^{-1}s^{-1}$)	K_D (mM)	T (K)
α-OMe-L-Fuc	$5 \cdot 10^4$	0.3	278
α-OMe-D-Man	10^6	4.5	298
α-D-Rha	10^6	3	298
β-D-Rha	10^6	3	298
α-L-Gal	10^6	0.5	298
β-L-Gal	10^6	1	298
Myo-inositol	10^6	3	298

Figure 16. NOE R-Factor calculated at several K_D in the 0.5-6 mM range for α/β-D-Rha, α/β-L-Gal and myo-inositol in the different ligand binding modes. MD ensembles comprising 200 snapshots for each sugar:lectin complex were used. The results shown are calculated at t_{sat} = 0.5 s. Ligand:protein ratios were adjusted to match the experimental ratios in each case, and the experimentally measured T1 for each sugar proton were used in the calculations. d) Estimated K_D values displaying the minimum NOE R-Factor for all studied DC-SIGN binders.

To consider the potential coexistence of multiple binding modes in solution, BM-Mixer calculations were conducted on the CORCEMA-calculated STDs. The results revealed that the majority of the best solutions involved a population of 80-100% binding poses belonging to the B1 pattern. This suggests minimal or no competition between binding modes for the same monosaccharide (Fig. 17).

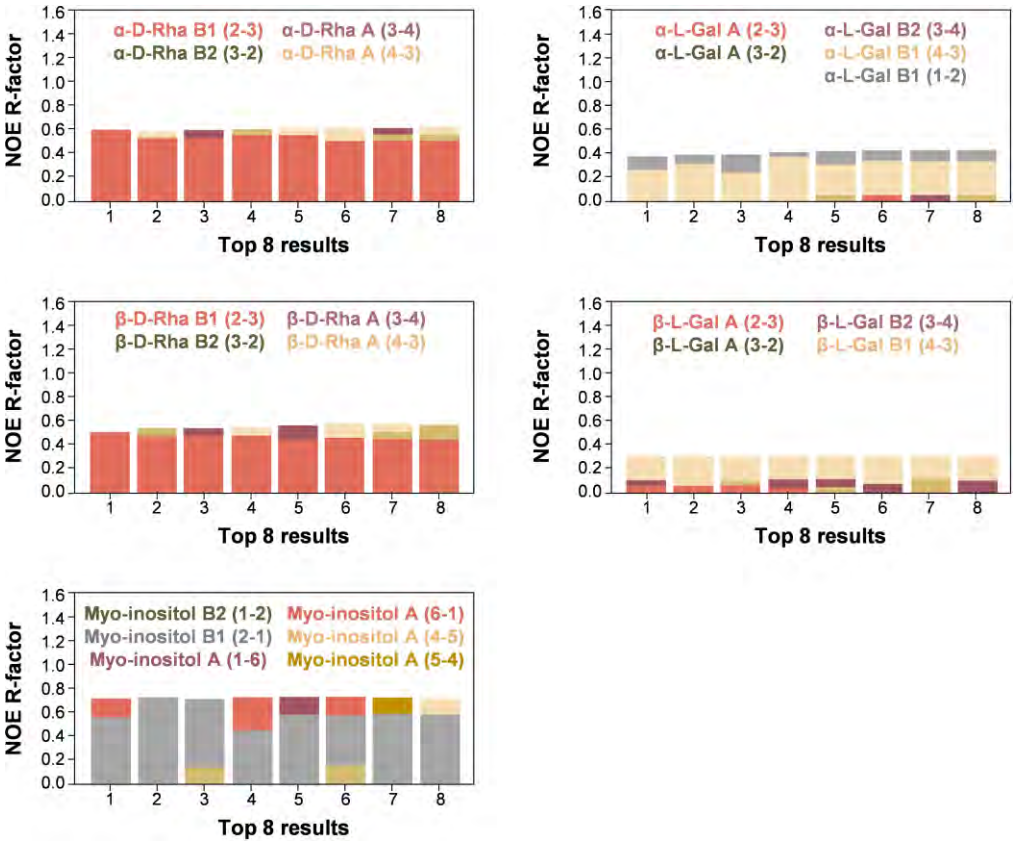


Figure 17. Top BM-Mixer solutions at t_{sat} = 0.5 s for α/β-D-Rha (theoretical K_D = 3 mM), α-L-Gal (theoretical K_D = 0.5 mM), β-L-Gal (theoretical K_D = 1 mM), and myo-inositol (theoretical K_D = 3 mM).

f. Ligand binding affinity by competition experiments using ^{19}F-NMR

Competition experiments were conducted to assess the *relative* binding affinity of the new ligands for DC-SIGN using α-D-Me-6-F-Man as mono-fluorinated spy molecule, using previously reported methodology [173]. In these experiments, increasing ligands concentrations ([I]) were introduced into a DC-SIGN sample in the presence of the spy molecule. At each [I]/[Spy] ratio, the transversal relaxation rate of the fluorine nucleuos, ^{19}F-$R_{2,obs}$, was calculated for each substrate (Fig. 18).

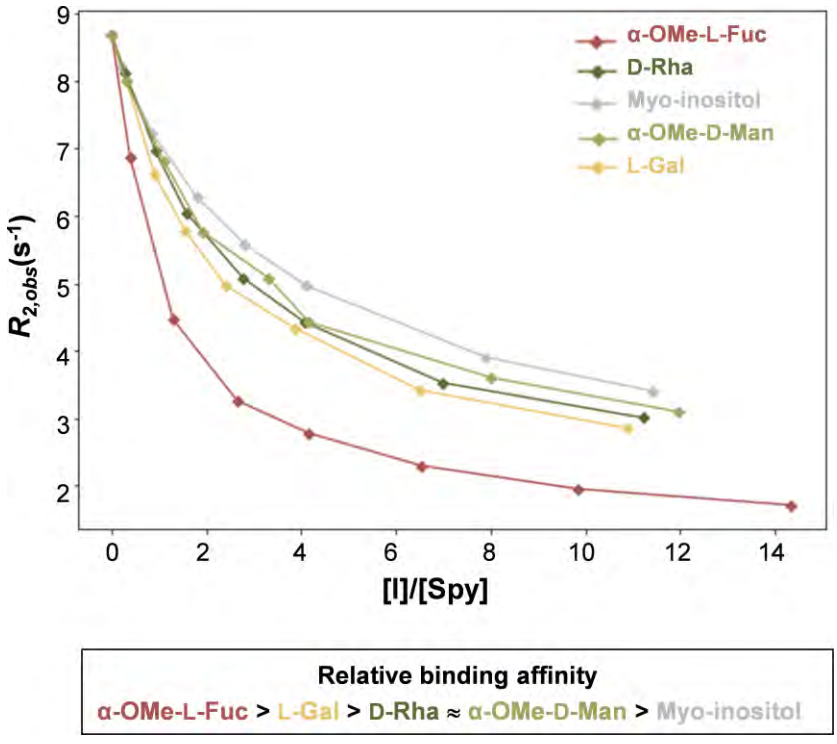


Figure 18. ^{19}F-$R_{2,obs}$ of the spy molecule α-OMe-6-F-Man for increasing amounts of α-OMe-D-Man, α-OMe-L-Fuc, D-Rha, L-Gal and myo-inositol at 298 K. The same initial solution of [α-OMe-6-F-Man] = 2.02 mM and [DC-SIGN (CRDs)] = 26.9 μM was used in each case.

These R_2 relaxation experiments results allow the ranking of the tested molecules based on their relative affinity for DC-SIGN, as shown in Fig. 18. In all cases, as the concentration of competing ligands increases, the observed relaxation rate of the spy molecule progressively returns to its original state. This phenomenon occurs because the bound spy molecule is displaced from the lectin binding site to a free state in the solution. Notably, α-OMe-L-Fuc exhibits a significantly more pronounced recovery of ^{19}F-$R_{2,obs}$ indicating its higher affinity compared to the other ligands, as suggested above, followed by L-Gal. D-Rha shows similar $R_{2,obs}$ recovery compared to α-OMe-D-Man, while myo-inositol is revealed as the lowest affinity binder, as it has the lowest impact on the fluorine relaxation rate of the spy molecule.

It is important to note that in the case of L-Gal and D-Rha, the ligand concentrations represent a sum of both α and β anomers in solution. Because it is impossible to distinguish the contribution of each individual anomer to the binding affinity in these relaxation experiments, the resulting observed affinity in these relaxation experiments reflects an apparent value for the anomeric mixture as a whole, potentially underestimating the affinity of any individual anomer. Only in the hypothetical scenario where both anomers were recognized by the lectin with the same K_D, would

the experiment accurately reflect the relative affinity of the receptor for L-Gal and D-Rha (either the α or β anomers). Based on previous STD experiments and MD simulations, this scenario is likely for D-Rha, where both anomers are expected to bind with similar affinity. In contrast, β-L-Gal is expected to bind DC-SIGN with lower affinity than the α anomer, thus resulting a decreased value for the observed affinity of the mixture in the ^{19}F-$R_{2,obs}$ NMR experiments.

One plausible explanation for the difference in affinity among different affinities of these ligands, which share all sharing the B1 structural pattern, may be the modulation of binding by Val351, as discussed earlier. From this perspective, α/β D-Rha would shows slightly higher affinity compared to D-Man due to its methyl group at C6, potentially engaging in better hydrophobic interactions with Val351 than the flexible and polar hydroxymethyl group counterpart in D-Man. Similarly, the endocyclic H1 C1H in α-OMe-L-Fuc and α-L-Gal, positioned in an equatorial disposition facing Val351, may form more favorable dispersion interactions. In the case of β-L-Gal, the lack of an equatorial C1H in the β-anomer might reduce potential hydrophobic contacts with Val351, as observed in MD simulations, thus leading to decreased affinity compared to the α anomer. Likewise, myo-inositol likely exhibits the lowest affinity due to its lack of the equatorial C1H necessary for van der Waals interactions with Val351, and more importantly, failing to meet the ideal $^{+}O_{ax}$-$^{-}O_{eq}$-O_{eq}-H_{eq} scaffold inferred above.

3. Conclusions

An in-depth understanding of the optimal minimal structural pattern recognized by the human lectin DC-SIGN has been performed through a combination of computational methods and STD-NMR experiments. This ideal minimal scaffold, $^{+}O_{ax}$-O_{eq}-O_{eq}-H_{eq}, serves as the basis for the recognition by DC-SIGN of six-membered cyclic polyhydroxylated ligands. Hence, several requirements must be met for a molecule to be recognized as a potential DC-SIGN binder. First, the specific chirality of the minimum binding epitope (the $C^{+}O - C^{-}O$ motif coordinating Ca^{2+}) is crucial. Once this condition is satisfied, the presence of an adjacent equatorial OH group followed by an equatorial H allows additional interactions with Glu354 (hydrogen bond) and Val351 (van der Waals), leading to increased stabilization and modulation of affinity towards the lectin. Notably, it has been demonstrated that L-Fuc and D-Man share a similar binding epitope, despite their different binding modes found in crystallographic structures when they are part of larger oligosaccharides.

Based on these findings, the interaction of α/β-L-Gal, α/β-D-Rha, and myo-inositol with DC-SIGN was explored for the first time. While myo-inositol exhibited lower affinity compared to α-OMe-D-Man, the α/β-D-Rha anomeric mixture displayed

significantly higher affinity for the lectin. The α/β-L-Gal mixture exhibited the highest affinity among the tested DC-SIGN ligands, α-OMe-D-Fuc being the best binder of all tested ligands. Our cluster QM calculations and MD simulations predicted, and STD-NMR experiments confirmed, that all these binders share the same binding epitope as L-Fuc and D-Man with DC-SIGN, although involving different hydroxyl groups within the six-membered ring. The insight gained into binding preferences and molecular interactions could have broader implications in the field of molecular recognition and drug design, potentially guiding future discovery campaigns.

4. Methods

a. PDB search, minimum binding epitope definition, and classification of binding modes

The crystallographic structures of DC-SIGN obtained from the Protein Data Bank (PDB) with IDs: 1SL4, 1SL5, 1K9I, , 2IT5, 2IT6, 2XR5, 2XR6, , 6GHV, 7NL6 and 7NL7 were carefully examined and compared using PyMOL [151].

b. QM cluster model building

The crystal structure of DC-SIGN in complex with lacto-*N*-fucopentaose III (LNFP III) (PDB ID: 1SL5) was chosen as the protein template for this study. For the cluster model, the Ca^{2+} atom and the polar residues surrounding it at the binding site, including Asn365, Asp366, Asn349, Glu347 and Glu354 were considered. To ensure the complete coordination sphere around the calcium atom and maintain correct capping, several modifications were made: i) the backbone NH of Asn365 was substituted with a hydrogen atom, ii) the backbone CO of Asp266 was capped with an *N*-methyl acetamide, iii) the side chains of Glu347 and Glu354 were truncated to acetic acid, and iv) the side chain of residue Asn349 was transformed into an acetamide group (Fig. 19). Additionally, the sidechain of Val351 was also considered.

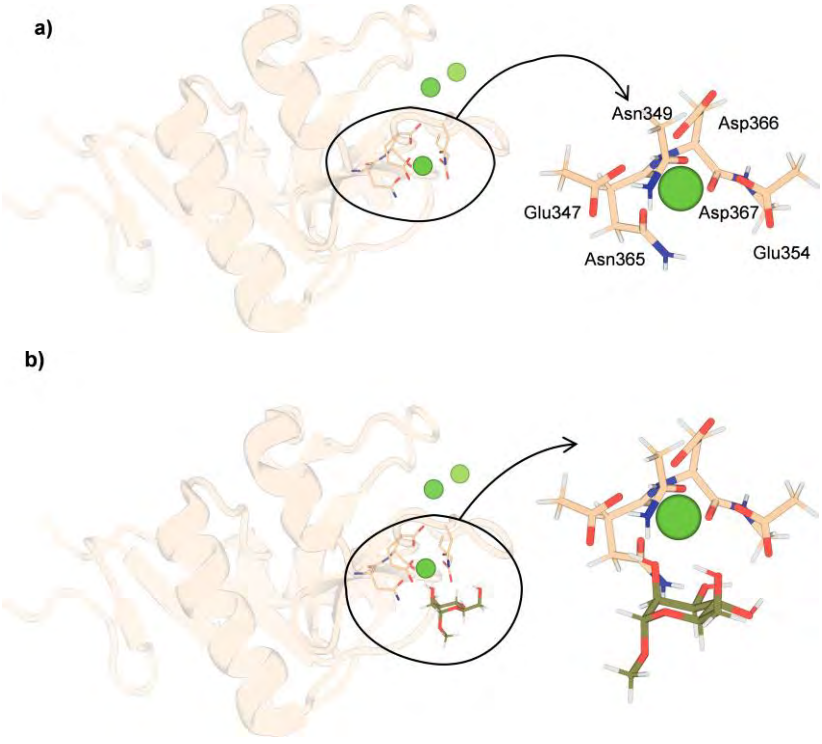


Figure 19. Cluster model of DC-SIGN's binding site (from crystallographic structure 1SL5) used for the ligand conformational search and binding enthalpy calculation. a) *apo* state (in brown sticks) (55 atoms included in the cluster model), b) example of a bound state with α-OMe-D-Man (in green sticks) (82 atoms included in the cluster model). Ca^{2+} atom is shown as a green sphere.

c. Conformational search

To explore the conformational space of both the free and protein-bound ligands, a Low-Mode Conformational Search (LMCS) [174] was conducted using Schrödinger Macromodel software [175]. The search involved 100000 Monte Carlo steps and focused on identifying all possible conformers. The analysis specifically examined the first 10 low-frequency modes of the system. Each Monte Carlo cycle started from the previous structure only if its energy fell within a 10 kcal mol^{-1} window. The OPLS3e force field was utilized for the conformational search, while water was employed as the solvent with the GB/SA (Generalized Born Surface Area) solvation model. The truncated Newton (TNCG) method was applied for the minimization process.

This method has demonstrated its effectiveness in identifying low-energy conformations for both cyclic and acyclic molecules [174]. The ligands were positioned within the binding site of the lectin, and a conformational search was performed for each of the binding poses (2-3, 3-2, 3-4, 4-3) that matched the predefined structural

binding patterns (A, B1 and B2). During the conformational search, all proteins atoms, excluding hydrogens, were kept fixed.

Additionally, the two hydroxyl groups of the ligand directly involved in Ca^{2+} coordination (Fig. 20) were kept in a fixed position to preserve the crystallographic binding geometry. By incorporating these constraints, the search was focused on exploring the conformational space of the ligands within the context of their interactions with the lectin.

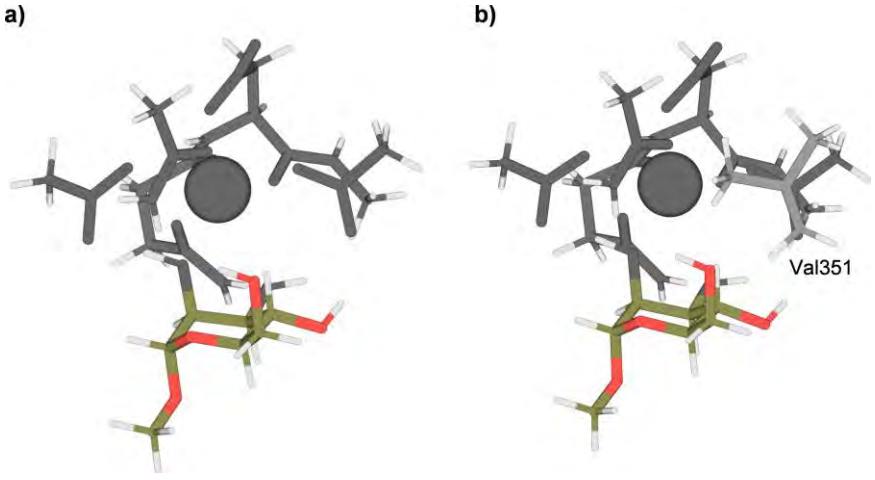


Figure 20. a) Example of a starting cluster model pose for the conformational search with *α*-OMe-D-Man in the B1 structural pattern. b) Extended model including Val351 (light gray). Atoms colored in gray were fixed during the calculation.

d. Quantum mechanical calculations

All unique conformers found through the LMCS were subjected to quantum mechanical geometry optimizations. Calculations were carried out with Gaussian 16 [176] using the M06-2X hybrid functional [177] and 6-31+G(d,p) basis set in combination with ultrafine integration grids. Bulk solvent effects in water considered implicitly through the IEF-PCM polarizable continuum model [178]. In these calculations, all protein atoms of the model except hydrogens and the *N*-methyl amide were frozen. Additionally, the C*α* atom of Val351 was fixed when this residue is considered in the calculations. All the energetically accessible conformers within a 3 kcal mol^{-1} range were subjected to additional frequency calculations. Frequency analyses were carried out at the same level used in the geometry optimizations, and the nature of the stationary points was determined in each case according to the appropriate number of negative eigenvalues of the Hessian matrix. Scaled frequencies were not considered. Enthalpies at 298 K (ΔH) were employed in the discussion on the

relative stabilities of the structures, as well as the lowest energy conformer for each calculated stationary point. Cartesian coordinates, electronic energies, entropies, enthalpies, Gibbs free energies, and the lowest frequencies of the lowest energy calculated structures are available in Tables 4 and 5.

Table 4. Table of energies, entropies and lowest frequencies of the calculated lowest energy structures in the unbound state[a].

Structure	E_{elec} (Hartree)	E_{elec} + ZPE (Hartree)	H (Hartree)	S (cal mol^{-1} K^{-1})	G (Hartree)	Lowest freq. (cm^{-1})	# of imag freq.
2H2O-gas1-wat	–152.814635	–152.768540	–152.761621	72.0	–152.795827	26.7	0
αOMe-D-Man (gg)	–726.257534	–726.029345	–726.014650	114.1	–726.068861	64.0	0
αOMe-D-Man (gt)	–726.257487	–726.029345	–726.014610	114.3	–726.068905	64.6	0
αOMe-L-Fuc	–651.061976	–650.838986	–650.825450	107.9	–650.876702	84.2	0
αOMe-L-Gal (gt)	–726.259285	–726.030597	–726.016161	112.7	–726.069699	76.2	0
αOMe-L-Gal (gg)	–726.259026	–726.030590	–726.015974	113.7	–726.069974	74.5	0

[a]Energy values calculated at the PCM(H_2O)M06-2X/6-31+G(d,p) level. 1 Hartree = 627.51 kcal mol^{-1}. Thermal corrections at 298.15 K.

Table 5. Table of energies, entropies and lowest frequencies of the calculated lowest energy structures in the bound state[a].

Without Val354

Structure	E_{elec} (Hartree)	E_{elec} + ZPE (Hartree)	H (Hartree)	S (cal mol^{-1} K^{-1})	G (Hartree)	Lowest freq. (cm^{-1})	# of imag freq.
model-H2O-complex	–2388.300631	–2387.945159	–2387.927731	134.4	–2387.991574	51.5	0
αOMe-D-Man23 (gg)	–2961.76338	–2961.227336	–2961.200839	177.4	–2961.28513	19.0	0
αOMe-D-Man32 (gt)	–2961.757597	–2961.221222	–2961.194270	181.4	–2961.280450	20.1	0
αOMe-D-Man34 (gg)	–2961.759575	–2961.223636	–2961.196666	181.5	–2961.282902	19.8	0
αOMe-D-Man43 (gg)	–2961.760766	–2961.224245	–2961.197598	178.4	–2961.282358	27.4	0
αOMe-L-Fuc-23	–2886.556885	–2886.027681	–2886.002299	173.6	–2886.084767	–4.9	0
αOMe-L-Fuc-32	–2886.558699	–2886.029127	–2886.003001	175.8	–2886.086527	27.8	0
αOMe-L-Fuc-34	–2886.559309	–2886.028408	–2886.002597	173.7	–2886.085132	23.8	0
αOMe-L-Fuc-43	–2886.564167	–2886.033263	–2886.007620	173.7	–2886.090151	20.8	0
αOMe-L-Gal-23 (gt)	–2961.756039	–2961.220181	–2961.193412	180.0	–2961.278946	19.6	0
αOMe-L-Gal-32 (gt)	–2961.756381	–2961.221028	–2961.194080	179.8	–2961.279497	24.9	0
αOMe-L-Gal-34 (tg)	–2961.761675	–2961.224605	–2961.198035	177.3	–2961.282281	24.7	0
αOMe-L-Gal-43 (tg)	–2961.766018	–2961.228943	–2961.202444	178.7	–2961.287359	17.9	0

Table 5 (cont.). Table of energies, entropies and lowest frequencies of the calculated lowest energy structures in the bound state[a].
With Val351

Structure	E_{elec} (Hartree)	E_{elec} + ZPE (Hartree)	H (Hartree)	S (cal mol^{-1} K^{-1})	G (Hartree)	Lowest freq. (cm^{-1})	# of imag freq.
model-H2O-complex	–2546.677879	–2546.192516	–2546.16948	161.3	–2546.246118	29.0	0
αOMe-D-Man23 (gt)	–3120.142694	–3119.477337	–3119.445022	203.3	–3119.541631	24.9	0
αOMe-D-Man32 (gt)	–3120.135531	–3119.469843	–3119.437229	207.4	–3119.535776	18.9	0
αOMe-D-Man34 (gg)	–3120.139629	–3119.473957	–3119.441693	202.9	–3119.538094	21.5	0
αOMe-D-Man43 (gg)	–3120.139742	–3119.473325	–3119.441278	202.0	–3119.537256	24.4	0
αOMe-L-Fuc-23	–3044.935476	–3044.276427	–3044.244711	204.3	–3044.341797	8.9	0
αOMe-L-Fuc-32	–3044.936480	–3044.277238	–3044.245546	200.6	–3044.340848	30.1	0
αOMe-L-Fuc-34	–3044.937432	–3044.276665	–3044.245247	199.3	–3044.339936	24.4	0
αOMe-L-Fuc-43	–3044.944907	–3044.284615	–3044.253428	198.7	–3044.347817	19.3	0
αOMe-L-Gal-23 (gt)	–3120.133954	–3119.468627	–3119.436081	207.6	–3119.534738	18.1	0
αOMe-L-Gal-32 (gt)	–3120.134183	–3119.469137	–3119.436533	206.0	–3119.534427	25.6	0
αOMe-L-Gal-34 (tg)	–3120.146674	–3119.479561	–3119.447748	200.1	–3119.542827	24.1	0
αOMe-L-Gal-43 (tg)	–3120.141344	–3119.474708	–3119.442768	200.3	–3119.537926	22.5	0

[a]Energy values calculated at the PCM(H_2O)M06-2X/6-31+G(d,p) level. 1 Hartree = 627.51 kcal mol^{-1}. Thermal corrections at 298.15 K.

e. Molecular dynamics simulations

Molecular dynamics (MD) simulations were run with Amber 20 [150] suite, using the *ff14SB* [152] and *GLYCAM 06j-1* [153] force fields for the protein and the carbohydrate ligands, respectively. Protein complexes were immersed in a box with 10 Å buffer of TIP3P [154] water molecules and neutralized by adding explicit Na^+ or Cl^- counterions. A two-stage optimization approach was performed. The first stage minimizes only the positions of solvent molecules and ions, and the second stage is an unrestrained minimization of all the atoms in the simulation cell. The systems were then heated by incrementing the temperature from 0 to 300 K under a constant pressure of 1 atm and periodic boundary conditions. Harmonic restraints of 10 kcal mol^{-1} were applied to the solute, and the Andersen temperature coupling scheme [155,179] was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Water molecules were treated with the SHAKE algorithm [156] such that the angle between the hydrogen atoms is kept fixed through the simulations. Long-range electrostatic effects were modelled using the particle mesh Ewald method [157]. An 8 Å cut-off was applied to Lennard-Jones interactions. Each system was equilibrated for 2 ns with 2 fs time step at a constant

volume and temperature of 300 K. Five independent production trajectories were then run for additional 200 ns under the same simulation conditions, leading to accumulated simulation times of 1 μs for each carbohydrate-lectin complexes studied, using the complete CRD structure with the ligand in the all the proposed binding poses.

The residence frequency of the ligands in the binding pocket were estimated based on the following criteria:

i) The distances between the Ca^{2+} and its two coordinating oxygen atoms in the ligand (< 3.5 Å for bound geometries)

ii) The distances between the same oxygen atoms of the ligand and the carboxylic carbon atom of the two glutamic acid residues of the binding site Glu345 and Glu347 (< 4 Å for bound geometries)

f. CORCEMA-ST calculations

The quantitative analysis of the STD-NMR data was accomplished by applying the complete relaxation and conformational exchange matrix theory as implemented in the MATLAB program CORCEMA-ST 3.8 [170]. The same conditions and parameters used in the STD experiments were conserved for each protein-ligand complex (including solvent, ligand and protein concentration, magnetic field strength, etc.). For parameters not directly measured (e.g., ligand and receptor correlation times), consistent values from the literature were applied [170,180].

As the experimental STD-NMR data were obtained in D_2O, calculations excluded exchangeable hydrogens (OH and NH). Additionally, experimental longitudinal relaxation times (T1) were measured and incorporated into the analysis. Generally, only well-resolved STD peaks were considered in the calculations.

For α/β-D-Rha, both isochronous H6-α and H6-β were incorporated, given their significance in the 2-3 binding pose. Due to a slight direct saturation of these protons in the on-resonance spectrum, their STD intensities were adjusted by subtracting the STD blank, experiment of α/β-D-Rha. This reference experiment was conducted under identical conditions but in the absence of DC-SIGN. For myo-inositol, the computed STD intensities of the isochronous proton signals (H1 with H3, and H4 with H6) were combined to facilitate comparison with the experimental STD intensities.

As the STD-NMR experiments were conducted at two distinct temperatures (278 K for α-OMe-L-Fuc and 298 K for the remaining ligand molecules), a protocol was devised to determine an optimal k_{on} value for CORCEMA-ST calculations at each temperature. The efficacy of the protocol was assessed using the α-OMe-D-Man and α-OMe-L-Fuc

complexes as benchmark systems, chosen for their extensively studied interaction with DC-SIGN and the availability of abundant experimental affinity data. For these carbohydrate-lectin complexes, simultaneous optimization of the association rate constant (k_{on}) and dissociation constant (K_D) in CORCEMA-ST calculations was conducted, with other input parameters kept fixed.

Initial steps involved extracting subsets of 50 randomly selected snapshots from MD trajectories of DC-SIGN complexes with α-OMe-D-Man and α-OMe-L-Fuc in various binding poses. CORCEMA-ST calculations were executed with K_D values ranging from 0.5 to 6 mM and k_{on} from 10^4 to 10^8 $M^{-1}s^{-1}$, within previously reported ranges. Three saturation times (0.5, 1, and 2 s) were considered. Subsequently, averaged STD values for each MD ensemble were calculated, and their alignment with experimental STD data was evaluated using the NOE-R Factor (Eq. 1):

$$NOE\ R-factor\ =\ \sqrt{\frac{\sum_{i=1}^{k}(STD_{exp,i}-\overline{STD}_{cal,i})^2}{\sum_{i=1}^{k}(STD_{exp,i})^2}} \qquad \text{(Equation 1)}$$

where $STD_{exp,i}$ and $STD_{cal,i}$ represent the experimental and averaged calculated STD intensities for proton *i*, and the sum extends to the *k*th ligand proton.

Moreover, these initial tests allowed an assessment of the self-consistency in CORCEMA-ST predictions concerning saturation time. The validation of CORCEMA-ST intensities using experimental STD at t_{sat} 0.5, 1, and 2 s allowed an examination of whether the best-ranked ensemble (with the lowest NOE-R Factor) remained consistent across the tested saturation time range. Thus, it was essential that for the (K_D, k_{on}) pair resulting in the minimum NOE R-Factor, this self-consistency validation criterion was met. In cases where this criterion was not fulfilled, the corresponding (K_D, k_{on}) combination was disregarded, and the second-best solution was re-evaluated for self-consistency across the saturation time range. This iterative process ensures greater confidence that the obtained results in the search for the best models explaining the experimental STDs are not contingent on specific user-selected input CORCEMA parameters.

The optimal k_{on} values were determined to be $5{\cdot}10^4$ $M^{-1}s^{-1}$ for α-OMe-L-Fuc and 10^6 $M^{-1}s^{-1}$ for α-OMe-D-Man:DC-SIGN complexes, yielding the lowest NOE R-Factors within the tested range. For consistency, k_{on} = 10^6 $M^{-1}s^{-1}$ was chosen for all remaining ligand-lectin systems, since the STD-NMR experiments were acquired at the same temperature (298 K). This selection is grounded in the reasonable assumption that, due to similar size and chemical features, the k_{on} values for all monosaccharides would be comparable at the same temperature. Consequently, variations in STD intensities

among sugars likely arise from differences in their respective k_{off}, influenced by specific lectin-sugar interactions, ultimately leading to distinct K_D values.

After optimizing k_{on} and K_D for α-OMe-L-Fuc and α-OMe-D-Man:DC-SIGN complexes, CORCEMA-ST calculations were extended to subsets of 200 snapshots from MD ensembles for each binding pose. Similarly, for α/β-L-Gal, α/β-D-Rha, and myo-inositol complexes, 200 snapshots from each MD trajectory were used. Four K_D values (0.5, 1, 3, and 6 mM) were computed for these complexes using the optimized k_{on} value of 10^6 $M^{-1}s^{-1}$.

The final CORCEMA-ST calculations used a saturation range of 0.7 ppm for α-OMe-D-Man, α/β-L-Gal, α-OMe-D-Gal, and myo-inositol, and 0.4 ppm for α-OMe-L-Fuc and D-Rhamnose. This accounts for varying attenuation applied for selective protein saturation in the STD-NMR experiments.

g. BM-MIXER calculations

For a given ligand:DC-SIGN complex, BM-mixer takes as input the STD data predicted with CORCEMA-ST for each binding mode. Subsequently, the program generates new STD datasets, referred to as 'binding-mode mixes', by randomly combining specific fractions of each of them. The program is designed to explore all possible combinations of relative contributions from the original STD-datasets within a specific search depth. This set of potential combinations (C) can be generalized for n binding modes with a search depth of k as:

$$C = \left\{(kx_1, kx_2, \dots, kx_{n-1}, 100 - \sum_{i=1}^{n-1} kx_i) \in \mathbb{N}^n \;\middle|\; 0 \le x_i \le \frac{100}{k}, \sum_{i=1}^{n-1} kx_i \le 100\right\}$$

For instance, consider a scenario with two binding modes labeled A and B, and a search depth of 20%. The set C of combinations is defined as follows:

$$C = \{(0,100), (20,80), (40,60), \dots, (100,0)\}.$$

Where the sum of the relative contributions from each binding mode always equals 100% (%A + %B = 100).

The NOE-R Factor for each mix in set C is calculated using Eq. 1. This process is repeated for a specific number of iterations, with the number set to 5 in the presented data. Subsequently, the averaged NOE-R Factors are computed, resulting in the outcomes depicted in Fig. 11c and Fig 16.

This work is part of the following article currently under revision:

Optimal Epitope Mapping and Discovery of New Ligands for DC-Sign Recognition. J. D. Martínez, R. Núñez-Franco, P. Valverde, J. Jiménez-Barbero, G. Jiménez-Osés, F. J. Cañada (*submitted manuscript*)

Chapter 5

Thermodynamic Stabilization of Human Frataxin

1. Introduction

Therapeutics based on proteins have benefits over small molecule drugs, providing higher potency and a more diverse range of functions, such as catalysis, signaling, and transport. Additionally, as they have evolved to perform highly specialized roles, they typically induce fewer side effects [181]. Since the first FDA approval of a recombinant protein, insulin, in 1982, approximately 350 protein-based therapeutics have been developed, spanning antibodies (Blinatumomab [182]), enzymes (Pegademase bovine [183]), coagulation factors, protein hormones, and cytokines, addressing a spectrum of conditions from leukemia to multiple sclerosis. Group I protein therapeutics, exemplified by insulin and pancreatic enzymes from pigs, utilize exogenous enzymes and regulatory proteins to counteract the impacts of deficiencies or malfunctions in natural proteins due to pathological mutations [184]. These developments show improved versatility and specificity, becoming integral components in contemporary medical treatment approaches.

The success of protein-based therapeutics is frequently limited by the stability of the biomolecular active principle. Numerous examples demonstrate that diminished thermodynamic stability *in vitro* often aligns with compromised protein homeostasis *in vivo* [185] due to early unfolding and subsequent degradation of the active species in the proteasome [186] or other protein clearance pathways [187]. This challenge can potentially be mitigated by developing protein variants that maintain stable folded conformations *at physiological temperatures*. For successful implementation of this *thermodynamic stabilization* strategy, it is imperative that the engineered variants closely mimic their natural counterparts to preserve the native function and avoid triggering immune responses [188].

Various strategies have been used to optimize protein-based therapeutics by engineering their physicochemical properties [181,182,191,183–190]. These strategies aim at regulating protein homeostasis, enhancing bioavailability, and extending the time they remain functional in the body. For instance, some strategies improve bioavailability by enlarging the hydrodynamic diameter of the protein. This reduces kidney filtration and is achieved through chemical modifications of the protein surface with polymer grafts like polyethylene glycol or by fusing them to larger, more soluble proteins.

In turn, thermodynamic stabilization and solubility can be achieved by engineering the amino acid sequence [181]. This methodology is especially advantageous as it bypasses chemical modifications, thereby enhancing reproducibility and yield in protein expression. This approach has been effectively used to modify, for instance, the pharmacokinetics of insulin. Engineering the amino acid sequence of insulin to alter its isoelectric point has produced the long-active variant (24 h) glargine [192],

where the slower absorption is due to increased precipitation resulting from an isoelectric point closer to physiological pH, as well as the faster acting variant glulisine, where shifting the isoelectric point in the opposite direction minimizes the formation of oligomers and facilitates absorption [193]. This approach of mutagenesis has been similarly applied to interferon β-1B, interleukin-2, and human growth hormone, among others [194–196].

Friedreich's ataxia, an autosomal recessive disorder manifesting between the ages of 10-15 years, is receptive to treatment with Group I protein-therapeutics. The disease manifests with progressive ataxia, loss of tendon reflexes, and dysarthria, with a prevalence of 1–2 per 50000 [197]. Patients diagnosed with this condition consistently exhibit diminished levels of functional frataxin, a compact iron-storage protein crucial for chaperoning ferrochelatase which serves as an allosteric modulator for iron-sulfur clusters biosynthesis. This activity is achieved by binding of frataxin to an assembly complex in the mitochondria, consisting of proteins NFS1, ISD11, ACP, and ISCU (Fig. 1). This deficiency consequently results in the impaired maturation of iron–sulfur cluster proteins, affecting mitochondrial respiratory complexes, Krebs cycle proteins, and proteins integral for DNA repair and replication [197].

Human frataxin undergoes maturation from a 210 amino acid precursor protein and, after proteolytic processing within the mitochondria, it is converted into its mature, operational form, constituting amino acids 81-210 [198]. Structurally, it includes a flexible N-terminus, a preserved antiparallel β-sheet integral for protein recognition, two α-helices, and a non-structured C-terminus [199]. Each frataxin unit can bind up to 7-10 Fe^{2+} ions via a conserved anionic surface (i.e., the acidic ridge, Fig. 1). When bound to Fe^{2+}, frataxin exhibits specificity in binding to protoporphyrin IX — which shares the same binding epitope as ferrochelatase — with micromolar affinity [200].

In over 96% of the patients, homozygous expansions of GAA trinucleotide repeat in intron 1 of the frataxin (FXN) gene are observed, triggering local chromatin changes leading to transcriptional repression of the gene (FXN mRNA levels reduced by 70-95%) [197]. In a minority of cases, patients with Friedreich's ataxia exhibit a GAA expansion on one allele and a *missense mutation* on the other. Over 10 missense mutations have been identified, including I154F and L198R, which lead to a reduction in frataxin levels, thereby affecting its stability, solubility, and function [201,202]. While no cure for the disease has been approved yet, there are indications suggesting that an increase in frataxin levels could potentially mitigate and revert the associated symptoms, opening avenues for gene therapy-based replacement approaches [203-205].

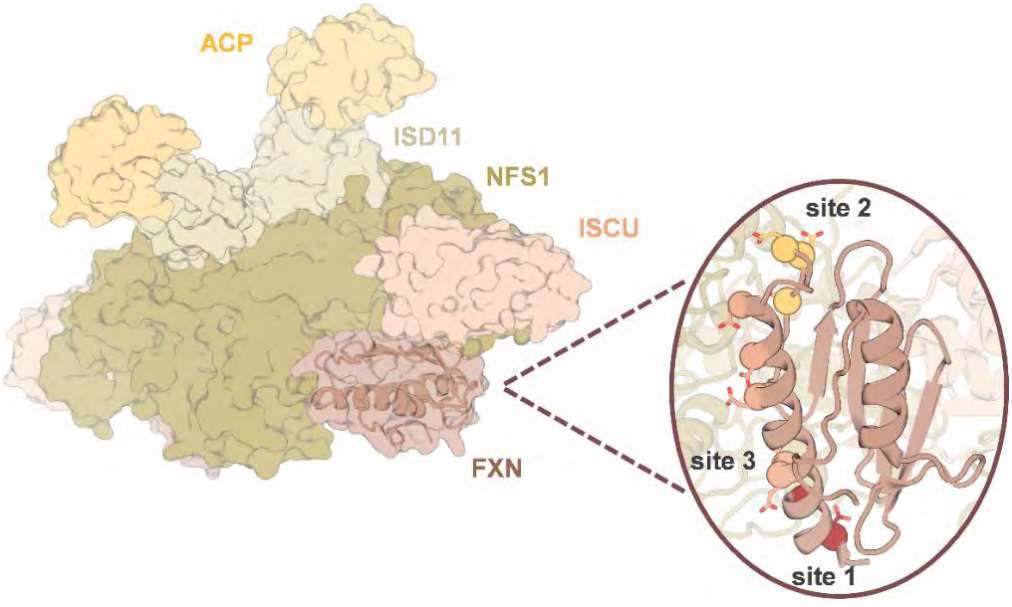


Figure 1. Cryo-EM structure of the frataxin-bound active human complex (PDB ID: 6NZU) [206]. The complex contains two copies of the NFS1-ISD11-ACP-ISCU-FXN hetero-pentamer. NFS1: mitochondrial cysteine desulfurase (colored in green); ISD11: LYR motif-containing protein 4 (colored in light brown); ACP: acyl carrier protein (colored in light orange); ISCU: iron-sulfur cluster assembly enzyme (colored in dark orange); FXN: frataxin (colored in dark red). A close-up view of frataxin is shown in the circle. The spheres represent the residues involved in iron binding, with each color representing one potential iron binding site, capable of binding more than one ion [207] (site 1: E92, E96, site2: E121, D122, D124, site3: E100, E101, D104, E108, E111, D112, D115).

Alternatively, protein replacement therapy emerges as a viable option, if frataxin can be successfully delivered to the mitochondria [197]. Currently, a Phase 2 human trial is underway to investigate the therapeutic potential of a fusion protein integrating FXN and CTI-1601 (Larimar Therapeutics, Inc. ClinicalTrials.gov Identifier: NCT05579691), which leverages the cell-penetrating capability of the trans-activator of transcription (TAT) peptide for targeted delivery [208,209]. To improve protein replacement therapy options, engineering the homeostasis of frataxin to increase its thermodynamic stability and solubility, can bolster its resistance to intracellular degradation and potentially its bioavailability, thus optimizing its therapeutic efficacy.

This study focuses on the structured region of mature human frataxin, spanning residues 91–210. Through computational design, engineered thermostable variants of both the wild-type frataxin and its pathological single mutants I154F and L198R are developed, highlighting their potential in protein replacement therapy.

Recent advances in Artificial Intelligence (AI) have produced tools such as AlphaFold [78], RoseTTAFold [79], and ESMFold [87], which are capable of predicting highly accurate three-dimensional protein models from sequences thus narrowing the sequence-structure gap [210] and introducing novel avenues for *in silico* protein design and engineering [211–213]. By integrating AlphaFold structure ensembles with Rosetta energy predictions, our group has developed a computational method to predict protein thermostability changes upon mutation. This method has been validated across diverse enzyme families engineered by directed evolution [214]. In this work we combine i) mutational hotspots selection from biological function and evolutionary data, ii) sequence sampling on these identified hotspots using ProteinMPNN [213]—a deep learning approach for sequence optimization of a given fold, and iii) our AlphaFold/Rosetta method for assessing thermostability to design stabilized, biologically active frataxin variants (Fig. 2).

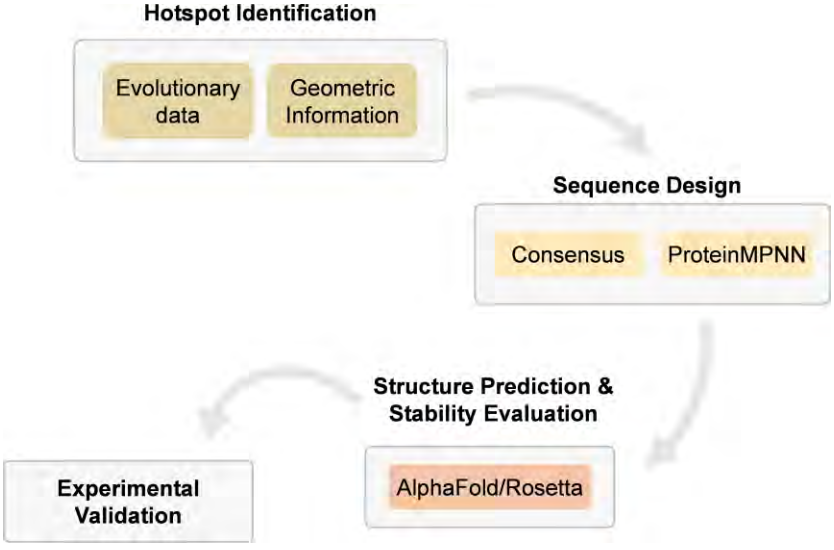


Figure 2. Design strategy for thermostabilizing human frataxin. The first panel illustrates the data guiding the design, including evolutionary and structural information. Then, sequence design options are presented, either relying on consensus information or sampling mutations with ProteinMPNN. Structures derived from the designed sequences are predicted and screened through our combined AlphaFold/Rosetta method to assess their thermostability. Finally, the top ranked designs are experimentally tested.

All designed variants have been expressed in *Escherichia coli* in high yields (see Methods), and their melting temperatures (T_m) measured. Among the 26 evaluated designs, all of them exhibited comparable or superior thermostability and unfolding reversibility than the natural sequence. Notably, the FXN-10 variant achieved a large thermostabilization of $\Delta T_m = +23$ °C relative to the wild-type. Additionally, this variant demonstrated enhanced thermodynamic stabilization at physiological temperatures,

leading to increased proteolytic resistance while retaining the ability to bind divalent cations and the FeS assembly complex, demonstrating its potential in protein replacement therapy.

2. Results and discussion

a. Design based on the consensus approach

Thermostability serves as an indicator of protein thermodynamic stability. Engineering this property is crucial as it seeks to broaden the temperature range at which proteins remain stable and functional. This enhancement is crucial for development of industrial and biomedical applications [215]. The consensus approach has yielded successful thermostabilization within enzyme families such as phytase-1, resulting in melting temperature increase above 20 °C through the cumulative effect of single mutations [216,217]. To set a baseline for the thermostabilization that can be achieved through evolutionary information, a consensus variant (FXN-01) was first designed using 7 mutational hotspots derived from a conservation analysis of the frataxin-enriched multiple sequence alignment named as MSA80 (alignment of sequences with at least 80% identity to human frataxin; positions with conservation below 60% selected as mutational hotspots; see Methods) and selecting the most abundant amino acid at each position.

The AlphaFold/Rosetta-based methodology [214] developed within our research group predicted the consensus variant FXN-01 to be more stable than the wild-type, and indeed it showed a T_m increase of + 6.1 °C (Tables 1 and 2). The F120P single mutation, located in a loop region, imparted the most pronounced thermostabilizing effect, while the remaining six mutations were nearly thermoneutral (variant FXN-02, ΔT_m = + 0.4 °C). The F120P mutation on its own (variant FXN-03) induced a substantial ΔT_m of + 5.1 °C. Aligned with these results, the strategic introduction of proline residues in flexible loops is a classic strategy for enhancing thermostability, frequently observed in thermophilic-mesophilic protein pairs [218]. Such proline locally increase protein rigidity in both the folded and denatured states [219], leading to augmented thermal stability by reducing the entropic penalty to folding.

Inspection of available X-ray and NMR structures of wild-type frataxin and Friedreich's Ataxia single mutants [199,206,220–222] reveals a flexible region (residues 136-141) interacting with Phe120 (Fig. 3a). Similarly, AlphaFold ensembles of wild-type frataxin depict significant flexibility with two distinct backbone conformations for residues 138 and 129 (Fig. 3a), highlighting AlphaFold's dynamic prediction capabilities [214,223]. In the ensemble of FXN-03, however, one conformation dominates, indicating that the F120P mutation reduces conformational

variability not only in the loop where it's located but also in the neighboring 137-140 region.

This selected conformation is predicted by Rosetta to be more energetically stable, particularly in FXN-03 (Fig. 3b).

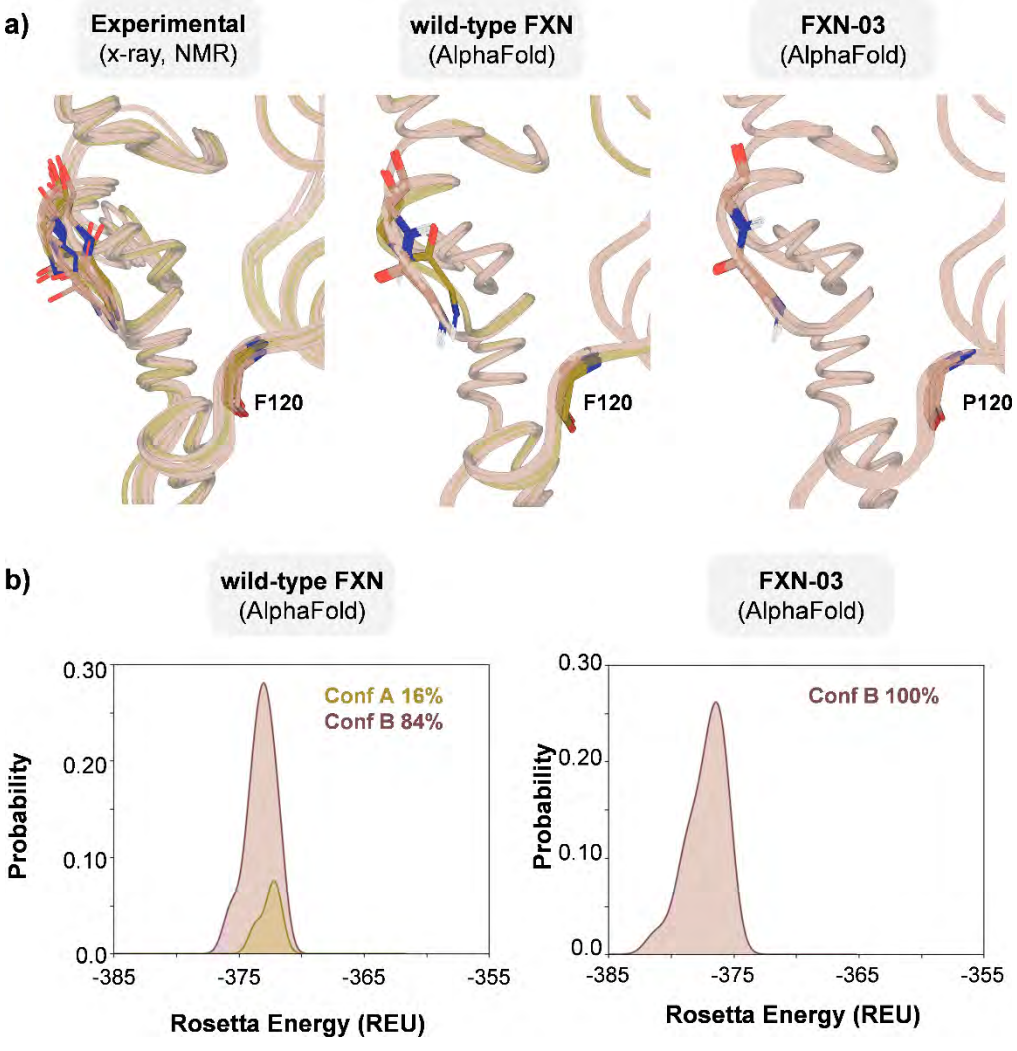


Figure 3. a) Flexible region between residues 137 to 140 in a) an overlay of NMR and X-ray structures (PDB ID: 1EKG, 3S4M, 3S5D, 3S5E, 3S5F, 3T3J, 3T3L, 3T3T, 3T3X, 1LY7, 6NZU) of different frataxin variants, and AlphaFold ensembles (25 structures) of wild-type frataxin and F120P single mutant FXN-03. While two backbone conformations are observed for residues 138 and 139 in both the experimental and AlphaFold predicted structures for wild-type frataxin, only one conformation is observed for FXN-03. b) Normalized kernel density estimates of the Rosetta energy distributions calculated for each variant in the two conformations. REU: Rosetta energy units.

Intrigued by the evolutionary significance of the highly conserved F120P mutation, a phylogenetic analysis was performed of the frataxin-enriched MSA80 alignment (see Methods). This analysis clearly pinpointed the presence of this substitution, alongside two additional mutations (S160T and T191S), acting as distinctive markers between different orders of mammals and amphibians/reptilians/birds (Figs. 4 and 13). Notably, the combination of these mutations into the triple mutant FXN-04, which our protocol predicted to be at least as stable as FXN-01, yielded a further T_m increase of + 8.6 °C compared to the wild-type.

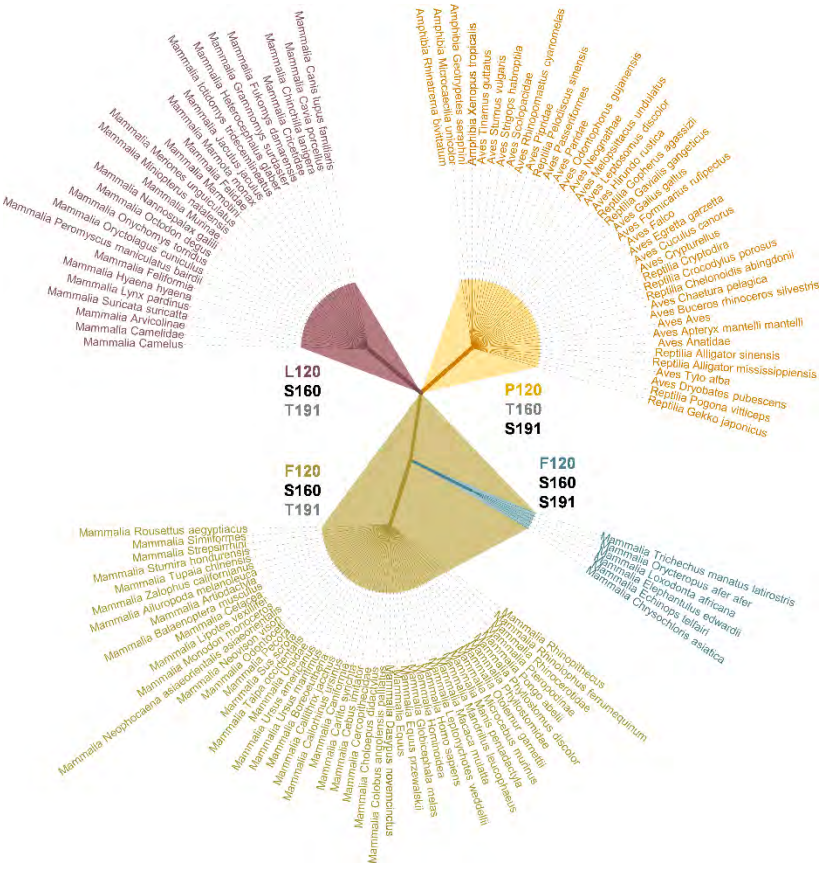


Figure 4. Unrooted phylogenetic tree of the sequences in a curated version of the MSA80 dataset, focusing on positions 120, 160 and 191. Branch lengths are ignored for clarity. Only branches with population >4% (highlighted with the corresponding colors in Table 10) are shown for clarity.

b. ProteinMPNN sequence design

Recently, the Baker Lab at the Institute for Protein Design, University of Washington, introduced a highly efficient deep learning-based protein sequence design method called ProteinMPNN. This method achieves exceptional protein sequence recovery on

native backbones, bypassing the need for extensive sidechain packing calculations [213]. As an alternative to the classic consensus approach, the same 7 mutational hotspots identified in the MSA80 dataset were sampled with ProteinMPNN (see Methods). This led to generation of three additional variants: FXN-05 and FXN-06 (each with 5 mutations), anticipated to be the most stable sequences with our AlphaFold/Rosetta protocol, and FXN-07 as the most frequently occurring solution (also with 5 mutations) (Tables 1 and 2). To our delight, all three designs exhibited increased T_m values compared to the wild-type FXN, ranging from + 2.5 °C to + 7.3 °C, thus either matching or surpassing the thermostabilization achieved by the consensus approach based solely on sequence conservation.

To explore the potential for enhanced thermostabilization through broader sequence sampling, ProteinMPNN was used to sample the larger sets of hotspots generated from the MSA70 and MSA60 sequence alignments. These alignments include sequences with a minimum identity of 70% and 60% to human frataxin, respectively. Within these alignments, 11 and 26 positions exhibit conservation levels below 60%, making them suitable for mutation (see Methods). This attempt yielded FXN-08 (9 mutations) and FXN-09 (16 mutations) as the variants predicted by our AlphaFold/Rosetta protocol to be the most stable within each pool. Of note, these variants outperformed all previous designs in terms of stability. With these variants, even larger increases in T_m over the wild-type were achieved (+ 11.9 °C and + 12.7 °C, respectively) (Tables 1 and 2). These findings strongly indicate that broader sequence design can indeed lead to improved thermostabilization. It is worth noting that ProteinMPNN selectively mutates a subset of available hotspots (i.e. not all positions marked as mutable are indeed modified) to achieve significant enhancements in target properties, as previously documented [213].

An alternate approach to use evolutionary data, without imposing arbitrary per-residue conservation thresholds, involves calculating the Shannon entropy per position [224] within a given multiple sequence alignment (MSA). Shannon entropy (SE) not only captures the conservation levels but also considers the broader distribution of amino acid identities across different sequences. Consequently, positions with higher SE values experience both more frequent variations and a to more diverse amino acids within the MSA. These high-SE positions are identified as mutational hotspots (see Methods).

Sampling the top 20 mutational hotspots identified through SE analysis on MSA80 using ProteinMPNN yielded FXN-10 (13 mutations), predicted to be the most stable variant within this set. Strikingly, FXN-10 displayed a substantial + 23.3 °C increase in T_m compared to wild-type (Tables 1 and 2). It should be emphasized that our computational approach efficiently yielded a superstable frataxin variant by selecting just 13 mutations from a relatively small mutational landscape (20 positions).

Moreover, this remarkable improvement in thermostability was achieved in a single design step, eliminating the need for multiple rounds of extensive mutation often required by methods like iterative saturation mutagenesis or directed evolution [225,226]. Furthermore, this approach achieved thermostabilization levels comparable to the most successful computational techniques to date [97,227,228], while significantly reducing the computational cost (less than 1 hour per variant on a mid-range workstation, see Table 12; calculations are trivially parallelizable across a computing cluster effectively allowing the design of hundreds of variants in less than a day).

Table 1. Summary of the thermostabilized designs for wild-type frataxin, with indication of the method used for hotspot selection, the number of mutational hotspots (N. hotspots), the approach followed to generate the mutations, the name of the variant, the number of mutations (N. mutations), the difference in melting temperature (ΔT_m), and the predicted unfolding free energy change ($\Delta\Delta G^{calc}$) (REU: Rosetta energy units). All quantities are referred to wild-type frataxin. The identity of the mutations for each variant are presented in Table 2. See Methods for the definition of multiple sequence alignments MSA80, MSA70 and MSA60.

Hotspot detection	N. hotspots	Mutation approach	Variant	N. mutations	ΔT_m (°C)	$\Delta\Delta G^{calc}_{mut,u}$ (REU)
Per-residue conservation ≤ 60% on MSA80	7	Conservation	FXN-01	7	6.1[a]	8.2
			FXN-02	6	0.4[a]	3.1
			FXN-03	1	5.1	4.1
			FXN-04	3	8.6	7.7
		ProteinMPNN	FXN-05	5	7.3	11.9
			FXN-06	5	4.8	10.6
			FXN-07	5	2.5	9.0
Per-residue conservation ≤ 60% on MSA70	11	ProteinMPNN	FXN-08	9	11.9	12.2
Per-residue conservation ≤ 60% on MSA60	26	ProteinMPNN	FXN-09	16	12.7	14.3
Shannon entropy (SE > 0.65) on MSA80	20	ProteinMPNN	FXN-10	13	23.3	12.7

Table 2. Number and identity of mutations for wild-type and designed frataxin variants.

Variant	N. mutations	Mutations
wild-type FXN	0	
FXN-01	7	F120P, S160T, K171R, A187S, A188T, T191S, K192T
FXN-02	6	S160T, K171R, A187S, A188T, T191S, K192T
FXN-03	1	F120P
FXN-04	3	F120P, S160T, T191S
FXN-05	5	F120P, S160T, K171T, A188T, T191S
FXN-06	5	F120P, S160T, K171S, A188K, T191S
FXN-07	5	F120P, S160T, K171E, A188K, T191S
FXN-08	9	R97K, Y118F, F120P, S160T, K171T, A188K, T191S, S202H, A204K
FXN-09	16	T93S, E108D, A114K, K116Q, Y118F, F120Q, E121P, S160T, K171S, N172S, A187S, A188E, A193L, K197P, L198I, S202H
FXN-10	13	R97K, A114K, Y118F, F120P, S129D, S160T, K171T, V180K, A187S, A188E, T191S, S202H, A204K

c. Rescue of pathological mutants I154F and L198R

Having successfully achieved thermostabilization of wild-type frataxin, the next goal was to rescue the thermostability of two pathological single mutants, FXN-I154F and FXN-L198R, both of which lead to substantial decreases in T_m by over 10 °C. In this context, consequences of pathological mutations in Friedreich's ataxia has been widely discussed, as distinct mutations result in varying disease phenotypes affecting protein levels, localization, and function [229].

Full-length frataxin comprises an unstructured region at its N-terminus including the initial 41 amino acids. Maturation into the functional form (residues 81-210) involves at least one intermediate state (residues 42-210) [229]. The I154F pathological mutation (ΔT_m = – 12.9 °C, Fig. 5) has been linked to decreased mature frataxin levels and hindered intermediate accumulation, without compromising its association with mitochondria [201]. This mutation also triggers precipitation upon iron binding due to increased flexibility [230]. In contrast, the L198R pathological mutation (ΔT_m = – 20.5 °C, Fig. 5), located in the C-terminal domain, introduces a positive charge in a nonpolar environment. This destabilizes the native state, leading to increased dynamics in the microsecond/millisecond timescale around the mutation [229,231,232].

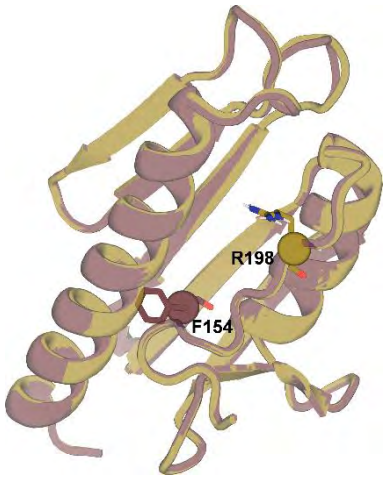


Figure 5. Overlay of AlphaFold models for FXN-I154F and FXN-L198R frataxin variants showing the location of the pathological mutations.

For each of these pathological variants, we generated eight designs by three different strategies:

i) Transferring one or the three stabilizing mutations identified in the phylogenetic analysis (F120P, S160T, T191S).
ii) Resampling the same 26 mutational hotspots sampled in the wild-type using ProteinMPNN.
iii) Transferring mutations from the most stable engineered wild-type variants (FXN-09 and FXN-10).

This resulted in designs FXN-11 to FXN-18 (derived from FXN-I154F) and FXN-19 to FXN-26 (derived from FXN-L198R) (Tables 3-4). Regarding the I154F mutation, five variants achieved T_m increases of at least + 20 °C, fully compensating for the destabilizing effect of the pathological mutation, and even surpassing wild-type stability. A similar trend was observed for the L198R pathological mutation, with four variants exhibiting ΔT_m > + 24 °C. Notably, the largest stabilization was accomplished by transferring mutations from FXN-10 (variants FXN-18 and FXN-26), although resampling with ProteinMPNN (variants FXN-15/16 and FXN-23/24) yielded comparable results in terms of both thermostability (ΔT_m + 20 to + 24 °C) and introduced mutations (96-98% sequence identity), as expected. Significantly, nearly all variants developed augmented stability without compromising folding reversibility; this is particularly notable for variants designed with ProteinMPNN, which in many cases even improved reversibility (see Methods, Fig. 16 and 18 and Table 16). This characteristic is crucial for enhancing protein homeostasis *in vivo*, as proteins that are

irreversibly unfolded tend to be more susceptible to aggregation and degradation. In summary, these results demonstrate the capability of this methodology to allosterically rescue the pathological single mutations through extensive substitutions in different regions of the protein.

Table 3. Summary of the thermostabilized designs for FXN-I154F and FXN-L198R with indication of the method used for hotspot selection, the number of mutational hotspots (N. hotspots), the method used to generate the mutations, the name of the variant, the number of mutations (N. mutations), the difference in melting temperature (ΔT_m), and the predicted unfolding free energy change ($\Delta\Delta G^{calc}$) (REU: Rosetta energy units). All quantities are referred to the corresponding parent pathological mutant FXN-I154F or FXN-L198R. The identity of the mutations for each variant are presented in Table 4. See Methods for the definition of multiple sequence alignments MSA80, MSA70 and MSA60.

Hotspot detection	N. hotspots	Mutation approach	Variant	N. mutations	ΔT_m (ºC)	$\Delta\Delta G^{calc}_{mut,u}$ (REU)
I154F + mutations from FXN-04			FXN-11	3	9.9	9.5
I154F + mutations from FXN-03			FXN-12	1	6.7	5.7
Per-residue conservation ≤ 60% on MSA60	26	ProteinMPNN	FXN-13	16	20.3	8.8
			FXN-14	15	21.1	9.0
Shannon entropy (SE > 0.65) on MSA80	20	ProteinMPNN	FXN-15	11	21.1	12.5
			FXN-16	11	20.5	13.3
I154F+ mutations from FXN-09			FXN-17	16	16.4	14.0
I154F+ mutations from FXN-10			FXN-18	13	23.4	15.2
L198R + mutations from FXN-04			FXN-19	3	9.3	2.6
L198R + mutations from FXN-03			FXN-20	1	4.1	1.7
Per-residue conservation ≤ 60% on MSA60	26	ProteinMPNN	FXN-21	12	19.8	10.5
			FXN-22	14	22.9	8.1
Shannon entropy (SE > 0.65) on MSA80	20	ProteinMPNN	FXN-23	13	23.7	11.7
			FXN-24	12	23.9	9.6
L198R+ mutations from FXN-09			FXN-25	15	16.6	15.1
L198R + mutations from FXN-10			FXN-26	13	24.1	12.3

Table 4. Number and identity of mutations for pathological mutants FXN-I154F, FXN-L198R and frataxin variants. Pathological mutation is shown in red.

Variant	N. mutations	Mutations
FXN-I154F	1	I154F
FXN-11	3+1	F120P, I154F, S160T, T191S
FXN-12	1+1	F120P, I154F
FXN-13	16+1	T93A, R97K, E108D, A114K, K116Q, Y118F, F120P, I154F, S160T, K171E, N172A, A188K, A193L, K197P, L198I, S202H, A204K
FXN-14	15+1	T93A, A114K, K116Q, Y118F, F120P, I154F, S160T, K171E, N172S, A188K, K192T, A193L, K197P, L198I, S202H, A204K
FXN-15	11+1	R97K, A114K, Y118F, F120P, S129D, I154F, S160T, K171S, A188K, T191S, S202H, A204K
FXN-16	11+1	R97K, A114K, Y118F, F120P, S129D, I154F, S160T, K171S, V180K, A188E, T191S, A204K
FXN-17	16+1	T93S, E108D, A114K, K116Q, Y118F, F120Q, E121P, I154F, S160T, K171S, N172S, A187S, A188E, A193L, K197P, L198I, S202H
FXN-18	13+1	R97K, A114K, Y118F, F120P, S129D, I154F, S160T, K171T, V180K, A187S, A188E, T191S, S202H, A204K
FXN-L198R	1	L198R
FXN-19	3+1	F120P, S160T, T191S, L198R
FXN-20	1+1	F120P, L198R
FXN-21	12+1	T93A, E108A, A114K, K116Q, Y118E, F120P, S160T, K171D, N172T, A188T, A193L, K197P, L198R
FXN-22	14+1	T93E, R97K, A114K, K116Q, Y118F, F120P, S160T, K171E, N172S, A188K, A193L, K197P, L198R, S202H, A204K
FXN-23	13+1	R97K, E108T, A114K, Y118F, F120P, S129D, S160T, K171S, V180K, A188E, T191S, L198R, S202H, A204K
FXN-24	12+1	R97K, E108K, A114K, Y118F, F120P, S129D, S160T, K171S, A188K, T191S, L198R, S202H, A204K
FXN-25	15+1	T93S, E108D, A114K, K116Q, Y118F, F120Q, E121P, S160T, K171S, N172S, A187S, A188E, A193L, K197P, L198R, S202H
FXN-26	13+1	R97K, A114K, Y118F, F120P, S129D, S160T, K171T, V180K, A187S, A188E, T191S, L198R, S202H, A204K

d. Physical origin of improved thermostability

The circular dichroism (CD) spectra of both wild-type frataxin and the highly stable FXN-10 variant were recorded at various temperatures (Fig. 6a and Methods).

Notably, at 75 °C, wild-type frataxin unfolds, while the FXN-10 variant maintains a secondary structure closely resembling that of the wild-type, even at this elevated temperature. This observation underscores that the mutations incorporated into FXN-10 do not significantly alter the native folded structure of the protein.

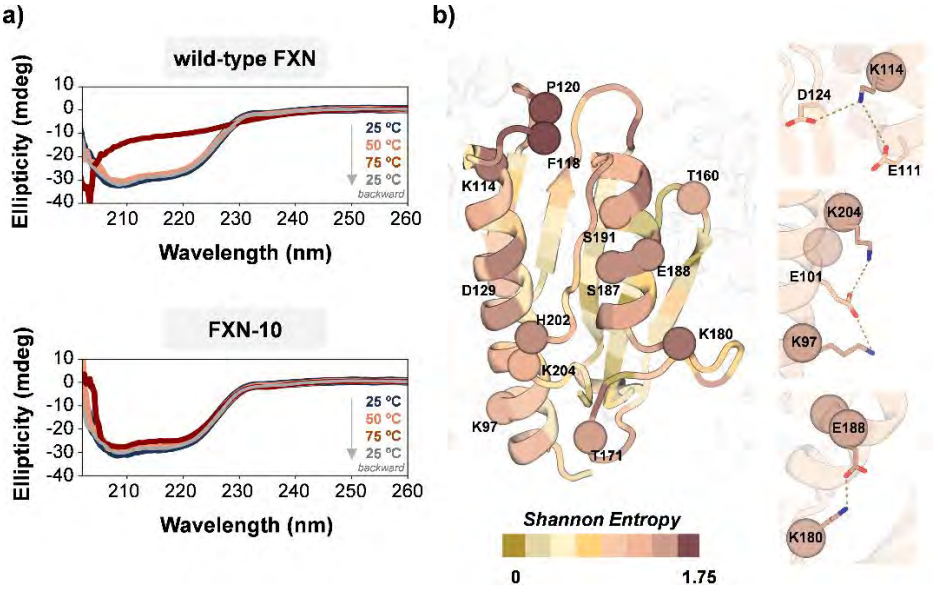
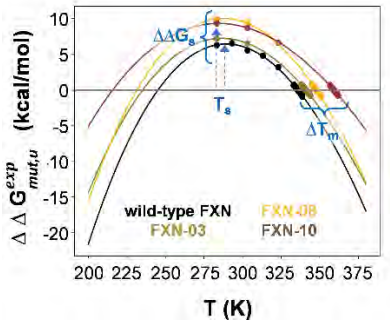


Figure 6. a) CD spectra at 25, 50, 75 °C (heating phase) and 25 °C (cooling phase) for wild-type frataxin and design FXN-10. b) AlphaFold model for FXN-10 colored according to Shannon entropy with indication of the mutated positions as spheres (left). Salt bridges introduced in thermostable design FXN-10 and not present in wild-type frataxin (right).

The most stable variants share many common mutations (8 between FXN-08 and FXN-10; 6 between FXN-09 and FXN-10), including hydrophobic-to-charged (A114K, A188E/K, A204K), hydrophobic-to-polar (A187S), charged-to-polar (K171T/S), substitutions to closely related amino acids (R97K, Y118F, S160T, T191S), interconversion of polar amino acids (S202H), and substitutions to proline (F120P, as discussed previously) (Tables 2 and 4). Such surface-level mutations involving charged amino acids or proline are frequently observed in thermophilic counterparts of mesophilic proteins [218]. In line with this, it is known that ProteinMPNN tends to introduce charged amino acids at lower sampling temperatures, which might contribute to the enhanced thermostability of designed sequences [213].

In fact, the five charged residues introduced in the most stable variant FXN-10 (Lys97, Lys114, Lys180, Glu188, and Lys204) establish a dense network of salt bridges absent in wild-type frataxin (Fig. 6b). This observation aligns with earlier discoveries that optimizing charge-charge interactions on the protein surface serves as a mechanism for stabilization [233–235]. Nevertheless, it is important to mention that the theoretical isoelectric point (pI = 5.4) and total charge at pH 7.4 (Z = – 8.1) of FXN-10 are nearly identical to those of the wild-type (pI = 5.2, Z = – 8.1) (Tables 14-16).

In order to gain a deeper understanding of the mechanisms underlying the achieved thermostabilization, we measured the protein stability curves [236] of selected variants through chemical denaturation experiments at different temperatures, focusing on the wild-type, FXN-03, FXN-08, and FXN-10 variants (Figs. 7 and 18). These variants showed similar temperatures of maximal stability (T_s = 11–16 °C) and raised stability curves at all temperatures – an ubiquitous phenomenon in the majority of mesophile-thermophile pairs [218,237–240] – with the FXN-08 and FXN-10 variants achieving a thermodynamic stabilization of $\Delta\Delta G_s \sim 3$ kcal mol^{-1}. While the stability curve of FXN-08 shows essentially the same curvature as the wild-type – indicating similar heat capacities of unfolding (ΔC_p) – it is shifted upwards due to a larger enthalpy of unfolding at the melting temperature (ΔH_m). On the other hand, variants FXN-03 and FXN-10 exhibit comparable ΔH_m and lower ΔC_p values relative to wild-type frataxin. Particularly, for the FXN-10 variant, which showcases the largest increase in T_m, ΔC_p decreases by 0.8 kcal mol^{-1}K^{-1}, resulting in a notable flattening of the curve, a stabilization mechanism observed in heat adaptation [237,239,240]. Significantly, the enhanced stability exhibited by these variants at physiological temperatures underscores the potential ability of this approach to improve homeostasis *in vivo*.



	T_m (ºC)	T_s (ºC)	ΔG_s (kcal mol^{-1})	ΔH_m (kcal mol^{-1})	ΔC_p (kcal mol^{-1} K^{-1})
WT	62.77 (0.06)	16.17	6.46	90.81 (1.42)	1.81 (0.06)
FXN-03	67.87 (0.04)	13.29	7.25	88.03 (1.33)	1.48 (0.05)
FXN-08	74.62 (0.05)	14.56	10.04	112.75 (2.01)	1.71 (0.06)
FXN-10	86.00 (0.05)	11.22	9.34	86.37 (0.91)	1.03 (0.03)

Figure 7. Stability curves measured for wild-type frataxin, FXN-03, FXN-08 and FXN-10, together with the melting temperature (T_m), temperature of maximum thermodynamic stability (T_s), maximum thermodynamic stability (ΔG_s), enthalpy of unfolding at T_m (ΔH_m), and heat capacity of unfolding (ΔC_p) derived from these curves. Numbers in parenthesis indicate the standard errors arising from the fitting. Note that FXN-08 is the most *thermodynamically stable* variant (i.e., largest $\Delta\Delta G_s$), while FXN-10 is the most *thermostable* one (i.e., largest ΔT_m).

e. Evaluation of *in silico* prediction accuracy

To assess the global accuracy and predictivity of our methodology, a comparison was made between the calculated ($\Delta\Delta G^{calc}$, Table 11) and the experimentally measured changes in thermostability for the set of variants designed in this work. As previously described, the curvature of the stability curve might not be the same across all variants; consequently, the generalized approximation that changes in melting temperatures and unfolding free energies are lineally related [236] (see Methods) cannot be applied

in this instance. On the other hand, a thorough determination of the thermodynamic stabilization ($\Delta\Delta G_s$) of the nearly 30 assayed variants from their stability curves is beyond the scope of this work. Thus, since the direct comparison between Rosetta and experimental energies is not possible, the correlation between $\Delta\Delta G^{calc}$ and ΔT_m values relative to wild-type frataxin was examined (Fig. 8). Overall, a strong correlation was observed between calculated and empirical stabilities within a T_m range extending over 40 °C, surpassing the predictability of widespread force-field and machine-learning-based methodologies [241].

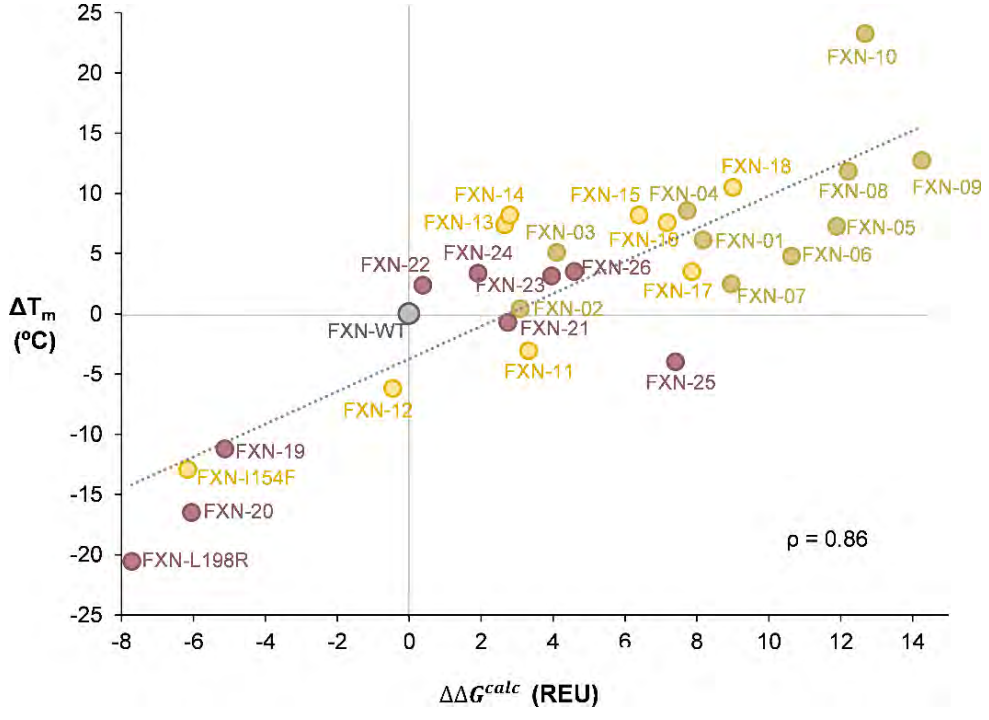


Figure 8. a) Correlation between theoretical $\Delta\Delta G^{calc}$ and experimental ΔT_m values for the 26 variants designed to stabilize wild-type frataxin (in green) and the two pathological mutants I154F (in yellow) and L198R (in red), using wild-type frataxin (in gray) as a reference. Positive and negative values indicate stabilization and destabilization, respectively. Each data point corresponds to a different protein variant. ρ: Pearson correlation coefficient. The dashed line indicates a linear regression with slope 1.35 and intercept –3.74.

Outliers from the generally good alignment are observed in designs FXN-10 and FXN-25, whose ΔT_m are under- and overestimated, respectively. Particularly, for FXN-10, the thermodynamic parameters derived from the stability curves suggest that the underestimation may arise from the flattening of the curve due to a lower ΔC_p, which differentially increases ΔT_m over $\Delta\Delta G_s$. Indeed, for the thermodynamically characterized variants, Rosetta energies display a better correlation with free energy

differences at the stability maximum than with those approximated from ΔT_m (Fig. 9). This suggests that our approach might be best suited for predicting changes in *thermodynamic stability* rather than *thermostability*.

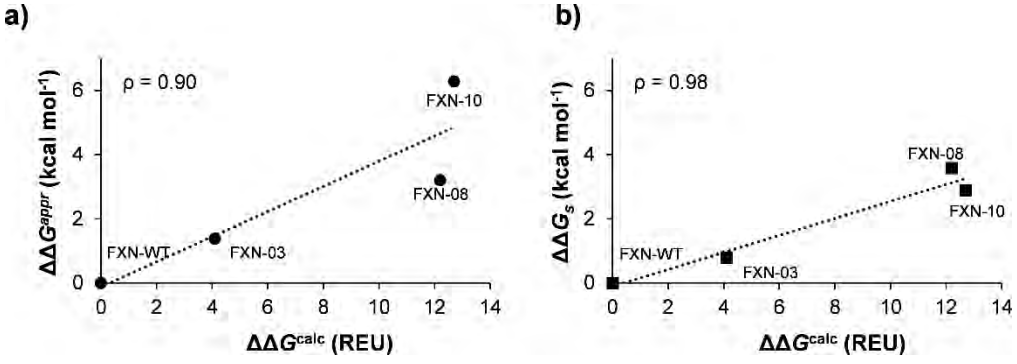


Figure 9. Correlation between theoretical $\Delta\Delta G^{calc}$ and experimental $\Delta\Delta G^{appr}$ (approximated from ΔT_m assuming parallel stability curves in the T_m region using Eq. 14, Table 16; ●) and $\Delta\Delta G_s$ (derived from the stabilization curves; ■) for designed variants FXN-03, FXN-08 and FXN-10, using wild-type frataxin as a reference. Positive values indicate stabilization. ρ: Pearson correlation coefficient. The dashed line indicates a linear regression.

Remarkably, when dissecting Rosetta energies, FXN-10 emerges as an outlier in terms of the different contributions to the computed values. More specifically, there were significant differences in the relative weight of Rosetta's *ref* term in the calculated energy of each variant compared to the wild-type. This term accounts for the contribution of the unfolded state to the unfolding free energy [242]; due to the prohibitive cost and susceptibility to errors of explicit calculations on the inherently complex unfolded state, the *ref* term was originally introduced in the Rosetta energy function to assign an structure-independent and empirically determined weight to each amino acid in a sequence to maximize native sequence recovery [242]. In FXN-10, this contribution surpasses those in other variants and dominates $\Delta\Delta G^{calc}$ (Fig. 10), pointing towards a significant role of the unfolded state in thermodynamic stabilization. Given that ΔC_p values are known to correlate with changes in the solvent accessible surface area between unfolded and folded states (ΔSASA) [243], we tentatively attriute the observed decrease in ΔC_p for FXN-10 to a smaller SASA in its unfolded state relative to the wild-type. This hypothesis is supported by the small SASA difference (~2%) calculated for the *folded* states of the two proteins (Table 11).

These results suggest that our AlphaFold/Rosetta-based methodology not only correctly predicts the direction and relative magnitude of thermostability changes due to mutation, but also qualitatively informs on the underlying biophysical mechanisms of thermostabilization.

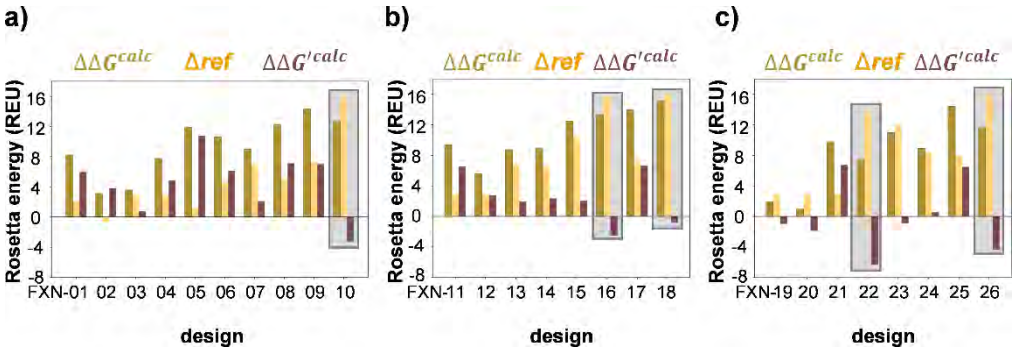


Figure 10. Calculated energies decomposition ($\Delta\Delta G'^{calc} = \Delta\Delta G^{calc} - \Delta ref$) for variants designed to stabilized a) wild-type frataxin, b) pathological mutation FXN-I154F, and c) pathological mutation mutant FXN-L198R. In each case the corresponding frataxin variant is used as a reference. Variants whose calculated stabilization is clearly dominated by the Δref term are highlighted in gray.

f. Stabilizing mutations confer improved resistance to proteolytic degradation

To determine how much the designed variants could enhance protein homeostasis in a biological context, their susceptibility to enzymatic degradation was assessed. Previous research on extremely thermophilic bacteria shows that enzymes derived from these organisms are more resistant to proteolysis than their mesophilic counterparts [244], suggesting that such resistance might be a common characteristic of thermostable proteins [245]. Indeed, resistance to proteolytic degradation has been used as a proxy of the thermodynamic folding stability of large protein libraries [246]. Consistent with this, the most thermostable frataxin variant developed in this study, FXN-10 (ΔT_m = + 23.3 °C), displayed significant resistance to degradation by trypsin compared to the wild-type (Fig. 11a and Table 17). This enhanced resistance to proteolytic degradation can extend the protein's lifetime in biological environments, making our designs attractive as candidate protein therapeutics.

g. Stabilizing mutations preserve biological function

In addition to being thermally and enzymatically stable, an engineered protein must maintain the relevant biological activity to be considered as a potential therapeutic. Through a series of NMR experiments developed in the Precision Medicine and Metabolism Lab at CIC bioGUNE [200], it was confirmed that the superstable variant FXN-10 retains the ability to bind Zn^{2+} ions (serving as a proxy for Fe^{2+}) and protoporphyrin IX (ppIX) which, when bound to metal-loaded frataxin shares the binding epitope with ferrochelatase (Figs. 19-21). FXN-10 was also able to engage in protein-protein interactions with the iron-sulfur assembly cluster. The chemical shift

perturbation (CSP) measured upon the addition of Zn^{2+} and ppIX indicated changes upon binding consistent with those identified for wild-type frataxin [200] (Fig. 11b, 22 and 23). Additionally, the interaction of frataxin with the iron-sulfur assembly complex reveals the appearance of a new set of signals (Fig. 11c, 24 and 25), consistent with the slow exchange regime of the interaction, in line with previous observations made for wild-type frataxin. These results validate the biological competency of the designed, stable variant.

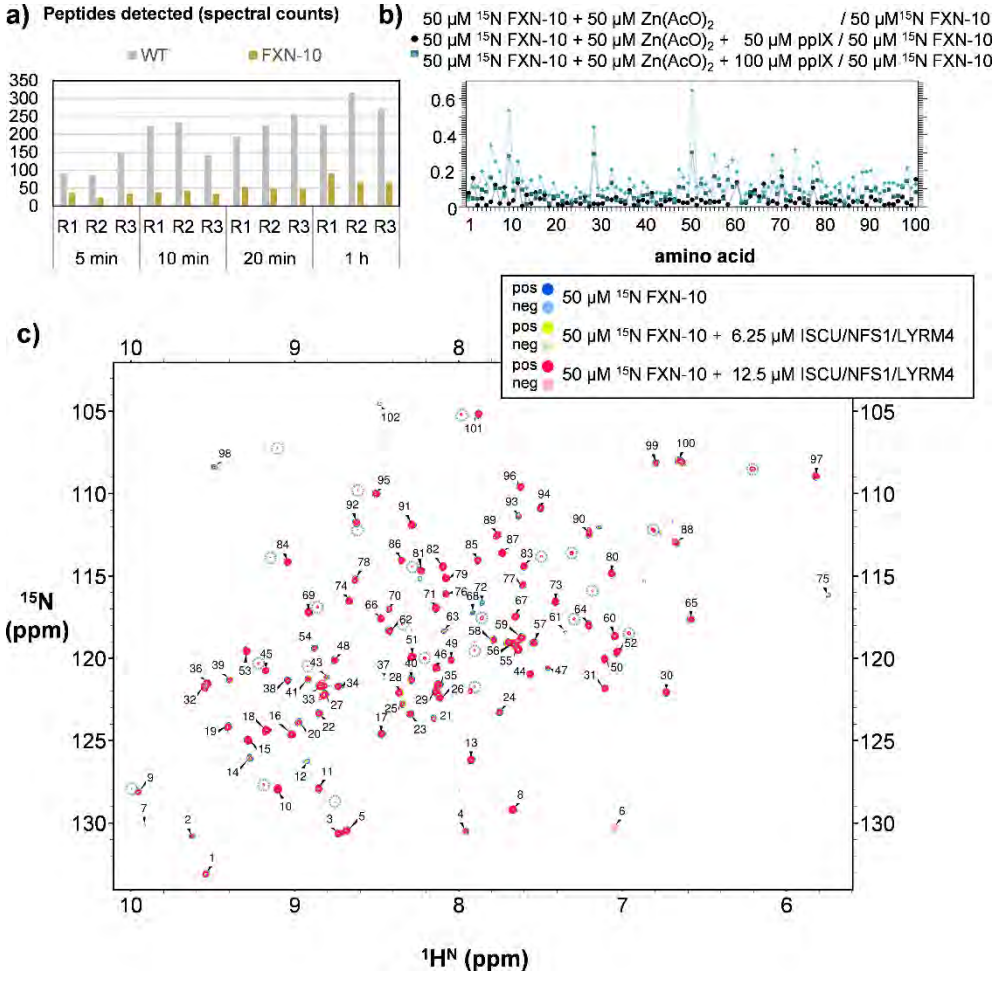


Figure 11. a) Protein stability assay by trypsin degradation and mass spectrometry analysis shows decreased proteolysis for engineered variant FXN-10 compared to wild-type frataxin. Spectral counts refer to peptide-spectrum match (PSM), i.e., peptides identified after protein digestion. R1, R2 and R2 stand for replicas 1, 2, and 3, respectively. b) CSP observed for u-^{15}N,^{13}C-labelled FXN-10 in the presence of Zn^{2+} and protoporphyrin IX (ppIX). c) 800 MHz 2D ^{15}N-sf-HMQC spectra of u-^{15}N,^{13}C-labelled FXN-10 in the presence of increasing amount of FeS assembly complex. Residue numbers are arbitrary due to lack of complete assignment of the protein.

3. Conclusions

Human frataxin variants with enhanced stability have been designed *in silico* by integrating evolutionary information for hotspot selection, sequence sampling with ProteinMPNN, and stability evaluation using AlphaFold ensembles and the Rosetta energy function. Remarkably, these predictions have been made in one single step, required minimal computational resources and achieved 100% success rate. Thus, all designed variants showed higher thermostability than the wild-type, to different degrees. The best design, FXN-10, whose stabilization is largely associated to a lower ΔC_p of unfolding as seen in heat adaptation mechanisms, exhibits remarkable resistance to proteolytic degradation. Moreover, it maintains the ability to bind metal ions and interact with the FeS assembly complex. These features make FXN-10 a prime candidate for protein replacement therapy targeting Friedreich's ataxia. Our rational approach might prove generalizable for designing proteins with improved properties for therapeutic uses and broader biotechnological applications.

4. Methods

a. Mutational hotspots selection

Different sets of mutable amino acids were identified from various conservation analyses conducted on a common multiple sequence alignment (MSA). This MSA was generated by the *jachmmer* [247] search of the wild-type target frataxin sequence (residues 91-210) against the UniRef90 database [248,249] at the initial stage of the AlphaFold version 2.3.0 structure prediction (database size: 140403594 sequences; execution flags: --F1 0.0005 --F2 5e-05 --F3 5e-07 --incE 0.0001 -E 0.0001 -N 1), and contains 1999 sequences. The distribution of sequence identities, computed after removing the gaps from the target frataxin sequence has a maximum at around 40% identity (Fig. 12). At a higher identity (above 60%), a collection of clusters predominantly comprises frataxin homologs, with sporadic occurrences of ferroxidase and phosphatidylinositol 4-phosphate 5-kinase (type-1 beta isoform) homologs (Table 5). Of note, ferroxidase activity has been reported for yeast frataxin [250,251].

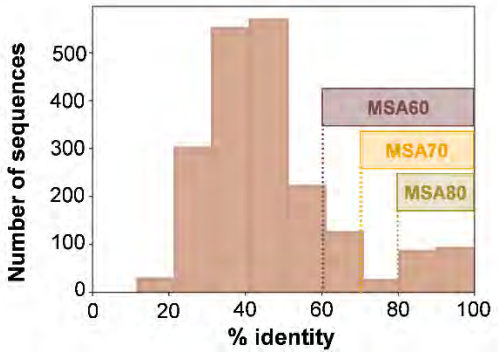


Figure 12. Distribution of sequences in the original Uniref90 multiple sequence alignment for human frataxin (1999 sequences), and derived subsets at 60% (MSA60), 70% (MSA70) and 80% (MSA80) identity, used for mutational hotspots identification.

To generate mutational pools of varying sizes, this MSA was further filtered by applying three different sequence identity cutoffs with respect to the target human frataxin sequence after removing gaps (60%, 70%, and 80%). This resulted in MSA60, MSA70, and MSA80, respectively (Tables 6-10). For each reduced alignment, positions with less than 60% conservation were chosen as mutational hotspots, guided by the principle that the least conserved positions are better suited for mutation without compromising structural integrity and the associated function [252–255]. Consequently, no mutational hotspots were allowed on the β-sheet plane, which is pivotal for binding to ISCU (iron-sulfur cluster assembly enzyme). Notably, the pathological mutation W155G located in this region has been identified to obliterate the desulfurase activity of the supercomplex responsible for iron–sulfur cluster assembly [256].

Table 5. Composition of the multiple sequence alignments used for identification of mutational hotspots. The different protein categories are extracted from their definition in the FASTA files downloaded from UniProt (https://www.uniprot.org).

Category	Uniref90	MSA60	MSA70	MSA80
Frataxin	576 (29%)	316 (94%)	196 (93%)	170 (94%)
Frataxin-like	12 (1%)	0 (0%)	0 (0%)	0 (0%)
Ferroxidase	1041 (52%)	2 (1%)	0 (0%)	0 (0%)
Iron-sulfur_cluster	75 (4%)	0 (0%)	0 (0%)	0 (0%)
Iron_donor_protein	82 (4%)	0 (0%)	0 (0%)	0 (0%)
Other_iron-related	9 (0%)	0 (0%)	0 (0%)	0 (0%)
Uncharacterized	103 (5%)	7 (2%)	7 (3%)	5 (3%)
Other	101 (5%)	10 (3%)	7 (3%)	5 (3%)
TOTAL	**1999 (100%)**	**335 (100%)**	**210 (100%)**	**180 (100%)**

A fourth collection of mutational hotspots was derived from the frataxin-enriched MSA80 dataset by evaluating each position of frataxin orthologs using Shannon entropy (SE) [224] and excluding gaps :

$$SE(position) = -\sum_i f_i \ln f_i \qquad \text{(Equation 1)}$$

Here, f_i is the frequency of each amino acid at the specified position. Variable positions return positive SE values, whereas fully conserved ones result in an SE value of zero. In our case, the 20 positions with the highest entropy (SE > 0.65) were arbitrarily selected, allowing a maximum of 15% of the sequence to mutate. Fourteen of these positions are shared with the hotspots identified from the MSA60 dataset using a per-residue conservation threshold of ≤ 60% (see Table 6 for a detailed comparison between the two approaches).

Table 6. Number of sequences and mutational hotspots obtained after applying different consecutive filters to the initial UniRef90 MSA obtained for the target human frataxin sequence (1999 sequences) during AlphaFold prediction of wild-type frataxin: 1) sequence identity and 2) individual amino acid conservation or Shannon entropy. Common hotspots across the different datasets are shown in red (found in the four mutation pools), orange (found in at least three of the mutation pools) and green (found through either per-residue conservation or Shannon entropy approaches).

Multiple sequence alignment	Sequence identity threshold	Number of sequences	Per-residue conservation threshold	N. of mut. hotspots	Mutational hotspots
MSA60	60%	335	60%	26	93, 94, **97**, 105, **108**, **114**, 116, **118**, **120**, 121, **140**, **160**, **171**, 172, 184, **187**, **188**, **190**, **192**, 193, 194, 197, 198, **202**, **204**, 208
MSA70	70%	210	60%	11	**97**, **118**, **120**, **160**, **171**, **187**, **188**, **191**, **192**, **202**, **204**
MSA80	80%	180	60%	7	**120**, **160**, **171**, **187**, **188**, **191**, **192**
Multiple sequence alignment	**Identity threshold**	**Number of sequences**	**Shannon entropy threshold**	**N. of mut. hotspots**	**Mutational hotspots**
MSA80	80%	180	0.65	20	**97**, **108**, **114**, **118**, **120**, 129, 134, **140**, 152,**160**, **171**, 179, 180, **187**, **188**, **190**, **191**, **192**, **202**, **204**

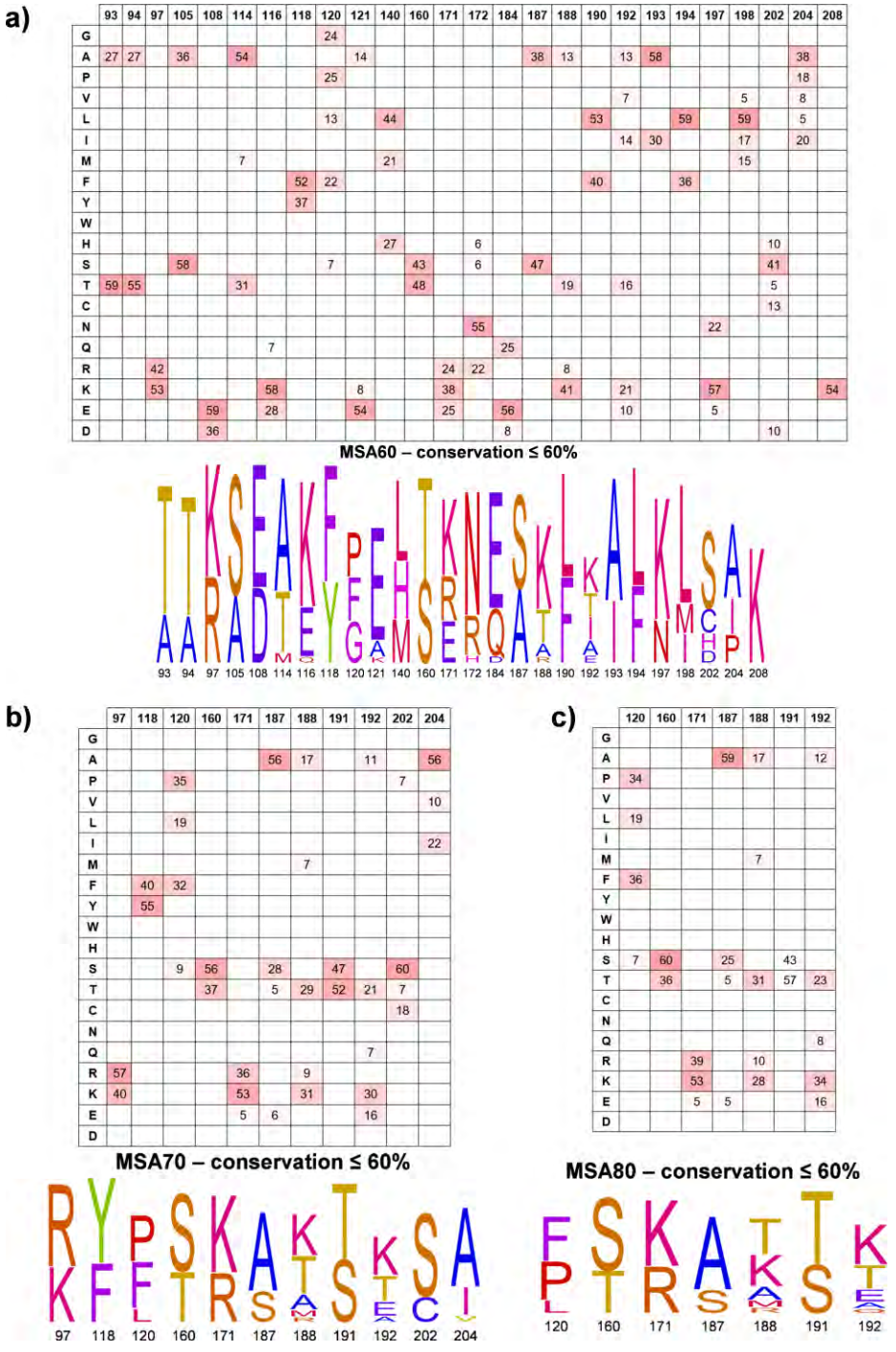


Figure 13. Per-residue conservation (*cons*) of the mutational hotspots derived from the a) MSA60 (335 sequences, conservation threshold 60%), b) MSA70 (210 sequences, conservation threshold 60%), and c) MSA80 (180 sequences, conservation threshold 60%), datasets sorted by amino acid type. For clarity, only values 5 ≤ *cons* ≤ 60% are shown. Cells are colored according to their conservation value (0%: white; 100%: red). A sequence logo representation of these mutational hotspots is shown below.

b. Phylogenetic analysis

First, a protein sequence search was performed using wild-type frataxin (residues 91-210) as a query and the Basic Local Alignment Search Tool (BLAST; https://blast.ncbi.nlm.nih.gov) on the non-redundant protein sequences (nr) database, excluding models (XM,XP), non-redundant RefSeq proteins (WP) and uncultured/environmental sample sequences with default parameters (*blastp* algorithm expect threshold 0.05, word size 5, max matches in a query range 0, matrix scoring BLOSUM62, gap costs of existence: 11 and extension 1, conditional compositional score matrix adjustment and no filters or masks). 994 sequences were found, which were filtered to ≥ 80% identity resulting in 389 sequences. A distance tree representation of these results (Blast Tree View; tree method: fast minimum evolution; max seq difference: 0.9; distance: Grishin protein; sequence label: Blast name) clearly shows evolutionary differences between amphibians/reptilians/birds and mammals characterized by mutations at positions 120 and 160 (Fig. 14).

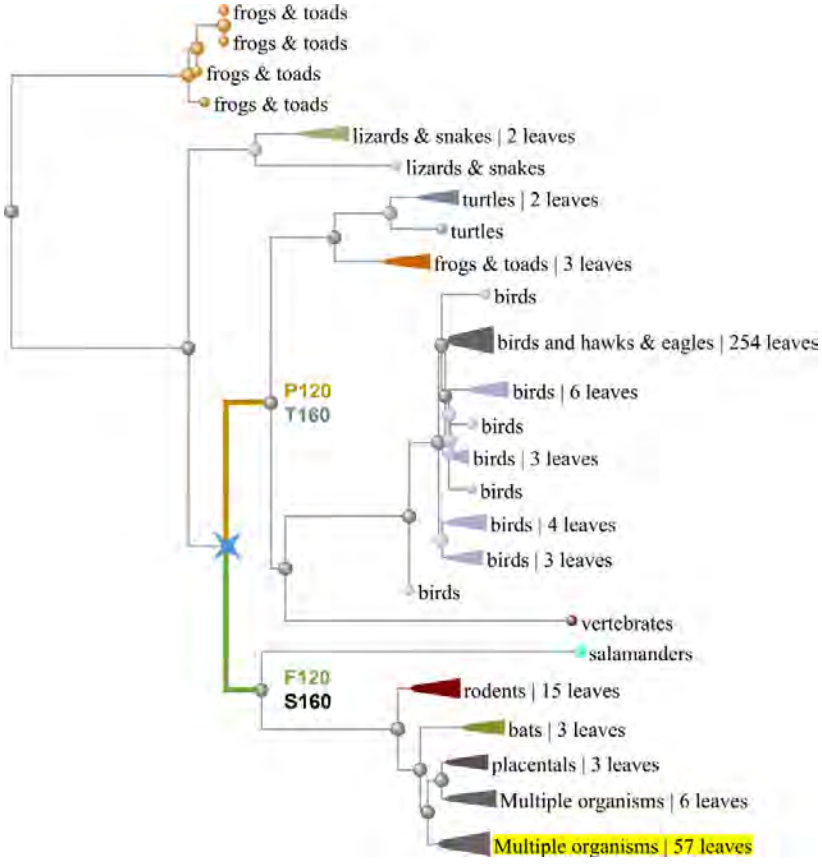


Figure 14. Blast tree view of sequences producing significant alignments to wild-type frataxin (contained in the cluster highlighted in yellow). The node characterized with mutations at positions 120 and 160 is highlighted with a blue star.

Second, the MSA80 dataset (180 sequences) was curated removing redundancies and sequences of unknown origin to yield 141 sequences, and their taxonomy annotated (kingdom, phylum, class, order, family, genus, species) using the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/tools/species-lookup) (Table 10). Sequences were clustered at positions 120, 160 and 191 and an approximately-maximum-likelihood phylogenetic tree was constructed using FastTree [257,258] and represented using Interactive Tree Of Life (iTOL; https://itol.embl.de) (Fig. 4). A clear partitioning is observed between different orders of mammals and amphibians/reptilians/birds (characterized by the F120P and S160T mutations with respect to human frataxin).

Table 10. Number of sequences and population of the clusters found within the curated MSA80 dataset (141 sequences) as a function of the amino acid identities at positions 120, 160 and 191, together with the taxonomic classes and orders of the species characterizing these clusters.

Position			Num. of. sequences	Population	Taxonomic class	Taxonomic order
120	160	191				
Phe	Ser	Thr	49	34.8%	Mammalia	Primates, carnivora, chiroptera, cetacea
Phe	Ser	Ser	6	4.3%	Mammalia	Afrosoricida, macroscelidea, proboscidea
Leu	Ser	Thr	27	19.1%	Mammalia	Rodentia, carnivora
Leu	Cys	Thr	2	1.4%	Mammalia	Rodentia
Ser	Ser	Ser	4	2.8%	Mammalia	Monotremata, diprotodontia
Ser	Ser	Thr	2	1.4%	Mammalia	Rodentia
Ser	Thr	Ser	3	2.1%	Mammalia	Didelphimorphia, diprotodontia
Ser	Thr	Thr	1	0.7%	Mammalia	Soricomorpha
Cys	Ser	Thr	5	3.5%	Mammalia	Chiroptera, cetacea
Arg	Ser	Thr	1	0.7%	Mammalia	Erinaceomorpha
Pro	Thr	Ser	40	28.4%	Aves	Passeriformes, galliformes, crocodylia, testudines
Pro	Ile	Ser	1	0.7%	Reptilia	Squamata

c. Sequence sampling with ProteinMPNN

Reference sequences were mutated simultaneously at all the selected hotspots for each mutation pool with ProteinMPNN [213] at three so-called temperatures (0.1, 0.2, and 0.3) excluding mutations to cysteine. In ProteinMPNN, the temperature is an hyperparameter that influences the amino acid compositional bias; lower temperatures have been shown to increase the frequency of charged over polar amino acids in the designs, leading to increased thermostability [213]. For each set of mutational hotspots, 30 sequences were generated (10 sequences per temperature) and

filtered selecting either the most frequently occurring ones or, more often, the sequences corresponding to the highest thermostabilization over the reference frataxin variant predicted with our AlphaFold/Rosetta-based protocol (see below).

d. Prediction of relative thermostability of designed variants

For each design, we calculated its relative thermostability ($\Delta\Delta G^{calc}$) defined as the difference in free energy of *unfolding* between each variant and the reference (either wild-type frataxin or the pathological single mutants I154F or L198R) using the approach described recently by our group [214]. Briefly, each sequence is subjected to 30 independent AlphaFold structure predictions (models 3-5, which do not use PDB templates, see Table 11 for a characterization of structure similarity) and each of the 90 generated models is scored using Rosetta's *minimize* application [111].

The average Rosetta energies of the 25 top scoring decoys of each protein represents its *folding* free energy (ΔG_f), in such a way that relative stability values are computed as $\Delta\Delta G^{calc} = -(\Delta G^{calc}_{design,f} - \Delta G^{calc}_{reference,f})$. Therefore, positive and negative $\Delta\Delta G^{calc}$ values indicate stabilization and destabilization, respectively.

For consistency, predictions on all variants were made using sequences excluding the N-terminal residues derived from the inserted affinity tags (i.e., starting at position 91) and the two last very flexible, negatively charged (Asp-Glu) C-terminal residues (i.e., finishing at position 208).

Table 11. Minimum average root-mean-square deviation of the atomic positions of the alpha carbons (RMSD-Cα), solvent-accessible surface area (SASA; probe radius 1.4 Å) and average Rosetta stability (ΔG_f^{calc}) in Rosetta energy units (REU) within the AlphaFold ensemble (top 25 decoys) for wild-type, pathological mutants and designed frataxin variants. Relative stability values ($\Delta\Delta G^{calc}$) are computed with respect to wild-type frataxin and pathological variants FXN-I154F or FXN-L198R.

Variant	RMSD-Cα (Å)	SASA (Å^2)	ΔG_f^{calc} (REU)	$\Delta\Delta G^{calc}$ (REU)[a]	$\Delta\Delta G^{calc}$ (REU)[b]	$\Delta\Delta G^{calc}$ (REU)[c]
wild-type FXN	0.1	6442 ± 29	–373.1 ± 1.1	0		
FXN-01	0.1	6480 ± 36	–381.3 ± 2.0	8.2		
FXN-02	0.1	6482 ± 28	–376.2 ± 1.2	3.1		
FXN-03	0.1	6454 ± 36	–377.2 ± 1.5	4.1		
FXN-04	0.1	6453 ± 34	–380.8 ± 1.9	7.7		
FXN-05	0.1	6456 ± 33	–385.0 ± 1.9	11.9		
FXN-06	0.1	6497 ± 32	–383.8 ± 1.7	10.6		
FXN-07	0.1	6537 ± 28	–382.1 ± 1.2	9		
FXN-08	0.1	6608 ± 23	–385.3 ± 1.5	12.2		
FXN-09	0.1	6213 ± 22	–387.4 ± 1.4	14.3		
FXN-10	0.1	6606 ± 23	–385.8 ± 1.5	12.7		

I154F	0.1	6495 ± 16	–367.0 ± 1.1	–6.2	0	
FXN-11	0.1	6514 ± 25	–376.4 ± 1.4	3.3	9.5	
FXN-12	0.1	6506 ± 16	–372.7 ± 1.8	–0.5	5.7	
FXN-13	0.1	6547 ± 19	–375.8 ± 2.6	2.7	8.8	
FXN–14	0.2	6460 ± 21	–375.9 ± 1.3	2.8	9	
FXN–15	0.1	6706 ± 21	–379.5 ± 1.2	6.4	12.5	
FXN–16	0.1	6652 ± 19	–380.3 ± 1.3	7.2	13.3	
FXN–17	0.2	6253 ± 19	–381.0 ± 0.7	7.9	14	
FXN–18	0.1	6648 ± 15	–382.1 ± 1.2	9	15.2	
L198R	0.11	6438 ± 34	–365.4 ± 1.5	–7.7		0
FXN–19	0.15	6421 ± 38	–368.0 ± 1.6	–5.1		2.6
FXN–20	0.15	6415 ± 37	–367.1 ± 2.1	–6.1		1.7
FXN–21	0.12	6277 ± 16	–375.9 ± 2.3	2.7		10.5
FXN–22	0.17	6601 ± 26	–373.5 ± 3.0	0.4		8.1
FXN–23	0.13	6496 ± 45	–377.1 ± 1.4	4		11.7
FXN–24	0.13	6623 ± 37	–375.0 ± 1.7	1.9		9.6
FXN–25	0.11	6236 ± 41	–380.5 ± 2.1	7.4		15.1
FXN–26	0.11	6561 ± 33	–377.7 ± 1.5	4.6		12.3

[a] Relative stability calculated with respect to wild-type FXN. [b] Relative stability calculated with respect to FXN-I154F. [c] Relative stability calculated with respect to FXN-L198R.

The most computationally demanding part of our ProteinMPNN/AlphaFold/Rosetta protein design protocol is the generation of the AlphaFold ensembles (models 3, 4, 5; 30 replicas; 90 predicted structures in total). Overall, designing and characterizing a mutant takes under an hour (around 40 minutes if precomputed MSA are reused), making this protocol attractive for protein engineering in the scale of hundreds of variants (Table 12).

Table 12. Timings for computational $\Delta\Delta G_{calc}$ prediction of a frataxin variant. Hardware: Intel Xeon Gold 6240R CPU 2.40GHz, Nvidia GeForce RTX 3090 GPU.

Type of calculation	Software	Hardware	Time
ProteinMPNN sampling	ProteinMPNN 1.0.1	CPU (1 core, 4 GB RAM memory) GPU (24 GB)	13 s
MSA generation	AlphaFold 2.3.0	CPU	20 m
Structure prediction	AlphaFold 2.3.0 models 3, 4, 5 30 replicas	CPU (1 core, 80 GB RAM memory) GPU (24 GB)	36 m
Minimization	Rosetta 3.13	CPU (1 core, 2 GB RAM memory)	1 m

e. Protein expression and purification

The recombinant plasmid pG-S21a (purchased from GenScript Biotech) encoding between restriction sites *NdeI* and *XhoI* for residues 91-210 of either the wild-type frataxin, the pathologic single mutants I154F and L198R or any of the 26 designed proteins bearing a N-terminal 6xHis tag, was transformed into BL21 (D3) *E. coli* competent cells, plated on Luria-Bertani (LB) broth-ampicillin agar plates and incubated overnight at 37 °C. A single colony from each plate was picked and then resuspended in an aqueous solution of 10 mL of LB broth (Lennox) and 10 μL of 50 mg/mL ampicillin followed by incubation at 37 ºC until the optical density at 600 nm (OD600) reached a value between 0.4-0.6. Induction was carried out by adding 10 μL of 0.5 M isopropyl β-D-1-thiogalactopyranoside (IPTG) and growth continued overnight at 25 °C. Initial designs FXN-01 and FXN-02 (and a non-6xHis-tagged version of wild-type frataxin) were expressed following the same protocol, but the encoded genes were equipped with a N-terminal 6xHis-GST tag. Cells were harvested by centrifugation at 5000 rpm for 20 min. The cell pellets were resuspended in 40 mL of lysis buffer (120 mM NaCl, 20 mM Tris pH 8.0, 2 mM imidazole, 1 mM protease inhibition cocktail PIC). The suspended cells were lysed by sonication (60% amplitude, 36 x 10 s bursts, with 20 s between each burst), and then clarified by centrifugation at 25000 rpm for 30 minutes at 4 °C. The soluble fraction was loaded onto 2 mL of Ni-NTA resin (Merck), cleaned with 10 mL of washing buffer (120 mM NaCl, 20 mM Tris pH 8.0) and eluted with 3.5 mL of high-imidazole buffer (120 mM NaCl, 20 mM Tris pH 8.0, 300 mM imidazole). The 6xHis-GST tag of designs FXN-01 and FXN-02 was cleaved by incubation with 2 IU of thrombin per mg of target protein at 4 °C overnight. The cleaved protein was separated from 6xHis-GST and the uncleaved protein by a second affinity purification step using the Ni-NTA resin. For subsequent imidazole removal the buffer of the pooled fractions was exchanged to 20 mM Tris pH 8.0 and 120 mM NaCl using PD-10 desalting columns packed with Sephadex G-25 resin (GE Healthcare). The elution volume was 3.5 mL. Protein purity (size: ~13.8 to 14.4 kDa) was monitored using SDS-PAGE (5–20% gradient gel) (Fig.15) and the concentrations were determined spectrophotometrically using an extinction coefficient ε in the 25440-26930 cm^{-1} M^{-1} range as determined by the ProtParam tool (https://web.expasy.org/protparam).

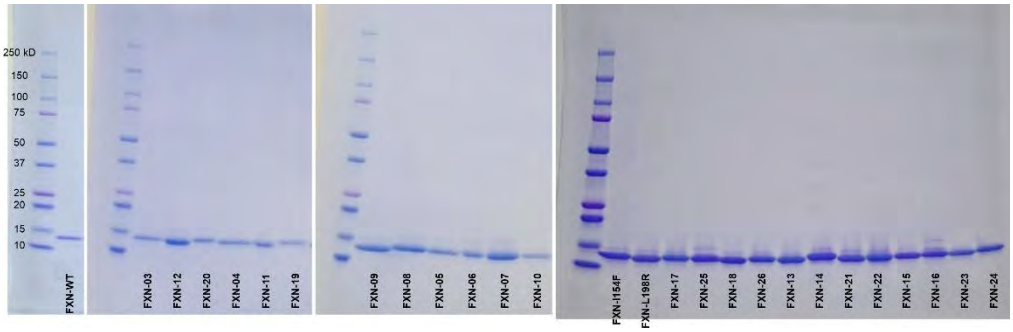


Figure 15. SDS-PAGE gels of expressed frataxin variants.

Table 13. Molecular weight (MW) , theoretical extinction coefficient at $\lambda = 280$ nm in water (E_{280}), isoelectric point (IP) and total charge at pH 7.4 (Z) (calculated with Isoelectric Point Calculator 2.0 [259]), and the differences in the number of charged amino acids (ΔN. charged) and prolines (ΔN. Pro) for wild-type and designed frataxin variants.

Variant	MW (Da)	E_{280} (M^{-1} cm^{-1})	IP	Z	ΔN. charged	ΔN. Pro
wild-type FXN	14240.7	26930	5.22	–8.1		
FXN-01	13770.17	26930	4.55	–9.1	-1	1
FXN-02	13820.23	26930	4.55	–9.1	-1	0
FXN-03	14190.64	26930	5.22	–8.1	0	1
FXN-04	14190.64	26930	5.22	–8.1	0	1
FXN-05	14193.59	26930	5.04	–9.1	-1	1
FXN-06	14206.63	26930	5.22	–8.1	0	1
FXN-07	14248.67	26930	5.08	–9.1	1	1
FXN-08	14283.81	25440	5.51	–7.1	1	1
FXN-09	14283.68	25440	5.02	–10.0	-2	2
FXN-10	14414.9	25440	5.39	–8.1	4	1

Table 14. Molecular weight (MW) , theoretical extinction coefficient at $\lambda = 280$ nm in water (E_{280}), isoelectric point (IP) and total charge at pH 7.4 (Z) (calculated with Isoelectric Point Calculator 2.0 [259]), and the differences in the number of charged amino acids (ΔN. charged) and prolines (ΔN. Pro) for pathological mutant FXN-I154F and frataxin variants.

Variant	MW (Da)	E_{280} (M^{-1} cm^{-1})	IP	Z	ΔN. charged	ΔN. Pro
FXN-I154F	14274.71	26930	5.22	–8.1		
FXN-11	14224.65	26930	5.22	–8.1	0	1
FXN-12	14224.65	26930	5.22	–8.1	0	1
FXN-13	14340.86	25440	5.22	–9.1	1	2
FXN-14	14371.83	25440	5.06	–10.1	0	2
FXN-15	14388.9	25440	5.52	–7.1	4	1
FXN-16	14368.82	25440	5.27	–8.2	4	1
FXN-17	14317.69	25440	5.02	–10.0	-2	2
FXN-18	14448.91	25440	5.39	–8.1	4	1

Table 15. Molecular weight (MW), theoretical extinction coefficient at λ = 280 nm in water (E_{280}), isoelectric point (IP) and total charge at pH 7.4 (Z) (calculated with Isoelectric Point Calculator 2.0 [259]), and the differences in the number of charged amino acids (ΔN. charged) and prolines (ΔN. Pro) for pathological mutant FXN- L198R and frataxin variants.

Variant	MW (Da)	E_{280} (M^{-1} cm^{-1})	IP	Z	ΔN. charged	ΔN. Pro
FXN-L198R	14283.72	26930	5.39	–7.1		
FXN-19	14233.66	26930	5.39	–7.1	1	1
FXN-20	14233.66	26930	5.39	–7.1	1	1
FXN-21	14197.58	25440	4.93	–10.0	0	2
FXN-22	14437.93	25440	5.24	–9.1	2	2
FXN-23	14399.89	25440	5.65	–6.1	4	1
FXN-24	14396.97	25440	5.92	–4.2	4	1
FXN-25	14326.71	25440	5.19	–9.0	-2	2
FXN-26	14457.92	25440	5.53	–7.1	4	1

f. Circular dichroism (CD) spectroscopy

The secondary structure of all frataxin variants was analyzed using CD spectroscopy. Protein samples were prepared at 5 μM concentration in buffer solution (20 mM HEPES pH 7.4, 120 mM NaCl) in a 2 mm quartz cuvette. CD spectra were recorded on a JASCO J-815 spectropolarimeter equipped with a Peltier temperature control unit at 25 °C. Data were collected from 200 to 300 nm with a 0.2 nm step size and a 2 nm bandwidth.

CD spectra were acquired first at 25 °C using freshly prepared proteins (blue lines in the plots below), and then after heating up to 90 °C and cooling back to 25 °C to test *unfolding/folding* reversibility (orange lines in Fig. 16).

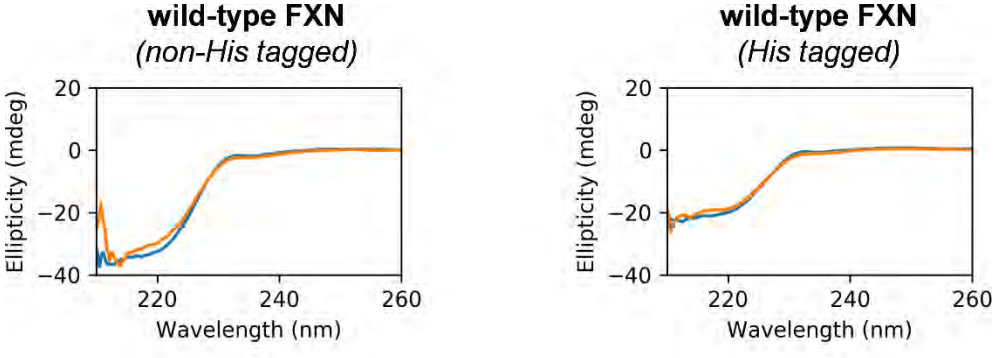


Figure 16. CD spectra of frataxin (FXN) variants (210-260 nm).

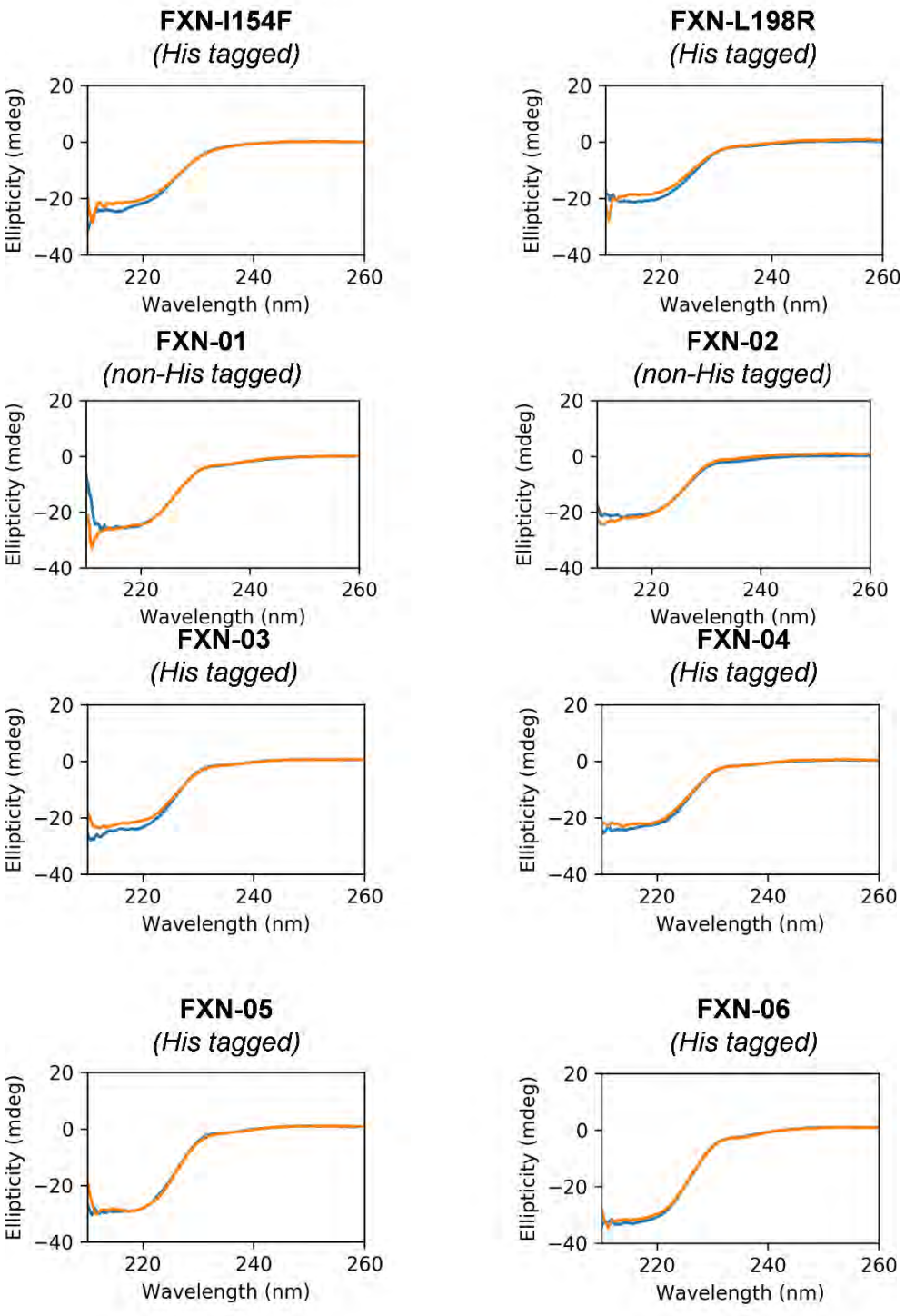


Figure 16 (cont). CD spectra of frataxin (FXN) variants (210-260 nm).

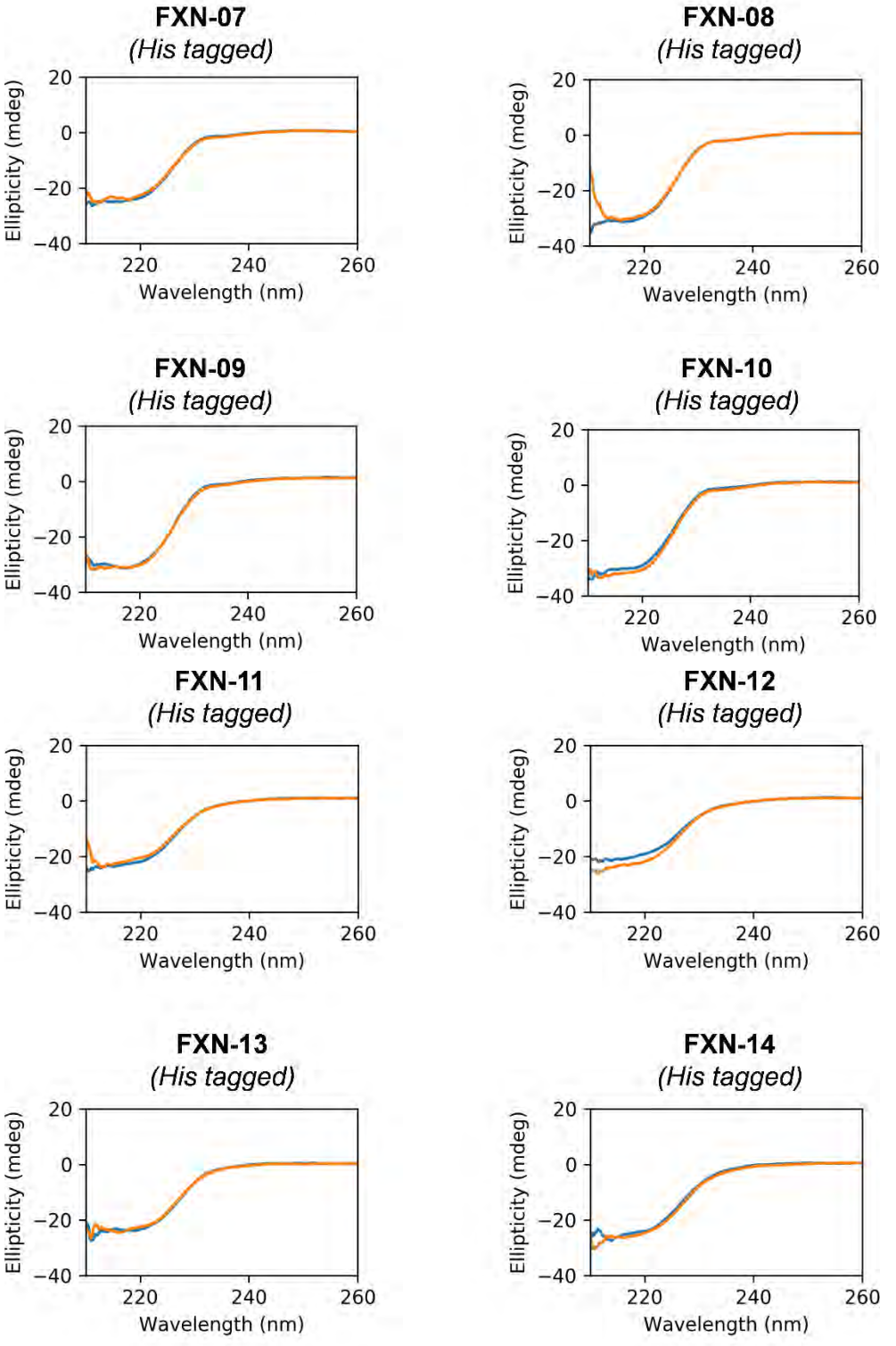


Figure 16 (cont). CD spectra of frataxin (FXN) variants (210-260 nm).

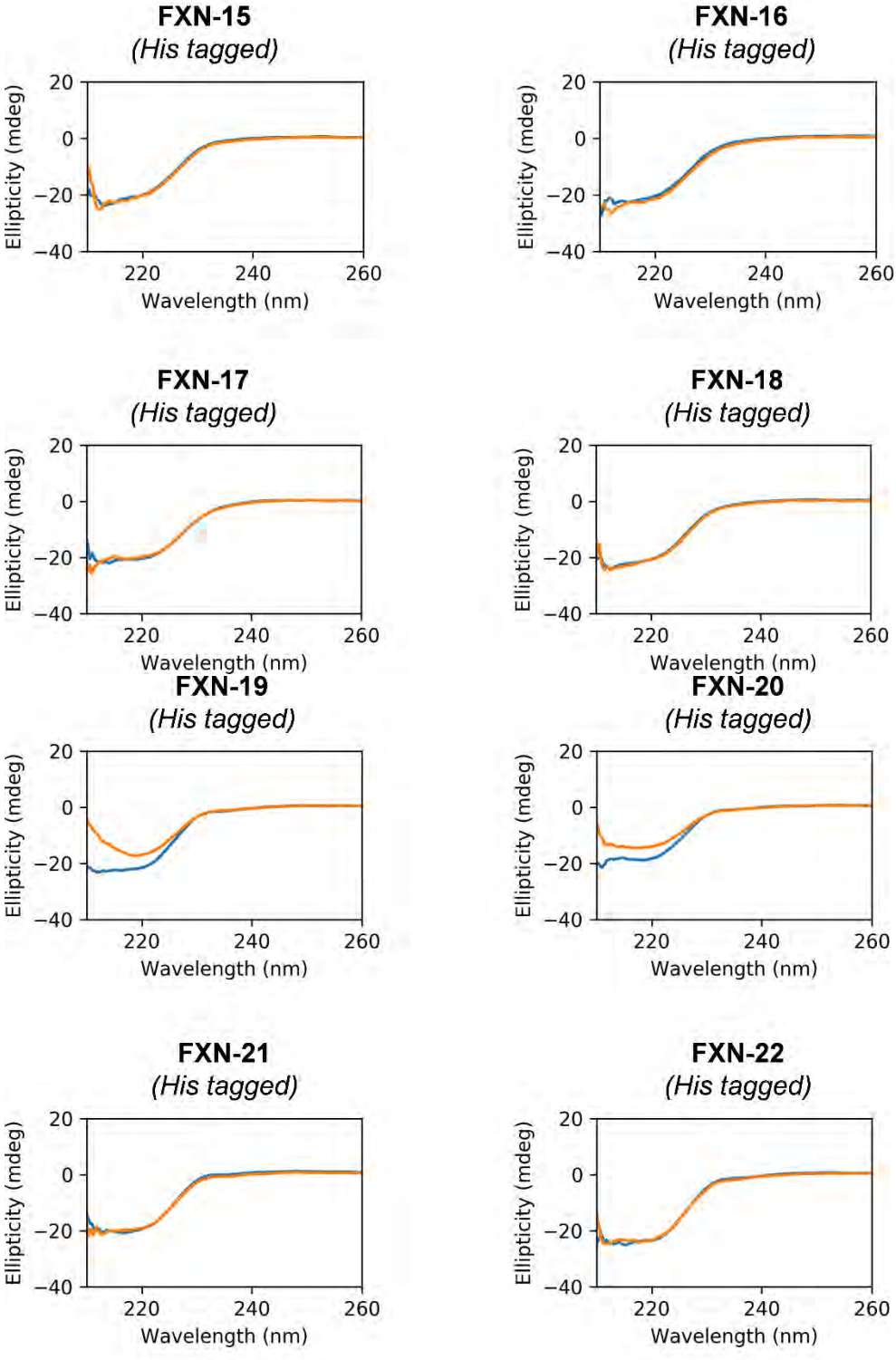


Figure 16 (cont). CD spectra of frataxin (FXN) variants (210-260 nm).

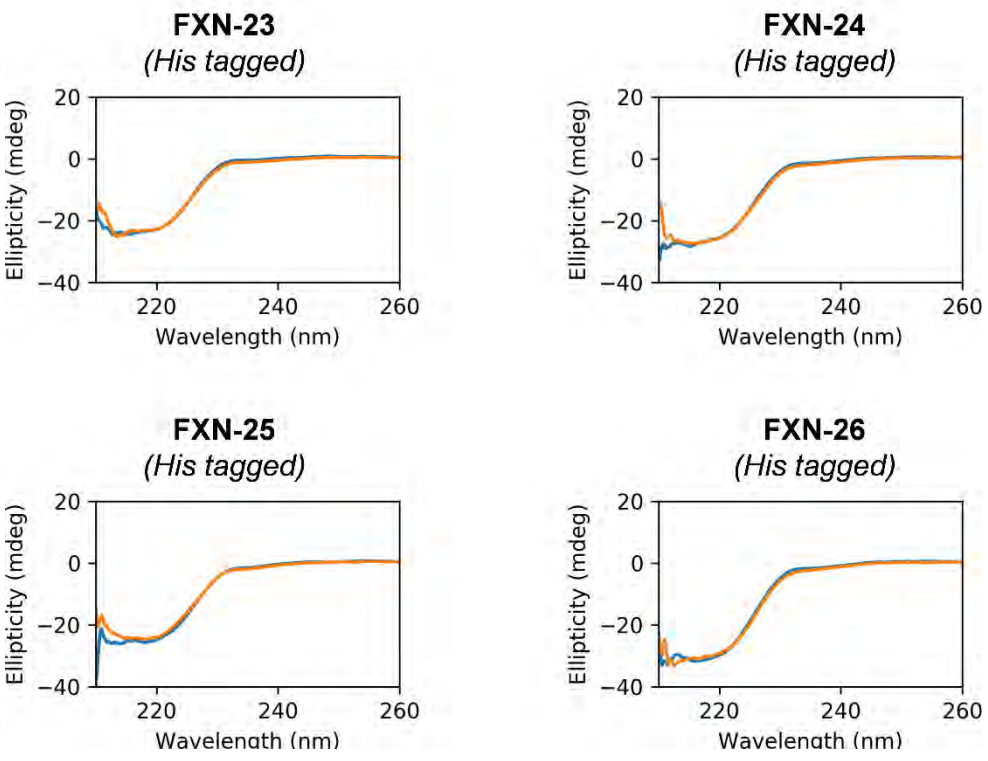


Figure 16 (cont). CD spectra of frataxin (FXN) variants (210-260 nm).

g. Melting temperature (T_m) measurement

Thermal stability was measured by monitoring the change in ellipticity at 222 nm as a function of temperature using a JASCO J-815 spectropolarimeter equipped with a Peltier temperature control unit. Protein samples were prepared at 5 μM concentration in buffer solution (20 mM HEPES pH 7.4, 120 mM NaCl) in a 2 mm quartz cuvette. The CD signal was monitored at a fixed wavelength (222 nm) while the temperature was increased from 35 °C to 90 °C at a rate of 1.2 °C/min. For the pathological FXN mutants I154F and L198R, temperature scans started at 15 °C.

For each variant, T_m values were computed by fitting ellipticity values (E) versus temperature to a two-state (*folded/unfolded*) model. The fitting was performed through a least-squares minimization of the error between the measured ellipticity values E and simulated ellipticity values E_{sim} obtained using the following equation:

$$E_{sim} = FF_{sim} \cdot (m_1 T + b_1) + (1 - FF_{sim}) \cdot (m_2 T + b_2) \qquad \text{(Equation 2)}$$

where FF_{sim} is the fraction of protein in the native (folded) state, T is the temperature, and the parameters m_1, b_1, m_2, and b_2 (fitting parameters) are the slopes and intercepts of the (linear) ellipticity in the native and denatured state, respectively (Fig. 17). Initial

guesses of m_1, b_1, m_2, and $b_2$were obtained from linear fitting of the first 20 ellipticity points (m_1, b_1) and the last 20 ellipticity points (m_2, b_2). $FF_{sim}$is defined as a parametric function derived from the equilibrium constant of folding (K_{sim}).

$$FF_{sim} = \frac{K_{sim}}{1 - K_{sim}} \quad \text{(Equation 3)}$$

In turn, the equilibrium constant is obtained from the free energy of unfolding (ΔG_{sim}):

$$K_{sim} = e^{\frac{\Delta G_{sim}}{RT}} \quad \text{(Equation 4)}$$

where $$\Delta G_{sim} = \Delta H_{sim} - T\Delta S_{sim} \quad \text{(Equation 5)}$$

and the enthalpy (ΔH_{sim}) and entropy (ΔS_{sim}) of unfolding, are calculated from guesses of the enthalpy of unfolding at the melting temperature (ΔH_{sim}), the change in specific heat capacity of unfolding (ΔC_p) and T_m, where $\Delta H_m$and T_m are the parameters to optimize and $\Delta c_p$is a constant.

$$\Delta H_{sim} = \Delta H_m + \Delta C_p(T - T_m) \quad \text{(Equation 6)}$$

$$\Delta S_{sim} = \Delta S_m + \Delta C_p \ln\frac{T}{T_m} \quad \text{(Equation 7)}$$

The initial guess values for ΔH_m and T_m were set to 48 kcal mol^{-1} and 333.15 K (60 °C), respectively. The value of ΔC_p is estimated from the number of residues of the protein [243].

Assuming a linear relationship between the number of residues (n) and the change in solvent accessible surface area of unfolding ($\Delta SASA$), the latter can be computed as:

$$\Delta SASA\ (\text{Å}^2) = -907 + 93 \cdot n \quad \text{(Equation 8)}$$

From $\Delta SASA$, ΔC_p can be estimated as:

$$\Delta C_p\ (cal\ mol^{-1}K^{-1}) = -251 + 0.19 \cdot \Delta SASA \quad \text{(Equation 9)}$$

For the 127 residues frataxin and its variants, and $\Delta SASA = 10904\ \text{Å}^2$ and $\Delta C_p =$ 1.82 kcal mol^{-1} K^{-1}.

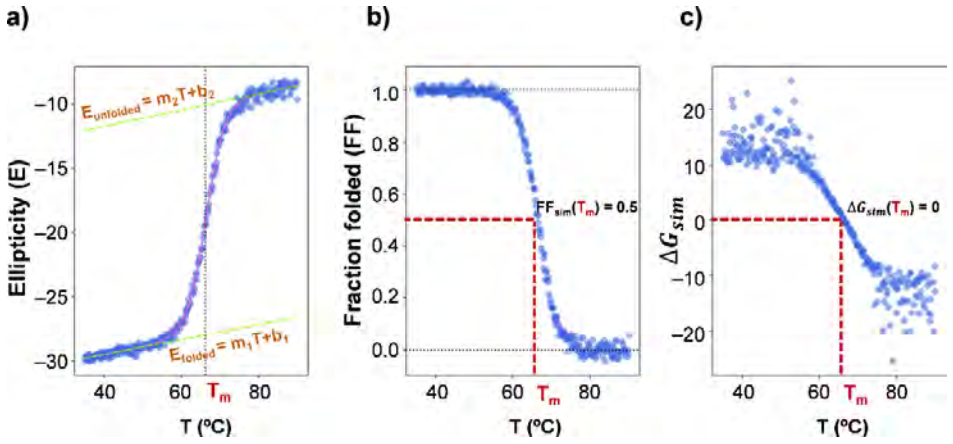


Figure 17. a) Evolution of the ellipticity as a function of temperature along thermal unfolding. Blue dots: measured E(T) values; magenta line: fitted E_{sim}(T). b) Evolution of the fraction of folded protein as a function of temperature along thermal unfolding. c) Evolution of the free energy of unfolding as a function of temperature along thermal unfolding.

In Fig. 15, T_m curves were represented considering for each variant the *heating* and *cooling* ellipticity values and linearly rescaling them to [0,1] range according to:

$$E_{norm} = \frac{E - E_{min}}{E_{max} - E_{min}} \qquad \text{(Equation 10)}$$

where E is the measured ellipticity, E_{min} the absolute minimum of E values considering the *heating* and *cooling* curves, and E_{max} the absolute maximum under the same conditions.

Folding reversibility percentages were computed from the fitted ellipticity values (Eq. 2) according to the following equation:

$$\text{reversibility (\%)} = \frac{E_{90}^{cool} - E_{35}^{cool}}{E_{90}^{heat} - E_{35}^{heat}} x\ 100 \qquad \text{(Equation 11)}$$

where E_{90}^{cool}, E_{35}^{cool}, E_{90}^{heat}, and E_{35}^{heat} are the ellipticities measured at 90 °C and 35 °C during *heating* (i.e., forward thermal unfolding) and *cooling* (reverse thermal folding).

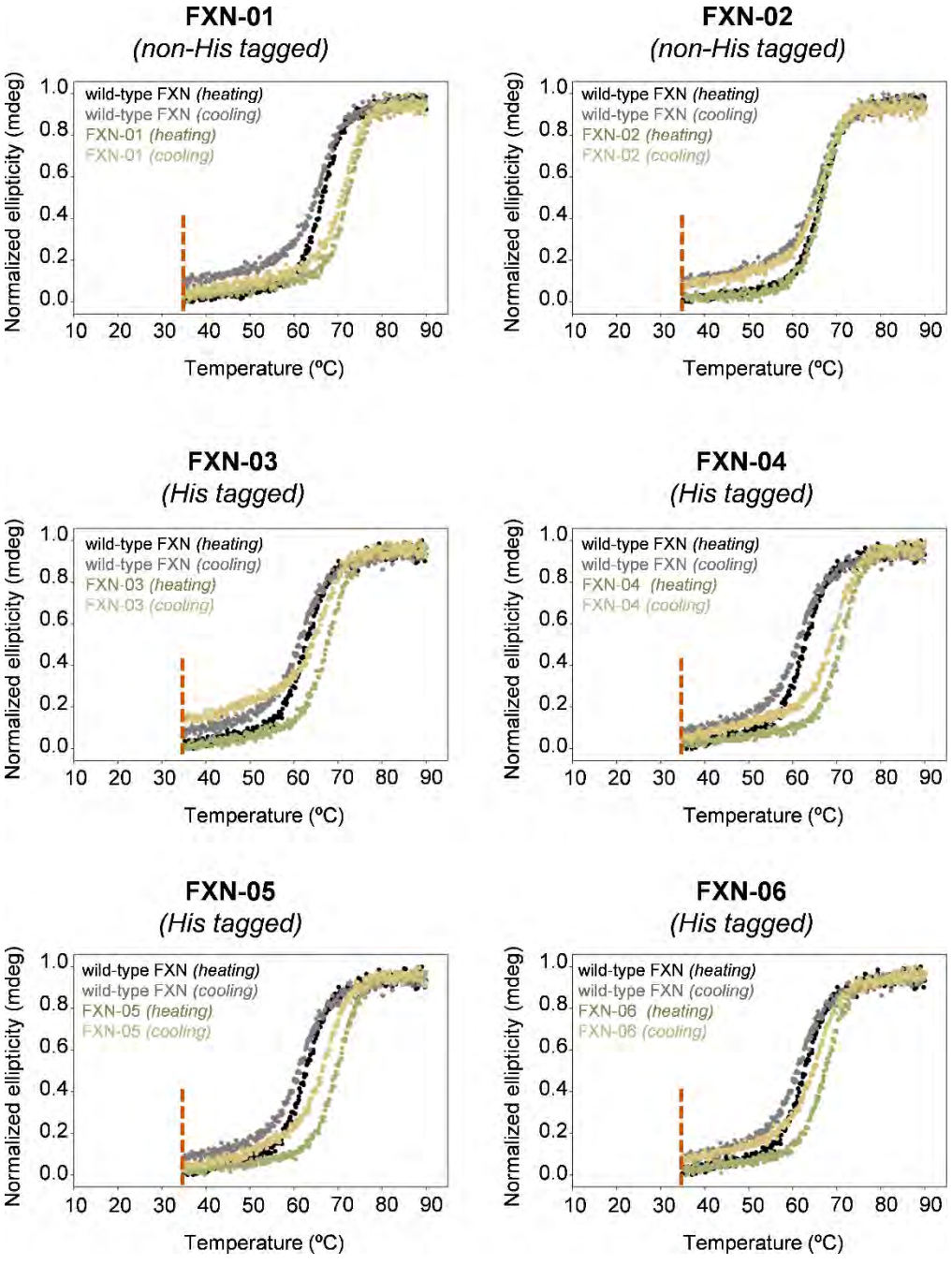


Figure 18. Thermal unfolding of frataxin (FXN) variants measured by circular dichroism (CD) spectroscopy. Reversibility is always measured at 35 °C for fair comparison across different variants (dashed red line).

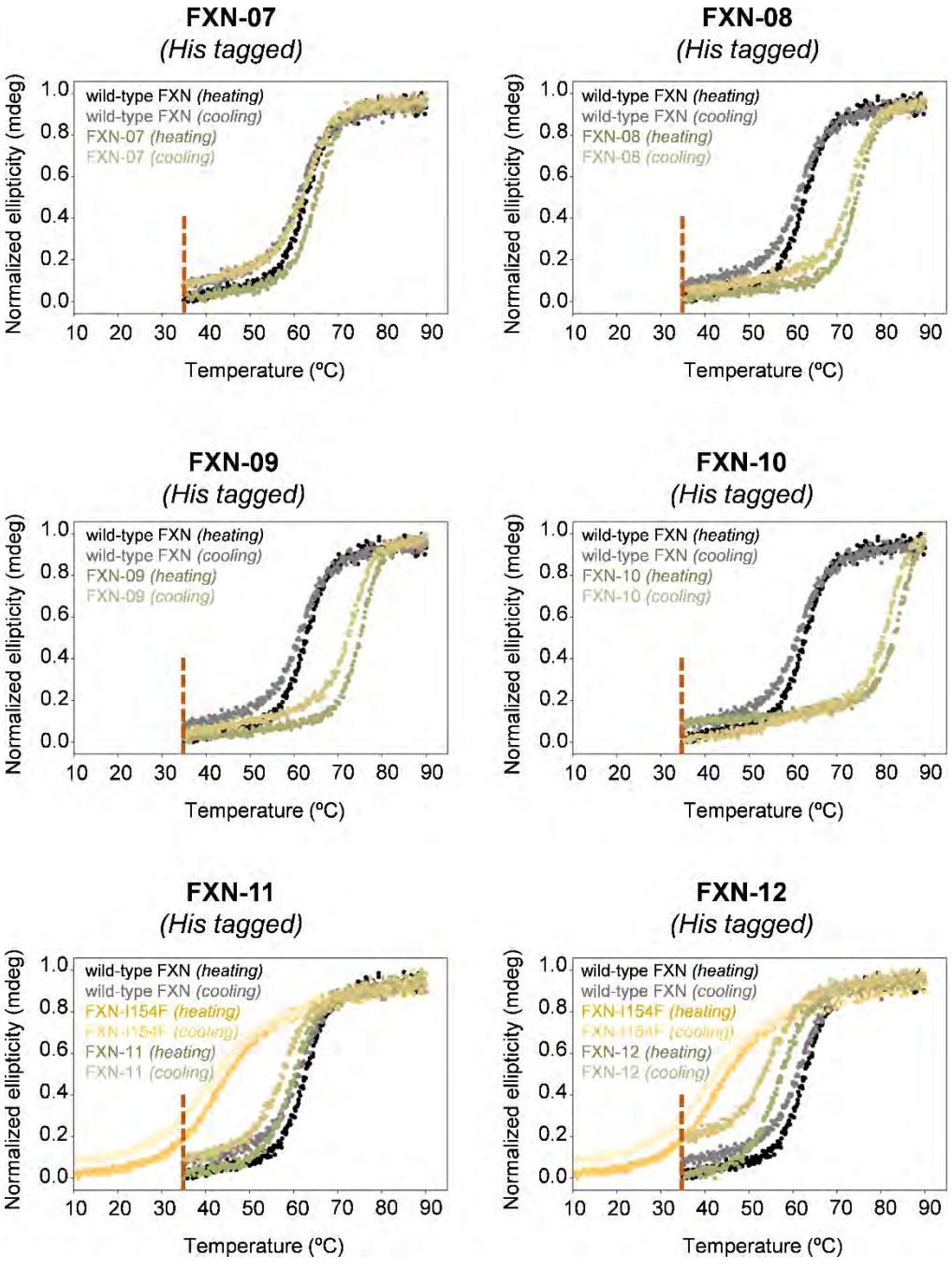


Figure 18 (cont). Thermal unfolding of frataxin (FXN) variants measured by circular dichroism (CD) spectroscopy. Reversibility is always measured at 35 °C for fair comparison across different variants (dashed red line).

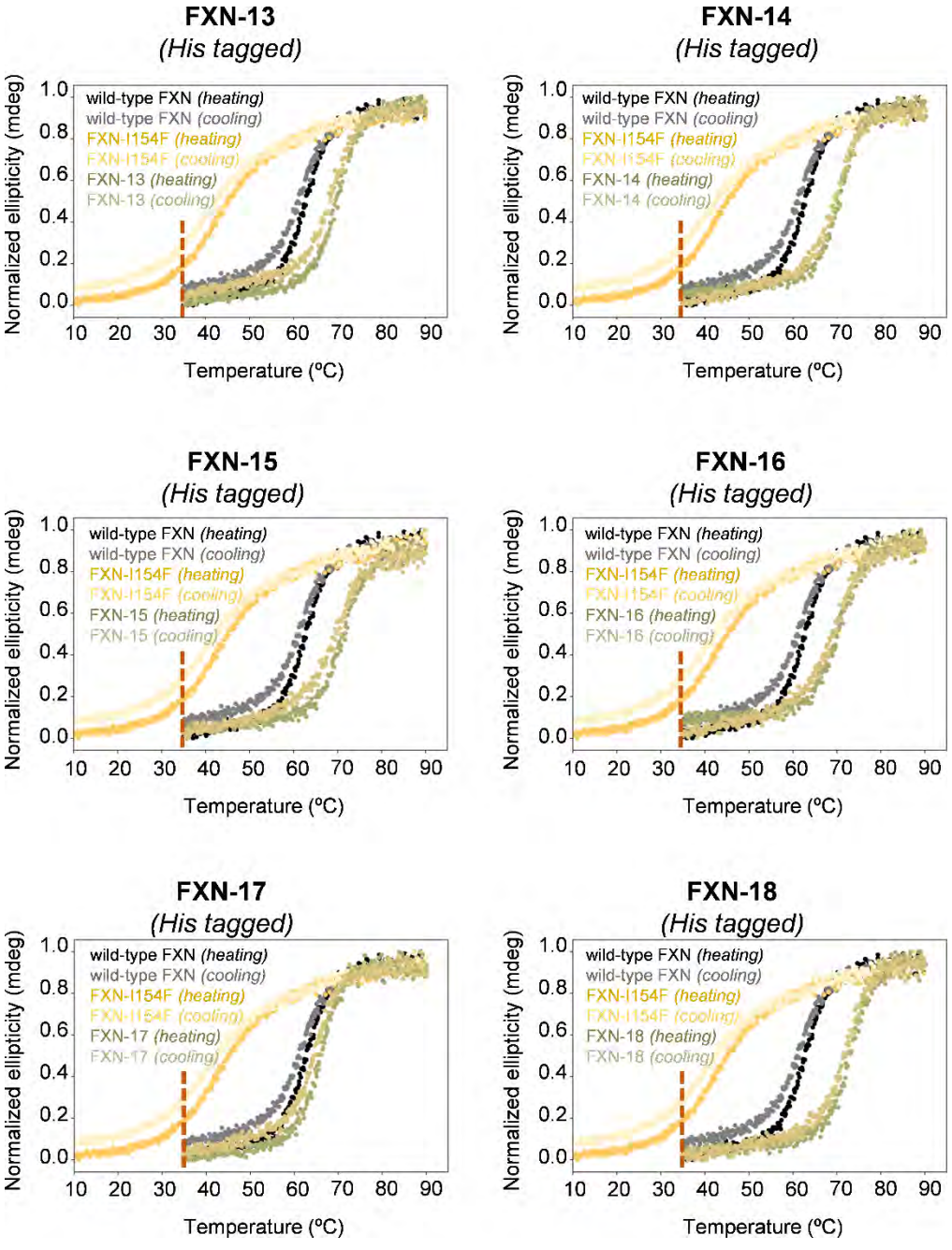


Figure 18 (cont). Thermal unfolding of frataxin (FXN) variants measured by circular dichroism (CD) spectroscopy. Reversibility is always measured at 35 °C for fair comparison across different variants (dashed red line).

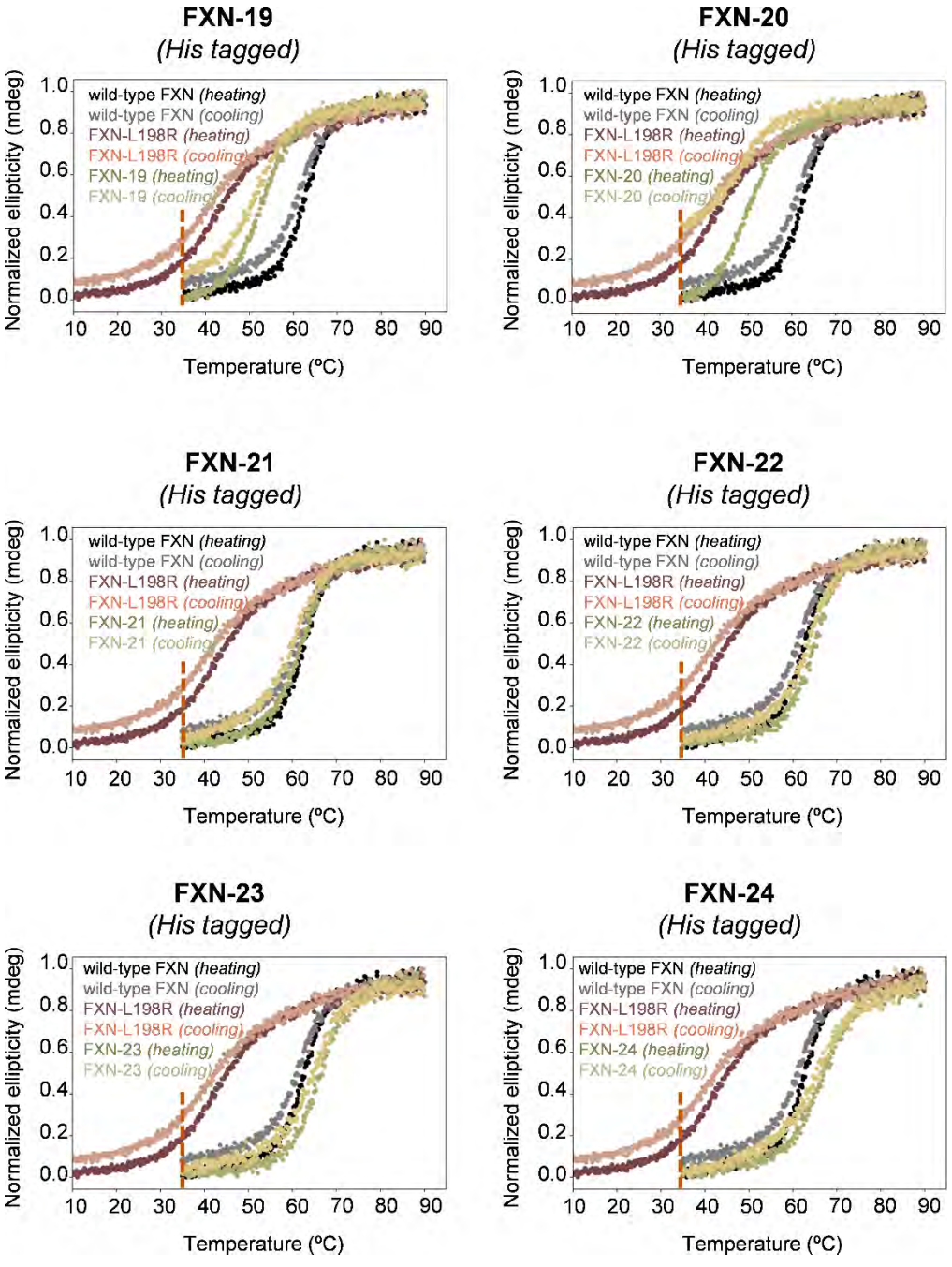


Figure 18 (cont). Thermal unfolding of frataxin (FXN) variants measured by circular dichroism (CD) spectroscopy. Reversibility is always measured at 35 °C for fair comparison across different variants (dashed red line).

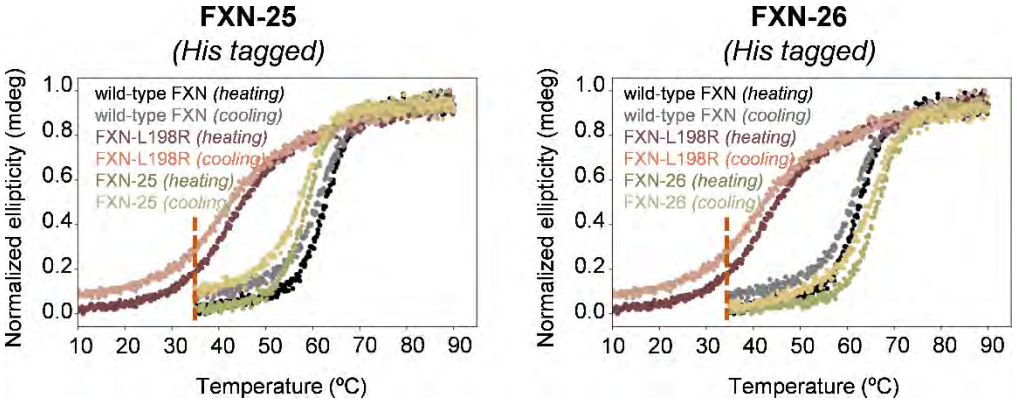


Figure 18 (cont). Thermal unfolding of frataxin (FXN) variants measured by circular dichroism (CD) spectroscopy. Reversibility is always measured at 35 °C for fair comparison across different variants (dashed red line).

h. Stability curve determination

Chemical denaturation experiments were performed to obtain the Gibbs-Helmholtz curves (ΔG_u vs T) of wild-type frataxin and FXN-03, FXN-08, and FXN-10. Protein samples were prepared at 1 μM concentration in buffer solution (20 mM HEPES pH 7.4, 120 mM NaCl) in a 10 mm quartz cuvette. Guanidinium hydrochloride (GdnHCl) was used as the denaturant. The concentration of GdnHCl was gradually increased from 0 to 4 M while monitoring the change in ellipticity in the 220-230 nm range using a JASCO J-815 spectropolarimeter equipped with a Peltier temperature control unit. ΔG_u values derived from chemical denaturation experiments were considered at five temperatures for wild-type frataxin (10, 20, 30, 40 and 50 °C), three for FXN-10 (10, 30, 50 °C), and two for FXN-08 and FXN-03 (10 and 30 °C). Ellipticity values (E) vs. denaturant concentration at each temperature were fitted using a least squares algorithm to a two-state model following Eq. 12.

$$E = FF \cdot (m_1[D] + b_1) + (1 - FF) \cdot (m_2[D] + b_2) \qquad \text{(Equation 12)}$$

where FF is the fraction of protein in the native (folded) state, $[D]$ the denaturant concentration, and parameters m_1, b_1, m_2, and b_2 the slopes and intercepts of the (linear) ellipticity in the native and denatured state, respectively. FF can be expressed as a function of the free energy of unfolding in the presence of the denaturant ΔG_{uD}, $FF = 1/(1 + e^{-\frac{\Delta GuD}{RT}})$, which is itself a function of the denaturant concentration $\Delta G_{uD}([D]) = \Delta G_u - m[D]$, where m is the denaturant slope and ΔG_u the free energy of unfolding at the given temperature in the absence of denaturant. Ellipticity data were simultaneously fitted optimizing i) m_1, b_1, m_2, and b_2, for each variant and temperature and ii) a common denaturant slope $m$using an in-house Python3 script, obtaining for each variant and temperature the extrapolated ΔG_u value.

Additional ΔG_u vs T datapoints were extracted from the CD spectra of thermal denaturation (i.e., T_m measurement) in a symmetric 6 °C range around the melting temperature (34 datapoints for wild-type frataxin, 31 for FXN-03, 32 for FXN-08, and 30 for FXN-10), where significant populations of the folded and unfolded states thus allowing the calculation of accurate ΔG_u values.

Finally, to obtain the stability curves the ΔG_u values were fitted to the Gibbs-Helmholtz equation (Eq. 13) with a least squares algorithm optimizing the ΔH_m ΔC_p and T_m parameters. ΔH_m is the enthalpy of unfolding at the melting temperature and ΔC_p the change in specific heat between denatured and native state.

$$\Delta G_u(T) = \Delta G_m - T\frac{\Delta H_m}{T_m} + \Delta C_p\left[T - T_m - T\ln(\frac{T}{T_m})\right] \qquad \text{(Equation 13)}$$

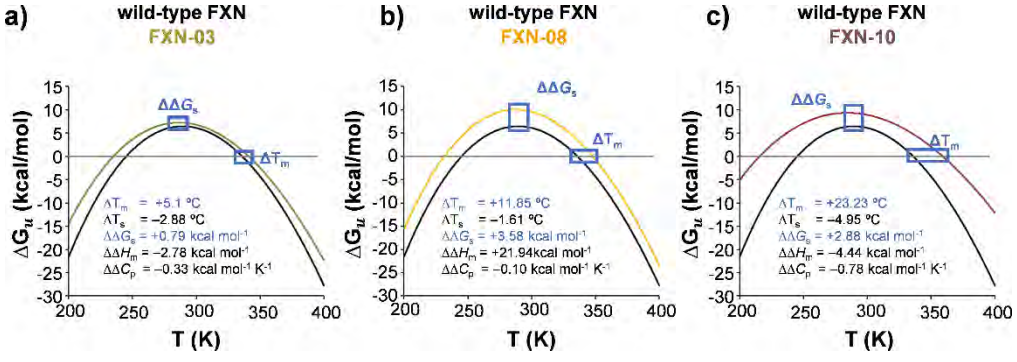


Figure 19. Fitted stability curves for wild-type frataxin (black), FXN-03 (green), FXN-08 (yellow) and FXN-10 (red), and changes in the thermodynamic parameters derived from them with respect to those of the wild-type. The differences between the *thermodynamic stabilization* ($\Delta\Delta G_s$) and *thermostabilization* (ΔT_m) measured for each designed variant with respect to the wild-type are highlighted as blue boxes. Note that FXN-08 is the most *thermodynamically stable* variant, while FXN-10 is the most *thermostable* one.

Under the assumption that the stability curves are parallel in the T_m region, approximated changes in stabilization for related variants ($\Delta\Delta G^{appr}$) can be obtained from a simple linear relationship using the ΔT_m, ΔH_m and T_m values derived from the stability curve of the wild-type (Table 16). This approximation assumes a common first derivative of $\Delta G_u(T)$ at the melting temperatures ($\Delta H_m/T_m = \Delta S_m$).

$$\Delta\Delta G^{appr} = \frac{\Delta H_m}{T_m}\Delta T_m \qquad \text{(Equation 14)}$$

Table 16. Experimentally measured *unfolding/folding* temperature (T_m), differences in *unfolding* T_m (ΔT_m), reversibility (rev.), and approximated thermostability changes ($\Delta\Delta G^{appr}$) for wild-type, pathological mutants and designed frataxin variants. $\Delta\Delta G^{appr}$ values were derived from ΔT_m values with respect to wild-type frataxin using Eq. 14.

Variant	$T_{m,u}$ (°C)[a]	$T_{m,f}$ (°C)[b]	rev. (%)	ΔT_m (°C)[a,c]	ΔT_m (°C)[a,d]	ΔT_m (°C)[a,e]	$\Delta\Delta G^{appr}$ (kcal mol^{-1})
wild-type FXN	62.8	61.4	92.9	0			0
FXN-01	72.6	71.8	97.5	+6.1[f]			1.7
FXN-02	66.9	66.5	93.1	+0.4[f]			0.1
FXN-03	67.9	66.5	86.3	5.1			1.4
FXN-04	71.3	69.9	96.6	8.6			2.3
FXN-05	70	67.6	97	7.3			2.0
FXN-06	67.6	65.5	94.7	4.8			1.3
FXN-07	65.2	63	93.6	2.5			0.7
FXN-08	74.6	73.5	96.4	11.9			3.2
FXN-09	75.5	73.3	99.2	12.7			3.4
FXN-10	86	82.7	111.8[g]	23.3			6.3
FXN-I154F	49.9	47.3	95.1	–12.9	0		–3.5
FXN-11	59.7	57.6	93	–3.1	9.9		–0.8
FXN-12	56.6	54.6	82.8	–6.2	6.7		–1.7
FXN-13	70.2	69.2	94.3	7.4	20.3		2.0
FXN-14	71	70.5	102.4[g]	8.2	21.1		2.2
FXN-15	71	69.9	99.3	8.2	21.1		2.2
FXN-16	70.4	69.5	106.9[g]	7.6	20.5		2.1
FXN-17	66.3	65.1	98.1	3.5	16.4		1.0
FXN-18	73.3	72.4	99.7	10.5	23.4		2.8
FXN-L198R	42.2	40	89	–20.5		0	–5.6
FXN-19	51.5	51.2	87.7	–11.2		9.3	–3.0
FXN-20	46.3	42.4	62.8	–16.5		4.1	–4.5
FXN-21	62.1	60.9	96.7	–0.7		19.8	–0.2
FXN-22	65.1	64	97.6	2.4		22.9	0.6
FXN-23	65.9	65	94.9	3.1		23.7	0.9
FXN-24	66.1	65	101.7[g]	3.4		23.9	0.9
FXN-25	58.8	58.3	91.1	–4.0		16.6	–1.18
FXN-26	66.3	65.5	97.9	3.5		24.1	1.0

[a] Measured during forward thermal *unfolding (u)* (i.e. heating). [b] Measured during reverse thermal *folding (f)* (i.e. cooling). [c] Calculated with respect to wild-type FXN. [d] Calculated with respect to FXN-I154F. [e] Calculated with respect to FXN-L198R. [f] The difference in melting temperature (ΔT_m) for these variants is calculated with respect to non-6xHis-tagged wild-type frataxin ($T_{m,u}$ = 66.5; $T_{m,f}$ = 65.9). [g] Reversibility values higher than 100% are due to numerical errors in the calculation of fitted ellipticities, and complete reversibility is assumed.

i. Proteolytic resistance assay

- **Mass spectrometry**

Wild-type frataxin and the superstable FXN-10 variant (three samples: R1, R2 and R3) were incubated with trypsin at 37 °C, at an enzyme:protein ratio of 1:100 for 5, 10, 20 and 60 min. The resulting peptides were desalted and resuspended in 0.1% formic acid using C18 stage tips (Millipore). Samples were analyzed in a hybrid trapped ion mobility spectrometry – quadrupole time of flight mass spectrometer (timsTOF Pro with PASEF, Bruker Daltonics) coupled online to a nanoElute liquid chromatograph (Bruker). Each sample (100 ng approx.) was directly loaded in a 15 cm Bruker nanoelute FIFTEEN C18 analytical column (Bruker) and resolved at 400 nl/min with a 100 min gradient. The column was heated to 50 °C using an oven.

- **Data analysis**

Database searching was performed using MASCOT 2.2.07 (Matrixscience) through Proteome Discoverer 1.4 (Thermo) against a Uniprot/Swissprot database consisting of *Homo sapiens* entries, including the frataxin variants of interest. The following parameters were adopted for the searches: carbamidomethylation of cysteines (C) as fixed modification and oxidation of methionines (M) as variable modifications, 20 ppm of peptide mass tolerance, 0.05 Da fragment mass tolerance and up to 2 missed cleavages. Spectral counts, that is, the number of spectra matching to a certain protein in each sample, were used for the assessment of frataxin cleavage levels.

Table 17. Protein stability assay by trypsin degradation and mass spectrometry analysis shows decreased proteolysis for the engineered variant FXN-10 compared to wild-type frataxin. *Spectral counts* refer to peptide-spectrum match (PSM), i.e. peptides identified after protein digestion. The data shown is the average over three replicas followed by its standard deviation.

Variant	Spectral counts			
	5 min	10 min	20 min	1h
wild-type FXN	109.3 ± 335.3	200 ± 46.6	224.0 ± 31.0	271.3 ± 45.0
FXN-10	31.7 ± 6.8	37.7 ± 10.1	51.0 ± 2.6	74.3 ± 13.6

j. Binding of frataxin variants to Zn^{2+}/ppIX and FeS assembly complex

- **Expression and purification of $^{15}N/^{13}C$ labelled proteins**

The recombinant plasmid pG-S21a (purchased from GenScript Biotech) encoding between restriction sites *NdeI* and *XhoI* for residues 91-210 of FXN-10 bearing a N-terminal 6xHis tag, was transformed into BL21 (D3) *E. coli* competent cells, plated on Luria-Bertani (LB) broth-ampicillin agar plates and incubated overnight at 37 °C. A

single colony from each plate was picked and then resuspended in an aqueous solution of 10 mL of LB broth (Lennox); this process was repeated in triplicate. Subsequently, the suspensions were incubated at 37 °C for 6-8 hours. Next, 200 μL of each of these preinocula were transferred to 200 mL of minimal M9 media (24 mM Na_2HPO_4, 11 mM KH_2PO_4, 4.3 mM NaCl, 2mM $MgSO_4$, 0.1 mM $CaCl_2$, 200 mg/L thiamine, 10 mg/L ampicillin, 10 mg/L biotin, and 50 mg/L ampicillin) supplemented with trace metal mix composition[260], along with 1 g/L $^{15}NH_4Cl$ (≥98 atom % ^{15}N, Sigma-Aldrich) for uniform ^{15}N labelling and 2 g/L D-glucose. These cultures were then incubated overnight at 37 °C. Subsequently, these precultures were added to 1.5 L of minimal M9 media and incubated again at 37 °C until the OD reached values between 0.4–0.6. Induction was carried out by adding 10 μL of 0.5 M isopropyl β-D-1-thiogalactopyranoside (IPTG) and growth continued overnight at 25 °C. The subsequent cell harvest and purification steps followed the protocol described in previous sections.

Uniform $^{15}N/^{13}C$ labelled human frataxin was prepared following the experimental procedure described previously by the Precision Medicine and Metabolism Lab at CIC bioGUNE [200] and used as a reference. Briefly, the recombinant plasmid pGS21a (purchased from GeneScript Biotech) encoding for residues D91 to A210 of wild- type human frataxin fused via a thrombin cleavage site to a N-terminal 6xHis-GST tag was transformed into *E. coli* BL21(DE3) competent cells, and expressed under control of the T7 promoter at 30 °C for about 18 h in minimal M9 media (24 mM Na_2HPO_4, 11 mM KH_2PO_4, 4.3 mM NaCl, 2mM $MgSO_4$, 0.1mM $CaCl_2$, 200 mg/L thiamine hydrochloride, 100 mg/L kanamycin) supplemented with trace metal mix composition [260], along with 1 g/L $^{15}NH_4Cl$ (≥98 atom % 15N, Sigma-Aldrich) for uniform ^{15}N labelling and 3 g/L D-glucose or, optionally, for the production of samples bearing uniform ^{13}C labelling with 2 g/L D-glucose-$^{13}C_6$ (≥99 atom % ^{13}C, Sigma-Aldrich) as the only nitrogen and carbon source, respectively.

For protein purification, the cell pellet was thawed and resuspended in lysis buffer (20 mM Tris pH 8, 120 mM NaCl, 2 mM imidazole, 10% glycerol, 0.1% triton-X100, and one tablet of cOmplete™EDTA-free protease inhibitor cocktail), and incubated for about 30 min on ice prior completing cell lysis by ultrasonication on ice for a total time of 2.5 min (Vibracell VC505 sonicator, 14 mm diameter probe). The cell debris was removed by ultracentrifugation at 60 kg. The soluble fraction was passed through a 0.22 μm filter and loaded onto 3 mL cobalt-charged NTA agarose column equilibrated with lysis buffer before extensive cleaning with washing buffer (20 mM Tris pH 8, 120 mM NaCl, 2 mM imidazole, and 10% glycerol) for at least 20 column volumes before elution of the protein from the column in one step in presence of 200 mM imidazole in the washing buffer. For subsequent imidazole removal the pooled fractions were desalted using a Sephadex™G-25 column equilibrated with 20 mM Tris pH 8, 120 mM NaCl, and 10% glycerol. Subsequently, the 6xHis-GST tag was cleaved by incubation

with thrombin at room temperature for 6 h applying 2 units of the enzyme per mg of target protein. Traces for uncleaved protein were removed by a second affinity chromatography step before loading the sample onto a Sd75/16/600 gel filtration column for final polishing. The apparent elution volume was 76.5 mL. The purity of the obtained protein (13.8 kDa) was monitored using SDS-PAGE (5%-20% gradient gel) and the concentration was determined spectrophotometrically (using an extinction coefficient ε of 26930 cm^{-1} M^{-1}) and by a Bradford assay.

- **Expression and purification of FeS assembly complex**

The FeS assembly complex prepared was composed of the iron-cluster assembly enzyme (ISCU, residues His36 to C-terminal Lys167, 15.0 kDa), human mitochondrial cysteine desulfurase (NFS1, residues Leu56 to C-terminal His457, 47.8 kDa, theoretical pI = 6.7), and human LYR motif-containing protein 4 (LYRM4, residue Arg6 to C-terminal Thr91, 10.6 kDa), all required for the *de novo* synthesis of iron-sulfur (Fe-S) clusters within mitochondria responsible for maturation of both, mitochondrial and cytoplasmic [2Fe-2S] and [4Fe-4S] proteins, respectively. The construct of ISCU was cloned into pET28a vector (purchased from GenScript Biotech), whereas LYRM4 and NFS1 were cloned into pCDFDuet™-1 plasmid (purchased from GenScript Biotech) containing two multiple cloning sites. Co-transformed BL21 *E. coli* cells for in vivo formation of the FeS assembly complex were obtained from sequential chemical transformation. Given the high affinity for the complex formation only the construct for cysteine desulfurase NFS1 was bearing a N-terminal His-tag (MGSSHHHHHH-, followed by a small linker and a TEV cleavage site SQDPNSSSG-ENLYFQ↓G-). Thus, during the initial affinity step both other components were captured and co-purified further when bound to that bait protein (providing a secondary affinity support). Under applied buffer conditions (50 mM HEPES Na at pH 8.0, 120 mM NaCl), the hydrodynamic radii of the complex (about 73.4 kDa) observed by gel filtration resembled the size of a di-trimeric complex (as observed in the orthorhombic unit cell of the complex crystal structure (PDB ID 6NZU). The calculated extinction coefficient for the di-trimer complex (2 x 655 residues, 146.2 kDa) is 2 x 57.760 M^{-1} cm^{-1}. Lysis of the cells were performed by a freeze and thaw cycle at -80 °C followed by sonication in the absence of any detergent in the buffer. Due to the basic character of LYR motif-containing protein 4 (LYRM4, pI = 10.7) and the iron-sulfur cluster assembly enzyme (ISCU, pI = 8.8), a digestion step with benzonase at 25 °C was mandatory to get rid of bound RNA/DNA contamination that would otherwise strongly interfere during the subsequent purification. After an initial affinity capture step using NTA resin charged with cobalt the concentrated elute was polished by gel filtration using a Sd200/16/600 column (preparative grade).

- **Sample preparation**

Frataxin variants were transferred into 50 mM HEPES Na pH 8.0, 120 mM NaCl, and 100 μM TCEP using PD-10 desalting columns packed with Sephadex G-25 resin (GE Healthcare) followed by concentration to about 300 μM using a 12 mL Vivaspin device (cutoff 10 kDa). Samples for NMR supplemented with 7% D_2O were inserted into regular 5 mm tubes. The final protein concentrations were determined spectrophotometrically at 280 nm, using extinction coefficients of 26930 M^{-1} cm^{-1} and 25440 M^{-1} cm^{-1} for the wild-type protein and the FXN-10 mutant, respectively.

Protoporphyrin IX (ppIX) was purchased from Frontier Scientific (Product Number: P562-9; CAS Number: 553-12-8).

- **NMR data acquisition**

For monitoring titration with increasing amounts of either $Zn(AcO)_2$/ppIX or FeS assembly complex, a series of 2D sofast HMQC experiments [261] with band-selective ^{H}N excitation for fast T_1 recovery were acquired at 298 K on a 600 MHz Bruker Avance III and a Bruker Avance III 800 MHz spectrometer equipped with a 5 mm TXI probe and a 5 mm TCI cryoprobe, respectively. The chemical shifts of the proton-bearing carbons and their attached protons of human wild-type frataxin and the FXN-10 variant were derived from 2D ^{13}C-HSQC spectra recorded with $^{1}J_{CH}$ matched adiabatic full passage (AFP) pulses and echo/anti-echo gradients for coherence selection. All experiments were acquired at 298 K. ^{1}H chemical shifts were directly referenced to added DSS (2,2-dimethyl-2-silapentane-5-sulphonic acid) and ^{15}N chemical shifts were referenced indirectly relative to ^{1}H using IUPAC ratios (https://bmrb.io/ref info/cshift.shtm). All NMR data were processed with NMRpipe [262] and analyzed with NMRFAM-Sparky [263].

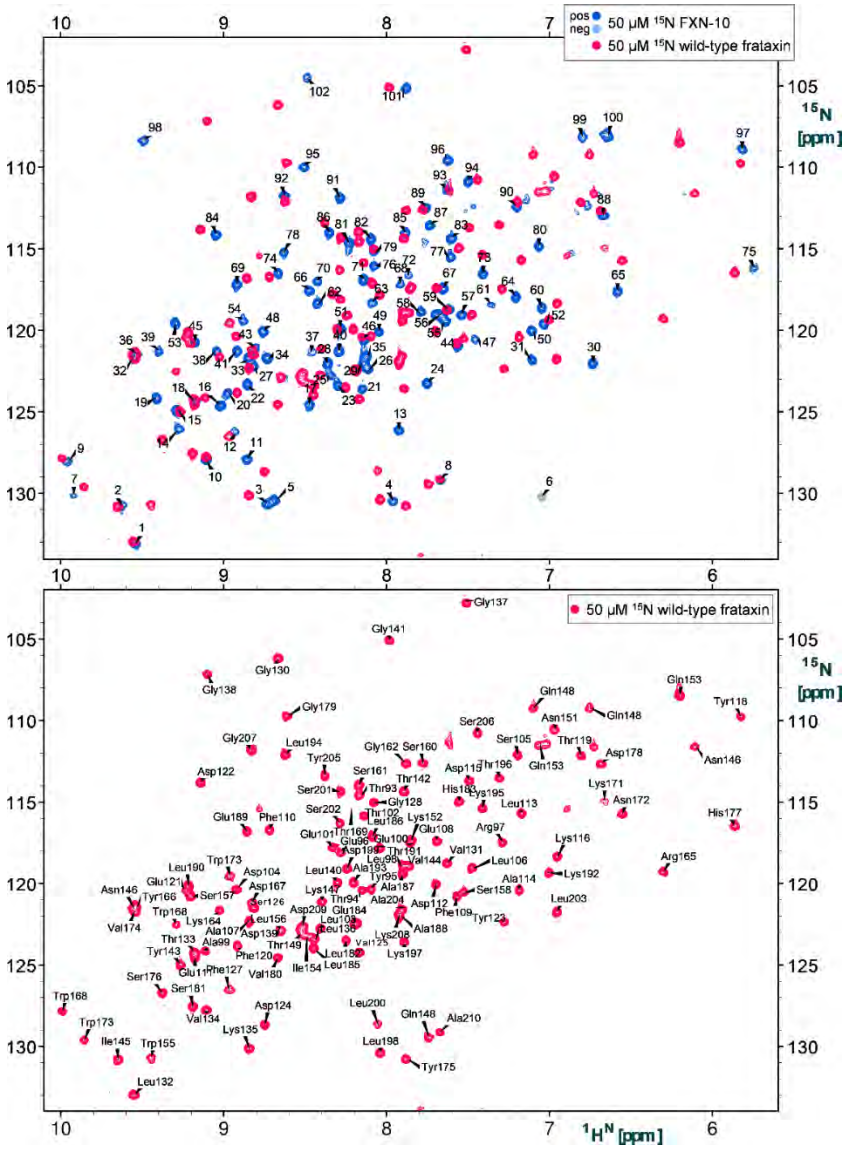


Figure 20. 600 MHz 2D ^{15}N-sf-HMQC spectra of u-^{15}N,^{13}C-labelled FXN-10 and u-^{15}N labelled wild-type frataxin. Top: overall, the chemical shift coordinates of both proteins are quite different due to presence of 13 single-residue mutations in FXN-10. Therefore, ^{1}H^{N}/^{15}N resonances are not unambiguously assigned for FXN-10 and all peaks are labelled with arbitrary numbers. Bottom: 2D ^{15}N-sf-HMQC spectra with backbone amide and sidechain amide assignments (for indol $H_{\varepsilon 1}/N_{\varepsilon}$, Asn $H_{\gamma 1,2}/N_{\gamma}$ and Gln $H_{\delta 1,2}/N_{\delta}$, respectively) of wild-type frataxin.

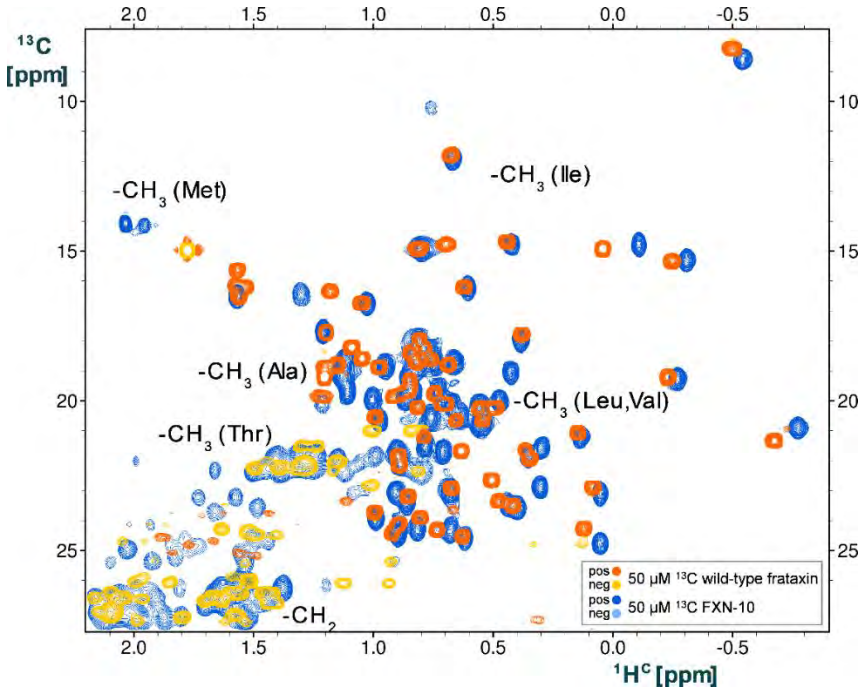


Figure 21. 800 MHz 2D ^{13}C-HSQC spectra of FXN-10 and constant time (ct) multiplicity-edited 2D ^{13}C-HSQC spectra of wild-type frataxin (methyl groups region). Despite the presence of 13 mutations in FXN-10, the resonances for the methyl groups superimpose reasonably well, suggesting that the packing of the protein core is very similar.

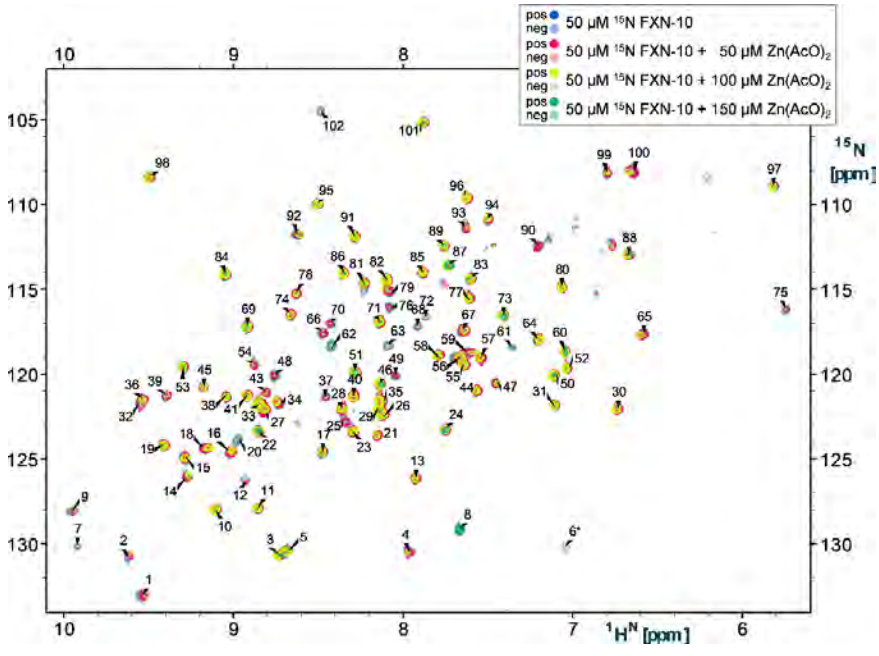


Figure 22. 600 MHz 2D 15N-sf-HMQC spectra of u-^{15}N,^{13}C-labelled FXN-10 (arbitrary residue numbers due to lack of assignment) in the presence of increasing Zn^{2+} concentrations.

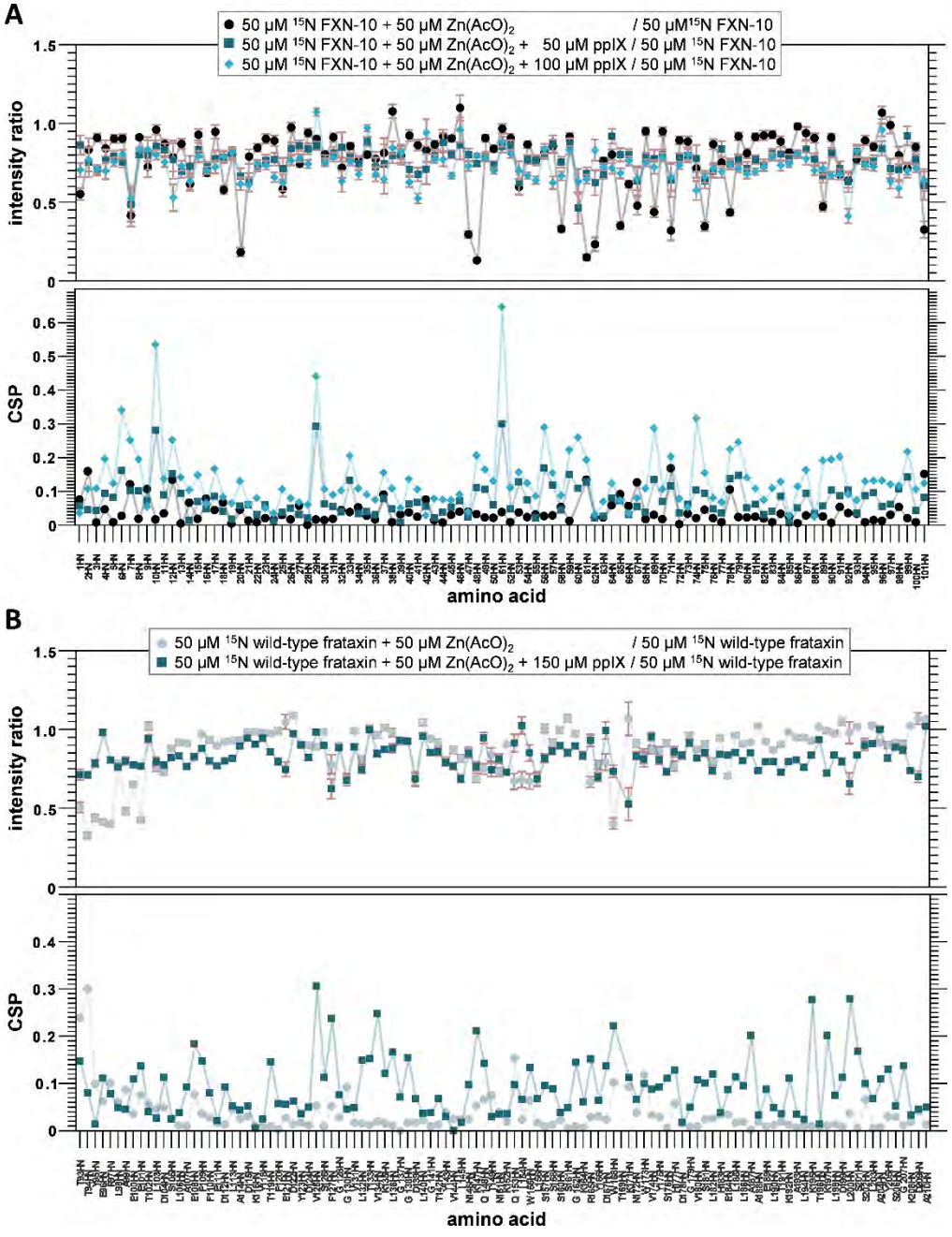


Figure 23. Signal intensity ratios and chemical shift perturbations (CSP) observed for a) u-^{15}N,^{13}C-labelled FXN-10 (arbitrary residue numbers due to lack of assignment), and b) u-^{15}N-labelled wild-type frataxin in the presence of Zn^{2+} and protoporphyrin IX (ppIX). Top: signal intensity ratios extracted from 2D ^{15}N-sf-HMQC spectra. Bottom: CSP plot shows the characteristic changes found at the C-terminus of helix α1 upon addition of a binder and indicates Zn^{2+}/ppIX binding.

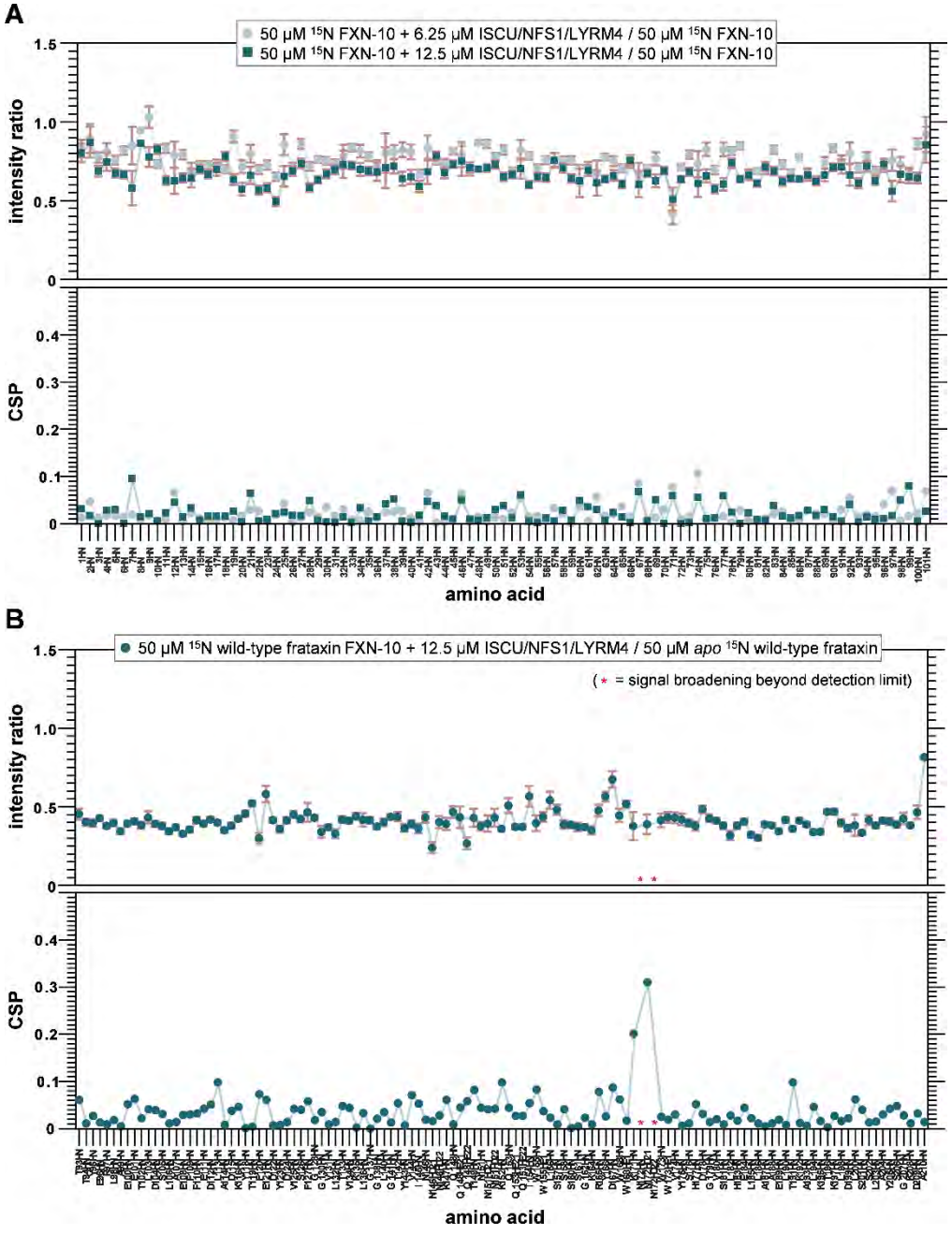


Figure 24. Signal intensity ratios and chemical shift perturbations (CSP) observed for a) u-^{15}N,^{13}C-labelled FXN-10 (arbitrary residue numbers due to lack of assignments), and b) u-15N-labelled wild-type frataxin in the presence of FeS assembly complex. Top: signal intensity ratios extracted from 2D ^{15}N-sf-HMQC spectra. Bottom: consistent with the slow exchange signature of the spectra, no changes of the chemical shifts are observed, except around residue K171 in the wild-type. This residue (substituted by Thr in FXN-10) stablishes transient interactions with N172 that might be differently populated in the presence of the FeS assembly complex.

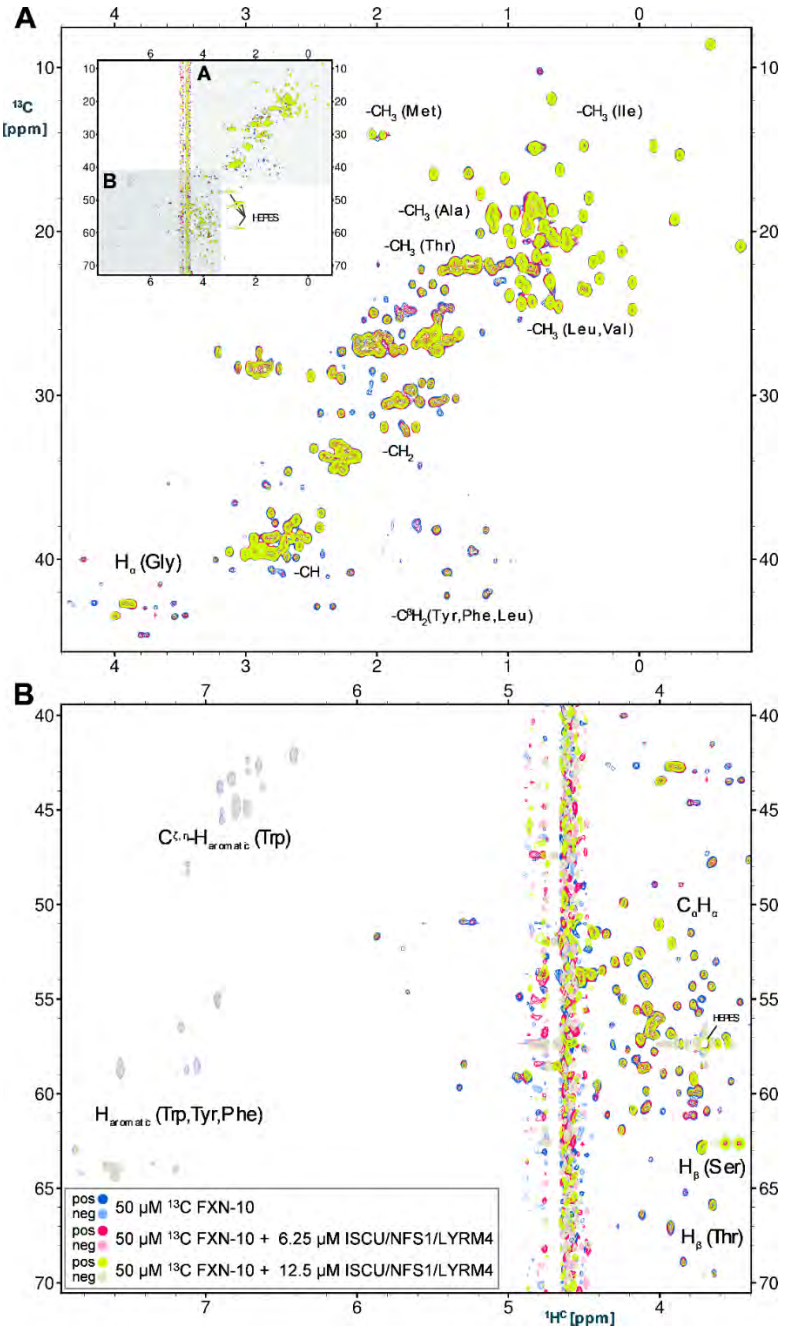


Figure 25. 800 MHz 2D non-constant time ^{13}C-HSQC spectra of u-^{15}N,^{13}C-labelled FXN-10 in the presence of increasing concentrations of the FeS assembly complex. Regions of typical chemical shifts for A) sidechain methine, methylene and methyl moieties as well as B) H_α and aromatic proton-bearing carbons, are displayed. The complete spectra are shown in the top left insert. The intensities of the C–$H_{x=1-3}$ resonances clearly decrease upon ISCU/NFS1/LYRM4 addition without any alteration of their chemical shift coordinates, evidencing binding to the FeS assembly complex.

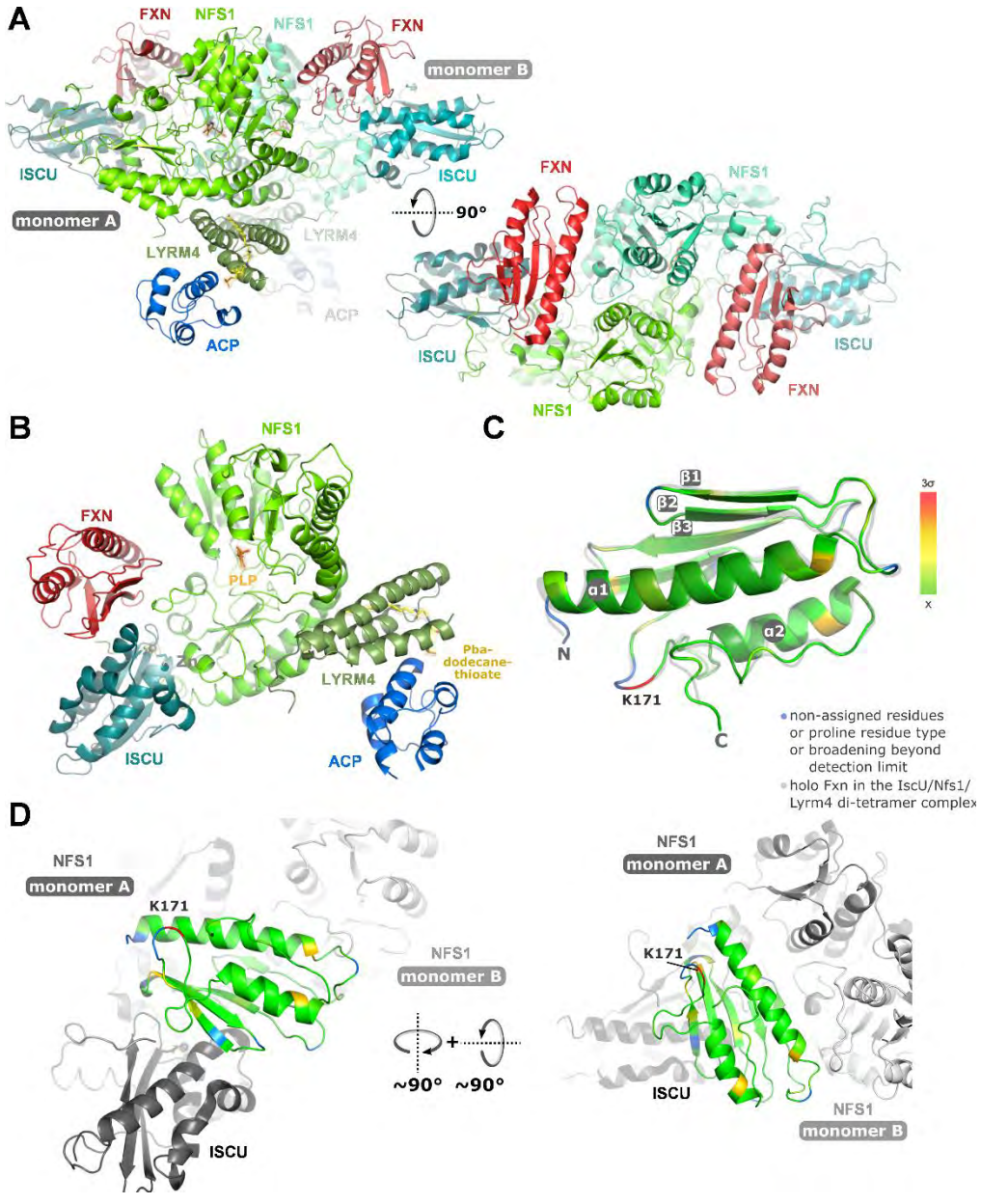


Figure 26. A) High resolution crystal structures of the homodimeric human FXN/NFS1/ISD11/ACP/ISCU-Zn^{2+} complex (PDB ID: 6NZU). Frataxin (FXN) is located between the ISCU and NFS1protomers. B) Each monomer contains a Zn^{2+} ion bound to ISCU, a pyridoxalphosphate (PLP) cofactor for the cysteine desulfurase NFS1, and a long chain fatty acid (40-phosphopantetheine) of ACP inserting into the helical center of the LYRM4 subunit. Mapping of backbone chemical shift perturbations (CSP) onto the NMR solution model (C) and to bound frataxin in the FeS assembly complex (D) locates only effected site around residue K171 located in the flexible loop β5/α2 facing outwards into the solvent. The color code is proportional to the standard deviation scale from the average value, as indicated in the bar legend (x: average value, in green).

This work has been published in the following article currently under revision:

Thermodynamic Stabilization of Human Frataxin. R. Núñez-Franco, A. Torres-Mozas, C. D. Navo, A. Schedlbauer, M. Azkargorta, I. Iloro, F. Elortza, G. Ortega, O. Millet, F. Peccati, G. Jiménez-Osés. *bioRxiv* 2023.09.08.556816 (*submitted manuscript*).

Chapter 6

Improving protein expression, stability and activity of tobacco etch virus (TEV) protease

1. Introduction

Evolution often prioritizes function over robustness in numerous naturally occurring proteins [264]. This tendency can lead to challenges such as reduced solubility, diminished thermal stability, and issues with expression in heterologous systems, ultimately impacting the quantity of functional protein [265,266]. The industrial use of numerous protein-based therapeutics and catalysts is often restricted due to their limited stability, highlighting the growing interest in protein stabilization research [267,268]. While traditional methods such as directed evolution are effective in enhancing certain protein characteristics, they can demand considerable resources and workforce [269,270]. In recent times, computational tools have been crafted to mirror the advantages of directed evolution with less need for hands-on experimentation [97,271–273]. One such tool, PROSS [97] (protein repair one-stop shop), combines evolutionary data with Rosetta's physics-driven energy calculations to remodel protein sequences based on their 3D structures. This has demonstrated improved solubility and thermal resistance in a range of native enzymes. Moreover, the rise of deep learning in protein modeling has paved the way for the development of novel natural protein variants. This includes the use of language models tailored for generating sequences specific to enzyme family or functions [273], convolutional neural networks that employ structural details to predict enhanced function mutations [272], and the introduction of simpler neural networks that guide combinatorial directed evolution processes [274].

Using deep learning-based methodologies for engineering protein sequences has proven highly effective in the production of novel proteins that exhibit enhanced levels of expression, solubility, and precision compared to initial designs [212,213,273]. A remarkable example is ProteinMPNN, recognized for its ability to generate stable sequences for designed backbones. Additionally, when applied to native backbones, it yields sequences that likely align more closely with intended structures than their original counterparts [213]. Therefore, ProteinMPNN has the potential to become a general-purpose tool to boost protein stability through sequence alterations. To validate this possibility, the widely recognized protease from the Tobacco etch virus (TEV) has been used as a benchmark protein.

TEV protease is extensively utilized in biotechnological applications to perform specific cleavage between glutamine and serine within its recognition sequence (ENLYFQ/S). This cleavage action serves to remove purification tags from recombinantly produced proteins. However, the practical use of TEV is often hindered by limited solubility, low thermostability, and suboptimal catalytic activity, leading to prolonged incubation times and incomplete cleavage. The core objective of this study is to leverage the capabilities of ProteinMPNN to optimize the stability of TEV protease, addressing its inherent limitations. The initial focus is on an autolysis-

resistant S219D variant [275] of TEV protease known as TEVd, which provides a reliable starting point for further enhancements.

ProteinMPNN uses a protein structure as an input to generate sequences that are predicted to idealize that given structure. The method relies solely on structural information and lacks access to functional insights. Thus, for ProteinMPNN-designed sequences to maintain their function, additional information is required. This work explores various approaches to preserve functionality throughout the design process, incorporating different levels of residue conservation and proximity to the active site in the mutation process.

This study has been developed as part of a research secondment at the Baker lab at the Institute for Protein Design (University of Washington).

2. Results and discussions

a. Solubility and activity enhancement

Overcoming TEV protease's solubility issues is vital for enhancing its use in biotechnological applications. In response to this, a series of TEV protease variants were designed using information on evolutionary conservation within the TEV protease family and ProteinMPNN as a primary tool. The main goal of this design campaign was to enhance solubility without compromising the inherent activity of the protease.

Different design sets were generated, preserving the identity of not only active site but also highly conserved residues (Fig. 1; see Methods). In short, 144 TEV protease variants were generated with ProteinMPNN by mutating non-active site residues located at least 7 Å away from the active site. Additionally, positions were ranked according to their conservation in the multiple sequence alignment, and the top 30, 50 or 70% most conserved ones were fixed during design. All of them were predicted by AlphaFold to adopt a TEV-like structure with high confidence (pLDDT scores of the variants > 87.5; pLDDT score of native variant = 90). These designs showed sequence identities ranging from 55 to 85% with respect to the native variant.

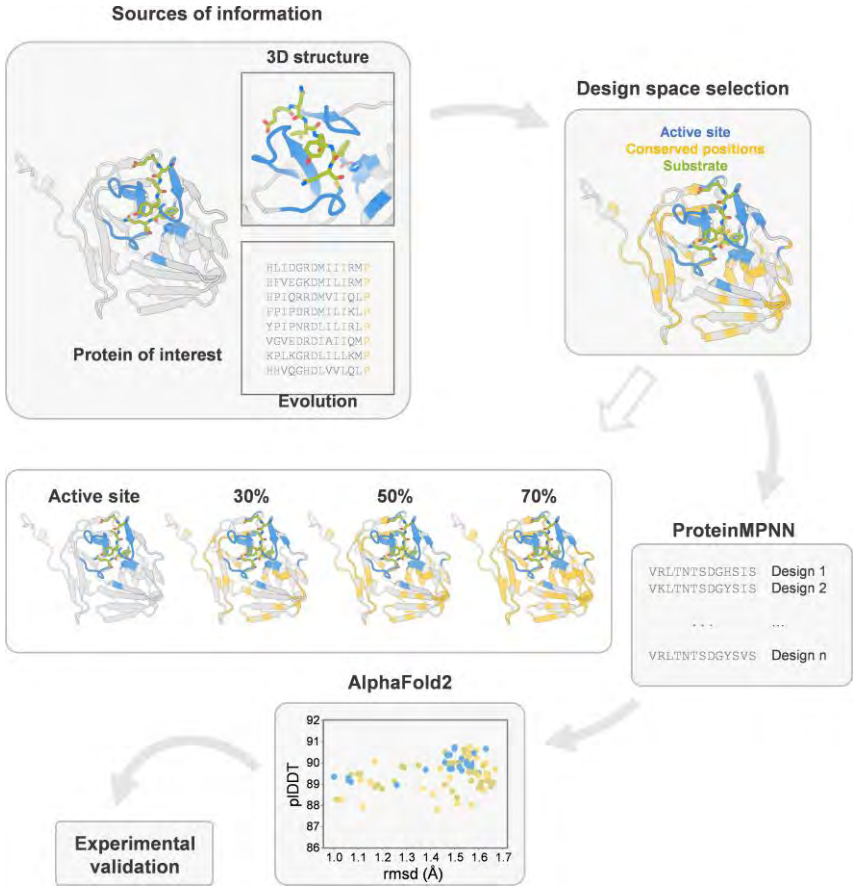


Figure 1. Design strategy for enhancing the expression and stability of the tobacco etch virus (TEV) protease. The first panel shows the data guiding the design, including the 3D structure of the protein and evolutionary insights derived from a multiple sequence alignment. The subsequent selection process for the design space prioritizes the preservation of enzyme native function. Ligand-binding site residues are selected based on their proximity to the ligand, using as reference the substrate-bound model of X-ray structure with PDB code 1LVM. The structure and fixed positions are then inputted into ProteinMPNN. The sequences are then predicted using AlphaFold, filtered and experimentally validated.

Synthetic genes containing ProteinMPNN-generated designs and several previously reported variants with improved stability used as controls, were introduced and expressed in *E. coli*. Following expression, the resulting proteins were purified using affinity and size exclusion chromatography (SEC) techniques. Among the 144 designs, 134 were successfully purified as monomers (Fig. 2a). Notably, designs featuring lower sequence conservation displayed higher soluble yields (Fig. 2b). In terms of soluble expression levels, 129 out 144 designs outperformed the parent variant TEVd (which had an average yield of 1 mg per liter of culture). The designs, on average, achieved a remarkable yield of 20.1 mg per liter of culture (Fig. 2c).

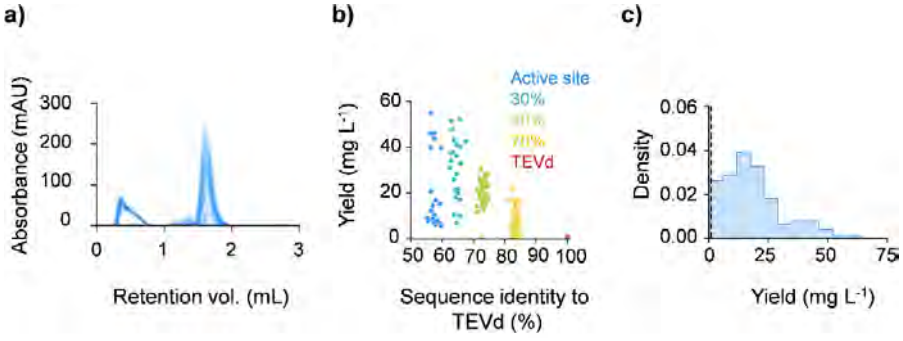


Fig 2. a) Size exclusion chromatography (SEC) absorbance profiles of 144 designed variants. 134 of 144 designs were successfully expressed and purified via SEC. b) Lower evolutionary constraints are linked to increased soluble expression levels. The legend denotes the regions that were kept fixed during the design based on sequence conservation, with active site residues being fixed in all designs. c) Histogram distribution of expression yields for the 144 designs. The black dashed line indicates the yield for the parent TEVd.

Catalytic activity was assessed using a previously described [270] coumarin derivative as a substrate, where 7-amino-4-trifluoromethylcoumarin was conjugated to the C-terminal of the substrate peptide Ac-ENLYFQ (Fig. 3a). Following protein purification, the enzyme was exposed to the peptide-coumarin substrate, and 64 designs displayed fluorescence progress curves exceeding the background, indicative of substrate turnover (Fig. 4). Despite designs generated without evolutionary constraints exhibited improved soluble expression compared to the parent sequence (Fig. 2b), they remained inactive towards the peptide substrate (Fig. 3b). Conversely, designs that demonstrated the highest activity were those designed with residues showing at least 50% conservation fixed.

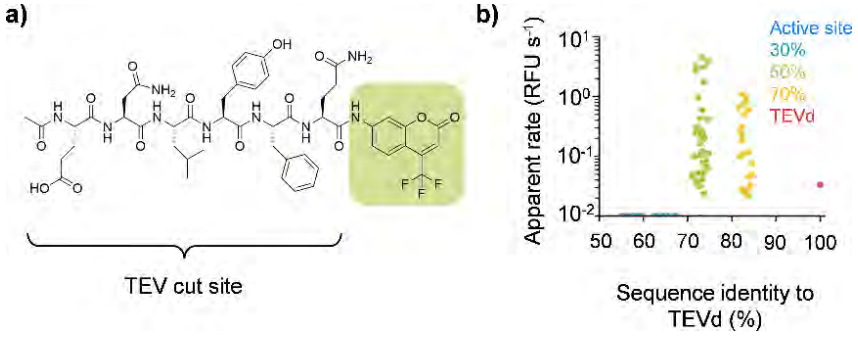


Figure 3. a) Structure of the peptide-coumarin substrate, AFC, used to assay proteolytic activity. b) Designs made with the active site and residues showing at least 50% conservation fixed during design exhibited the highest catalytic activity. Raw apparent rate is reported in relative fluorescent units (RFU) per second.

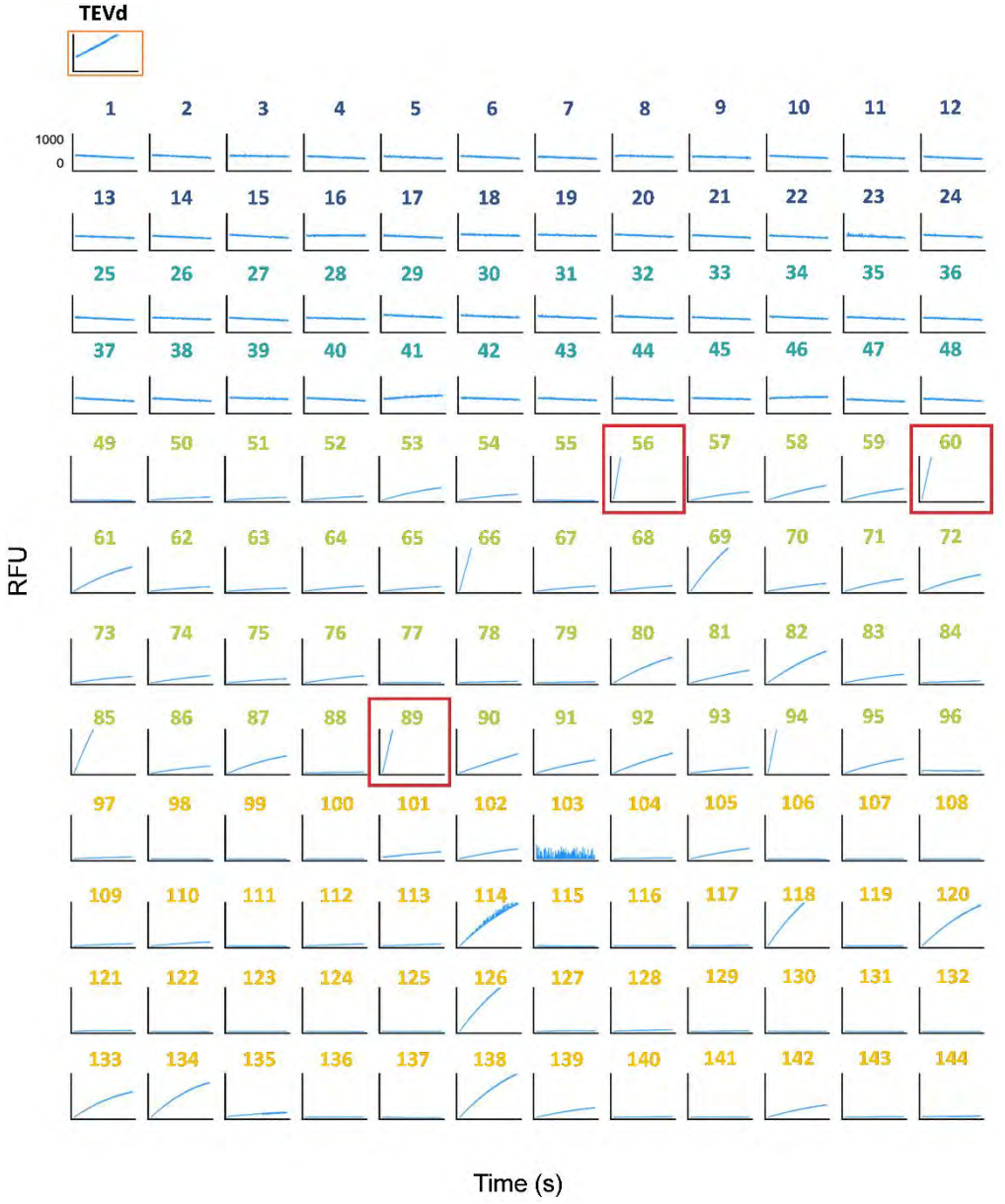


Figure 4. Initial screening of proteolytic activity on a fluorescent reporter substrate. Designs generated by fixing active site only residues (in blue), and active site plus residues conserved at least 30% (in dark green), 50% (in light green) or 70% (in yellow) were assayed. Pure protein was normalized to 500 nM and assayed against a single concentration of substrate of 10μM AFC, in an initial screen for catalytic turnover.

A detailed kinetic analysis was conducted on three highly active designs from the 50% conservation set – named hyperTEV56 (59 mutations), hyperTEV60 (62 mutations) and hyperTEV89 (58 mutations) – alongside the parent sequence TEVd [275]. The designs exhibited up to 26-fold enhanced catalytic efficiencies (k_{cat}/K_M) in comparison to TEVd (Table 1, Fig. 5).

Table 1. Kinetic parameters for three designed and parent TEV variants. Kinetic parameters were derived from a cleavage assay with the fluorescent peptide-coumarin substrate. Uncertainties are standard deviations of values calculated from fitting three technical replicates.

Variant	k_{cat} (min^{-1})	K_m (μM)	k_{cat}/K_m ($\mu M^{-1}\ min^{-1}$)
hyperTEV56	0.0106 ± 0.0005	1.4 ± 0.2	0.00770
hyperTEV60	0.0140 ± 0.0020	1.4 ± 0.4	0.01000
hyperTEV89	0.0050 ± 0.0001	2.0 ± 1.0	0.00240
TEVd	0.0023 ± 0.0003	6.0 ± 3.0	0.00039

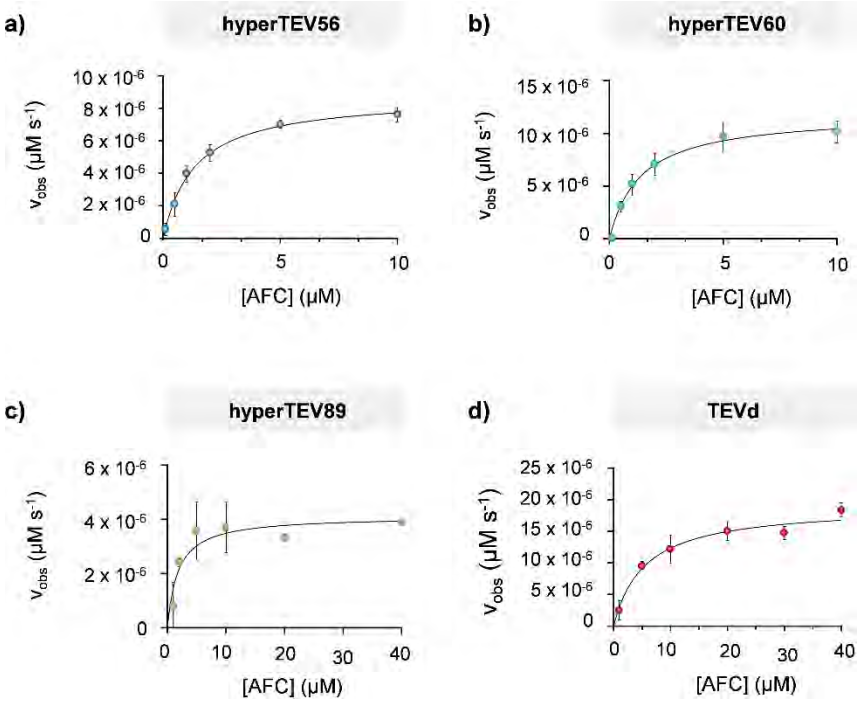


Figure 5. Michaelis-Menten plots for three designed TEV variants and parent TEVd. Error bars represent standard deviations from three technical replicates.

Next, the most promising designs were evaluated using a fusion protein substrate to test their efficiency in tag removal. Designs hyperTEV56, hyperTEV60, and hyperTEV89 were compared with a selection of previously engineered TEV protases [275–279]. Proteins were incubated at 30 °C with the fusion protein substrate MBP-TEVcs-FKBP-EGFP, where MBP is maltose-binding protein, TEVcs is the TEV peptide cleavage site (ENLYFQ↓G), FKBP is FK506-binding protein, and EGFP is enhanced green fluorescent protein. Proteolysis was estimated by observing the

increase of the cleaved product through SDS-PAGE (Fig. 6a). Designs hyperTEV56 and hyperTEV60 showed notably superior cleavage rates than the parent TEVd, resulting in 50% cleaved product roughly after 4 hours, whereas TEVd achieved a similar yield in 24 hours (Fig. 6a and b). These designs also surpassed other known TEV variants, with results showing 30% turnover for superTEV, 15% for TEV1Δ, and 50% for S219V after a 24-hour incubation period (Fig. 6b). Straight-line fitting of the product accumulation and substrate depletion reveal catalytic efficiencies that corroborate those determined in the peptide assay (Fig. 6c). The gains in catalytic efficiency are due to both increase in k_{cat}, which could reflect a higher fraction of enzyme in a catalytic competent state, and decrease in K_M, which could be related to better substrate binding properties.

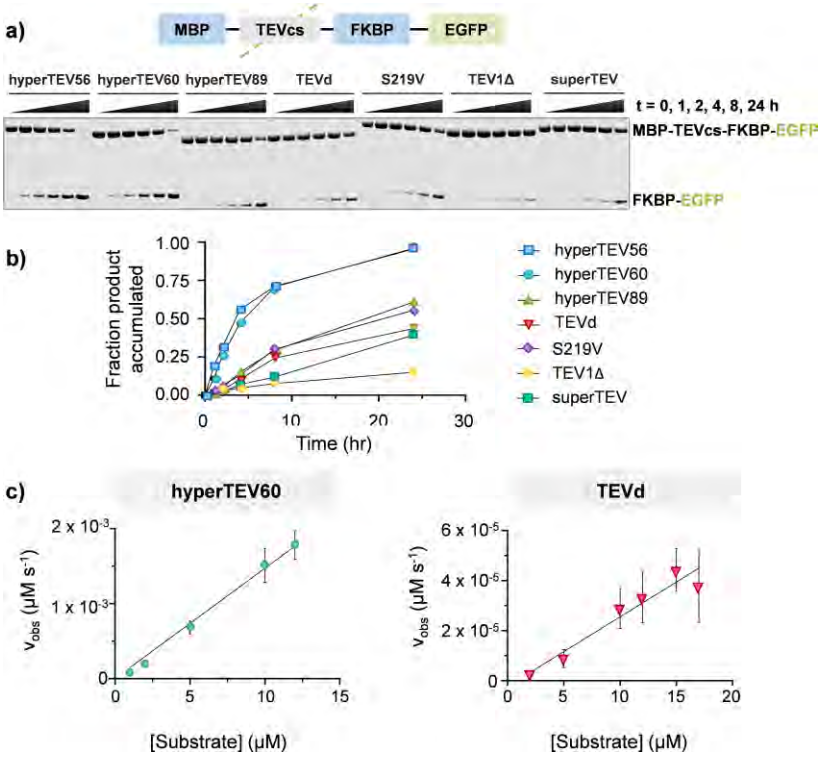


Figure 6. a) Diagram of TEV substrate (top) and product fluorescent gel image of TEV cleavage reactions at various time points (bottom). b) Plot of accumulated product normalized to fluorescent intensity of uncleaved substrate over time. Fluorescence intensity was quantified with ImageJ software. Designs hyperTEV56 and hyperTEV60 show increased turnover rate compared to previously reported variants. c) Straight-line fit for initial turnover rates in gel assay for hyperTEV60 and TEVd. Curves were fitted from monitoring substrate conversion for hyperTEV60 and production accumulation for TEVd. Error bars represent standard deviation from three technical replicates.

b. Thermal stability

An approximate melting temperature of 84 °C for hyperTEV60 was measured by circular dichroism (CD) spectroscopy analysis, which is 40 °C higher than that of TEVd (Fig. 7a and c). To the best of our knowledge, this temperature surpasses any previously reported TEV variant's melting temperature. To delve deeper into the stability of the designed variant, TEVd and the most active design, hyperTEV60, were subjected to incubation at 30 °C for various time intervals prior to their use in the aforementioned peptide-coumarin cleavage assay. Following a 4-hour incubation period, hyperTEV60 retains 90% of its initial cleavage activity, while TEVd's activity dropped to just 15% of its original level (Fig. 7b). These findings suggest a substantial enhancement in benchtop stability. Collectively, these results underscore the capability of sequence design using ProteinMPNN to enhance the benchtop stability, thermostability, and catalytic activity of natural enzymes.

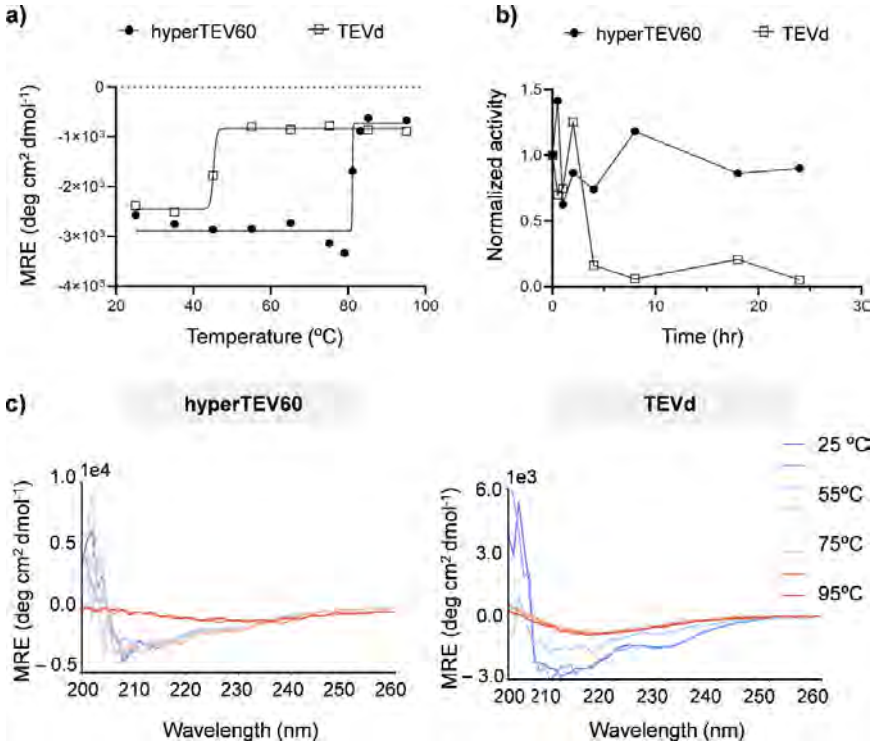


Figure 7. a) CD melting temperature plots of designed and parent TEV (signal reported in molar residue ellipticity, MRE). b) Benchtop stability comparison of parent TEVd and designed variant hyperTEV60 assessed as the activity (normalized to the value at t = 0) measured after being incubated at 30 °C for a given time. c) CD spectra of hyperTEV60 and TEVd over a temperature gradient from 25 °C to 95 °C indicates enhanced resistance to thermal unfolding in ProteinMPNN design hyperTEV60. CD signal is reported in molar residue ellipticity (MRE).

c. Molecular dynamics simulations analysis

While catalytic and substrate-binding residues remained unchanged during ProteinMPNN design, notable improvements in catalytic efficiency with both peptide and protein substrates were observed. Mutations away from the active site can modulate catalytic activity by stabilizing catalytically productive conformations [280,281] or inducing global conformational shifts [282]. To investigate if the stabilization of functional conformations contributed to enhanced activity, microsecond molecular dynamics (μs-MDs) simulations were carried out on TEV-peptide complexes. These simulations revealed increased rigidity in loop regions throughout the designs compared to TEVd (Fig. 8 and 11a). Such backbone rigidification, especially in areas not directly involved in substrate binding, might be associated not only with the measured higher thermal stability, but also with allosteric enhancement of substrate binding, supported by the 2- to 3-fold lower K_M values measured for the designed variants (Table 1). The region spanning residues 115 to 125 (region 1 in Fig. 8) showed a qualitative correlation between rigidity and activity; hyperTEV60 exhibited the highest rigidity, while TEVd and a non-active design showed increased flexibility (Fig. 8 and 11a).

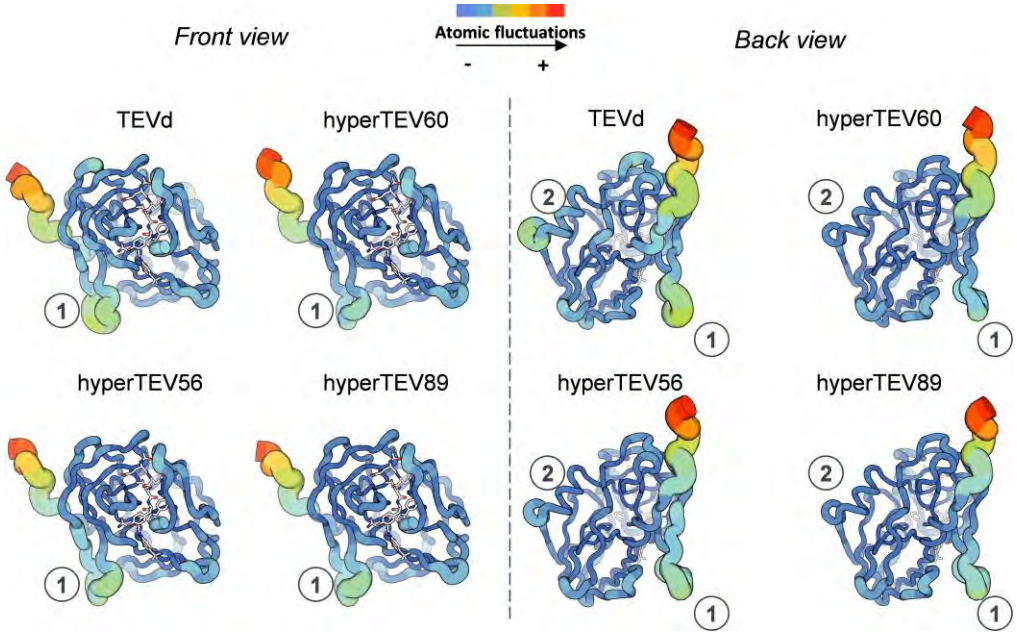
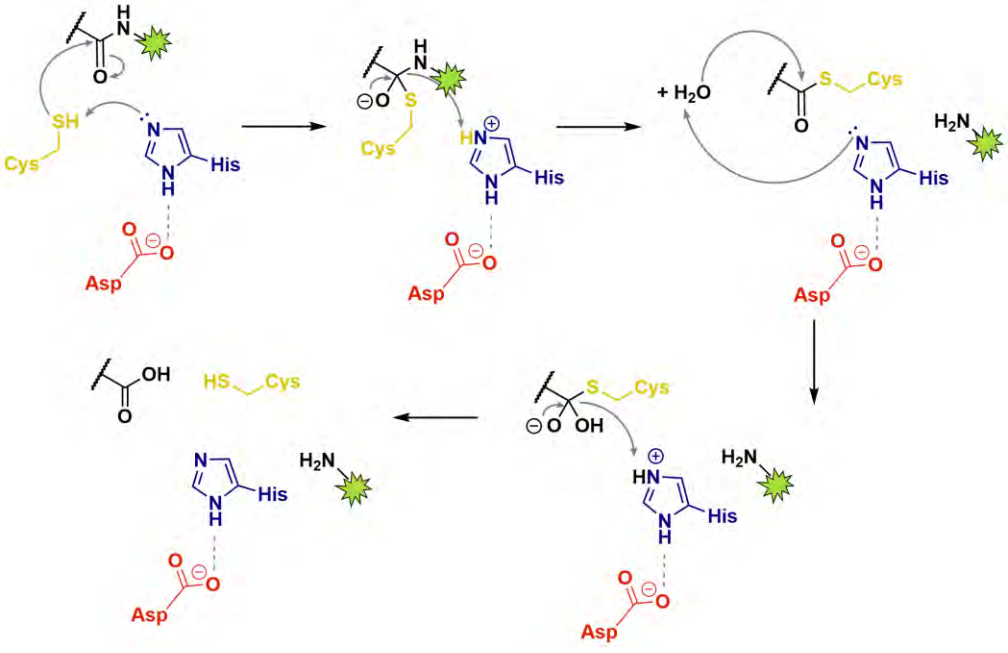


Figure 8. Visualization of the atomic root-mean-square fluctuations (RMSF) of designed and native TEV variants. Rigidified regions with respect to TEVd are marked with numbered circles.

The catalytic mechanism of TEV involves a well-studied cysteine protease mechanism (Scheme 1). TEV protease operates through a catalytic triad consisting of cysteine (Cys151), histidine (His46), and aspartic acid (Asp81) residues. In this reaction, the cysteine residue acts as a nucleophile, initiating the cleavage by forming a bond with the carbon atom of the peptide bond in the substrate. Histidine plays a pivotal role, functioning as a general base to facilitate this nucleophilic attack with assistance of deprotonated aspartic acid. Simultaneously, the backbone NH groups of residues Gly149 and Cys151 form the oxyanion hole that stabilize the negative charge developed in the carbonyl oxygen at the transition state and subsequent tetrahedral intermediate in the first step, thus aiding in the efficient cleavage of the peptide bond. Collapse of this intermediate leads to release of the C-terminal product fragment and a covalent thioester enzyme-substrate complex. The second half of the reaction involves subsequent release of the N-terminal product fragment in the deacylation step, in which a water molecule acts as a nucleophile to hydrolyze the acyl-enzyme intermediate and regenerate the catalyst. This mechanism allows TEV protease to efficiently and specifically cleave peptide bonds in target substrates upon recognition of the ENLYFQ↓G sequence (the ↓ symbol denotes the cleavage site).



Scheme 1. Mechanism of peptide bond cleavage catalyzed by TEV protease. The green star represents the fluorescent probe.

To assess the structural integrity of the TEV designs, specific catalytic distances (Fig. 9) were measured during the MD simulations. These distance measurements provided insights into the structural dynamics and catalytic competence of the TEV protease variants. Interestingly, all the designed variants exhibited a reduced amount of catalytically competent Cys-His dyad conformations compared to native TEV protease (TEVd). However, a notable exception was observed in the case of hyperTEV60, which displayed an increased occurrence of such competent conformations, correlating with its enhanced catalytic efficiency (k_{cat}), as depicted in Figure 10.

These differences offer insights into how ProteinMPNN might facilitate activity increase without explicitly introducing function-enhancing design elements. It is also plausible that a primary contributor to the improved k_{cat} in hyperTEV60 is the increased fraction of the protein existing in a catalytically competent state on a broader structural scale, including not only the catalytic triad but also a more global conformational context within the protein, not directly observable in the timescale of MD simulations (10 μs). This global perspective underscores the multifaceted nature of enzyme design and its capacity to optimize enzymatic function through intricate structural adaptations.

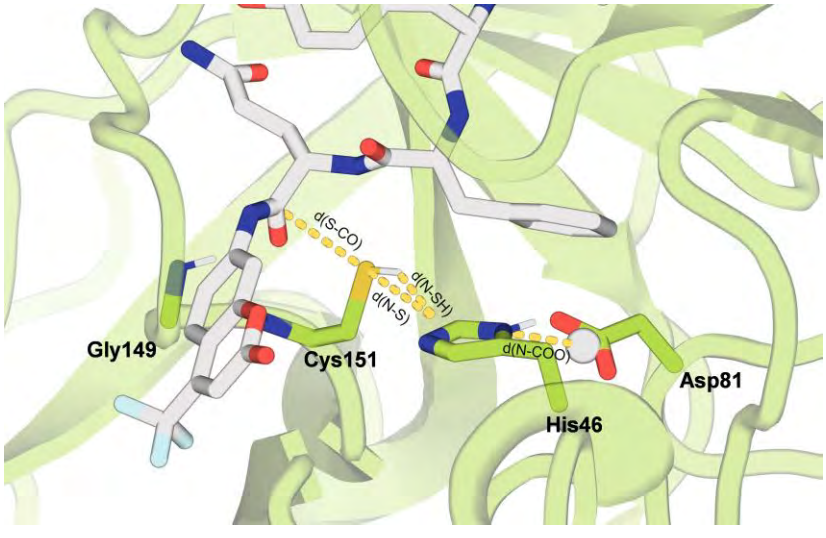


Figure 9. Key distances within the TEV catalytic triad (in green sticks) as seen in a selected snapshot from a MD simulation of TEVd protease. Peptide substrate is shown in gray sticks. The grey sphere represents the center of mass of the carboxylate oxygens.

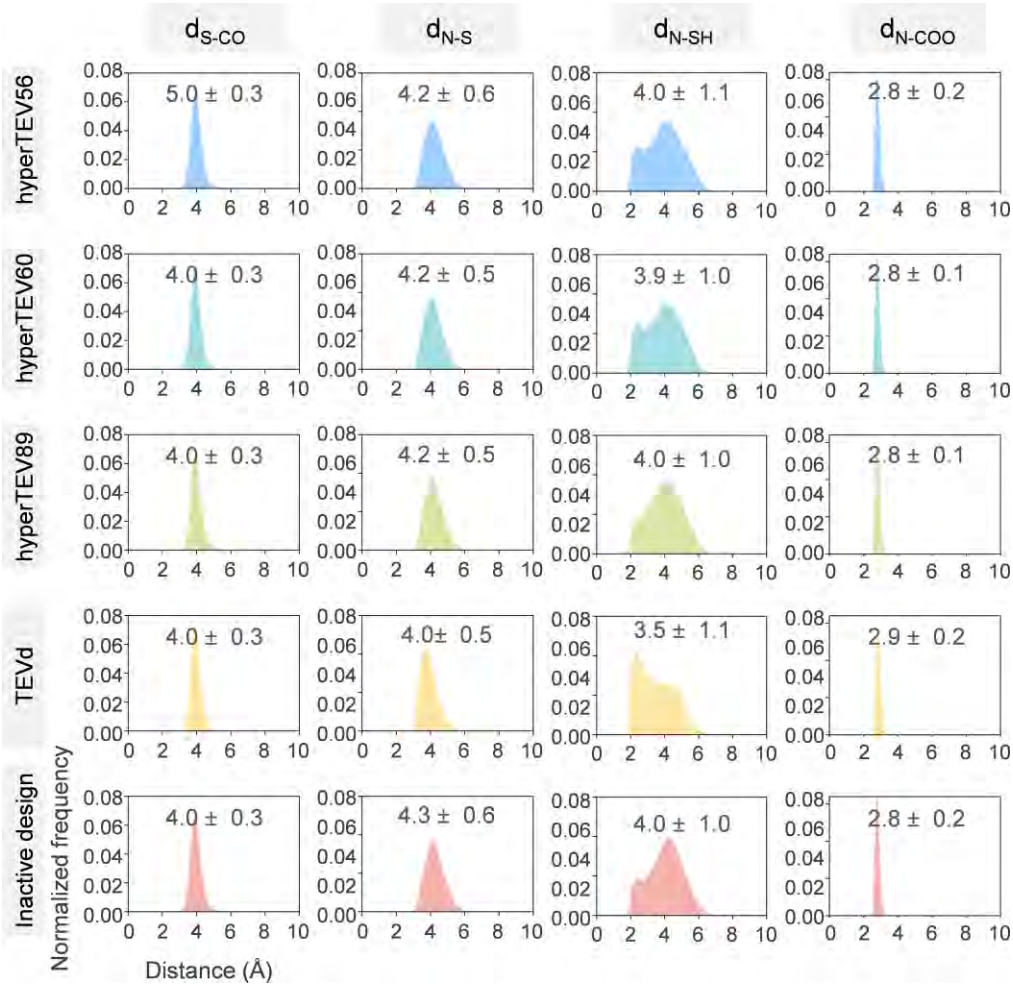


Figure 10. Histogram distributions of catalytic distances obtained from MD simulations on designed and parent TEV protease variants.

d. AlphaFold ensemble analysis

To further validate the observations from MD simulations and gain insight into the structural implications of the designs, an analysis using AlphaFold ensemble predictions was carried out. For many protein types, including globular proteins and protein complexes, it has been established that pLDDT scores from AlphaFold are highly consistent with RMSF profiles from MD simulations [283]. In this context, the per-residue pLDDT analysis provided an in-depth examination of structural confidence across the sequence. Consistent patterns of structural rigidity and flexibility, similar to those observed in the MD simulations (Fig. 11), were apparent.

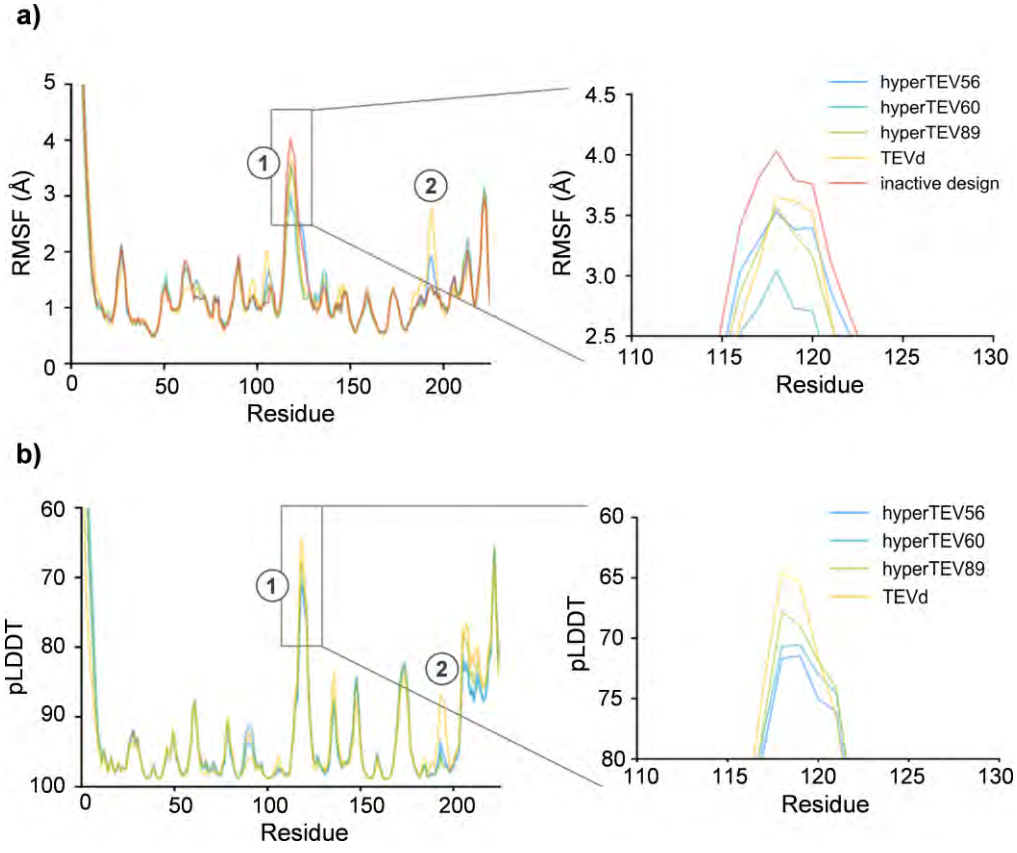


Figure 11. Root-mean-square fluctuations (RMSF) of designs in region 1 shows a positive qualitative correlation between activity and rigidification, with TEVd and a design inactive on the peptide substrate showing the highest flexibly in this region. c) Per-residue pLDDT values from AlphaFold ensemble prediction exhibit similar trends of increased rigidification in more active designs.

3. Conclusions

In this work, we demonstrate the efficacy of ProteinMPNN, guided by sequence and structural information, in enhancing the expression, stability, and function of native proteins. For the TEV protease, several variants were identified with improved soluble yield and thermostability compared to the native protein. Notably, the most optimized TEV protease designs exhibited superior apparent catalytic efficiency on both peptide and protein substrates compared to the parent enzyme and other documented variants. Although the ideal number of residues to conserve for maintaining (and potentially augmenting) function might vary and require empirical determination, the simplicity of this method combined with the computational efficiency and user-friendliness of ProteinMPNN makes the process straightforward.

In alignment with the results shown in the previous chapter, where in silico designs yielded variants like FXN-10 with exceptional thermostability and resistance to proteolytic degradation, the presented methodology based on ProteinMPNN for sequence design and AlphaFold for structure prediction proves its efficacy. The success of this approach, both in TEV protease and frataxin cases, emphasizes the efficiency, simplicity, and robustness of utilizing ProteinMPNN and deep learning methods in protein design. Importantly, our approach drastically reduces the number of variants that need experimental validation, making it a more efficient alternative to traditional high-throughput screenings. These advancements in protein design broaden potential applications, both in enhancing biotechnological innovations and in facilitating the development of specialized therapeutic approaches. For industry, enhanced proteins, such as the designed TEV variants, have the potential to revolutionize synthetic processes in an environmentally friendly manner. Simultaneously, these developments present opportunities for innovative medical treatments, as demonstrated by potential therapies for conditions such as Friedreich's ataxia.

4. Methods

a. Fixed residue selection

Several approaches were tested to maintain functionality throughout the design process. Initially, and in order to preserve the catalytic machinery and substrate-binding site of the enzyme, the amino acid identities of the first shell functional positions were kept constant. Active site residues were identified based on their proximity to the substrate in the ligand-bound crystal structure of the autolysis resistant S129D variant (PDB ID: 1LVM), specifically those with backbone atoms within 7 Å or sidechain atoms within 6 Å of the substrate. As a second step, the residues most conserved within the protein family were also held fixed, considering that even residues distal to the active site can make significant contributions to function [281] This conservation was determined through multiple sequence alignments (MSAs). The MSA was constructed using four iterative HHblits searches [84] against the UniRef30 database (accessed June 30, 2020) at E-value cutoffs of 1e-50, 1e-30, 1e-10, and 1e-4, and the final results was filtered for 90% identity redundancy, 50% coverage, and 30% minimum query identity. From this alignment, amino acid frequencies at each position were determined, highlighting the most prevalent amino acid at each location. Positions were ranked based on the conservation of the predominant amino acid, and the top segments (30%, 50%, and 70%) of these positions were retained during the sequence design process.

b. ProteinMPNN sequence design and structure prediction

For each set of fixed residues, sequence design on a fixed backbone was carried out using ProteinMPNN. The crystallographic structure of TEVd (PDB ID: 1LVM) was used as structural input to ProteinMPNN, with the active site and conserved residues being excluded from the design. The amino acid cysteine was not allowed to be introduced during the design process. Throughout the design, three temperature settings (0.1, 0.2, and 0.3) were explored to balance native sequence recovery and sequence diversity. Sequence generation was accomplished using a ProteinMPNN model that was calibrated with a noise threshold of 0.2 Å on protein backbone training sets. A total of 144 sequences were produced under different constraints:

- 24 variants with only the active site residues fixed (183 mutational hotspots)
- 24 variants with both the active site and the top 30% most conserved residues fixed (155 mutational hotspots)
- 48 variants with both the active site and the top 50% most conserved residues fixed (94 mutational hotspots)
- 48 variants with both the active site and the top 70% most conserved residues fixed (51 mutational hotspots)

After design, the sequences underwent structural prediction using AlphaFold [78], using model 3 with 6 iterative recycling steps. Predictions for both the new designs and the native TEV proved to be of low confidence when relying solely on a single sequence with few recycling iterations. It became evident that utilizing MSAs for structural templating was crucial for precise predictions. To generate MSAs of each design for structure prediction, the MSA of the parent sequence or designed variants were used. The AlphaFold models for all the designed sequences exhibited Cα RMSD < 2.0 Å and pLDDT scores > 85.0. They were also predicted to maintain essential structural components within their active site, leading to their subsequent selection for experimental testing.

c. Expression and purification of TEV designs

Double-stranded DNA fragments encoding the designs (codon-optimized for bacterial expression) were purchased from Integrated DNA Technologies (IDT) as eBlocks™ Gene Fragments. Following the Golden Gate cloning protocol [212], the DNA fragments encoding design sequences and including overhangs suitable for a BsaI restriction digest were cloned into a custom pET29b(+) target vector containing lethal ccdb gene, and C-terminal SNAC42 and hexahistidine tags (#191551, Addgene). This yielded final expressed sequences as: MSHHHHHHSG<design>GS. Vectors

containing TEV designs were transformed into *E. coli* BL21(DE3) by heat shock. DNA was incubated on ice with competent cells for 30 minutes, followed by 10 second heat shock at 42 °C, and 2-minute incubation on ice. 100 μL rich medium (super optimal broth with catabolite repression) was added to transformed cells and samples were incubated at 37 °C, 1050 rpm on a Heidolph shaker for 1 hour. Entire transformations were transferred to 900 μL of TBM-5052 autoinduction expression medium containing 50 μg/mL kanamycin. Expression cultures were incubated at 37 °C, 1050 rpm for 20 hours. Pellets were harvested by centrifugation at 4,000 g for 10 minutes and lysed with BPER lysis reagent containing 6.25 Units/mL benzonase (4 uL / 40 mL at 250 U/μL), 0.1 mg/mL lysozyme, and 1 mM PMSF. Lysate was collected by centrifugation at 4000 xg for 20 minutes and applied to Ni-NTA resin that was equilibrated with wash buffer (20 mM Tris-HCl, 300 mM NaCl, 25 mM imidazole, pH 8.0). The resin was washed with 25 column volumes (CV) of wash buffer. Protein was eluted with 250 μL of elution buffer (20 mM Tris-HCl, 300 mM NaCl, 540 mM imidazole, pH 8.0) and further purified via size exclusion chromatography (SEC) in an S75 5/150 GL increase column (GE Healthcare). Protein collected from SEC was normalized to 1 μM where possible.

In scale-up experiments, 50 mL cultures of TBM-5052 autoinduction media with 50 μg/mL kanamycin were inoculated with a scrape of transformed competent cells from glycerol stock and grown at 37 °C, 200 rpm for 20 hours. Cells were harvested by centrifugation at 10000 xg for 10 minutes, resuspended in 30 mL of wash buffer (20 mM Tris-HCl, 300 mM NaCl, 25 mM imidazole, pH 8.0) containing 0.01 mg/mL Dnase, 0.1 mg/mL lysozyme, and a protease inhibitor tablet (Thermo Scientific Pierce), and lysed by sonication. Lysate was collected via centrifugation at 18000 xg for 40 minutes and applied to Ni-NTA resin that was equilibrated with wash buffer. The resin was washed with 30 CV of wash buffer. Protein was eluted with 4 mL of elution buffer and concentrated to 1 mL in a 3 kDa protein concentrator (Millipore Sigma). Concentrated protein was purified by SEC as described above.

d. Expression and purification of MBP-TEVcs-FKBP-EGFP construct

The protease substrate FKBP-EGFP was cloned into an *E. coli* expression vector containing an N-terminal maltose binding protein (MBP), a TEV protease recognition site, and a C-terminal His-6 tag. The FKBP-EGFP coding sequence was obtained from Addgene #106924, with a 4X GGS linker between FKBP and EGFP. Vector containing the protease substrate was transformed into *E. coli* BL21(DE3) by heat shock. Cells were transferred to 4 0.5 L LB medium cultures with 50 μg/mL kanamycin and incubated at 37 °C, 200 rpm until optical density reached 0.5 AU, at which point expression was induced with 1 mM IPTG. Temperature was reduced to 18 °C and cells were incubated for an additional 18 hours. Cells were harvested by centrifugation at

10000 xg for 10 minutes, resuspended in 30 mL of wash buffer (20 mM Tris-HCl, 300 mM NaCl, 25 mM imidazole, pH 8.0) containing 0.01 mg/mL Dnase, 0.1 mg/mL lysozyme, and a protease inhibitor tablet (Thermo Scientific Pierce), and lysed by sonication. Lysate was collected via centrifugation at 18000 xg for 40 minutes and applied to Ni-NTA resin that was equilibrated with wash buffer. The resin was washed with 30 column volumes (CV) of wash buffer. Protein was eluted with elution buffer until resin no longer appeared yellow and concentrated to 1 mL in a 3 kDa protein concentrator (Millipore Sigma). Concentrated protein was purified by SEC as described above.

e. Kinetic characterization of designed proteases

Designs were initially screened for activity on a peptide-coumarin conjugate substrate (WuXi) of the TEV recognition sequence (ENLYFQ) fused to a fluorescent coumarin derivative (7-amino-4-trifluoromethylcoumarin). The N-terminus of the peptide bears an acetyl modification and the C-terminus is conjugated to the coumarin group via an amide bond. Initial activity screen was performed in 50 mM Tris-HCl, 50 mM NaCl, pH 8.0 buffer containing freshly prepared 2 mM DTT. Reactions contained 500 nM protein and 10 μM substrate at a total volume of 30 μL. Protein and substrate were rapidly mixed and monitored for fluorescence at excitation 400 nm, emission 492 nm at room temperature (RT) for 5 hours in a BioTek Synergy Neo2 microplate reader.

For detailed kinetic characterization, reactions were performed in 50 mM Tris-HCl pH 8.0 containing 50 mM NaCl, 1 mM EDTA, and freshly prepared 2 mM DTT. For TEV redesigns, reactions contained 50 nM protein and substrate concentration ranging from 0.1 μM to 10 μM at a total volume of 30 μL. Protein and substrate were rapidly mixed and monitored for fluorescence at excitation 400 nm, emission 492 nm at RT for 2 hours in a BioTek Synergy Neo2 microplate reader. Fluorescent signal was converted to concentration of cleaved coumarin product using a calibration curve of 7-amino-4-trifluoromethylcoumarin. Reactions were performed in triplicate and each technical replicate was separately fitted to a Michaelis Menten model. Expressed uncertainty in k_{cat} and K_M is the standard deviation between technical replicates.

f. Screening of designed proteases on fusion protein MBP-TEVcs-FKBP-EGFP

Reactions were performed in 50 mM Tris-HCl, 50 mM NaCl, 1 mM EDTA, pH 8.0 buffer containing freshly prepared 2 mM DTT. Reactions contained 60 nM protein and substrate concentrations ranging from 2 μM to 17 μM. Reactions were incubated at 30 °C and at 0, 1, 2, 4, 8, and 24 hours, 10 μL aliquots were quenched in 10 μL of 2X Laemmli loading buffer and subsequently frozen in liquid nitrogen. Samples were analyzed by SDS-PAGE and imaged for EGFP fluorescence at 488 nm on a LI-COR

Odyssey M imager. Band intensities were quantified with ImageJ software and converted to concentration using a standard curve prepared from known amounts of cleaved substrate with fluorescence gel imaging. A straight-line fit was applied to the initial velocities using GraphPad Prism. Points represent the averages of three technical replicates and error bars represent the standard deviations.

g. Benchtop stability characterization of TEV redesigns

Samples of purified enzyme were incubated at 30 °C for 0.5, 1, 2, 4, 8, 18, or 24 hours before being used in the previously described peptide-coumarin cleavage assay. The activity of samples was defined as initial rate of turnover and normalized to initial rate at incubation of t = 0 hours.

h. Circular dichroism spectroscopy

To determine secondary structure and thermostability of the designs, far-ultraviolet circular dichroism (CD) measurements were carried out on a JASCO J-1500 instrument using a 1 mm pathlength cuvette. Samples of purified protein were prepared at 1.0 mg/mL in 20 mM sodium phosphate, 50 mM potassium fluoride. The temperature of the sample was scanned from 25 °C to 95 °C with full spectrum scans from 190 nm to 260 nm performed after each 10 degree increment. The signal at 216 nm was plotted over the temperature gradient and fitted to a Boltzmann sigmoidal curve with GraphPad Prism 9. T_m values were calculated from the inflection point.

i. Molecular Dynamics simulations

Structures generated with AlphaFold [78] were used as starting geometries. For the protein-substrate complexes, substrate peptide was superimposed onto AlphaFold structures using the crystallographic structure of catalytically active TEV protease (PDB ID: 1LVM) as a template. Simulations were carried out with AMBER 20 [150] implemented with the ff14SB force field for the protein and substrate peptide, and the general Amber force field (GAFF2) [284] for the substrate peptide C-terminal fluorescent probe (7-amino-4-(trifluoromethyl)coumarin). Parameters were generated with the antechamber module of AMBER, combining ff14SB and GAFF2 force fields and with partial charges set to fit the electrostatic potential generated with HF/631G(d) using the RESP method [285]. The charges were calculated according to the Merz-Singh-Kollman scheme using Gaussian 16 [176]. Catalytic histidine residue (H46) was modeled in its Nδ1-H tautomeric state (corresponding to residue name HID in Amber). Initial structures were neutralized with either Na^+ or Cl^- ions and set at the center of a cubic TIP3P [286] water box with a buffering distance between solute and box of 10 Å.

A two-stage geometry optimization approach was performed. The first stage minimizes only the positions of solvent molecules and ions, and the second stage is an

unrestrained minimization of all the atoms in the simulation cell. The system was then heated by incrementing the temperature from 0 to 300 K under a constant pressure of 1 atm and periodic boundary conditions (PBC). Harmonic restraints of 10 kcal mol^{-1} were applied to the solute, and the Andersen temperature coupling scheme [155,179] was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Water molecules were treated with the SHAKE algorithm [287] such that the angle between the hydrogen atoms is kept fixed through the simulations. Long-range electrostatic effects were modeled using the particle mesh Ewald method [157]. An 8 Å cut-off was applied to Lennard-Jones interactions. The system was equilibrated for 2 ns with a 2 fs time step at a constant volume and temperature of 300 K. Ten independent production trajectories were then run for additional 1000 ns under the same simulation conditions, leading to accumulated simulation times of 10 μs for each system. Root mean square (rms) fluctuations and interatomic distance analyses were carried out with the *cpptraj* module of AMBER.

This work has been published in the following article:

Improving Protein Expression, Stability, and Function with ProteinMPNN. K. H. Sumida, R. Núñez-Franco, I. Kalvet, S. J. Pellock, B. I. M. W., L. F. Milles, J. Dauparas, J. Wang, Y. Kipnis, N. Jameson, A. Kang, J. De La Cruz, B. Sankaran, A. K. Bera, G. Jiménez-Osés, D. Baker. *J. Am. Chem. Soc.* **2024**, (*article ASAP*)

Chapter 7

Conclusions

The main conclusions derived from this Doctoral Thesis are summarized below:

1. A detailed study of the binding between carbohydrates of different complexities (blood group antigens and monosaccharides) and human galectins (*h*-Gal) through Molecular Dynamics (MD) simulations has revealed the very different characteristics of such protein-glycan complexes, particularly allosteric communication networks and binding site hydration. While homodimeric *h*-Gal-1 and, to a lesser extent, *h*-Gal-7 show significant long-range dynamic effects extending from the binding sites to very distant regions of the proteins – corroborated by relaxation NMR spectroscopy – lectins with spatially separated carbohydrate recognition domains (CRD) such as chimeric galectin *h*-Gal-3 and tandem galectins *h*-Gal-4 and *h*-Gal-8 do not show this behavior despite the high similarities existing among their binding sites. Additionally, hydration around the bound ligand has been modeled to be also strikingly different, particularly in the case of *h*-Gal-3. Such differences might be at the origin of the markedly different behavior observed experimentally for those galectins in terms of binding enthalpies and entropies, and the compensation between them determining affinity towards carbohydrates.

2. The weak (micro- to millimolar affinity) and highly dynamic binding observed experimentally between lectins – both galectins and DC-SIGN – and simple carbohydrates has been corroborated computationally by MD simulations in the microsecond scale. Hence, unbinding events have been frequently observed during the simulations, which makes mandatory the propagation of multiple trajectories instead of single, very long ones as it is common practice nowadays in molecular recognition studies.

3. A minimal ligand binding epitope has been found computationally, and verified experimentally, for the recognition of carbohydrate ligands to DC-SIGN lectin. Such minimal motif, present in natural monosaccharide binders D-mannose and L-fucose, is characterized by three hydroxyl groups arranged in a very particular axial-equatorial-equatorial orientation in space, and was discovered to exist also in L-galactose and D-rhamnose. These findings provide a rationale for designing new inhibitors for DC-SIGN with enhanced properties.

4. The use of multiple binding modes, as well as structural ensembles derived from MD simulations, enhance the capabilities of the NMR-based CORCEMA method to interpret STD information. This approach provides a broader view of carbohydrate binding to lectins than the use of single structural models since, as verified through both experimental and computational studies, such molecular recognition event is frequently weak and intrinsically dynamic in nature.

5. The combined used of evolutionary information and deep-learning models for sequence design (ProteinMPNN) and structure prediction (AlphaFold) has demonstrated to be exceedingly successful in improving the properties of very different proteins, such as human frataxin and TEV protease. The rational selection of mutable positions by imposing evolutionary (i.e. discarding highly conserved amino acids) and functional constraints based on structure (i.e. forbidding mutation at the surroundings of the active site or certain binding motifs), virtually guarantees success when sampling sequence with ProteinMPNN, as verified by the high confidence achieved by the AlphaFold models and, most importantly, by experimental assays. With an average mutation rate of only 10% across the entire amino acid sequence for human frataxin and 27% for TEV protease, which involves substitution in around 60% of the allowed mutational hotspots in both cases, ProteinMPNN has demonstrated an unparalleled accomplishment in fold idealization through sequence design.

6. Our approach for protein design has achieved impressive improvements in protein expression yield and stability (both thermodynamic and proteolytic, relevant for homeostasis), which are crucial for their potential use as biotechnological tools or therapeutic agents. Moreover, and although not imposed explicitly as a driving force or used as a selection method, the catalytic efficiency of TEV has been improved, and the native biological activity of frataxin has been preserved. Hence, our rational deep-learning based strategy constitutes a promising approach for protein design and engineering, complementing powerful high-throughput techniques such as random mutagenesis or directed evolution.

Appendix

Collaborations

This chapter compiles all the articles that were published in collaboration with other research groups during the doctoral studies.

1. **Precise Installation of Diazo-Tagged Side-Chains on Proteins to Enable In Vitro and In-Cell Site-Specific Labeling** (*Bioconjugate Chemistry.* **2020**, *31*, 1604-1610)

Abstract: The chemistry of diazo compounds has generated a huge breadth of applications in the field of organic synthesis. Their versatility combined with their tunable reactivity, stability, and chemoselectivity makes diazo compounds desirable reagents for chemical biologists. Here, we describe a method for the precise installation of diazo handles on proteins and antibodies in a mild and specific approach. Subsequent 1,3-cycloaddition reactions with strained alkynes enable both bioimaging through an in-cell "click" reaction and probing of the cysteine proteome in cell lysates. The selectivity and efficiency of these processes makes these suitable reagents for chemical biology studies.

Specifically, my contribution was to perform quantum mechanical calculations of activation barriers and frontier molecular orbitals involved in the reactions. These calculations provided insights into a 1,3-dipolar cycloaddition reaction involving strained cyclooctynes with diazo compounds.

2. **Unravelling the Time Scale of Conformational Plasticity and Allostery in Glycan Recognition by Human Galectin-1 (***Phys. Chem. Chem. Phys.* **2022**, *24(4)*, 1965-1973)


NMR Spectroscopy

Unravelling the Time Scale of Conformational Plasticity and Allostery in Glycan Recognition by Human Galectin-1

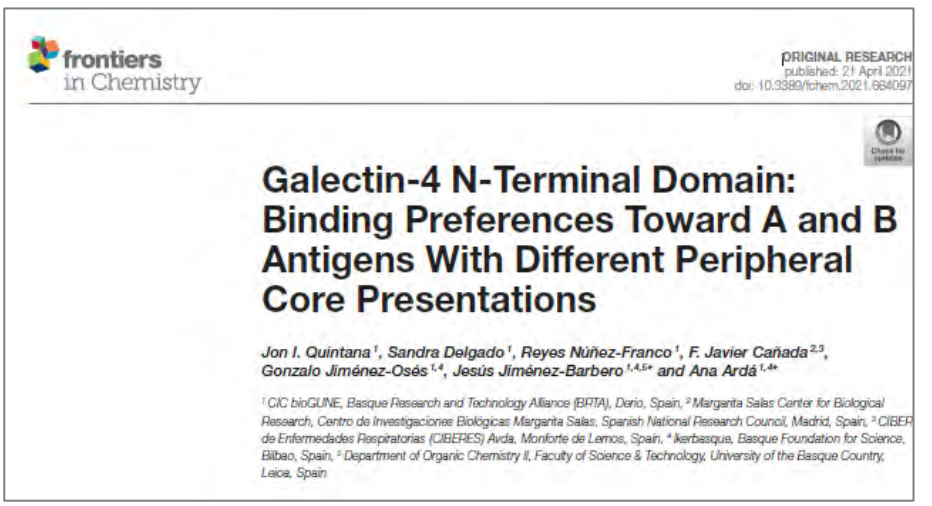
Sara Bertuzzi,[a] Ana Gimeno,[a] Reyes Núñez-Franco,[a] Ganeko Bernardo-Seisdedos,[a] Sandra Delgado,[a] Gonzalo Jiménez-Osés,[a] Oscar Millet,[a] Jesús Jiménez-Barbero,*[a, b, c] and Ana Ardá*[a]

Dedicated to the memory of Doctor Kilian Muñiz

Abstract: The interaction of human galectin-1 with a variety of oligosaccharides, from di-(N-acetyllactosamine) to tetra-saccharides (blood B type-II antigen) has been scrutinized by using a combined approach of different NMR experiments, molecular dynamics (MD) simulations, and isothermal titration calorimetry. Ligand- and receptor-based NMR experiments assisted by computational methods allowed proposing three-dimensional structures for the different complexes, which explained the lack of enthalpy gain when increasing the chemical complexity of the glycan. Interestingly, and independently of the glycan ligand, the entropy term does not oppose the binding event, a rather unusual feature for protein-sugar interactions. CLEANEX-PM and relaxation dispersion experiments revealed that sugar binding affected residues far from the binding site and described significant changes in the dynamics of the protein. In particular, motions in the microsecond-millisecond timescale in residues at the protein dimer interface were identified in the presence of high affinity ligands. The dynamic process was further explored by extensive MD simulations, which provided additional support for the existence of allostery in glycan recognition by human galectin-1.

My contribution in this work involved the analysis of allosteric communication through microsecond molecular dynamic simulations. These calculations aided at understanding dynamic processes at the molecular level, shedding light on allostery mechanisms and their implications in carbohydrate binding by lectins.

3. **Galectin-4 N-terminal domain binding preferences toward A and B antigens with different peripheral core presentations** (*Front. Chem.*, **2021**, *9*, 193)


frontiers in Chemistry

ORIGINAL RESEARCH
published: 21 April 2021
doi: 10.3389/fchem.2021.664097

Galectin-4 N-Terminal Domain: Binding Preferences Toward A and B Antigens With Different Peripheral Core Presentations

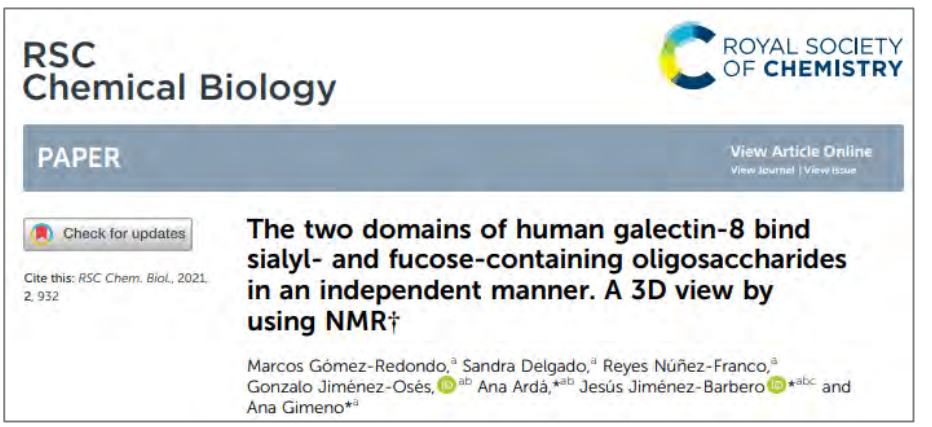
Jon I. Quintana[1], Sandra Delgado[1], Reyes Núñez-Franco[1], F. Javier Cañada[2,3], Gonzalo Jiménez-Osés[1,4], Jesús Jiménez-Barbero[1,4,5]* and Ana Ardá[1,4]*

[1] CIC bioGUNE, Basque Research and Technology Alliance (BRTA), Derio, Spain, [2] Margarita Salas Center for Biological Research, Centro de Investigaciones Biológicas Margarita Salas, Spanish National Research Council, Madrid, Spain, [3] CIBER de Enfermedades Respiratorias (CIBERES) Avda, Monforte de Lemos, Spain, [4] Ikerbasque, Basque Foundation for Science, Bilbao, Spain, [5] Department of Organic Chemistry II, Faculty of Science & Technology, University of the Basque Country, Leioa, Spain

Abstract: The tandem-repeat Galectin-4 (Gal-4) contains two different domains covalently linked through a short flexible peptide. Both domains have been shown to bind preferentially to A and B histo blood group antigens with different affinities, although the binding details are not yet available. The biological relevance of these associations is unknown, although it could be related to its attributed role in pathogen recognition. The presentation of A and B histo blood group antigens in terms of peripheral core structures differs among tissues and from that of the antigen-mimicking structures produced by pathogens. Herein, the binding of the N-terminal domain of Gal-4 toward a group of differently presented A and B oligosaccharide antigens in solution has been studied through a combination of NMR, isothermal titration calorimetry (ITC), and molecular modeling. The data presented in this paper allow the identification of the specific effects that subtle chemical modifications within this antigenic family have in the binding to the N-terminal domain of Gal-4 in terms of affinity and intermolecular interactions, providing a structural-based rationale for the observed trend in the binding preferences.

I contributed to this study by performing MD to provide insights into the mechanisms and structural dynamics of carbohydrate binding to the N-terminal carbohydrate recognition of human galectin 4.

4. **The two domains of human galectin-8 bind sialyl-and fucose-containing oligosaccharides in an independent manner. A 3D view by using NMR** (*RSC Chem. Biol.*, **2021**, *2*, 932-941)


RSC Chemical Biology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *RSC Chem. Biol.*, 2021, 2, 932

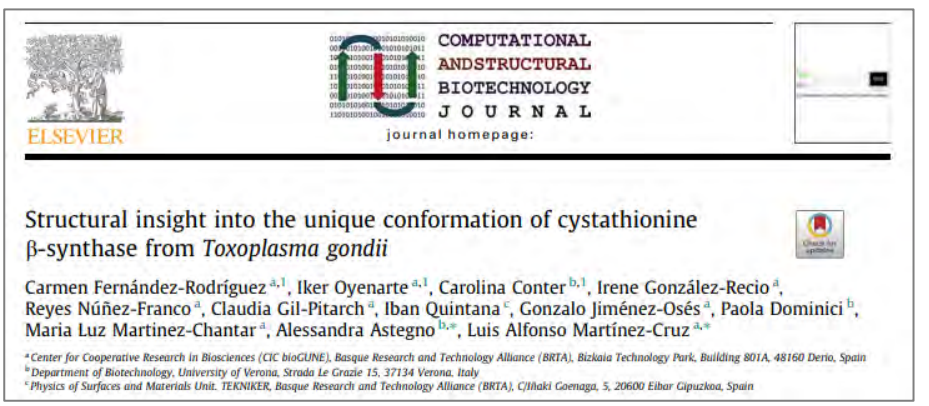
The two domains of human galectin-8 bind sialyl- and fucose-containing oligosaccharides in an independent manner. A 3D view by using NMR†

Marcos Gómez-Redondo,[a] Sandra Delgado,[a] Reyes Núñez-Franco,[a] Gonzalo Jiménez-Osés,[ab] Ana Ardá,*[ab] Jesús Jiménez-Barbero[abc] and Ana Gimeno*[a]

Abstract: The interaction of human galectin-8 and its two separate N-terminal and C-terminal carbohydrate recognition domains (CRD) to their natural ligands has been analysed using a synergistic combination of experimental NMR and ITC methods, and molecular dynamics simulations. Both domains bind the minimal epitopes N-acetyllactosamine (1) and Galβ1–3GalNAc (2) in a similar manner. However, the N-terminal and C-terminal domains show exquisite and opposing specificity to bind either Neu5Ac- or Fuc-containing ligands, respectively. Moreover, the addition of the high-affinity ligands specific for one of the CRDs does not make any effect on the binding at the alternative one. Thus, the two CRDs behave independently and may simultaneously target different molecular entities to promote clustering through the generation of supramolecular assemblies.

In this work, I performed MD simulations to refine a model for *h*-Gal-8 full length and explore its structure and dynamics. The simulations revealed interactions of the peptide linker with different domains of the protein. Simulations also captured ligand unbinding events, thereby providing insights into binding affinity dynamics, which correlated with experimental dissociation constants.

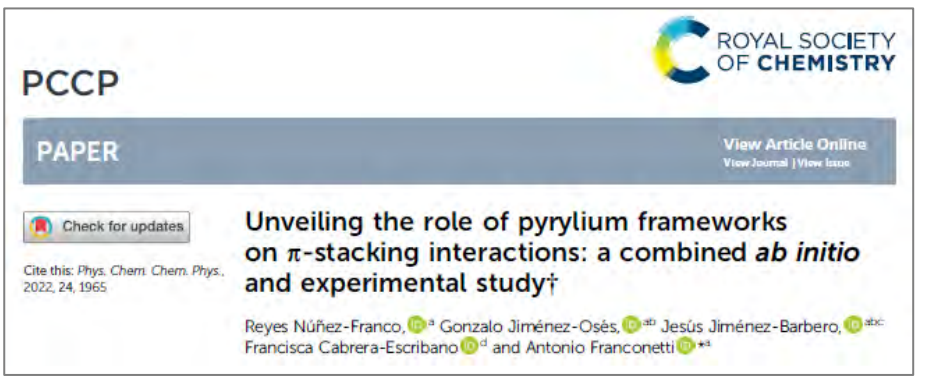
5. **Structural insight into the unique conformation of cystathionine β-synthase from Toxoplasma gondii** (*Comput Struct Biotechnol J.*, **2021**, *19*, 3542-3555)



Abstract: Cysteine plays a major role in the redox homeostasis and antioxidative defense mechanisms of many parasites of the phylum Apicomplexa. Of relevance to human health is Toxoplasma gondii, the causative agent of toxoplasmosis. A major route of cysteine biosynthesis in this parasite is the reverse transsulfuration pathway involving two key enzymes cystathionine β-synthase (CBS) and cystathionine γ-lyase (CGL). CBS from T. gondii (TgCBS) catalyzes the pyridoxal-5′-phosphate-dependent condensation of homocysteine with either serine or O-acetylserine to produce cystathionine. The enzyme can perform alternative reactions that use homocysteine and cysteine as substrates leading to the endogenous biosynthesis of hydrogen sulfide, another key element in maintaining the intracellular redox equilibrium. In contrast with human CBS, TgCBS lacks the N-terminal heme binding domain and is not responsive to S-adenosylmethionine. Herein, we describe the structure of a TgCBS construct that lacks amino acid residues 466-491 and shows the same activity of the native protein. TgCBS Δ466-491 was determined alone and in complex with reaction intermediates. A complementary molecular dynamics analysis revealed a unique domain organization, similar to the pathogenic mutant D444N of human CBS. Our data provides one missing piece in the structural diversity of CBSs by revealing the so far unknown three-dimensional arrangement of the CBS-type of Apicomplexa. This domain distribution is also detected in yeast and bacteria like Pseudomonas aeruginosa. These results pave the way for understanding the mechanisms by which TgCBS regulates the intracellular redox of the parasite, and have far-reaching consequences for the functional understanding of CBSs with similar domain distribution.

In this work, I performed MD simulations to investigate the catalytic behavior of TgCBS. A principal component analysis (PCA) was performed on the MD simulations to uncover significant structural changes within the homodimeric protein, such as longitudinal and lateral displacements, as well as rotational transformations of Bateman domains within the protein. These results contribute to a deeper understanding of CBS enzyme dynamics and its catalytic mechanisms with implications for therapeutic interventions.

6. **Unveiling the role of pyrylium frameworks on π-stacking interactions a combined ab initio and experimental study** (*Phys. Chem. Chem. Phys.* **2022**, *24(4)*, 1965-1973)


PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Phys. Chem. Chem. Phys.*, 2022, 24, 1965

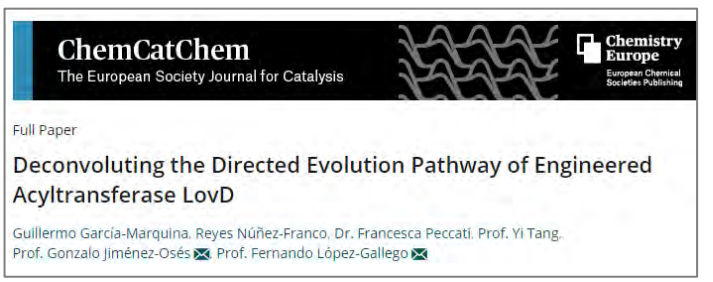
Unveiling the role of pyrylium frameworks on π-stacking interactions: a combined *ab initio* and experimental study†

Reyes Núñez-Franco, [a] Gonzalo Jiménez-Osés, [ab] Jesús Jiménez-Barbero, [abc] Francisca Cabrera-Escribano [d] and Antonio Franconetti [*a]

Abstract: A multidisciplinary study is presented to shed light on how pyrylium frameworks, as π–hole donors, establish π–π interactions. The combination of CSD analysis, computational modelling (ab intitio, DFT and MD simulations) and experimental NMR spectroscopy data provides essential information on the key parameters that characterize these interactions, opening new avenues for further applications of this versatile heterocycle.

My contribution in this work involved performing MD simulations to explore the structure and dynamics of a pyrylium tetrafluoroborate in DMSO, complementing quantum mechanical calculations and NMR experiments

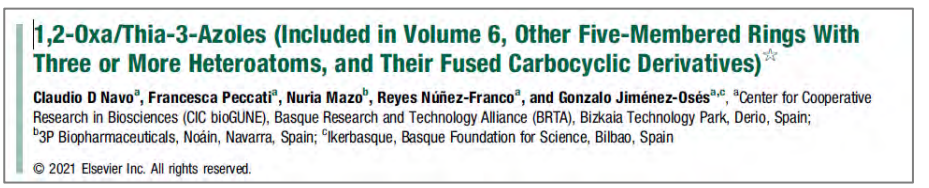
7. **Deconvoluting the directed evolution pathway of engineered acyltransferase LovD** (*Chem. Cat. Chem*, **2022**, *14(4)*)



Abstract: Pharmaceutical industry is progressively replacing the chemical synthesis of cholesterol-lowering agents by enzymatic processes. The directed evolution of acyltransferase LovD was a breakthrough in the synthesis of simvastatin, although little is known about how the in vitro evolution path raised up an engineered variant (LovD9) with excellent biocatalytic properties (high catalytic efficiency and stability under reaction conditions). In this study, we unveil how different mutation clusters scattered across LovD9 primary sequence specifically contribute to enhance both enzyme kinetics and stability. To this aim, simvastatin synthetic and hydrolytic activities, kinetic parameters and thermostability of several engineered variants were assessed. Through a rational combination of those clusters of mutations, we generated the variant LovD−BuCh2 whose catalytic efficiency is around 90 % of that obtained with LovD9 but with 15 less mutations. Supported by molecular dynamics simulations, this work demonstrates the cumulative effect of mutations at both the active site and the substrate entrance channel to enhance binding of the acyl donor and speed up the acyl transfer step from the acyl-enzyme complex to monacolin J acid, while simultaneously minimizing detrimental side-reaction pathways, substrate inhibition and increasing thermostability.

My contribution in this study can be found in the MD simulation section, where the enhanced catalytic activity of LovD variants resulting from directed evolution is explored. Previous MD analysis revealed improved activity in certain variants due to the broader motions of catalytic Tyr188. This was absent in crystallography. The study extended this by evaluating different variant ability to transfer the α-dimethylbutyryl group to acyl acceptor MJA in their acyl-enzyme complex forms. Models, derived from crystal structures, underwent MD simulations, revealing intramolecular interactions optimizing substrate orientation. LovD6, a highly evolved variant, maintained an optimal distance for efficient binding to MJA. MD studies also uncovered structural changes caused by evolution and elucidated the impact of specific mutations on catalytic efficiency. Evolution-induced mutations narrowed the substrate entrance channel, reducing water exposure and enhancing product formation.

8. **1, 2-OxaThia-3-Azoles** (*Reference Module in Chemistry, Molecular Sciences and Chemical Engineering*, **2021)**



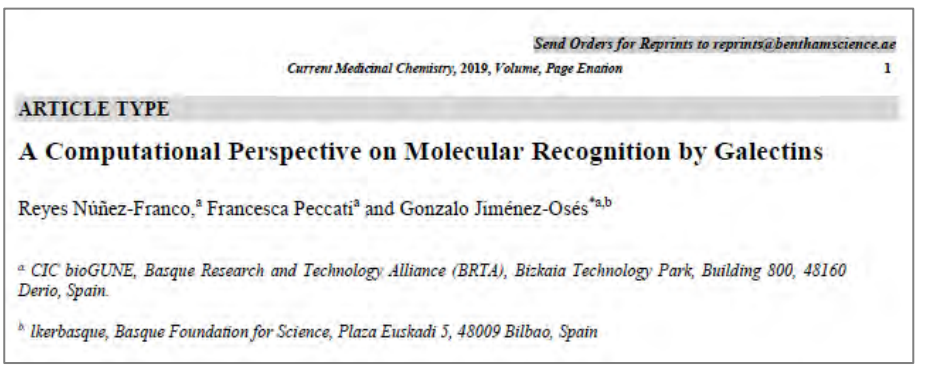
1,2-Oxa/Thia-3-Azoles (Included in Volume 6, Other Five-Membered Rings With Three or More Heteroatoms, and Their Fused Carbocyclic Derivatives)☆

Claudio D Navo[a], Francesca Peccati[a], Nuria Mazo[b], Reyes Núñez-Franco[a], and Gonzalo Jiménez-Osés[a,c], [a]Center for Cooperative Research in Biosciences (CIC bioGUNE), Basque Research and Technology Alliance (BRTA), Bizkaia Technology Park, Derio, Spain; [b]3P Biopharmaceuticals, Noáin, Navarra, Spain; [c]Ikerbasque, Basque Foundation for Science, Bilbao, Spain

In this article, the chemistry of 1,2,3-dithiazoles and 1,2,3-oxathiazoles is covered from 2009 to 2018 inclusive. 1,2,5-Oxathiazoles have not been investigated during this period. The chemistry of 4-chloro-1,2,3-dithiazolo-5-imines, -5-ones, -5-thiones, and -5-ylidenes that can readily obtained from Appel salt (4,5-dichloro-1,2,3-dithiazolium chloride) has been greatly expanded with many contributions from the Koutentis and Rakitin groups. The structure and electronic properties of bis-dithiazolyl radicals have been extensively investigated by the combination of many experimental and theoretical techniques. The reactions of cyclic sulfimidates and sulfamidates have been profusely investigated, achieving a dominant position in the preparation of a- and b-amino acids, amino alcohols, amines and fused bicycles with high control of chemo-, regio- and diastereoselectivity.

My contribution to this book chapter was to review all the recent literature regarding 1,2,3-dithiazoles, and write the corresponding section.

9. A Computational Perspective on Molecular Recognition by Galectins (*Curr Med Chem.*, **2022**, *29(7)*, 1219-1231)



Send Orders for Reprints to reprints@benthamscience.ae
Current Medicinal Chemistry, 2019, Volume, Page Enation 1

ARTICLE TYPE

A Computational Perspective on Molecular Recognition by Galectins

Reyes Núñez-Franco,[a] Francesca Peccati[a] and Gonzalo Jiménez-Osés[*a,b]

[a] CIC bioGUNE, Basque Research and Technology Alliance (BRTA), Bizkaia Technology Park, Building 800, 48160 Derio, Spain.

[b] Ikerbasque, Basque Foundation for Science, Plaza Euskadi 5, 48009 Bilbao, Spain

Abstract: This article presents an overview of recent computational studies dedicated to the analysis of binding between galectins and small-molecule ligands. We first present a summary of the most popular simulation techniques adopted for calculating binding poses and binding energies and then discuss relevant examples reported in the literature for the three main classes of galectins (dimeric, tandem, and chimera). We show that simulation of galectin-ligand interactions is a mature field that has proven invaluable for completing and unraveling experimental observations. Future perspectives to further improve the accuracy and cost-effectiveness of existing computational approaches will involve the development of new schemes to account for solvation and entropy effects, which represent the main current limitations to the accuracy of computational results.

My contribution to this review was to summarize theoretical methods commonly used to model galectin-ligand complexes, particularly MD simulations, and write the manuscript.

10. Distal Mutations Shape Substrate-Binding Sites during Evolution of a Metallo-Oxidase into a Laccase (*ACS Catalysis*, **2022**, *12*, 5022-5035)

ABSTRACT: Laccases are in increasing demand as innovative solutions in the biorefinery fields. Here, we combine mutagenesis with structural, kinetic, and *in silico* analyses to characterize the molecular features that cause the evolution of a hyperthermostable metallo-oxidase from the multicopper oxidase family into a laccase (k_{cat} 273 s^{-1} for a bulky aromatic substrate). We show that six mutations scattered across the enzyme collectively modulate dynamics to improve the binding and catalysis of a bulky aromatic substrate. The replacement of residues during the early stages of evolution is a stepping stone for altering the shape and size of substrate-binding sites. Binding sites are then fine-tuned through high-order epistasis interactions by inserting distal mutations during later stages of evolution. Allosterically coupled, long range dynamic networks favor catalytically competent conformational states that are more suitable for recognizing and stabilizing the aromatic substrate. This work provides mechanistic insight into enzymatic and evolutionary molecular mechanisms and spots the importance of iterative experimental and computational analyses to understand local-to-global changes.

My specific contribution to this research involved computational aspects, including conducting docking studies and MD simulations. Additionally, I carried out analyses related to allosteric regulation using protein residue network analysis. This computational framework allowed for a comprehensive exploration of the enzyme behavior, offering valuable insights into its structural dynamics, substrate binding, and allosteric interactions. The integration of these computational techniques contributed significantly to the overall understanding of the enzymatic and evolutionary processes described in the study.

11. Structures of the Inhibitory Receptor Siglec-8 in Complex with a High-Affinity Sialoside Analogue and a Therapeutic Antibody (*JACS Au*, **2023**, *3(1)*, 204–215)


JACS Au

pubs.acs.org/jacsau

Article

Structures of the Inhibitory Receptor Siglec-8 in Complex with a High-Affinity Sialoside Analogue and a Therapeutic Antibody

Maria Pia Lenza, Unai Atxabal, Corwin Nycholat, Iker Oyenarte, Antonio Franconetti, Jon Imanol Quintana, Sandra Delgado, Reyes Núñez-Franco, Carmen Teresa Garnica Marroquín, Helena Coelho, Luca Unione, Gonzalo Jiménez-Oses, Filipa Marcelo, Mario Schubert, James C. Paulson, Jesús Jiménez-Barbero,* and June Ereño-Orbea*

Abstract: Human sialic acid binding immunoglobulin-like lectin-8 (Siglec-8) is an inhibitory receptor that triggers eosinophil apoptosis and can inhibit mast cell degranulation when engaged by specific monoclonal antibodies (mAbs) or sialylated ligands. Thus, Siglec-8 has emerged as a critical negative regulator of inflammatory responses in diverse diseases, such as allergic airway inflammation. Herein, we have deciphered the molecular recognition features of the interaction of Siglec-8 with the mAb lirentelimab (2C4, under clinical development) and with a sialoside mimetic with the potential to suppress mast cell degranulation. The three-dimensional structure of Siglec-8 and the fragment antigen binding (Fab) portion of the anti-Siglec-8 mAb 2C4, solved by X-ray crystallography, reveal that 2C4 binds close to the carbohydrate recognition domain (V-type Ig domain) on Siglec-8. We have also deduced the binding mode of a high-affinity analogue of its sialic acid ligand (9-N-napthylsufonimide-Neu5Ac, NSANeuAc) using a combination of NMR spectroscopy and X-ray crystallography. Our results show that the sialoside ring of NSANeuAc binds to the canonical sialyl binding pocket of the Siglec receptor family and that the high affinity arises from the accommodation of the NSA aromatic group in a nearby hydrophobic patch formed by the N-terminal tail and the unique G–G′ loop. The results reveal the basis for the observed high affinity of this ligand and provide clues for the rational design of the next generation of Siglec-8 inhibitors. Additionally, the specific interactions between Siglec-8 and the N-linked glycans present on the high-affinity receptor FcεRIα have also been explored by NMR.

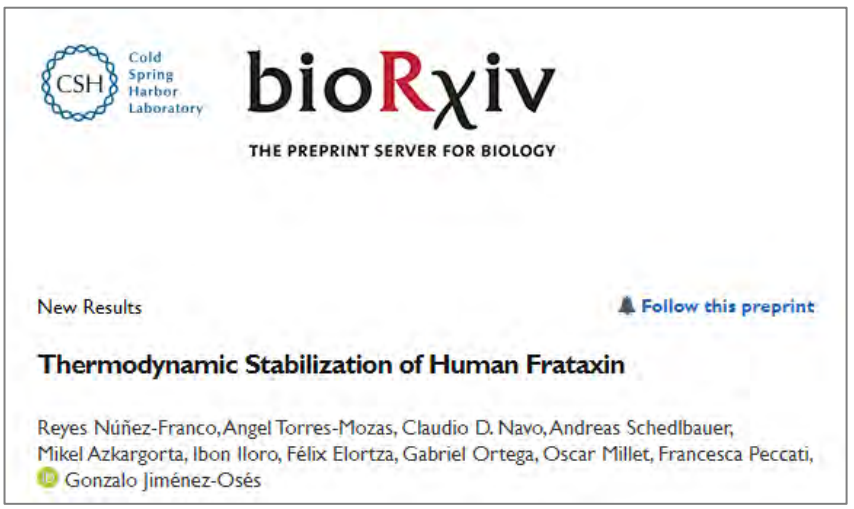
In this study I contributed with MD simulations to assess the conformational stability of the complex formed between Siglec-8 and NSANeu5Ac. The simulations, covering 500 ns, confirmed the stability of the crystallographic binding pose. Key polar contacts, salt bridges involving the sugar sulfate group, and hydrogen bonds were maintained, aligning with the NMR data obtained for sulfated ligands. The presence of the sulfate group was found to be essential for stabilizing interactions within the canonical sialic acid binding site of Siglec-8. Additionally, removing the sulfate group led to disrupted interactions and increased motion of Neu5Ac. This MD analysis validated the crystallographic interactions and provided insights into the ligand-receptor binding dynamics.

12. Expanding the Substrate Scope of Acyltransferase LovD9 for the Biosynthesis of Statin Analogues (*Chem. Eur. J.*, **2023**, *29(42)*)

Abstract: This study identifies new acyl donors for manufacturing statin analogues through the acylation of monacolin J acid by the laboratory evolved acyltransferase LovD9. Vinyl and p-nitrophenyl esters have emerged as alternate substrates for LovD9-catalyzed acylation. While vinyl esters can reach product yields as high as the ones obtained by α-dimethyl butyryl-S-methyl-3-mercaptopropionate (DMB-SMMP), the thioester for which LovD9 was evolved, p-nitrophenyl esters display a reactivity even higher than DMB-SMMP for the first acylation step yet the acylation product yield is lower. The reaction mechanisms were elucidated through quantum mechanics (QM) calculations.

In the context of this study, I conducted Quantum Mechanical calculations on the mechanism underlying catalytic Ser76 acylation with various thioester and ester surrogate models. A significant shift was observed in the acylation mechanism from the conventional stepwise process observed with thioesters and alkyl/vinyl esters to a concerted mechanism in the case of *p*-nitrophenyl (*p*NP) esters. These calculations not only shed light on the unique reactivity of *p*NP esters but also studied different acyl donors, including thioesters, as novel substrates for the biosynthesis of statins catalyzed by engineered LovD9.

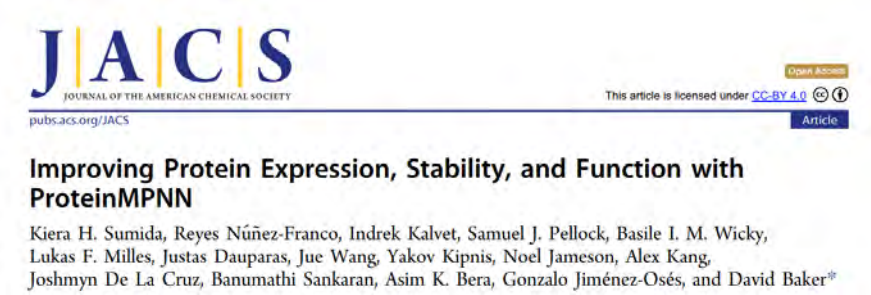
13. Thermodynamic Stabilization of Human Frataxin



Abstract: Recombinant proteins and antibodies are routinely used as drugs to treat prevalent diseases such as diabetes or cancer, while enzyme replacement and gene therapies are the main therapeutic intervention lines in rare diseases. In protein-based therapeutics, optimized in vivo stability is key as intrinsic denaturation and intracellular proteostatic degradation will limit potency, particularly in treatments requiring a sustained action, while clearance mechanisms may limit the amount of circulating protein. In vivo stability is ultimately correlated with the intrinsic thermodynamic stability of the biomolecule, but this is difficult to optimize because it often goes at the expense of reducing protein activity. Here, we have used in silico engineering approaches to thermodynamically stabilize human frataxin, a small mitochondrial protein that acts as an allosteric activator for the biosynthesis of Fe-S clusters, whose genetically-driven impairment results in a rare disease known as Friedreich ataxia. Specifically, we developed an efficient thermostability engineering computational approach that combines information on amino acid conservation, the Rosetta energy function, and two recent artificial intelligence tools – AlphaFold and ProteinMPNN – to produce thermodynamically stabilized variants of human frataxin. Such protein variants rescued the large destabilization exerted by well-known pathological mutations, with an increase over 20 °C in the melting temperature and a thermodynamic stabilization of more than 3 kcal·mol^{-1} at the physiological temperature. This stability surplus is translated into an enhanced resistance to proteolysis, while maintaining the protein fully functional. This case-study highlights the power of our combined computational approach to generate optimized variants, adequate for protein-based therapeutics.

In this work, I performed the computational design and energetic evaluation of frataxin variants, together with the experimental characterization of designs, including expression, purification and thermodynamic characterization.

14. Improving Protein Expression, Stability, and Function with ProteinMPNN



Abstract: Natural proteins are highly optimized for function but are often difficult to produce at a scale suitable for biotechnological applications due to poor expression in heterologous systems, limited solubility, and sensitivity to temperature. Thus, a general method that improves the physical properties of native proteins while maintaining function could have wide utility for protein-based technologies. Here, we show that the deep neural network ProteinMPNN, together with evolutionary and structural information, provides a route to increasing protein expression, stability, and function. For both myoglobin and tobacco etch virus (TEV) protease, we generated designs with improved expression, elevated melting temperatures, and improved function. For TEV protease, we identified multiple designs with improved catalytic activity as compared to the parent sequence and previously reported TEV variants. Our approach should be broadly useful for improving the expression, stability, and function of biotechnologically important proteins.

My contribution in this study was related to the TEV design campaign. I computationally designed part of the variants and characterized them experimentally, including expression, stability and catalytic activity.

References

[1] A View Of The Hydrophobic Effect. N. T. Southall, K. A. Dill, A. D. J. Haymet, *J. Phys. Chem. B* **2002**, *106*, 521–533.

[2] Water In Protein Hydration And Ligand Recognition. M. Maurer, C. Oostenbrink, *J. Mol. Recognit.* **2019**, *32*, DOI 10.1002/JMR.2810.

[3] Enthalpy-Entropy Compensation In Biomolecular Recognition: A Computational Perspective. F. Peccati, G. Jiménez-Osés, *ACS Omega* **2021**, *6*, 11122–11130.

[4] Energetics Of Ligand-Induced Conformational Flexibility In The Lactose Permease Of Escherichia Coli. Y. Nie, I. Smirnova, V. Kasho, H. R. Kaback, *J. Biol. Chem.* **2006**, *281*, 35779.

[5] Water Networks Contribute To Enthalpy/Entropy Compensation In Protein-Ligand Binding. B. Breiten, M. R. Lockett, W. Sherman, S. Fujita, M. Al-Sayah, H. Lange, C. M. Bowers, A. Heroux, G. Krilov, G. M. Whitesides, *J. Am. Chem. Soc.* **2013**, *135*, 15579–15584.

[6] Mechanism Of The Hydrophobic Effect In The Biomolecular Recognition Of Arylsulfonamides By Carbonic Anhydrase. P. W. Snyder, J. Mecinović, D. T. Moustakas, S. W. Thomas, M. Harder, E. T. Mack, M. R. Lockett, A. Héroux, W. Sherman, G. M. Whitesides, *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 17889–17894.

[7] Compensating Enthalpic And Entropic Changes Hinder Binding Affinity Optimization. V. Lafont, A. A. Armstrong, H. Ohtaka, Y. Kiso, L. Mario Amzel, E. Freire, *Chem. Biol. Drug Des.* **2007**, *69*, 413–422.

[8] Congeneric But Still Distinct: How Closely Related Trypsin Ligands Exhibit Different Thermodynamic And Structural Properties. T. Brandt, N. Holzmann, L. Muley, M. Khayat, C. Wegscheid-Gerlach, B. Baum, A. Heine, D. Hangauer, G. Klebe, *J. Mol. Biol.* **2011**, *405*, 1170–1187.

[9] Applying Thermodynamic Profiling In Lead Finding And Optimization. G. Klebe, *Nat. Rev. Drug Discov.* **2015**, *14*, 95–110.

[10] The Roles Of Water In The Protein Matrix: A Largely Untapped Resource For Drug Discovery. F. Spyrakis, M. H. Ahmed, A. S. Bayden, P. Cozzini, A. Mozzarelli, G. E. Kellogg, *J. Med. Chem.* **2017**, *60*, 6781–6828.

[11] Just Add Water! The Effect Of Water On The Specificity Of Protein-Ligand Binding Sites And Its Potential Application To Drug Design. J. E. Ladbury, *Chem. Biol.* **1996**, *3*, 973–980.

[12] Analysis Of Ligand-Bound Water Molecules In High-Resolution Crystal Structures Of Protein-Ligand Complexes. Y. Lu, R. Wang, C. Y. Yang, S. Wang, *J. Chem. Inf. Model.* **2007**, *47*, 668–675.

[13] Under Water's Influence. G. Hummer, *Nat. Chem. 2010 211* **2010**, *2*, 906–907.

[14] Entropy-Enthalpy Compensation: Role And Ramifications In Biomolecular Ligand Recognition And Design. J. D. Chodera, D. L. Mobley, **2013**, *42*, 121–142.

[15] Ligand Binding Introduces Significant Allosteric Shifts In The Locations Of Protein Fluctuations. A. Kumar, R. L. Jernigan, *Front. Mol. Biosci.* **2021**, *8*, 733148.

[16] Weighted Implementation Of Suboptimal Paths (WISP): An Optimized Algorithm And Tool For Dynamical Network Analysis. A. T. Van Wart, J. Durrant, L. Votapka, R. E. Amaro, *J. Chem. Theory Comput.* **2014**, *10*, 511–517.

[17] Cell Surface Carbohydrates In Adhesion And Migration. S. B. Oppenheimer, *Integr. Comp. Biol.* **1978**, *18*, 13–23.

[18] Fine Tuning Of Cell Signals By Glycosylation. K. Furukawa, Y. Ohkawa, Y. Yamauchi, K. Hamamura, Y. Ohmi, K. Furukawa, *J. Biochem.* **2012**, *151*, 573–578.

[19] Pathogen-Host Protein-Carbohydrate Interactions As The Basis Of Important Infections. K. A. Karlsson, *Adv. Exp. Med. Biol.* **2001**, *491*, 431–443.

[20] The Sweet Side Of Immune Evasion: Role Of Glycans In The Mechanisms Of Cancer Progression. A. F. F. R. Nardy, L. Freire-de-Lima, C. G. Freire-de-Lima, A. Morrot, *Front. Oncol.* **2016**, *6*, 178805.

[21] Protein-Carbohydrate Interactions, And Beyond …. K. De Schutter, E. J. M. Van Damme, *Molecules* **2015**, *20*, 15202.

[22] Protein-Carbohydrate Complexes: Binding Site Analysis, Prediction, Binding Affinity And Molecular Dynamics Simulations. K. Veluraja, N. R. S. Shanmugam, J. J. Blessy, R. A. Jeyaram, B. Lalithamaheswari, M. M. Gromiha, *Protein Interact. Comput. Methods, Anal. Appl.* **2020**, 299–332.

[23] Lectins As Promising Therapeutics For The Prevention And Treatment Of HIV And Other Potential Coinfections. M. Mazalovska, J. C. Kouokam, *Biomed Res. Int.* **2018**, *2018*.

[24] Drug Targeting To The Colon With Lectins And Neoglycoconjugates. T. Minko, *Adv. Drug Deliv. Rev.* **2004**, *56*, 491–509.

[25] Protein-Carbohydrate Interactions As Part Of Plant Defense And Animal Immunity. K. De Schutter, E. J. M. Van Damme, *Mol. 2015, Vol. 20, Pages 9029-9053* **2015**, *20*, 9029–9053.

[26] Molecular Recognition In C-Type Lectins: The Cases Of DC-SIGN, Langerin, MGL, And L-Sectin. P. Valverde, J. D. Martínez, F. J. Cañada, A. Ardá, J. Jiménez-Barbero, *ChemBioChem* **2020**, *21*, 2999–3025.

[27] Current Status On Therapeutic Molecules Targeting Siglec Receptors. M. P. Lenza, U. Atxabal, I. Oyenarte, J. Jiménez-Barbero, J. Ereño-Orbea, *Cells 2020, Vol. 9, Page 2691* **2020**, *9*, 2691.

[28] Galectins At A Glance. L. Johannes, R. Jacob, H. Leffler, *J. Cell Sci.* **2018**, *131*.

[29] Concept, Strategy And Realization Of Lectin-Based Glycan Profiling. J. Hirabayashi, *J. Biochem.* **2008**, *144*, 139–147.

[30] The Challenges Of Glycan Recognition With Natural And Artificial Receptors. S. Tommasone, F. Allabush, Y. K. Tagger, J. Norman, M. Köpf, J. H. R. Tucker, P. M. Mendes, *Chem. Soc. Rev.* **2019**, *48*, 5488–5505.

[31] Clusters, Bundles, Arrays And Lattices: Novel Mechanisms For Lectin–Saccharide-Mediated Cellular Interactions. C. F. Brewer, M. C. Miceli, L. G. Baum, *Curr. Opin. Struct. Biol.* **2002**, *12*, 616–623.

[32] Sweet Spots In Functional Glycomics. J. C. Paulson, O. Blixt, B. E. Collins, *Nat. Chem. Biol. 2006 25* **2006**, *2*, 238–248.

[33] Synthetic Multivalent Ligands In The Exploration Of Cell-Surface Interactions. L. L. Kiessling, J. E. Gestwicki, L. E. Strong, *Curr. Opin. Chem. Biol.* **2000**, *4*, 696–703.

[34] Lectins: Carbohydrate-Specific Proteins That Mediate Cellular Recognition. H. Lis, N. Sharon, *Chem. Rev.* **1998**, *98*, 637–674.

[35] Multivalent Glycoconjugates As Anti-Pathogenic Agents. A. Bernardi, J. Jiménez-Barbero, A. Casnati, C. De Castro, T. Darbre, F. Fieschi, J. Finne, H. Funken, K. E. Jaeger, M. Lahmann, T. K. Lindhorst, M. Marradi, P. Messner, A. Molinaro, P. V. Murphy, C. Nativi, S. Oscarson, S. Penadés, F. Peri, R. J. Pieters, O. Renaudet, J. L. Reymond, B. Richichi, J. Rojo, F. Sansone, C. Schäffer, W. Bruce Turnbull, T. Velasco-Torrijos, S. Vidal, S. Vincent, T. Wennekes, H. Zuilhof, A. Imberty, *Chem. Soc. Rev.* **2013**, *42*, 4709–4727.

[36] Heteromultivalent Glycooligomers As Mimetics Of Blood Group Antigens. K. S. Bücher, P. B. Konietzny, N. L. Snyder, L. Hartmann, *Chemistry* **2019**, *25*, 3301–3309.

[37] Glycomimetics Versus Multivalent Glycoconjugates For The Design Of High Affinity Lectin Ligands. S. Cecioni, A. Imberty, S. Vidal, *Chem. Rev.* **2015**, *115*, 525–561.

[38] CuAAC Synthesis Of Resorcin[4]Arene-Based Glycoclusters As Multivalent Ligands Of Lectins. Z. H. Soomro, S. Cecioni, H. Blanchard, J. P. Praly, A. Imberty, S. Vidal, S. E. Matthews, *Org. Biomol. Chem.* **2011**, *9*, 6587–6597.

[39] Functional Glyco-Nanogels For Multivalent Interaction With Lectins. J. S. J. Tang, S. Rosencrantz, L. Tepper, S. Chea, S. Klöpzig, A. Krüger-Genge, J. Storsberg, R. R. Rosencrantz, *Mol. 2019, Vol. 24, Page 1865* **2019**, *24*, 1865.

[40] Structural Basis Of Lectin-Carbohydrate Recognition. W. I. Weis, K. Drickamer, *Annu. Rev. Biochem.* **1996**, *65*, 441–473.

[41] Structural Characterization Of The DC-SIGN-LewisX Complex. K. Pederson, D. A. Mitchell, J. H. Prestegard, *Biochemistry* **2014**, *53*, 5700–5709.

[42] Lectin-Carbohydrate Interactions: Different Folds, Common Recognition Principles. S. Elgavish, B. Shaanan, *Trends Biochem. Sci.* **1997**, *22*, 462–467.

[43] Lectins: Carbohydrate-Specific Proteins That Mediate Cellular Recognition. H. Lis, N. Sharon, *Chem. Rev.* **1998**, *98*, 637–674.

[44] Constructing A Molecular Model Of The Interaction Between Antithrombin III And A Potent Heparin Analog. P. D. J. Grootenhuis, C. A. A. van Boeckel, *J. Am. Chem. Soc.* **1991**, *113*, 2743–2747.

[45] Computational Modeling Of The Sugar-Lectin Interaction. D. Neumann, C. M. Lehr, H. P. Lenhof, O. Kohlbacher, *Adv. Drug Deliv. Rev.* **2004**, *56*, 437–457.

[46] Bergenin: A Computationally Proven Promising Scaffold For Novel Galectin-3 Inhibitors. R. S. Jayakody, P. Wijewardhane, C. Herath, S. Perera, *J. Mol. Model.* **2018**, *24*.

[47] Computational Methods For Calculation Of Protein-Ligand Binding Affinities In Structure-Based Drug Design. Z. Dutkiewicz, *Phys. Sci. Rev.* **2022**, *7*, 933–968.

[48] QM And QM/MM Approaches To Evaluating Binding Affinities. K. E. Shaw, C. J. Woods, A. J. Mulholland, *Burger's Med. Chem. Drug Discov.* **2010**, 725–752.

[49] Ensembles Are Required To Handle Aleatoric And Parametric Uncertainty In Molecular Dynamics Simulation. M. Vassaux, S. Wan, W. Edeling, P. V. Coveney, *J. Chem. Theory Comput.* **2021**, *17*, 5187–5197.

[50] Rapid, Accurate, Precise And Reproducible Ligand-Protein Binding Free Energy Prediction. S. Wan, A. P. Bhati, S. J. Zasada, P. V. Coveney, *Interface Focus* **2020**, *10*.

[51] Molecular Simulations Of Carbohydrates And Protein-Carbohydrate Interactions: Motivation, Issues And Prospects. E. Fadda, R. J. Woods, *Drug Discov. Today* **2010**, *15*, 596–609.

[52] Involvement Of Water In Carbohydrate-Protein Binding: Concanavalin A Revisited. R. Kadirvelraj, B. L. Foley, J. D. Dyekjær, R. J. Woods, *J. Am. Chem. Soc.* **2008**, *130*, 16933–16942.

[53] Molecular Dynamics Study Of Pseudomonas Aeruginosa Lectin-II Complexed With Monosaccharides. N. K. Mishra, P. Kulhánek, L. Šnajdrová, M. Petřek, A. Imberty, J. Koča, *Proteins* **2008**, *72*, 382–392.

[54] Contribution Of Ligand Desolvation To Binding Thermodynamics In A Ligand–Protein Interaction. N. Shimokhina, A. Bronowska, S. W. Homans,

Angew. Chemie Int. Ed. **2006**, *45*, 6374–6376.

[55] Energetics Of Lectin-Carbohydrate Binding. A Microcalorimetric Investigation Of Concanavalin A-Oligomannoside Complexation. B. A. Williams, M. C. Chervenak, E. J. Toone, *J. Biol. Chem.* **1992**, *267*, 22907–22911.

[56] Binding Sub-Site Dissection Of A Carbohydrate-Binding Module Reveals The Contribution Of Entropy To Oligosaccharide Recognition At "Non-Primary" Binding Subsites. A. L. Van Bueren, A. B. Boraston, *J. Mol. Biol.* **2004**, *340*, 869–879.

[57] Carbohydrate-Protein Recognition: Molecular Dynamics Simulations And Free Energy Analysis Of Oligosaccharide Binding To Concanavalin A. R. A. Bryce, I. H. Hillier, J. H. Naismith, *Biophys. J.* **2001**, *81*, 1373–1388.

[58] Investigations On The Binding Specificity Of β-Galactoside Analogues With Human Galectin-1 Using Molecular Dynamics Simulations. J. Jino Blessy, N. R. Siva Shanmugam, K. Veluraja, M. Michael Gromiha, *J. Biomol. Struct. Dyn.* **2021**, *40*, 10094–10105.

[59] Thermodynamic And Structural Aspects Of Molecular Recognition In Mannose-Binding Protein Complexes: A Theoretical Study Over HRP-ArtinM Association. A. A. do E. Santo, G. T. Feliciano, *J. Mol. Model.* **2021**, *27*, 1–9.

[60] Recognition Of Selected Monosaccharides By Pseudomonas Aeruginosa Lectin II Analyzed By Molecular Dynamics And Free Energy Calculations. N. K. Mishra, Z. Kříž, M. Wimmerová, J. Koča, *Carbohydr. Res.* **2010**, *345*, 1432–1441.

[61] A Priori Evaluation Of Aqueous Polarization Effects Through Monte Carlo QM-MM Simulations. J. Gao, X. Xia, *Science.* **1992**, *258*, 631–635.

[62] Recent Developments Of The Quantum Chemical Cluster Approach For Modeling Enzyme Reactions. P. E. M. Siegbahn, F. Himo, *J. Biol. Inorg. Chem.* **2009**, *14*, 643–651.

[63] Do Quantum Mechanical Energies Calculated For Small Models Of Protein-Active Sites Converge. L. Hu, J. Eliasson, J. Heimdal, U. Ryde, *J. Phys. Chem. A* **2009**, *113*, 11793–11800.

[64] Carbohydrate–Protein Interactions. S. Pérez, I. Tvaroška, in *Adv. Carbohydr. Chem. Biochem.*, Academic Press, **2014**, *71*, 9–136.

[65] Protein Coagulation And Its Reversal : The Preparation Of Insoluble Globin, Soluble Globin And Heme. M. L. Anson, A. E. Mirsky, *J. Gen. Physiol.* **1930**, *13*, 469.

[66] Conformation Changes Of Proteins. R. Lumry, H. Eyring, *J. Phys. Chem.* **1954**, *58*, 110–120.

[67] The Kinetics Of Formation Of Native Ribonuclease During Oxidation Of The

Reduced Polypeptide Chain.. C. B. Anfinsen, E. Haber, M. Sela, F. H. White, *Proc. Natl. Acad. Sci. U. S. A.* **1961**, *47*, 1309–1314.

[68] Principles That Govern The Folding Of Protein Chains. C. B. Anfinsen, *Science.* **1973**, *181*, 223–230.

[69] Experimental And Theoretical Aspects Of Protein Folding. C. B. Anfinsen, H. A. Scheraga, *Adv. Protein Chem.* **1975**, *29*, 205–300.

[70] Effective Energy Functions For Protein Structure Prediction. T. Lazaridis, M. Karplus, *Curr. Opin. Struct. Biol.* **2000**, *10*, 139–145.

[71] The Protein Folding Problem. K. A. Dill, S. B. Ozkan, M. S. Shell, T. R. Weikl, **2008**, *37*, 289–316.

[72] How To Fold Graciously. C. Levinthal, *Mossbauer Spectrosc. Biol. Syst. Proc.* **1969**, *67*, 22–24.

[73] Funnels, Pathways, And The Energy Landscape Of Protein Folding: A Synthesis. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, P. G. Wolynes, *Proteins Struct. Funct. Bioinforma.* **1995**, *21*, 167–195.

[74] Dominant Forces In Protein Folding. K. A. Dill, *Biochemistry* **1990**, *29*, 7133–7155.

[75] Do All Backbone Polar Groups In Proteins Form Hydrogen Bonds?. P. J. Fleming, G. D. Rose, *Protein Sci.* **2005**, *14*, 1911.

[76] Advances In Protein Structure Prediction And Design. B. Kuhlman, P. Bradley, *Nat. Rev. Mol. Cell Biol. 2019 2011* **2019**, *20*, 681–697.

[77] Exploring The Structure And Function Paradigm. O. C. Redfern, B. Dessailly, C. A. Orengo, *Curr. Opin. Struct. Biol.* **2008**, *18*, 394–402.

[78] Highly Accurate Protein Structure Prediction With AlphaFold. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nat.* **2021**, *596*, 583–589.

[79] Accurate Prediction Of Protein Structures And Interactions Using A Three-Track Neural Network. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. Dustin Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. Van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. Christopher Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, *Science.* **2021**, *373*, 871–876.

[80] Sequence-Structure-Function Relationships In The Microbial Protein Universe. J. Koehler Leman, P. Szczerbiak, P. D. Renfrew, V. Gligorijevic, D. Berenberg, T. Vatanen, B. C. Taylor, C. Chandler, S. Janssen, A. Pataki, N. Carriero, I. Fisk, R. J. Xavier, R. Knight, R. Bonneau, T. Kosciolek, *Nat. Commun.* **2023**, *14*, 1–11.

[81] Advances In Homology Protein Structure Modeling. Z. Xiang, *Curr. Protein Pept. Sci.* **2006**, *7*, 217–227.

[82] Gapped BLAST And PSI-BLAST: A New Generation Of Protein Database Search Programs. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

[83] Profile Hidden Markov Models.. S. R. Eddy, *Bioinformatics* **1998**, *14*, 755–763.

[84] HHblits: Lightning-Fast Iterative Protein Sequence Searching By HMM-HMM Alignment. M. Remmert, A. Biegert, A. Hauser, J. Söding, *Nat. Methods 2011 92* **2011**, *9*, 173–175.

[85] Physics-Based Protein-Structure Prediction Using A Hierarchical Protocol Based On The UNRES Force Field: Assessment In Two Blind Tests. S. Ołziej, C. Czaplewski, A. Liwo, M. Chinchio, M. Nanias, J. A. Vila, M. Khalili, Y. A. Arnautova, A. Jagielska, M. Makowski, H. D. Schafroth, R. Kaźmierkiewicz, D. R. Ripoll, J. Pillardy, J. A. Saunders, Y. K. Kang, K. D. Gibson, H. A. Scheraga, *Proc. Natl. Acad. Sci.* **2005**, *102*, 7547–7552.

[86] Blind Test Of Physics-Based Prediction Of Protein Structures. M. Scott Shell, S. Banu Ozkan, V. Voelz, G. A. Wu, K. A. Dill, *Biophys. J.* **2009**, *96*, 917–924.

[87] Evolutionary-Scale Prediction Of Atomic-Level Protein Structure With A Language Model. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, *Science.* **2023**, *379*, 1123–1130.

[88] High-Resolution De Novo Structure Prediction From Primary Sequence. R. Wu, F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, Z. Wu, Q. Xie, B. Berger, J. Ma, J. Peng, *bioRxiv* **2022**, 2022.07.21.500999.

[89] Principles Of Protein Stability And Their Application In Computational Design. A. Goldenzweig, S. J. Fleishman, *Annu. Rev. Biochem.* **2018**, *87*, 105–129.

[90] Arming Yourself For The In Silico Protein Design Revolution. S. P. Walker, V. V. B. Yallapragada, M. Tangney, *Trends Biotechnol.* **2021**, *39*, 651–664.

[91] Automated Selection Of Stabilizing Mutations In Designed And Natural Proteins. B. Borgo, J. J. Havranek, *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 1494–1499.

[92] Design, Structure And Stability Of A Hyperthermophilic Protein Variant. S. M. Malakauskas, S. L. Mayo, *Nat. Struct. Biol. 1998 56* **1998**, *5*, 470–475.

[93] A Large Scale Test Of Computational Protein Design: Folding And Stability Of Nine Completely Redesigned Globular Proteins. G. Dantas, B. Kuhlman, D. Callender, M. Wong, D. Baker, *J. Mol. Biol.* **2003**, *332*, 449–460.

[94] FireProt: Energy- And Evolution-Based Computational Design Of Thermostable Multiple-Point Mutants. D. Bednar, K. Beerens, E. Sebestova, J. Bendl, S. Khare, R. Chaloupkova, Z. Prokop, J. Brezovsky, D. Baker, J. Damborsky, *PLOS Comput. Biol.* **2015**, *11*, e1004556.

[95] Protein Stability: Computation, Sequence Statistics, And New Experimental Methods. T. J. Magliery, *Curr. Opin. Struct. Biol.* **2015**, *33*, 161–168.

[96] The Consensus Concept For Thermostability Engineering Of Proteins. M. Lehmann, L. Pasamontes, S. F. Lassen, M. Wyss, *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* **2000**, *1543*, 408–415.

[97] Automated Structure- And Sequence-Based Design Of Proteins For High Bacterial Expression And Stability. A. Goldenzweig, M. Goldsmith, S. E. Hill, O. Gertman, P. Laurino, Y. Ashani, O. Dym, T. Unger, S. Albeck, J. Prilusky, R. L. Lieberman, A. Aharoni, I. Silman, J. L. Sussman, D. S. Tawfik, S. J. Fleishman, *Mol. Cell* **2016**, *63*, 337–346.

[98] Progress In Computational Protein Design. S. M. Lippow, B. Tidor, *Curr. Opin. Biotechnol.* **2007**, *18*, 305–311.

[99] Recapitulation And Design Of Protein Binding Peptide Structures And Sequences. V. D. Sood, D. Baker, *J. Mol. Biol.* **2006**, *357*, 917–927.

[100] Affinity Enhancement Of An In Vivo Matured Therapeutic Antibody Using Structure-Based Computational Design. L. A. Clark, P. A. Boriack-Sjodin, J. Eldredge, C. Fitch, B. Friedman, K. J. M. Hanf, M. Jarpe, S. F. Liparoto, Y. Li, A. Lugovskoy, S. Miller, M. Rushe, W. Sherman, K. Simon, H. Van Vlijmen, *Protein Sci.* **2006**, *15*, 949.

[101] Engineered Antibody Fc Variants With Enhanced Effector Function. G. A. Lazar, W. Dang, S. Karki, O. Vafa, J. S. Peng, L. Hyun, C. Chan, H. S. Chung, A. Eivazi, S. C. Yoder, J. Vielmetter, D. F. Carmichael, R. J. Hayes, B. I. Dahiyat, *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 4005–4010.

[102] Main Structural Targets For Engineering Lipase Substrate Specificity. S. H. Albayati, M. Masomian, S. N. H. Ishak, M. S. B. M. Ali, A. L. Thean, F. B. M. Shariff, N. D. B. M. Noor, R. N. Z. R. A. Rahman, *Catal. 2020, Vol. 10, Page 747* **2020**, *10*, 747.

[103] Expanding The Range Of Substrate Acceptance Of Enzymes: Combinatorial Active-Site Saturation Test. M. T. Reetz, M. Bocola, J. D. Carballeira, D. Zha, A. Vogel, *Angew. Chemie Int. Ed.* **2005**, *44*, 4192–4196.

[104] O-/ N-/ S-Specificity In Glycosyltransferase Catalysis: From Mechanistic

Understanding To Engineering. D. Teze, J. Coines, F. Fredslund, K. D. Dubey, G. N. Bidart, P. D. Adams, J. E. Dueber, B. Svensson, C. Rovira, D. H. Welner, *ACS Catal.* **2021**, *11*, 1810–1815.

[105] Computational Redesign Of Endonuclease DNA Binding And Cleavage Specificity. J. Ashworth, J. J. Havranek, C. M. Duarte, D. Sussman, R. J. Monnat, B. L. Stoddard, D. Baker, *Nature* **2006**, *441*, 656–659.

[106] Rational Design Of New Binding Specificity By Simultaneous Mutagenesis Of Calmodulin And A Target Peptide. D. F. Green, A. T. Dennis, P. S. Fam, B. Tidor, A. Jasanoff, *Biochemistry* **2006**, *45*, 12547–12559.

[107] OptZyme: Computational Enzyme Redesign Using Transition State Analogues. M. J. Grisewood, N. P. Gifford, R. J. Pantazes, Y. Li, P. C. Cirino, M. J. Janik, C. D. Maranas, *PLoS One* **2013**, *8*, e75358.

[108] Computational Enzyme Engineering Pipelines For Optimized Production Of Renewable Chemicals. M. Scherer, S. J. Fleishman, P. R. Jones, T. Dandekar, E. Bencurova, *Front. Bioeng. Biotechnol.* **2021**, *9*, 673005.

[109] Mechanism-Guided Computational Design Of ω-Transaminase By Reprograming Of High-Energy-Barrier Steps. L. Yang, K. Zhang, M. Xu, Y. Xie, X. Meng, H. Wang, D. Wei, *Angew. Chemie Int. Ed.* **2022**, *61*, e202212555.

[110] Design Of A Novel Globular Protein Fold With Atomic-Level Accuracy. B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, D. Baker, *Science.* **2003**, *302*, 1364–1368.

[111] Rosetta3: An Object-Oriented Software Suite For The Simulation And Design Of Macromolecules. A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y. E. A. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, P. Bradley, *Methods Enzymol.* **2011**, *487*, 545–574.

[112] Automated Design Of Efficient And Functionally Diverse Enzyme Repertoires. O. Khersonsky, R. Lipsh, Z. Avizemer, Y. Ashani, M. Goldsmith, H. Leader, O. Dym, S. Rogotner, D. L. Trudeau, J. Prilusky, P. Amengual-Rigo, V. Guallar, D. S. Tawfik, S. J. Fleishman, *Mol. Cell* **2018**, *72*, 178-186.e5.

[113] Computationally Designed Hyperactive Cas9 Enzymes. P. D. Vos, G. Rossetti, J. L. Mantegna, S. J. Siira, A. P. Gandadireja, M. Bruce, S. A. Raven, O. Khersonsky, S. J. Fleishman, A. Filipovska, O. Rackham, *Nat. Commun. 2022 131* **2022**, *13*, 1–11.

[114] Computational Redesign Of Enzymes For Regio- And Enantioselective Hydroamination. R. Li, H. J. Wijma, L. Song, Y. Cui, M. Otzen, Y. Tian, J. Du, T. Li, D. Niu, Y. Chen, J. Feng, J. Han, H. Chen, Y. Tao, D. B. Janssen, B. Wu,

Nat. Chem. Biol. 2018 147 **2018**, *14*, 664–670.

[115] Computation-Aided Engineering Of Cytochrome P450 For The Production Of Pravastatin. M. A. Ashworth, E. Bombino, R. M. De Jong, H. J. Wijma, D. B. Janssen, K. J. McLean, A. W. Munro, *ACS Catal.* **2022**, *12*, 15028–15044.

[116] Coupling Protein Side-Chain And Backbone Flexibility Improves The Re-Design Of Protein-Ligand Specificity. N. Ollikainen, R. M. de Jong, T. Kortemme, *PLOS Comput. Biol.* **2015**, *11*, e1004335.

[117] Stereodivergent Protein Engineering Of A Lipase To Access All Possible Stereoisomers Of Chiral Esters With Two Stereocenters. J. Xu, Y. Cen, W. Singh, J. Fan, L. Wu, X. Lin, J. Zhou, M. Huang, M. T. Reetz, Q. Wu, *J. Am. Chem. Soc.* **2019**, *141*, 7934–7945.

[118] Computationally Supported Inversion Of Ketoreductase Stereoselectivity. E. Delgado-Arciniega, H. J. Wijma, C. Hummel, D. B. Janssen, *ChemBioChem* **2023**, *24*, e202300032.

[119] Characterization Of A Helical Protein Designed From First Principles. L. Regan, W. F. Degrado, *Science.* **1988**, *241*, 976–978.

[120] The Design Of A Four-Helix Bundle Protein. W. F. DeGrado, L. Regan, S. P. Ho, *Cold Spring Harb. Symp. Quant. Biol.* **1987**, *52*, 521–526.

[121] Repacking Protein Cores With Backbone Freedom: Structure Prediction For Coiled Coils.. P. B. Harbury, B. Tidor, P. S. Kim, *Proc. Natl. Acad. Sci.* **1995**, *92*, 8408–8412.

[122] Essentials Of De Novo Protein Design: Methods And Applications. E. Marcos, D. A. Silva, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1374.

[123] Improvements To Robotics-Inspired Conformational Sampling In Rosetta. A. Stein, T. Kortemme, *PLoS One* **2013**, *8*, e63090.

[124] Accurate De Novo Design Of Hyperstable Constrained Peptides. G. Bhardwaj, V. K. Mulligan, C. D. Bahl, J. M. Gilmore, P. J. Harvey, O. Cheneval, G. W. Buchko, S. V. S. R. K. Pulavarti, Q. Kaas, A. Eletsky, P. S. Huang, W. A. Johnsen, P. J. Greisen, G. J. Rocklin, Y. Song, T. W. Linsky, A. Watkins, S. A. Rettie, X. Xu, L. P. Carter, R. Bonneau, J. M. Olson, E. Coutsias, C. E. Correnti, T. Szyperski, D. J. Craik, D. Baker, *Nat. 2016 5387625* **2016**, *538*, 329–335.

[125] Sub-Angstrom Accuracy In Protein Loop Reconstruction By Robotics-Inspired Conformational Sampling. D. J. Mandell, E. A. Coutsias, T. Kortemme, *Nat. Methods* **2009**, *6*, 551–552.

[126] Design Of Structurally Distinct Proteins Using Strategies Inspired By Evolution. T. M. Jacobs, B. Williams, T. Williams, X. Xu, A. Eletsky, J. F. Federizon, T. Szyperski, B. Kuhlman, *Science* **2016**, *352*, 687–690.

[127] Computational Design Of An Enzyme Catalyst For A Stereoselective Bimolecular Diels-Alder Reaction. J. B. Siegel, A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St.Clair, J. L. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk, F. E. Michael, D. Baker, *Science* **2010**, *329*, 309.

[128] Kemp Elimination Catalysts By Computational Enzyme Design. D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, D. Baker, *Nat. 2008 4537192* **2008**, *453*, 190–195.

[129] De Novo Computational Design Of Retro-Aldol Enzymes. L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, D. Hilvert, K. N. Houk, B. L. Stoddard, D. Baker, *Science.* **2008**, *319*, 1387–1391.

[130] De Novo Protein Design By Deep Network Hallucination. I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera, F. DiMaio, L. Carter, C. M. Chow, G. T. Montelione, D. Baker, *Nat.* **2021**, *600*, 547–552.

[131] Scaffolding Protein Functional Sites Using Deep Learning. J. Wang, S. Lisanza, D. Juergens, D. Tischer, J. L. Watson, K. M. Castro, R. Ragotte, A. Saragovi, L. F. Milles, M. Baek, I. Anishchenko, W. Yang, D. R. Hicks, M. Expòsit, T. Schlichthaerle, J.-H. Chun, J. Dauparas, N. Bennett, B. I. M. Wicky, A. Muenks, F. DiMaio, B. Correia, S. Ovchinnikov, D. Baker, *Science* **2022**, *377*, 387–394.

[132] Robust Deep Learning Based Protein Sequence Design Using ProteinMPNN. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, *Science.* **2022**, *378*, 49-56.

[133] De Novo Design Of Protein Structure And Function With RFdiffusion. J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, *Nat.* **2023**, 1–3.

[134] Joint Generation Of Protein Sequence And Structure With Rosettafold Sequence Space Diffusion. S. L. Lisanza, J. M. Gershon, S. Tipps, L. Arnoldt, S. Hendel, J. N. Sims, X. Li, D. Baker, *bioRxiv* **2023**, 2023.05.08.539766.

[135] Galectin-4 N-Terminal Domain: Binding Preferences Toward A And B Antigens With Different Peripheral Core Presentations. J. I. Quintana, S. Delgado, R. Núñez-Franco, F. J. Cañada, G. Jiménez-Osés, J. Jiménez-Barbero, A. Ardá, *Front. Chem.* **2021**, *9*, 193.

[136] Minimizing The Entropy Penalty For Ligand Binding: Lessons From The

Molecular Recognition Of The Histo Blood-Group Antigens By Human Galectin-3. A. Gimeno, S. Delgado, P. Valverde, S. Bertuzzi, M. A. Berbís, J. Echavarren, A. Lacetera, S. Martín-Santamaría, A. Surolia, F. J. Cañada, J. Jiménez-Barbero, A. Ardá, *Angew. Chemie Int. Ed.* **2019**, *58*, 7268–7272.

[137] Unravelling The Time Scale Of Conformational Plasticity And Allostery In Glycan Recognition By Human Galectin-1. S. Bertuzzi, A. Gimeno, R. Núñez-Franco, G. Bernardo-Seisdedos, S. Delgado, G. Jiménez-Osés, O. Millet, J. Jiménez-Barbero, A. Ardá, *Chem. - A Eur. J.* **2020**, *26*, 15643–15653.

[138] The Two Domains Of Human Galectin-8 Bind Sialyl- And Fucose-Containing Oligosaccharides In An Independent Manner. A 3D View By Using NMR. M. Gómez-Redondo, S. Delgado, R. Núñez-Franco, G. Jiménez-Osés, A. Ardá, J. Jiménez-Barbero, A. Gimeno, *RSC Chem. Biol.* **2021**, *2*, 932–941.

[139] Galectins. R. D. Cummings, F.-T. Liu, G. A. Rabinovich, S. R. Stowell, G. R. Vasta, *Carbohydrates Chem. Biol.* **2022**, *4–4*, 625–647.

[140] Distal Mutations Shape Substrate-Binding Sites During Evolution Of A Metallo-Oxidase Into A Laccase. V. Brissos, P. T. Borges, R. Núñez-Franco, M. F. Lucas, C. Frazão, E. Monza, L. Masgrau, T. N. Cordeiro, L. O. Martins, *ACS Catal.* **2022**, *12*, 5022–5035.

[141] In Silico Study Of Allosteric Communication Networks In GPCR Signaling Bias. A. Morales-Pastor, F. Nerín-Fonz, D. Aranda-García, M. Dieguez-Eceolaza, B. Medel-Lacruz, M. Torrens-Fontanals, A. Peralta-García, J. Selent, *Int. J. Mol. Sci.* **2022**, *23*, 7809.

[142] Grid Inhomogeneous Solvation Theory: Hydration Structure And Thermodynamics Of The Miniature Receptor Cucurbit[7]Uril. C. N. Nguyen, T. Kurtzman Young, M. K. Gilson, *J. Chem. Phys.* **2012**, *137*, 973–980.

[143] Quantifying The Role Of Water In Protein–Carbohydrate Interactions. S. M. Tschampel, R. J. Woods, *J. Phys. Chem. A* **2003**, *107*, 9175.

[144] Water Structuring Properties Of Carbohydrates, Molecular Dynamics Studies On 1,5-Anhydro-D-Fructose. J. Behler, D. W. Price, M. G. B. Drew, *Phys. Chem. Chem. Phys.* **2001**, *3*, 588–601.

[145] Solvation Thermodynamic Mapping Of Molecular Surfaces In AmberTools: GIST. S. Ramsey, C. Nguyen, R. Salomon-Ferrer, R. C. Walker, M. K. Gilson, T. Kurtzman, *J. Comput. Chem.* **2016**, *37*, 2029–2037.

[146] Sialyl Lewisx: A "Pre-Organized Water Oligomer"?. F. P. C. Binder, K. Lemme, R. C. Preston, B. Ernst, *Angew. Chemie Int. Ed.* **2012**, *51*, 7327–7331.

[147] The Mean Hydration Of Carbohydrates As Studied By Normalized Two-Dimensional Radial Pair Distributions. C. Andersson, S. B. Engelsen, *J. Mol. Graph. Model.* **1999**, *17*, 101–105.

[148] Water Sculpts The Distinctive Shapes And Dynamics Of The Tumor-Associated Carbohydrate Tn Antigens: Implications For Their Molecular Recognition. I. A. Bermejo, I. Usabiaga, I. Compañón, J. Castro-López, A. Insausti, J. A. Fernández, A. Avenoza, J. H. Busto, J. Jiménez-Barbero, J. L. Asensio, J. M. Peregrina, G. Jiménez-Osés, R. Hurtado-Guerrero, E. J. Cocinero, F. Corzana, *J. Am. Chem. Soc.* **2018**, *140*, 9952–9960.

[149] New Insights Into α-GalNAc-Ser Motif: Influence Of Hydrogen Bonding Versus Solvent Interactions On The Preferred Conformation. F. Corzana, J. H. Busto, G. Jiménez-Osés, J. L. Asensio, J. Jiménez-Barbero, J. M. Peregrina, A. Avenoza, *J. Am. Chem. Soc.* **2006**, *128*, 14640–14648.

[150] AMBER 2020, D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F.Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York and P.A. Kollman, **2020**, University of California, San Francisco.

[151] The PyMOL Molecular Graphics System, Version 2.4. Schrödinger LLC.

[152] Ff14SB: Improving The Accuracy Of Protein Side Chain And Backbone Parameters From Ff99SB. J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

[153] GLYCAM06: A Generalizable Biomolecular Force Field. Carbohydrates. K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R. Daniels, B. L. Foley, R. J. Woods, *J. Comput. Chem.* **2008**, *29*, 622–655.

[154] Comparison Of Simple Potential Functions For Simulating Liquid Water. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1998**, *79*, 926.

[155] Molecular Dynamics Simulations At Constant Pressure And/Or Temperature. H. C. Andersen, *undefined* **1980**, *72*, 2384–2393.

[156] Settle: An Analytical Version Of The SHAKE And RATTLE Algorithm For Rigid Water Models. S. Miyamoto, P. A. Kollman, *J. Comput. Chem.* **1992**, *13*, 8.

[157] Particle Mesh Ewald: An N·log(N) Method For Ewald Sums In Large Systems. T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* **1998**, *98*, 10089.

[158] A Density-Based Algorithm For Discovering Clusters In Large Spatial Databases With Noise. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, **1996**.

[159] PTRAJ And CPPTRAJ: Software For Processing And Analysis Of Molecular Dynamics Trajectory Data. D. R. Roe, T. E. Cheatham, *J. Chem. Theory Comput.* **2013**, *9*.

[160] A Density-Based Algorithm For Discovering Clusters In Large Spatial Databases With Noise. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, in *Proc. 2nd Int. Conf. Knowl. Discov. Data Min.*, **1996**. 226-131.

[161] C-Type Lectins. R. D. Cummings, R. P. McEver, *Essentials Glycobiol. Cold Spring Harb. Lab. Press* **2009**.

[162] DC-SIGN, A Dendritic Cell–Specific HIV-1-Binding Protein That Enhances Trans-Infection Of T Cells. T. B. H. Geijtenbeek, D. S. Kwon, R. Torensma, S. J. Van Vliet, G. C. F. Van Duijnhoven, J. Middel, I. L. M. H. A. Cornelissen, H. S. L. M. Nottet, V. N. KewalRamani, D. R. Littman, C. G. Figdor, Y. Van Kooyk, *Cell* **2000**, *100*, 587–597.

[163] Dendritic Cells And C-Type Lectin Receptors: Coupling Innate To Adaptive Immune Responses. S. J. Van Vliet, J. J. García-Vallejo, Y. Van Kooyk, *Immunol. Cell Biol.* **2008**, *86*, 580–587.

[164] Recent Trends In Quantum Chemical Modeling Of Enzymatic Reactions. F. Himo, *J. Am. Chem. Soc.* **2017**, *139*, 6780–6786.

[165] Quantum Chemical Modeling Of Enzyme Active Sites And Reaction Mechanisms. F. Himo, *Theor. Chem. Acc.* **2006**, *116*, 232–240.

[166] Structural Basis For Distinct Ligand-Binding And Targeting Properties Of The Receptors DC-SIGN And DC-SIGNR. Y. Guo, H. Feinberg, E. Conroy, D. A. Mitchell, R. Alvarez, O. Blixt, M. E. Taylor, W. I. Weis, K. Drickamer, *Nat. Struct. Mol. Biol.* **2004**, *11*, 591–598.

[167] Molecular Basis Of The Differences In Binding Properties Of The Highly Related C-Type Lectins DC-SIGN And L-SIGN To Lewis X Trisaccharide And Schistosoma Mansoni Egg Antigens*. E. Van Liempt, A. Imberty, C. M. C. Bank, S. J. Van Vliet, Y. Van Kooyk, T. B. H. Geijtenbeek, I. Van Die, **2004**, *279*, 32.

[168] Molecular Insights Into DC-SIGN Binding To Self-Antigens: The Interaction With The Blood Group A/B Antigens. P. Valverde, S. Delgado, J. D. Martínez, J. B. Vendeville, J. Malassis, B. Linclau, N. C. Reichardt, F. J. Cañada, J. Jiménez-Barbero, A. Ardá, *ACS Chem. Biol.* **2019**, *14*, 1660–1671.

[169] Mono- And Di-Fucosylated Glycans Of The Parasitic Worm S. Mansoni Are Recognized Differently By The Innate Immune Receptor DC-SIGN. A. D. Srivastava, L. Unione, M. A. Wolfert, P. Valverde, A. Ardá, J. Jiménez-Barbero, G. J. Boons, *Chemistry* **2020**, *26*, 15605.

[170] Complete Relaxation And Conformational Exchange Matrix (CORCEMA) Analysis Of Intermolecular Saturation Transfer Effects In Reversibly Forming

Ligand–Receptor Complexes. V. Jayalakshmi, N. R. Krishna, *J. Magn. Reson.* **2002**, *155*, 106–118.

[171] The Interaction Of Fluorinated Glycomimetics With DC-SIGN: Multiple Binding Modes Disentangled By The Combination Of NMR Methods And MD Simulations. J. D. Martínez, A. S. Infantino, P. Valverde, T. Diercks, S. Delgado, N.-C. Reichardt, A. Ardá, F. J. Cañada, S. Oscarson, J. Jiménez-Barbero, *Pharmaceuticals* **2020**, *13*, 179.

[172] Unraveling Sugar Binding Modes To DC-SIGN By Employing Fluorinated Carbohydrates. J. D. Martínez, P. Valverde, S. Delgado, C. Romanò, B. Linclau, N. C. Reichardt, S. Oscarson, A. Ardá, J. Jiménez-Barbero, F. J. Cañada, *Molecules* **2019**, *24*, 2337.

[173] The Interaction Of Fluorinated Glycomimetics With DC-SIGN: Multiple Binding Modes Disentangled By The Combination Of NMR Methods And MD Simulations. J. D. Martínez, A. S. Infantino, P. Valverde, T. Diercks, S. Delgado, N. C. Reichardt, A. Ardá, F. J. Cañada, S. Oscarson, J. Jiménez-Barbero, *Pharm. 2020, Vol. 13, Page 179* **2020**, *13*, 179.

[174] Low Mode Search. An Efficient, Automated Computational Method For Conformational Analysis: Application To Cyclic And Acyclic Alkanes And Cyclic Peptides. I. Kolossváry, W. C. Guida, *J. Am. Chem. Soc.* **1996**, *118*, 5011–5019.

[175] Schrödinger Release 2022-3: MacroModel, Schrödinger. **2021**

[176] Gaussian 16, Revision C.01, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian, Inc., Wallingford CT, **2016**

[177] The M06 Suite Of Density Functionals For Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, And Transition Elements: Two New Functionals And Systematic Testing Of Four M06-Class Functionals And 12 Other Function. Y. Zhao, D. G. Truhlar, *Theor. Chem. Acc.* **2008**, *120*, 215–241.

[178] Continuous Surface Charge Polarizable Continuum Models Of Solvation. I. General Formalism. G. Scalmani, M. J. Frisch, *J. Chem. Phys.* **2010**, *132*.

[179] The Role Of Long Ranged Forces In Determining The Structure And Properties Of Liquid Water. T. A. Andrea, W. C. Swope, H. C. Andersen, *J. Chem. Phys.* **1983**, *79*, 4576–4584.

[180] A Microscale Protein NMR Sample Screening Pipeline. P. Rossi, G. V. T. Swapna, Y. J. Huang, J. M. Aramini, C. Anklin, K. Conover, K. Hamilton, R. Xiao, T. B. Acton, A. Ertekin, J. K. Everett, G. T. Montelione, *J. Biomol. NMR* **2010**, *46*, 11–22.

[181] Engineering Protein-Based Therapeutics Through Structural And Chemical Design. S. B. Ebrahimi, D. Samanta, *Nat. Commun.* **2023**, *14*, 1–11.

[182] FDA Approval: Blinatumomab. D. Przepiorka, C. W. Ko, A. Deisseroth, C. L. Yancey, R. Candau-Chacon, H. J. Chiu, B. J. Gehrke, C. Gomez-Broughton, R. C. Kane, S. Kirshner, N. Mehrotra, T. K. Ricks, D. Schmiel, P. Song, P. Zhao, Q. Zhou, A. T. Farrell, R. Pazdur, *Clin. Cancer Res.* **2015**, *21*, 4035–4039.

[183] Pegademase Bovine (PEG-ADA) For The Treatment Of Infants And Children With Severe Combined Immunodeficiency (SCID). C. Booth, H. B. Gaspar, *Biologics* **2009**, *3*, 349.

[184] Protein Therapeutics: A Summary And Pharmacological Classification. B. Leader, Q. J. Baca, D. E. Golan, *Nat. Rev. Drug Discov. 2007 71* **2008**, *7*, 21–39.

[185] The Two Faces Of Protein Misfolding: Gain- And Loss-Of-Function In Neurodegenerative Diseases. K. F. Winklhofer, J. Tatzelt, C. Haass, *EMBO J.* **2008**, *27*, 336–349.

[186] Therapeutic Potential Of Proteasome Inhibitors In Congenital Erythropoietic Porphyria. J. M. Blouin, Y. Duchartre, P. Costet, M. Lalanne, C. Ged, A. Lain, O. Millet, H. De Verneuil, E. Richard, *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 18238–18243.

[187] Adapting Proteostasis For Disease Intervention. W. E. Balch, R. I. Morimoto, A. Dillin, J. W. Kelly, *Science.* **2008**, *319*, 916–919.

[188] Drug Development: Longer-Lived Proteins. S. Kontos, J. A. Hubbell, *Chem. Soc. Rev.* **2012**, *41*, 2686–2695.

[189] Recombinant Factor VIIa (Eptacog Alfa): A Pharmacoeconomic Review Of Its Use In Haemophilia In Patients With Inhibitors To Clotting Factors VIII Or IX. K. A. Lyseng-Williamson, G. L. Plosker, *Pharmacoeconomics* **2007**, *25*, 1007–1029.

[190] Insulin Detemir: A Review Of Its Use In The Management Of Diabetes Mellitus. G. M. Keating, *Drugs* **2012**, *72*, 2255–2287.

[191] Current Perspectives On Interferon Beta-1b For The Treatment Of Multiple

Sclerosis. M. Marziniak, S. Meuth, *Adv. Ther.* **2014**, *31*, 915.

[192] Insulin Glargine: A Systematic Review Of A Long-Acting Insulin Analogue. F. Wang, J. M. Carabino, C. M. Vergara, *Clin. Ther.* **2003**, *25*, 1541–1577.

[193] Clinical Pharmacokinetics And Pharmacodynamics Of Insulin Glulisine. R. H. A. Becker, A. D. Frick, *Clin. Pharmacokinet.* **2008**, *47*, 7–20.

[194] Protein Folding, Misfolding, Stability, And Aggregation: An Overview. R. M. Murphy, A. M. Tsai, *Misbehaving Proteins Protein (Mis)Folding, Aggregation, Stab.* **2006**, 3–13.

[195] Tuning Intracellular Homeostasis Of Human Uroporphyrinogen III Synthase By Enzyme Engineering At A Single Hotspot Of Congenital Erythropoietic Porphyria. F. ben Bdira, E. González, P. Pluta, A. Laín, A. Sanz-Parra, J. M. Falcon-Perez, O. Millet, *Hum. Mol. Genet.* **2014**, *23*, 5805–5813.

[196] Stabilizing The CH2 Domain Of An Antibody By Engineering In An Enhanced Aromatic Sequon. W. Chen, L. Kong, S. Connelly, J. M. Dendle, Y. Liu, I. A. Wilson, E. T. Powers, J. W. Kelly, *ACS Chem. Biol.* **2016**, *11*, 1852–1861.

[197] Therapeutic Prospects For Friedreich's Ataxia. S. Zhang, M. Napierala, J. S. Napierala, *Trends Pharmacol. Sci.* **2019**, *40*, 229–233.

[198] In Vivo Maturation Of Human Frataxin. I. Condò, N. Ventura, F. Malisan, A. Rufini, B. Tomassini, R. Testi, *Hum. Mol. Genet.* **2007**, *16*, 1534–1540.

[199] Crystal Structure Of Human Frataxin. D. P. Sirano, R. Shigeta, Y. I. Chi, M. Ristow, S. E. Shoelson, *J. Biol. Chem.* **2000**, *275*, 30753–30756.

[200] Protoporphyrin IX Binds To Iron(II)-Loaded And To Zinc-Loaded Human Frataxin. G. Bernardo-Seisdedos, A. Schedlbauer, T. Pereira-Ortuzar, J. M. Mato, O. Millet, *Life* **2023**, *13*, 222.

[201] Selected Missense Mutations Impair Frataxin Processing In Friedreich Ataxia. E. Clark, J. S. Butler, C. J. Isaacs, M. Napierala, D. R. Lynch, *Ann. Clin. Transl. Neurol.* **2017**, *4*, 575–584.

[202] Missense Mutations Linked To Friedreich Ataxia Have Different But Synergistic Effects On Mitochondrial Frataxin Isoforms. H. Li, O. Gakh, D. Y. Smith, W. K. Ranatunga, G. Isaya, *J. Biol. Chem.* **2013**, *288*, 4116–4127.

[203] Rapid And Complete Reversal Of Sensory Ataxia By Gene Therapy In A Novel Model Of Friedreich Ataxia. F. Piguet, C. de Montigny, N. Vaucamps, L. Reutenauer, A. Eisenmann, H. Puccio, *Mol. Ther.* **2018**, *26*, 1940–1952.

[204] Prevention And Reversal Of Severe Mitochondrial Cardiomyopathy By Gene Therapy In A Mouse Model Of Friedreich's Ataxia. M. Perdomini, B. Belbellaa, L. Monassier, L. Reutenauer, N. Messaddeq, N. Cartier, R. G. Crystal, P. Aubourg, H. Puccio, *Nat. Med. 2014 205* **2014**, *20*, 542–547.

[205] Frataxin Restoration In The Nervous System: Possibilities For Gene Therapy. D. R. Lynch, E. Kichula, H. Lin, *Mol. Ther.* **2018**, *26*, 1880–1882.

[206] Structure Of The Human Frataxin-Bound Iron-Sulfur Cluster Assembly Complex Provides Insight Into Its Activation Mechanism. N. G. Fox, X. Yu, X. Feng, H. J. Bailey, A. Martelli, J. F. Nabhan, C. Strain-Damerell, C. Bulawa, W. W. Yue, S. Han, *Nat. Commun. 2019 101* **2019**, *10*, 1–8.

[207] Mapping Iron Binding Sites On Human Frataxin: Implications For Cluster Assembly On The ISU Fe-S Cluster Scaffold Protein. J. Huang, E. Dizin, J. A. Cowan, *J. Biol. Inorg. Chem.* **2008**, *13*, 825–836.

[208] P55 - Characterization Of In Vivo Disposition Of CTI-1601: A Mitochondria Targeted Therapy For Friedreich's Ataxia. E. Gonzalez, E. Wagner, N. Mess, A. Wang, M. Payne, D. Bettoun, E. Ottinger, B. Rup, X. Xu, *Drug Metab. Pharmacokinet.* **2020**, *35*, S38.

[209] Cardiovascular Research In Friedreich Ataxia: Unmet Needs And Opportunities. C. Research, F. Ataxia, R. M. Payne, *Basic to Transl. Sci.* **2022**, *7*, 1267–1283.

[210] AlphaFold Protein Structure Database: Massively Expanding The Structural Coverage Of Protein-Sequence Space With High-Accuracy Models. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Zídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, *Nucleic Acids Res.* **2022**, *50*, D439–D444.

[211] Scaffolding Protein Functional Sites Using Deep Learning. J. Wang, S. Lisanza, D. Juergens, D. Tischer, J. L. Watson, K. M. Castro, R. Ragotte, A. Saragovi, L. F. Milles, M. Baek, I. Anishchenko, W. Yang, D. R. Hicks, M. Expòsit, T. Schlichthaerle, J. H. Chun, J. Dauparas, N. Bennett, B. I. M. Wicky, A. Muenks, F. DiMaio, B. Correia, S. Ovchinnikov, D. Baker, *Science.* **2022**, *377*, 387–394.

[212] Hallucinating Symmetric Protein Assemblies. B. I. M. Wicky, L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A. K. Bera, D. Baker, *Science* **2022**, *378*, 56-61.

[213] Robust Deep Learning–Based Protein Sequence Design Using ProteinMPNN. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, *Science (80-.).* **2022**, *378*, 49–56.

[214] Accurate Prediction Of Enzyme Thermostabilization With Rosetta Using AlphaFold Ensembles. F. Peccati, S. Alunno-Rufini, G. Jiménez-Osés, *J. Chem. Inf. Model.* **2023**, *63*, 898–909.

[215] Protein Thermostability Engineering. H. P. Modarres, M. R. Mofrad, A. Sanati-Nezhad, *RSC Adv.* **2016**, *6*, 115252–115270.

[216] Consensus Sequence Design As A General Strategy To Create Hyperstable, Biologically Active Proteins. M. Sternke, K. W. Tripp, D. Barrick, *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *166*, 11275–11284.

[217] The Consensus Concept For Thermostability Engineering Of Proteins: Further Proof Of Concept. M. Lehmann, C. Loch, A. Middendorf, D. Studer, S. F. Lassen, L. Pasamontes, A. P. G. M. Van Loon, M. Wyss, *Protein Eng. Des. Sel.* **2002**, *15*, 403–411.

[218] How Do Thermophilic Proteins Deal With Heat?. S. Kumar, R. Nussinov, *Cell. Mol. Life Sci.* **2001**, *58*, 1216–1233.

[219] Context And Force Field Dependence Of The Loss Of Protein Backbone Entropy Upon Folding Using Realistic Denatured And Native State Ensembles. M. C. Baxa, E. J. Haddadian, A. K. Jha, K. F. Freed, T. R. Sosnick, *J. Am. Chem. Soc.* **2012**, *134*, 15929–15936.

[220] Friedreichs Ataxia Variants I154F And W155R Diminish Frataxin-Based Activation Of The Iron-Sulfur Cluster Assembly Complex. C. L. Tsai, J. Bridwell-Rabb, D. P. Barondeau, *Biochemistry* **2011**, *50*, 6478–6487.

[221] Structure - Function Analysis Of Friedreich's Ataxia Mutants Reveals Determinants Of Frataxin Binding And Activation Of The Fe - S Assembly Complex. J. Bridwell-Rabb, A. M. Winn, D. P. Barondeau, *Biochemistry* **2011**, *50*, 7265–7274.

[222] Towards A Structural Understanding Of Friedreich's Ataxia: The Solution Structure Of Frataxin. G. Musco, G. Stier, B. Kolmerer, S. Adinolfi, S. Martin, T. Frenkiel, T. Gibson, A. Pastore, *Structure* **2000**, *8*, 695–707.

[223] SPEACH_AF: Sampling Protein Ensembles And Conformational Heterogeneity With Alphafold2. R. A. Stein, H. S. McHaourab, *PLOS Comput. Biol.* **2022**, *18*, e1010483.

[224] A Mathematical Theory Of Communication. C. E. Shannon, *Bell Syst. Tech. J.* **1948**, *27*, 379–423.

[225] Thermostable Bacillus Subtilis Lipases: In Vitro Evolution And Structural Insight. S. Ahmad, M. Z. Kamal, R. Sankaranarayanan, N. M. Rao, *J. Mol. Biol.* **2008**, *381*, 324–340.

[226] Deconvoluting The Directed Evolution Pathway Of Engineered Acyltransferase LovD. G. García-Marquina, R. Núñez-Franco, F. Peccati, Y. Tang, G. Jiménez-Osés, F. López-Gallego, *ChemCatChem* **2022**, *14*, e202101349.

[227] Stable And Functionally Diverse Versatile Peroxidases Designed Directly From Sequences. S. Barber-Zucker, V. Mindel, E. Garcia-Ruiz, J. J. Weinstein, M.

Alcalde, S. J. Fleishman, *J. Am. Chem. Soc.* **2022**, *144*, 3564–3571.

[228] PROSS 2: A New Server For The Design Of Stable And Highly Expressed Protein Variants. J. J. Weinstein, A. Goldenzweig, S. Y. Hoch, S. J. Fleishman, *Bioinformatics* **2021**, *37*, 123–125.

[229] Frataxin Structure And Function. I. H. Castro, M. F. Pignataro, K. E. Sewell, L. D. Espeche, M. G. Herrera, M. E. Noguera, L. Dain, A. D. Nadra, M. Aran, C. Smal, M. Gallo, J. Santos, *Subcell. Biochem.* **2019**, *93*, 393–438.

[230] Dynamics, Stability And Iron-Binding Activity Of Frataxin Clinical Mutants. A. R. Correia, C. Pastore, S. Adinolfi, A. Pastore, C. M. Gomes, *FEBS J.* **2008**, *275*, 3680–3690.

[231] Human Frataxin Folds Via An Intermediate State. Role Of The C-Terminal Region. S. E. Faraj, R. M. González-Lebrero, E. A. Roman, J. Santos, *Sci. Reports 2016 61* **2016**, *6*, 1–16.

[232] The Alteration Of The C-Terminal Region Of Human Frataxin Distorts Its Structural Dynamics And Function. S. E. Faraj, E. A. Roman, M. Aran, M. Gallo, J. Santos, *FEBS J.* **2014**, *281*, 3397–3419.

[233] Protein Stability And Surface Electrostatics: A Charged Relationship. S. S. Strickler, A. V. Gribenko, A. V. Gribenko, T. R. Keiffer, J. Tomlinson, T. Reihle, V. V. Loladze, G. I. Makhatadze, *Biochemistry* **2006**, *45*, 2761–2766.

[234] To Charge Or Not To Charge?. J. M. Sanchez-Ruiz, G. I. Makhatadze, *Trends Biotechnol.* **2001**, *19*, 132–135.

[235] Contribution Of Surface Salt Bridges To Protein Stability: Guidelines For Protein Engineering. G. I. Makhatadze, V. V. Loladze, D. N. Ermolenko, X. F. Chen, S. T. Thomas, *J. Mol. Biol.* **2003**, *327*, 1135–1148.

[236] Protein Stability Curves. W. J. Becktel, J. A. Schellman, *Biopolymers* **1987**, *26*, 1859–1877.

[237] A Thermodynamic Comparison Of Mesophilic And Thermophilic Ribonucleases H†. J. Hollien, S. Marqusee, *Biochemistry* **1999**, *38*, 3831–3836.

[238] Increasing Protein Stability: Importance Of ΔCp And The Denatured State. H. Fu, G. Grimsley, J. M. Scholtz, C. N. Pace, *Protein Sci.* **2010**, *19*, 1044–1052.

[239] Thermodynamic Stability And Folding Of Proteins From Hyperthermophilic Organisms. K. A. Luke, C. L. Higgins, P. Wittung-Stafshede, *FEBS J.* **2007**, *274*, 4023–4033.

[240] Lessons In Stability From Thermophilic Proteins. A. Razvi, J. M. Scholtz, *Protein Sci.* **2006**, *15*, 1569–1578.

[241] Evaluating Protein Engineering Thermostability Prediction Tools Using An Independently Generated Dataset. P. Huang, S. K. S. Chu, H. N. Frizzo, M. P.

Connolly, R. W. Caster, J. B. Siegel, *ACS Omega* **2020**, *5*, 6487–6493.

[242] The Rosetta All-Atom Energy Function For Macromolecular Modeling And Design. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J. Gray, *J. Chem. Theory Comput.* **2017**, *13*, 3031–3048.

[243] Denaturant M Values And Heat Capacity Changes: Relation To Changes In Accessible Surface Areas Of Protein Unfolding. J. K. Myers, C. Nick Pace, J. Martin Scholtz, *Protein Sci.* **1995**, *4*, 2138–2148.

[244] A Correlation Between Protein Thermostability And Resistance To Proteolysis.. R. M. Daniel, D. A. Cowan, H. W. Morgan, M. P. Curran, *Biochem. J.* **1982**, *207*, 641.

[245] The Structural Stability Of A Protein Is An Important Determinant Of Its Proteolytic Susceptibility In Escherichia Coli. D. A. Parsell, R. T. Sauer, *J. Biol. Chem.* **1989**, *264*, 7590–7595.

[246] Mega-Scale Experimental Analysis Of Protein Folding Stability In Biology And Design. K. Tsuboyama, J. Dauparas, J. Chen, E. Laine, Y. Mohseni Behbahani, J. J. Weinstein, N. M. Mangan, S. Ovchinnikov, G. J. Rocklin, *Nat. 2023 6207973* **2023**, *620*, 434–444.

[247] A New Generation Of Homology Search Tools Based On Probabilistic Inference.. S. R. Eddy, *Genome Inform.* **2009**, *23*, 205–211.

[248] UniRef: Comprehensive And Non-Redundant UniProt Reference Clusters. B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, C. H. Wu, *Bioinformatics* **2007**, *23*, 1282–1288.

[249] UniRef Clusters: A Comprehensive And Scalable Alternative For Improving Sequence Similarity Searches. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, *Bioinformatics* **2015**, *31*, 926–932.

[250] The Ferroxidase Activity Of Yeast Frataxin. S. Park, O. Gakh, S. M. Mooney, G. Isaya, *J. Biol. Chem.* **2002**, *277*, 38589–38595.

[251] Yeast Frataxin Sequentially Chaperones And Stores Iron By Coupling Protein Assembly With Iron Oxidation. S. Park, O. Gakh, H. A. O'Neill, A. Mangravita, H. Nichol, G. C. Ferreira, G. Isaya, *J. Biol. Chem.* **2003**, *278*, 31340–31351.

[252] Deciphering The Message In Protein Sequences: Tolerance To Amino Acid Substitutions. J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, R. T. Sauer, *Science (80-.).* **1990**, *247*, 1306–1310.

[253] Predicting Enzyme Function From Protein Sequence. J. Minshull, J. E. Ness, C. Gustafsson, S. Govindarajan, *Curr. Opin. Chem. Biol.* **2005**, *9*, 202–209.

[254] An Improved General Amino Acid Replacement Matrix. S. Q. Le, O. Gascuel, *Mol. Biol. Evol.* **2008**, *25*, 1307–1320.

[255] A Thermodynamic Model Of Protein Structure Evolution Explains Empirical Amino Acid Substitution Matrices. C. Norn, I. André, D. L. Theobald, *Protein Sci.* **2021**, *30*, 2057–2068.

[256] Conformational Stability, Dynamics And Function Of Human Frataxin: Tryptophan Side Chain Interplay. L. D. Espeche, K. E. Sewell, I. H. Castro, L. Capece, M. F. Pignataro, L. Dain, J. Santos, *Arch. Biochem. Biophys.* **2022**, *715*, 109086.

[257] FastTree: Computing Large Minimum Evolution Trees With Profiles Instead Of A Distance Matrix. M. N. Price, P. S. Dehal, A. P. Arkin, *Mol. Biol. Evol.* **2009**, *26*, 1641–1650.

[258] FastTree 2 – Approximately Maximum-Likelihood Trees For Large Alignments. M. N. Price, P. S. Dehal, A. P. Arkin, *PLoS One* **2010**, *5*, e9490.

[259] IPC 2.0: Prediction Of Isoelectric Point And PKa Dissociation Constants. L. P. Kozlowski, *Nucleic Acids Res.* **2021**, *49*, W285–W292.

[260] Protein Production By Auto-Induction In High-Density Shaking Cultures. F. W. Studier, *Protein Expr. Purif.* **2005**, *41*, 207–234.

[261] SOFAST-HMQC Experiments For Recording Two-Dimensional Deteronuclear Correlation Spectra Of Proteins Within A Few Seconds. P. Schanda, E. Kupĉe, B. Brutscher, *J. Biomol. NMR* **2005**, *33*, 199–211.

[262] NMRPipe: A Multidimensional Spectral Processing System Based On UNIX Pipes. F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, A. Bax, *J. Biomol. NMR* **1995**, *6*, 277–293.

[263] NMRFAM-SPARKY: Enhanced Software For Biomolecular NMR Spectroscopy. W. Lee, M. Tonelli, J. L. Markley, *Bioinformatics* **2015**, *31*, 1325–1327.

[264] Structural Bases Of Stability-Function Tradeoffs In Enzymes. B. M. Beadle, B. K. Shoichet, *J. Mol. Biol.* **2002**, *321*, 285–296.

[265] Protein Stability: Computation, Sequence Statistics, And New Experimental Methods. T. J. Magliery, *Curr. Opin. Struct. Biol.* **2015**, *33*, 161–168.

[266] Protein Recovery From Inclusion Bodies Of Escherichia Coli Using Mild Solubilization Process. A. Singh, V. Upadhyay, A. K. Upadhyay, S. M. Singh, A. K. Panda, *Microb. Cell Fact.* **2015**, *14*.

[267] Engineering Functional Thermostable Proteins Using Ancestral Sequence Reconstruction. R. E. S. Thomson, S. E. Carrera-Pacheco, E. M. J. Gillam, *J. Biol. Chem.* **2022**, *298*.

[268] Current Perspectives On Stability Of Protein Drug Products During Formulation, Fill And Finish Operations. N. Rathore, R. S. Rajan, *Biotechnol. Prog.* **2008**, *24*, 504–514.

[269] Directed Evolution: Past, Present And Future. R. E. Cobb, R. Chao, H. Zhao, *AIChE J.* **2013**, *59*, 1432–1440.

[270] Directed Evolution: Bringing New Chemistry To Life. F. H. Arnold, *Angew. Chem. Int. Ed. Engl.* **2018**, *57*, 4143–4148.

[271] Discovery Of Novel Gain-Of-Function Mutations Guided By Structure-Based Deep Learning. R. Shroff, A. W. Cole, D. J. Diaz, B. R. Morrow, I. Donnell, A. Annapareddy, J. Gollihar, A. D. Ellington, R. Thyer, *ACS Synth. Biol.* **2020**, *9*, 2927–2935.

[272] Machine Learning-Aided Engineering Of Hydrolases For PET Depolymerization. H. Lu, D. J. Diaz, N. J. Czarnecki, C. Zhu, W. Kim, R. Shroff, D. J. Acosta, B. R. Alexander, H. O. Cole, Y. Zhang, N. A. Lynd, A. D. Ellington, H. S. Alper, *Nat.* **2022**, *604*, 662–667.

[273] Large Language Models Generate Functional Protein Sequences Across Diverse Families. A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, N. Naik, *Nat. Biotechnol.* **2023**, *41*, 1099–1106.

[274] Machine Learning-Assisted Directed Protein Evolution With Combinatorial Libraries. Z. Wu, S. B. Jennifer Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 8852–8858.

[275] Tobacco Etch Virus Protease: Mechanism Of Autolysis And Rational Design Of Stable Mutants With Wild-Type Catalytic Proficiency. R. B. Kapust, J. Tözsér, J. D. Fox, D. E. Anderson, S. Cherry, T. D. Copeland, D. S. Waugh, *Protein Eng.* **2001**, *14*, 993–1000.

[276] A Combined Approach To Improving Large-Scale Production Of Tobacco Etch Virus Protease. P. G. Blommel, B. G. Fox, *Protein Expr. Purif.* **2007**, *55*, 53–68.

[277] Structural Basis For The Substrate Specificity Of Tobacco Etch Virus Protease. J. Phan, A. Zdanov, A. G. Evdokimov, J. E. Tropea, H. K. Peters, R. B. Kapust, M. Li, A. Wlodawer, D. S. Waugh, *J. Biol. Chem.* **2002**, *277*, 50564–50572.

[278] Directed Evolution Improves The Catalytic Efficiency Of TEV Protease. M. I. Sanchez, A. Y. Ting, *Nat. Methods* **2020**, *17*, 167–174.

[279] Screening, Large-Scale Production And Structure-Based Classification Of Cystine-Dense Peptides. C. E. Correnti, M. M. Gewe, C. Mehlin, A. D. Bandaranayake, W. A. Johnsen, P. B. Rupert, M. Y. Brusniak, M. Clarke, S. E. Burke, W. De Van Der Schueren, K. Pilat, S. M. Turnbaugh, D. May, A. Watson, M. K. Chan, C. D. Bahl, J. M. Olson, R. K. Strong, *Nat. Struct. Mol. Biol.* **2018**, *25*,

270–278.

[280] How Directed Evolution Reshapes The Energy Landscape In An Enzyme To Boost Catalysis. R. Otten, R. A. P. Pádua, H. A. Bunze, V. Nguyen, W. Pitsawong, M. Patterson, S. Sui, S. L. Perry, A. E. Cohen, D. Hilvert, D. Kern, *Science* **2020**, *370*, 1442–1446.

[281] The Role Of Distant Mutations And Allosteric Regulation On LovD Active Site Dynamics. G. Jiménez-Osés, S. Osuna, X. Gao, M. R. Sawaya, L. Gilson, S. J. Collier, G. W. Huisman, T. O. Yeates, Y. Tang, K. N. Houk, *Nat. Chem. Biol.* **2014**, *10*, 431–436.

[282] Effects Of Point Mutation On Enzymatic Activity: Correlation Between Protein Electronic Structure And Motion In Chorismate Mutase Reaction. T. Ishida, *J. Am. Chem. Soc.* **2010**, *132*, 7104–7118.

[283] AlphaFold2 Models Indicate That Protein Sequence Determines Both Structure And Dynamics. H.-B. Guo, A. Perminov, S. Bekele, G. Kedziora, S. Farajollahi, V. Varaljay, K. Hinkle, V. Molinero, K. Meister, C. Hung, P. Dennis, N. Kelley-Loughnane, R. Berry, *Sci. Reports.* **2022**, *12*, 10696.

[284] Development And Testing Of A General Amber Force Field. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[285] A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints For Deriving Atomic Charges: The RESP Model. C. I. Bayly, P. Cieplak, W. D. Cornell, P. A. Kollman, *J. Phys. Chem.* **1993**, *97*, 10269–10280.

[286] Comparison Of Simple Potential Functions For Simulating Liquid Water. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926–935.

[287] Settle: An Analytical Version Of The SHAKE And RATTLE Algorithm For Rigid Water Models. S. Miyamoto, P. A. Kollman, *J. Comput. Chem.* **1992**, *13*, 952–962.