UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

PhD Course in Molecular and Cellular Biology
XXXVII Cycle
Academic Year 2023-2024

Integrative Modelling for the Characterisation of the Structure and Dynamics of Biomolecules

Scientific advisor: Prof. Carlo Camilloni

Federico Ballabio
R13343

Table of Contents

1 — Abstract

This work explores the benefits of integrating computational structural biology techniques with experimental data to overcome the inherent limitations of each approach. Integrative modelling provides a more comprehensive understanding of the structure, dynamics, and function of complex biomolecular systems. Experimental data can guide computational methods, improving accuracy, reducing limits and approximations, and validating results, while computational techniques inspire new experimental designs and provide explanations for observed phenomena. In this work, I have reported on some personal contributions that exemplify the application of computational techniques in the context of integrative modelling, with a particular focus on three main manuscripts. The first paper describes the development of a small-angle scattering model for the *in silico* reconstruction of scattering intensities from atomic coordinates during molecular dynamics (MD) simulations. This model can be coupled with restraining strategies, such as metainference, to generate conformational ensembles at atomistic resolution in agreement with the experimental data using MD simulations. This approach was used to determine the closed state conformations of the human gelsolin, a plasma protein. The second manuscript investigates the inactivation mechanism of the human olfactory receptor OR51E2. MD simulations revealed that calcium ions play a key role in stabilising the inactive state of the protein. This study integrates different computational techniques to propose a novel molecular mechanism of receptor inactivation and provides a rationale for future experimental validation. The third manuscript explores the molecular basis of a pale green phenotype in the barley population *TM2490*, linked to a mutation in magnesium chelatase subunit I, an enzyme involved in the chlorophyll synthesis pathway. AI-based structural modelling and molecular docking studies elucidate how this mutation might affect ATP binding, providing insights for future crop breeding strategies aimed at improving photosynthetic efficiency.

2 — Introduction

2.1 Introduction to Biomolecules

2.1.1 Definition and Role in Cellular Functions and Life Processes

Biomolecules are the building blocks of life and play a central role in the structure, function and regulation of cells and tissues in all living organisms. These organic molecules include proteins, nucleic acids (DNA and RNA), carbohydrates and lipids. Among these, deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins are not only directly associated with the storage, transmission and execution of genetic information, but are also involved in various cellular processes such as regulation, signalling and catalysis.

Deoxyribonucleic Acid (DNA)

DNA is a biomolecule that acts as the genetic blueprint for an organism. It consists of two strands that coil around each other to form a double helix, a structure first described by James Watson and Francis Crick in 1953.[1] Each strand is composed of simpler molecules, called nucleotides, which contain a sugar group, a phosphate group, and a nitrogenous base. The sequence of the nucleobases (adenine, thymine, cytosine, and guanine) encodes the genetic information necessary for the development, functioning, growth, and reproduction of all living organisms and many viruses.[2]

The primary function of DNA is to serve as a repository for genetic information. This information is used in the synthesis of proteins, which are responsible for most cellular functions.[3] The ability of DNA to replicate allows genetic information to be transferred from cell to cell and from parent to offspring.[4] Mutations in the DNA sequence can lead to the emergence of new variants that can either cause disease or facilitate evolutionary adaptations, illustrating its dual role in maintaining stability and facilitating variability.[5]

Ribonucleic Acid (RNA)

RNA plays a major role in the translation of the genetic code from DNA into proteins, a process that is fundamental to all forms of life.[6] This molecule exhibits structural similarities to DNA but typically exists as a single strand.[7] Furthermore, it contains the pentose sugar ribose, which has a hydroxyl moiety at the carbon in position 2, in contrast to the hydrogen atom in deoxyribose in DNA. Additionally, the base thymine is substituted by the base uracil, which lacks a methyl group on the carbon atom at position 5. The synthesis of RNA molecules from DNA templates is called transcription.[8]

There are several types of RNA, each serving different functions within the cell:

- **Messenger RNA (mRNA):** Responsible for the transfer of genetic information from DNA to the ribosomes, the macromolecular complexes responsible for protein synthesis. It acts as a template for the assembly of amino acids into proteins, based on the sequence of codons (combinations of three nucleotides) it contains.[9,10]

- **Transfer RNA (tRNA):** It assists in the delivery of the appropriate amino acid to the ribosome during the process of protein synthesis. The tRNA molecule, by associating to the codon sequence on the mRNA through its anticodon, it allows the addition of amino acids to the growing polypeptide chain.[11] The standard genetic code uses 64 codons to encode the 20 standard amino acids and the stop signals for translation. Not all of these codons require unique tRNAs because of the phenomenon known as 'wobble base pairing', which allows a single tRNA to recognise multiple codons.[12,13]

- **Ribosomal RNA (rRNA):** Structural component of ribosomes which, together with ribosomal proteins, are the sites of protein synthesis. The rRNA molecules ensure proper alignment of mRNA and tRNAs and catalyse the formation of peptide bonds.[14]

- **Regulatory RNAs:** Include microRNAs (miRNAs[15]) and small interfering RNAs (siRNAs[16]), which are involved in gene regulation by interfering with the translation and stability of specific mRNAs. These molecules can promote mRNA degradation, prevent translation initiation, and even influence chromatin structure to modulate gene expression and silence genes at the transcriptional level.[17]

The versatility of RNA goes far beyond its role as an intermediary between DNA and proteins. Its capabilities include genetic regulation,[18,19] structural functions and even catalytic activity, as in ribozymes.[20,21] RNA also serves as the genetic material in many viruses[22] and is involved in intracellular signalling, highlighting its diverse functions in biological processes.

Proteins

Proteins are the biochemical ”workhorses” of the cell, performing a wide range of functions necessary for the survival and proliferation of living organisms. These biomolecules consist of one or more polypeptide chains, which are polymers of amino acids covalently linked together by peptide bonds. The specific sequence of amino acids in a protein determines its three-dimensional structure, which in turn determines its function.[6,23]

Proteins have many functions in the cell, including:

- **Enzymatic Activity:** It refers to the ability of proteins, called enzymes, to catalyse biochemical reactions, ensuring compatibility with the timescales required to sustain life. For example, DNA polymerase catalyses the synthesis of new strands of DNA during cell replication, while lactase cleaves lactose, a disaccharide sugar, into glucose and galactose, two monosaccharides, aiding the digestion of dairy products. Enzymes lower the activation energy required for reactions, making biological processes occur at specific rates and in a controlled manner, providing proper timing and coordination in the biological context.[6,24]

- **Structural Support:** Structural proteins provide support and shape to cells, tissues, and even viruses. For example, in multicellular organisms, collagen strengthens and stabilises connective tissue, providing elasticity and resilience,[25] while actin and tubulin form the cytoskeleton of cells, giving them shape and enabling movement.[26,27] In viruses, capsid proteins form a protective shell around the viral genome, ensuring structural integrity and aiding infection of the host.[28]

- **Transport and Storage:** Certain proteins are involved in the transport of molecules. For example, haemoglobin in red blood cells binds and transports oxygen from the lungs to tissues throughout the body, ensuring efficient oxygen delivery to support cellular respiration.[29] Other proteins, such as ferritin, have a storage function by sequestering iron in a safe, bioavailable form, maintaining iron homeostasis and preventing toxicity.[30]

- **Signalling and Communication:** Proteins are essential in cellular signalling pathways, mediating communication within and between cells. Receptors on the cell surface detect external signals, such as hormones and neurotransmitters, and initiate a cascade of intracellular events leading to a cellular response. For example, insulin receptors on the membrane of muscle, fat and liver cells respond to insulin levels, facilitating the regulation of glucose uptake and maintaining blood sugar balance.[6]

- **Immune Response:** Proteins have many roles in the immune response. For example, antibodies are specialised proteins that recognise and bind to specific antigens (molecules found on the surface of pathogens, foreign substances or abnormal cells such as cancerous or damaged cells), leading to their neutralisation or marking them for destruction.[31] The preoteins of the complement system collaborate to directly target and degrade invading pathogens.[32] In addition, cytokines act as signalling proteins that help coordinate the immune response, directing the activation and movement of immune cells to sites of infection, injury, or cellular abnormality.[33]

- **Gene Regulation:** Gene regulation is mediated by several classes of proteins. Transcription factors bind to specific DNA sequences and control the activation or repression of target genes.[34] Other proteins, such as histone modifiers,[35] chromatin remodelers,[36] and RNA-binding proteins,[37] regulate gene expression by altering chromatin structure or influencing mRNA stability. Together, these proteins coordinate development, cellular differentiation, and responses to environmental stimuli, allowing cells to adapt, grow, repair, and maintain homeostasis.

The Interconnected Role of DNA, RNA, and Proteins

The central dogma of molecular biology describes the flow of genetic information within a biological system: DNA → RNA → Protein.[38] DNA provides long-term storage of genetic information, which is transcribed into RNA, the intermediate that directs the synthesis of proteins. The proteins then perform the functions necessary for the cell to survive, grow and reproduce. This flow of information is possible because biomolecules do not work in isolation. Their functions are closely interconnected. This network of interactions and feedbacks enables cells to respond dynamically to changes in their environment, to adapt to new conditions and, ultimately, to sustain and perpetuate life.[6,24]

2.1.2 Historical Perspective

The study and comprehension of biomolecules has experienced a profound transformation over the past two centuries, in parallel with advances in scientific methods and technologies. The journey from the initial identification to the description of the structure and function of biomolecules in contemporary science has been marked by a series of landmark discoveries.

Early Recognition and the Identification of Nucleic Acids

The earliest investigations into the molecular basis of life began in the 19th century. In 1869, Friedrich Miescher, a Swiss biochemist, first identified a novel phosphorus-containing substance in the nuclei of white blood cells, which he described as ″nuclein″. This substance was later identified as DNA. Miescher's discovery marked the beginning of the field of molecular biology, although the role of DNA in heredity was not yet understood.[39,40]

The identification of nucleic acids as carriers of genetic information took several decades. In the early 20th century, Phoebus Levene made a significant contributions to the field of molecular biology by identifying the components of nucleic acids: the sugar, phosphate, and nitrogenous bases. He also distinguished between RNA and DNA, although he incorrectly postulated that DNA was made up of equal amounts of bases.[41] Despite this inaccuracy, Levene's work laid the foundation for understanding the basic components of nucleic acids.

The Role of DNA in Heredity

The function of DNA as the hereditary material was established through a series of experiments conducted in the mid-20th century. The work of Frederick Griffith in 1928 provided the first indication of the role of DNA in heredity through his transformation experiments with *Streptococcus pneumoniae*, where he observed that a ″transforming principle″ could transfer genetic traits between bacterial strains.[42] However, it was not until 1944 that Oswald Theodore Avery, Colin Munro MacLeod, and Maclyn McCarty identified DNA as this ″transforming principle″, demonstrating conclusively that DNA carries genetic information.[43]

This finding was further confirmed in 1952 by the Hershey-Chase experiment, led by Alfred Day Hershey and Martha Cowles Chase. Using bacteriophages labelled with radioactive isotopes, they showed that DNA, not protein, was the genetic material transferred from viruses to bacterial cells. Together, these experiments shifted the scientific consensus, solidifying the recognition of DNA as the molecule responsible for heredity.[44]

The Discovery of the DNA Double Helix

The structure of DNA was elucidated in 1953, representing a milestone in the field of molecular biology. James Watson and Francis Crick,[1] using experimental data from Rosalind Franklin and Maurice Wilkins, proposed the double helix model of DNA.[45,46] This model, which is characterised by two antiparallel strands wound around each other with complementary base pairing (adenine with thymine, cytosine with guanine), provided a structural basis for understanding the replication and transmission of genetic information.

The model proposed by Watson and Crick suggested a mechanism by which genetic information could be copied and transmitted across generations. The complementary nature of the base pairs indicated that each strand could serve as a template for the synthesis of a new complementary strand during DNA replication. This breakthrough laid the groundwork for a molecular understanding of the principles of inheritance initially proposed by Gregor Mendel in the 19th century, thereby establishing a direct link between the concept of genes and a physical structure.[6,47]

RNA and the Central Dogma

Following the discovery of the DNA double helix, research efforts have also focused on understanding the function of RNA in gene expression. The identification of mRNA by François Jacob and Jacques Lucien Monod in 1961 helped to understand the process by which genetic information is transferred from DNA to protein.[48] This discovery helped to formulate the central dogma of molecular biology.

The central dogma elucidated the function of RNA as an intermediary that transcribes genetic information from DNA and translates it into proteins. The subsequent identification of tRNA and rRNA, and their respective functions in translation, provided further insight into the process by which the genetic code is decoded and proteins are synthesised.

Advancements in Protein Structure Determination

In parallel with advances in nucleic acid research, significant progress was made in understanding the structure of proteins. The first protein structures were determined using X-ray crystallography in the 1950s, with the structures of myoglobin and haemoglobin being described by John Cowdery Kendrew and Max Ferdinand Perutz, respectively.[49,50] These studies revealed the complex three-dimensional shapes that proteins can adopt and highlighted the importance of structural biology in understanding protein function. The realisation that proteins are not just static structures but dynamic entities capable of adopting multiple conformations shifted the focus to understanding the relationship between protein dynamics and function. Techniques such as nuclear magnetic resonance spectroscopy and small-angle scattering allowed dynamic aspects to be studied in addition to structure.

The Rise of Integrative Modelling and Computational Biology

The late 20th and early 21st centuries witnessed a revolution in the field of molecular biology with the advent of computational methods. The development of quantum and molecular mechanics simulations, molecular docking approaches, bioinformatics tools, and artificial intelligence applied to biology has transformed the ability to study biomolecules.[51–55] Coupling these techniques with experimental data enhances the ability to explore the structure and dynamics of biomolecules, and provides insights into the structure-function relationships.

Integrative modelling, which combines data from multiple experimental sources with computational methods, has emerged as a powerful tool in structural biology. By providing a more complete and accurate picture of the molecular behaviour, these approaches are improving the understanding of biological processes, including those related to disease mechanisms and drug discovery.

Historical details of computational biology are discussed in section 2.5.3.

2.2 Structure of Biomolecules

2.2.1 Primary to Quaternary Structures

Before understanding how biomolecules perform their different functions in biological systems, it is necessary to know how they are structurally organised. Proteins and nucleic acids exhibit a hierarchy of structural organisation, ranging from simple linear chains to complex three-dimensional architectures. The specific functions of the biomolecules are intimately connected to their structural organisation at each level of complexity.

Primary Structure

The primary structure of a biomolecule refers to its linear sequence of monomeric units. In the case of proteins, this sequence consists of amino acids linked by peptide bonds to form a polypeptide chain. The order of the amino acids is determined by the genetic code encoded in the DNA, and it is this specific sequence that ultimately determines the higher order structures and functions of the protein.[23] For nucleic acids such as DNA and RNA, the primary structure refers to the sequence of nucleotides, each of which consists of a sugar, a phosphate group and a nitrogenous base.[24]

The primary structure contains the information necessary for the molecule to fold into its functional form. Any alteration in the primary sequence, such as a point mutation in DNA or a substitution of an amino acid in a protein, can have a significant impact on the function of the molecule, potentially leading to dysfunctional proteins and disease states.[56] The importance of primary structure is underlined by the principle of sequence specificity, whereby a single change in the sequence can alter the properties of the molecule and its interactions with other molecules.

Secondary Structure

Secondary structure refers to the regular, repeating folding patterns within a polypeptide chain that result from hydrogen bonding between the backbone amides. In proteins, the two most common types of secondary structure are the alpha-helix and the beta-sheet:

- The **alpha-helix** is a right-handed coiled structure, stabilised by hydrogen bonds between the carbonyl oxygen of one amino acid and the amide hydrogen of another amino acid four residues down the chain. This structure is commonly found in the transmembrane regions of proteins and serves as a structural scaffold in many proteins.[57–59]

- The **beta-sheet** consists of beta strands that are aligned next to each other, forming a sheet-like arrangement. These strands can be oriented in parallel or antiparallel configurations, and the structure is stabilised by hydrogen bonds between carbonyl oxygens and amide hydrogens on adjacent strands. Beta-sheets provide significant stability to the protein and are commonly found in the core regions of globular proteins.[60]

The formation of secondary structures in proteins is primarily driven by interactions involving the backbone. Although the backbone is chemically identical in all proteins (N-Cα-C-O), the local environment created by the sequence of side chains and their interactions influences which secondary structures are stabilised.[61]

In nucleic acids, secondary structure includes elements such as double-stranded helices in DNA, hairpins, loops and bulges in RNA.[62,63] The secondary structures are stabilised by hydrogen bonds between complementary bases. Similar to proteins, the secondary structure of RNA is closely linked to its function.[64] For instance, considering tRNA, the secondary structure allows the molecule to interact with mRNA and ribosomes during protein synthesis.[11]

Tertiary Structure

The tertiary structure of a biomolecule refers to its overall three-dimensional shape formed by the spatial arrangement of secondary structural elements and their side chains or bases. In proteins, this level of structure results from various interactions, including hydrogen bonds, ionic bonds, hydrophobic interactions and van der Waals forces, as well as disulphide bridges between cysteine residues.[65] The tertiary structure is stabilised by the interactions between the side chains of the amino acids, which can be hydrophobic or hydrophilic, charged or uncharged, influencing how the protein folds and interacts with its environment.[66,67]

Tertiary structures create specific geometric shapes and chemical environments necessary for the biological function of the protein. The three-dimensional conformation of a protein determines its ability to interact with other molecules, including substrates, ligands, and other proteins. For example, the active site of an enzyme, responsible for catalysing biochemical reactions, is directly shaped by the tertiary structure of the protein. Alterations to this structure can result in changes or loss of function.[68]

In nucleic acids, tertiary structures include the supercoiling[69] of DNA and the complex folding of RNA molecules into globular shapes.[64] These structures are essential for functions such as the compact packaging of DNA in the nucleus and the catalytic activities of ribozymes.[18]

Quaternary Structure

Quaternary structure is the highest level of organisation and refers to the assembly of multiple polypeptide chains (subunits) into a functional, multi-subunit complex. These subunits may be identical or different, and they interact through non-covalent interactions, such as hydrogen bonds, ionic interactions, and hydrophobic effects, as well as covalent bonds like disulphide bonds. The quaternary structure is stabilised by the same types of interactions that govern the tertiary structure but involves more complex inter-subunit associations. The quaternary structure allows regulation of activity through co-operative binding and allosteric effects, where binding of a molecule to one subunit can influence the activity of the entire complex.[70]

Quaternary structures are less common in nucleic acids, but are exemplified by the packaging of DNA with histone proteins to form nucleosomes,[71,72] the basic units of chromatin structure in eukaryotic cells. This organisation allows DNA to be efficiently compacted (although each human cell is only a few micrometres in diameter, it contains approximately 2 metres of DNA[6]) and is involved in the regulation of gene expression.[73]

Importance of Structure in Biological Molecules

The structural organisation of biomolecules from the primary to the quaternary level is crucial because structure dictates function. The specific three-dimensional shapes of proteins and nucleic acids enable these molecules to perform their biological roles with high specificity and efficiency. In proteins, shape determines their ability to interact with other molecules, catalyse chemical reactions and transmit signals within and between cells. For nucleic acids, structure is key to their ability to store and transfer genetic information and to participate in the regulation of gene expression.

2.2.2 Techniques for Structural Determination

Determining the three-dimensional structures of biomolecules is an essential step in understanding their function and interaction within biological systems. Several experimental techniques have been developed for structural determination, each with its own advantages and limitations. The most prominent of these are X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy (cryo-EM) and electron paramagnetic resonance (EPR) spectroscopy. These methods provide detailed insights into the atomic and molecular arrangements of biomolecules.

X-ray Crystallography

X-ray crystallography is a powerful and widely used technique for determining the atomic structure of biomolecules, particularly proteins and nucleic acids. The process involves crystallising the biomolecule of interest and then irradiating it with X-ray beams. The X-rays are diffracted by the electrons in the crystal, producing a diffraction pattern that can be analysed to reconstruct a three-dimensional electron density map. From this map, the positions of the individual atoms within the molecule can be deduced with high precision.[74]

- Pros: The main advantage of X-ray crystallography is its ability to produce high-resolution structural data, often at atomic resolution. This precision allows the detailed architecture of biomolecules to be visualised, including the positioning of side chains and interactions with ligands or other molecules. X-ray crystallography has been used to determine the structures of a wide range of biological macromolecules, from small peptides to large protein complexes, providing essential insights into their function and mechanism.[75,76]
- Cons: Despite its strengths, X-ray crystallography has several limitations. A major challenge is the need for high-quality crystals, which can be difficult to obtain, particularly for large, flexible or membrane-associated proteins.[77] The crystallisation process can also introduce artefacts, potentially leading to structures that do not fully represent the molecule in its native state.[78] In addition, the technique generally provides a static picture of the molecule, with limited information about conformational flexibility or transient structural states.

Nuclear Magnetic Resonance Spectroscopy

Nuclear Magnetic Resonance (NMR) spectroscopy is another important tool for structural determination and is particularly suited to the study of biomolecules in solution, which more closely mimics physiological conditions. This technique exploits the magnetic properties of certain atomic nuclei, most commonly hydrogen (^{1}H), but also carbon (^{13}C), nitrogen (^{15}N), and phosphorus (^{31}P). When placed in a strong magnetic field and subjected to radio frequency pulses, these nuclei resonate at characteristic frequencies, providing detailed information about their chemical environment.[79]

NMR spectroscopy yields structural information by analysing various parameters, including chemical shifts, which indicate the electronic environment surrounding a nucleus, and scalar J coupling constants, which provide information about the bonding between nuclei. In addition, nuclear Overhauser effects (NOEs) measure the through-space interactions between nuclei that are close together but not necessarily bound.[80] By studying these interactions, researchers can infer the relative distances and angles between atoms, which are crucial for determining three-dimensional structure.[81,82]

- Pros: One of the key advantages of NMR is the ability to study biomolecules in a solution state, which more closely resembles a physiological environment than the crystalline state required for X-ray crystallography. This capability makes NMR particularly valuable for the study of small to medium-sized proteins and nucleic acids, including those that are difficult to crystallise. NMR can provide information on secondary and tertiary structure and, in favourable cases, can achieve near-atomic resolution.[83]

- Cons: NMR spectroscopy has inherent limitations, primarily related to the size of the biomolecule under study. As the molecular weight increases, the complexity of the NMR spectra also increases, making it challenging to resolve overlapping signals and interpret the data. This limitation typically restricts the use of NMR to biomolecules smaller than 50 kDa, although advances in NMR techniques and isotopic labelling have extended this range.[84] In addition, NMR requires relatively large amounts of highly purified samples, which can be a limitation for certain biomolecules.[85]

Cryo-Electron Microscopy

Cryo-electron microscopy is an increasingly popular technique for structural determination, particularly of large macromolecular complexes. Cryo-EM involves the rapid freezing of samples to cryogenic temperatures. This rapid freezing prevents the formation of ice crystals, thereby preserving the sample in a near-native hydrated state without the artefacts that can be introduced by crystallisation or chemical fixation.[86] The sample is then analysed using an electron microscope: electrons are passed through the sample and the scattered electrons are collected to produce high-resolution two-dimensional images. The images represent different orientations of the molecule. Thousands to hundreds of thousands of these 2D projections are then aligned computationally and combined using advanced image processing algorithms to reconstruct a detailed three-dimensional model of the molecule.[87,88]

- Pros: Cryo-EM is particularly useful for studying large complexes that are often difficult to crystallise. The method has advanced significantly in recent years, with improvements in electron detectors and image processing algorithms enabling near-atomic resolution for many biomolecules. Cryo-EM allows the visualisation of large and heterogeneous assemblies, providing insights into their structural organisation and interactions.[89]
- Cons: While cryo-EM has many strengths, it also has limitations. The resolution achieved can vary depending on the quality of the sample preparation, the stability of the complex, and the imaging conditions. Achieving high resolution often requires a large number of images and advanced computational resources, which can be time consuming and resource intensive. In addition, cryo-EM is generally less suitable for small proteins (typically below 100 kDa) due to the lower contrast and signal-to-noise ratio in the resulting images.[90–92]

Electron Paramagnetic Resonance Spectroscopy

Electron paramagnetic resonance (EPR) spectroscopy is a technique used to study molecules containing unpaired electrons, such as free radicals and transition metal complexes. EPR detects the energy transitions of unpaired electrons as they absorb microwave radiation in the presence of an external magnetic field, providing detailed information about the electronic environment.[93]

In structural biology, EPR is often used in combination with a method known as site-directed spin labelling (SDSL),[94,95] in which specific amino acid residues within a protein are systematically replaced by cysteine residues, which can then be chemically modified to attach a paramagnetic probe, such as a nitroxide spin label. These spin labels introduce unpaired electrons into the protein, making it EPR-active.

EPR spectroscopy can then measure the interactions between the spin labels to provide information about the distance and relative orientation between the labelled sites.

- Pros: EPR allows to investigate the structural organisation of biomolecules that contain paramagnetic centres or that can be modified to include spin labels. It can provide distance measurements in the range of 1.5 to 8 nanometres.[96,97] This makes EPR particularly useful for studying larger protein complexes and for gaining insight into the spatial arrangement of specific sites within a molecule.

- Cons: The main limitation of EPR is its dependence on the presence of unpaired electrons, which are not naturally present in most biological molecules. The introduction of spin labels can sometimes alter the native structure or function of the protein, potentially affecting the accuracy of the structural data obtained. In addition, EPR typically provides less detailed structural information than X-ray crystallography or NMR spectroscopy and often requires the use of complementary techniques to build comprehensive structural models.[98]

Final thoughts on Structural Determination Techniques

Each of these structural determination techniques provides unique insights into the architecture of biomolecules. Understanding their strengths and limitations allows the selection of the appropriate method for specific structural questions. Integrating data from multiple techniques can provide a more complete picture of biomolecular structures, which is an important step in the process of characterising their function and interactions.

2.3 Structure-Function Relationship

2.3.1 Concept of Structure-Function Relationship

The relationship between structure and function is a fundamental principle in molecular biology that provides the basis for understanding the role of biomolecules within biological systems: as discussed earlier, this concept can be summarised as "structure dictates function". The specific three-dimensional arrangement of atoms within a biomolecule directly determines its biological activity and interactions. The shape, charge distribution, hydrophobicity and flexibility of a molecule determine how it recognises and interacts with other molecules, catalyses chemical reactions and performs its functions within the cell. To understand the role and function of a biomolecule, it is necessary to consider how its structure affects its activity, as changes in structure often lead to alterations or loss of function, potentially resulting in disease states or dysfunctional biological processes.[6]

Enzymes and Catalytic Function

Enzymes are one of the best examples of structure-function relationships. These biological catalysts accelerate chemical reactions by lowering the activation energy required. The active site of an enzyme, typically a small pocket or groove, is uniquely shaped to bind specific substrates. The precise arrangement of amino acid residues within the active site facilitates proper substrate positioning, stabilizes the transition state, and enables the conversion of substrates into products.[99]

A classic example is the enzyme lysozyme, which is able to hydrolyse the β-(1,4)-glycosidic bonds in peptidoglycans, structural components of the bacterial cell walls.[100] The structure of lysozyme contains a deep cleft shaped to accommodate the peptidoglycan layer of bacterial cell walls. The active site residues of lysozyme interact specifically with the substrate, facilitating the cleavage of glycosidic bonds and ultimately leading to the lysis of bacterial cells. Detailed knowledge of the structure of lysozyme, obtained by X-ray crystallography, has provided insights into its catalytic mechanism and revealed how specific interactions and conformations enable its function.[101]

DNA Structure and Genetic Information Storage

The double helix structure of DNA is another example of the structure-function paradigm.[6] The specific arrangement of the two antiparallel strands and the complementary base pairing not only stabilise the DNA molecule, but also ensure the faithful replication and storage of genetic information. The uniform DNA diameter and the helical twist enable it to store vast amounts of information in a compact form that fits within the confines of a cell nucleus.

In addition, understanding the structural features of DNA has led to insights into the mechanisms of replication, transcription and repair. The recognition of specific DNA sequences by proteins such as transcription factors and DNA polymerases is based on molecular interactions determined by the structure of DNA. The helical shape and major and minor grooves of DNA allow specific protein domains to recognise base sequences without unwinding the helix, illustrating how structural conformation is intricately linked to its role in genetic regulation.[102]

Hemoglobin and Oxygen Transport

Haemoglobin, the oxygen-carrying protein in red blood cells, is a clear example of the importance of quaternary structure for function. Haemoglobin consists of four subunits, each containing a haem group, a heterocyclic organic ring called porphyrin which coordinates an iron ion capable of binding an oxygen molecule. The cooperative binding of oxygen is a direct consequence of the quaternary structure of haemoglobin. When a haem group binds oxygen, it induces a conformational change in the haemoglobin molecule that increases the oxygen affinity of the remaining haem groups. This cooperative binding mechanism, which is essential for efficient oxygen transport and release, is made possible by the specific arrangement and interaction of the haemoglobin subunits.[103]

Structural studies of haemoglobin have also shed light on several pathological conditions. For example, in sickle cell anaemia, a single amino acid substitution in the beta chain of haemoglobin (glutamic acid to valine) causes haemoglobin molecules to aggregate into fibrous structures, distorting the red blood cells into a sickle shape. This structural change impairs the ability of haemoglobin to transport oxygen and leads to the clinical symptoms of the disease.[104] Understanding the structural basis of haemoglobin function and its pathological variants highlights the importance of structure in both normal physiology and disease states.

Antibodies and Immune Response

Antibodies are specialised proteins of the adaptive immune system that recognise and bind to specific antigens, such as pathogens, foreign substances, or damaged cells, and mark them for destruction. The ability of antibodies to selectively bind to a wide range of antigens is due to their hypervariable region, the structure of which can adapt to different shapes and chemical properties.[105] The structure of the variable region, formed by loops known as complementarity determining regions (CDRs), determines the specificity and affinity of the antibody for its antigen. The structural understanding of antibody-antigen interactions has not only provided insights into immune recognition, but has also facilitated the design of therapeutic antibodies.[106]

G Protein-Coupled Receptors and Signal Transduction

G protein-coupled receptors (GPCRs) are a large family of membrane proteins involved in cellular signalling. These receptors transduce the signal into the cell upon binding of extracellular ligands, triggering a variety of physiological responses. The structure of GPCRs is characterised by seven transmembrane alpha helices.[107]

Structure determination of GPCRs has revealed how ligand binding to the extracellular side induces conformational changes that are transmitted across the membrane. These changes lead to the activation of intracellular G proteins, triggering downstream signalling pathways.[108] The detailed understanding of GPCR structures has driven drug development, making this family of receptors a common target for pharmaceuticals.[109] The design of drugs that can specifically modulate GPCR activity is guided by insights into their structural conformation and ligand binding mechanisms.[110]

Final thoughts on Structure-Function Relationship

The previous examples support the concept that the structure of a molecule determines its function. By determining the three-dimensional arrangements of atoms and their corresponding specific interactions, structural biology provides insights into how molecules perform their activity. Understanding structural details can provide the basis for elucidating the mechanisms underlying biological processes, designing drugs and developing therapeutic strategies. This knowledge can be used to explore how dynamic properties further modulate function.

2.3.2 Limitations of Solely Structural Approaches

While the determination of biomolecular structures has been essential to the understanding of biological function, it is widely recognised that structure alone often does not provide a complete picture of how biomolecules function. The traditional view that the static three-dimensional shape of a molecule completely determines its function is limited by the complexity and dynamic nature of biological systems.[111] Although high-resolution structures obtained by techniques such as X-ray crystallography and cryo-electron microscopy have provided invaluable insights, these static snapshots represent only one aspect of biomolecular behaviour. In many cases, understanding the full range of the function of a moleculer requires consideration of its dynamic properties and conformational flexibility.[112]

Conformational Flexibility and Function

A major limitation of purely structural approaches is the inability to capture the dynamic nature of biomolecules. Proteins and nucleic acids do not adopt a single, rigid conformation, but exist as ensembles of multiple conformations in equilibrium. Conformational changes and, more generally, the intrinsic dynamics of the molecule are required to exert their biological function. For example, enzymes often undergo significant structural rearrangements upon substrate binding, a phenomenon known as induced fit.[113]

In addition to induced fit, enzymes can also sample different conformations, and the substrate selectively binds and stabilises the conformation most favourable for catalysis. This adaptability, described as conformational selection model, allows the enzyme to create an optimal environment for catalysis, increasing reaction rates and specificity.[114,115] However, a static structural snapshot of the enzyme, whether in its unbound or bound form, may not fully capture the dynamic conformational changes required for its function. While structural studies of the bound enzyme can provide valuable insights into the active site and substrate interactions, they often miss the dynamic processes that occur during catalysis. Similarly, proteins with allosteric sites show how structural changes in one region of a protein can reverberate and influence activity at a distant location.[116] Allosteric regulation is widely diffuse in biological processes, such as the regulation of metabolic pathways and signal transduction.[117]

Limitations in Capturing Transient States

Another limitation of static structural approaches is their inability to capture transient intermediates. Many biological processes, such as enzyme catalysis, signal transduction, and even protein folding itself, involve short-lived intermediates that determine the outcome of these processes. These intermediates are typically difficult to observe using traditional structural techniques because they exist only briefly and at low concentrations relative to stable conformations.[118]

For example, protein folding involves the formation of transient intermediates that can lead either to correctly folded, functional proteins or to misfolded, non-functional or toxic species.[119] Misfolding can lead to conditions such as Alzheimer and Parkinson diseases where protein aggregates are a hallmark.[120,121] Traditional structural methods can reveal the structure of the fully folded or aggregated states, but provide limited information about the intermediates that are crucial for understanding the folding pathway and the points at which misfolding occurs.[56]

Structural Heterogeneity in Macromolecular Complexes

Biomolecular complexes often exhibit structural heterogeneity, where different subunits or domains can adopt multiple conformations simultaneously or sequentially. This heterogeneity can be essential for the function of large macromolecular assemblies such as ribosomes, spliceosomes and viral capsids, which perform complex, multi-step processes. Static structural methods may only capture one or a few of these conformational states, potentially leading to an incomplete or misleading understanding of how these complexes function.[122]

For example, the ribosome undergoes many conformational changes during translation, including interactions with mRNA, tRNAs and various translation factors. Cryo-EM studies have provided snapshots of different states of the ribosome,[123] but to capture the full range of motions and transitions involved in translation, the dynamic nature of these interactions must be taken into account. Relying solely on static structures may miss the coordinated and sequential steps of protein synthesis.

Challenges in Drug Design and Functional Modulation

Drug design requires an understanding of the dynamic aspects of protein targets in order to develop effective therapeutics. Traditional drug design approaches often focus on the structure of the active site of a protein in a static conformation. However, many drug targets, particularly those involved in signalling pathways, are inherently flexible and exist in multiple conformations. Drugs that stabilise specific active or inactive states of a target protein can lead to more effective and selective therapies.[124]

The Role of Environmental Factors

Biomolecular function is also influenced by the local environment, including pH, ionic strength and the presence of other macromolecules. These factors can induce conformational changes or alter the stability of certain states. Structural studies conducted in isolation may not fully represent conditions within a living cell, where the crowded environment and interactions with other cellular components can significantly affect molecular structure and function. This phenomenon, known as the ”molecular crowding effect”, together with the other environmental factors, underlines the importance of studying biomolecules under conditions that closely mimic and reproduce physiological or relevant environments to gain accurate insights into their behaviour.[125]

Limitations in Understanding Disordered Regions

A major drawback of traditional structural approaches is the difficulty in capturing the conformational properties of intrinsically disordered regions (IDRs) or IDR-like regions of proteins. These regions, which lack a stable, well-defined structure, are highly flexible and often play a critical role in protein function. IDRs are involved in diverse biological processes such as signal transduction, molecular recognition and regulation by adopting multiple conformations depending on their interaction partners.[126,127] The dynamic nature of these regions allows them to participate in a wide range of interactions and functional states, which cannot be fully appreciated using static snapshots provided by methods such as X-ray crystallography or cryo-electron microscopy.[128] Without the ability to observe these flexible regions, structural methods miss important functional aspects of proteins, leading to an incomplete understanding of their role in cellular processes.

2.4 Dynamics of Biomolecules

2.4.1 Importance of Dynamics of Biomolecules

Biomolecules are dynamic entities, constantly undergoing conformational changes and fluctuations. Unlike the determination of the static structures, which provide a snapshot of a biomolecule, like a single movie frame, the study of the dynamics means taking into account the motions that a molecule experiences over time. These motions can range from rapid, local fluctuations of atomic positions to large-scale conformational changes involving entire domains or multiple subunits. Without considering the dynamic behaviour of biomolecules, it is not possible to fully understand their ability to interact with other molecules, respond to environmental changes and carry out their biological functions effectively.[112]

Conformational Flexibility and Function

As discussed in the previous section, conformational flexibility refers to the ability of a biomolecule to adopt multiple conformations or shapes, essential for exerting its function. This flexibility allows molecules to interact with different partners, bind to different substrates and participate in complex regulatory networks. Proteins can undergo conformational changes that expose or conceal binding sites, allowing them to regulate interactions with other proteins. These dynamic changes are extremely important, for example, in signalling pathways, where proteins must respond rapidly to cellular signals by changing their conformation to modulate their activity and interactions.[129] Another example is the tumour suppressor protein p53, which has a dynamic behaviour that allows it to bind to different DNA sequences and interact with multiple protein partners to regulate a wide range of cellular processes, including DNA repair, cell cycle arrest and apoptosis.[130]

Role of Dynamics in Protein Folding and Stability

Protein folding is another area that relies heavily on dynamics. To become functional, proteins must adopt specific three-dimensional structures, a process that involves exploring conformational space to reach the energetically favourable folded state. The dynamic nature of folding allows proteins to navigate different pathways and overcome kinetic barriers to achieve their correct conformations.[56] An example of this is the role of molecular chaperones, which assist in protein folding by interacting with partially folded states and facilitating their proper folding. These dynamic interactions prevent aggregation and enhance protein stability within the cellular environment.[131]

Intrinsically Disordered Regions and Their Functional Roles

IDRs and intrinsically disordered proteins (IDPs), although being common in the proteomes of all the living kingdoms,[132] are extreme cases of dynamic biomolecules. Unlike proteins with well-defined structures, IDRs and IDPs lack a stable tertiary structure under physiological conditions.[133] Instead, they exist as flexible, dynamic ensembles of conformations. This intrinsic disorder allows them to perform a wide range of interactions and functions, often serving as hubs in cellular signalling networks.[134]

IDRs can bind to multiple partners with high specificity and low affinity, enabling them to act as molecular scaffolds or hubs that bring together different signalling molecules. The dynamic nature of IDRs allows them to adopt different conformations depending on their interaction partners, facilitating the regulation of complex cellular processes.[135]

Nucleic Acid Dynamics and Function

The biological functions of nucleic acids are strictly related to their dynamic properties. Although DNA is often represented as a static double helix, it undergoes several dynamic processes, including bending, twisting and unwinding. Processes such as replication, transcription and repair depend on these dynamic movements.[136] The flexibility of DNA allows it to wrap around histone proteins to form nucleosomes and higher order chromatin structures that regulate gene expression and access to genetic information. RNA molecules, particularly non-coding RNAs, exhibit considerable conformational flexibility, enabling them to perform a wide range of functions. The dynamic nature of RNA allows it to respond to environmental signals by changing its conformation, thereby modulating its activity and interactions.[137]

2.4.2 Experimental Techniques for Studying Dynamics

To study the dynamics of molecules, a variety of experimental techniques have been developed to provide detailed information about the motions and conformational changes that biomolecules undergo. These techniques include nuclear magnetic resonance (NMR) relaxation, Förster resonance energy transfer (FRET), time-resolved spectroscopy and small-angle scattering (SAS). Each of these techniques offers unique insights into different aspects of molecular dynamics, ranging from local atomic motions to large-scale conformational changes.

Nuclear Magnetic Resonance Relaxation

NMR relaxation is a powerful technique for studying biomolecular dynamics, particularly at the atomic level. In NMR spectroscopy, relaxation refers to the process by which nuclear spins return to their equilibrium state after being perturbed by a radiofrequency pulse. The rates of relaxation, known as the longitudinal and transverse relaxation times, are sensitive to the motions of the nuclei and their interactions with the local magnetic environment.[138] By analysing these relaxation rates, it is possible to gain insight into the timescales and amplitudes of atomic motions within biomolecules.[139]

NMR relaxation provides valuable information about both fast (picosecond to nanosecond) and slow (microsecond to millisecond) dynamics. Fast dynamics are often associated with local motions such as bond vibrations, side-chain rotations and loop flexibility. Slow dynamics can reflect larger conformational changes such as domain movements, folding/unfolding events or ligand interactions.[140] This technique can also identify regions of a protein that exhibit different dynamic behaviours, such as rigid core regions versus flexible surface loops, providing a comprehensive picture of the dynamic landscape of the molecule.[141]

Förster Resonance Energy Transfer

FRET is a technique used to study the distance and orientation between two chromophores, typically called donor and acceptor.[142] When the donor chromophore is excited by a specific wavelength of light, it can non-radiatively transfer energy to the acceptor chromophore if they are in close proximity (typically within 1-10 nm). The efficiency of this energy transfer is highly sensitive to the distance between the donor and acceptor, making FRET an excellent tool for studying molecular interactions and conformational changes.[143]

FRET is particularly valuable for studying the dynamics of biomolecules in real time and in living cells.[144] It can be used to monitor conformational changes, protein-protein interactions and the assembly or disassembly of macromolecular complexes. By labelling specific sites within a molecule with donor and acceptor fluorophores, it is possible to observe changes in FRET efficiency that correspond to changes in distance, revealing dynamic processes such as the opening and closing of enzyme active sites, the folding of proteins, or the conformational changes of nucleic acids during processes such as transcription and replication.[145]

Time-Resolved Spectroscopies

Time-resolved spectroscopy includes a range of techniques that allow the temporal evolution of biomolecular processes to be recorded. These techniques involve monitoring changes in the absorption, emission or scattering of light by a sample following a rapid perturbation, such as a laser pulse. The ability to measure changes on timescales ranging from femtoseconds to milliseconds makes time-resolved spectroscopy ideal for studying fast dynamic processes that are otherwise difficult to detect.[146]

Time-resolved spectroscopy provides insight into the mechanisms of biochemical reactions, conformational transitions and energy transfer processes. For example, time-resolved fluorescence spectroscopy can monitor the dynamics of protein folding by observing changes in the fluorescence of tryptophan residues or labelled probes.[147] Time-resolved infrared and Raman spectroscopy can be used to study changes in secondary structure and hydrogen bonding during enzyme catalysis or protein-ligand binding. By capturing the temporal sequence of events, these techniques can reveal intermediates and transition states, providing a deeper understanding of the pathways and mechanisms underlying biomolecular function.[148]

Small-Angle Scattering

Small-angle scattering (SAS) techniques, including small-angle X-ray scattering (SAXS) and small-angle neutron scattering (SANS), are powerful tools for studying the overall shape, size and conformational flexibility of biomolecules in solution.[149] In these techniques, a beam of monochromatic X-rays (photons) or neutrons is directed at a sample and the scattered radiation is measured at small angles relative to the incident beam. The scattering pattern resulting from this interaction provides information about the distribution of electron density (in SAXS) or nuclear density (in SANS), which can be used to reconstruct the three-dimensional shape and structural properties of the sample.

SAXS is based on the scattering of X-rays by electrons in the sample. It provides insight into the electron density distribution within biomolecules. The technique is particularly useful for studying biological macromolecules such as proteins, nucleic acids and their complexes in their native solution state. SAXS data can be used to determine the radius of gyration (a measure of the compactness of the molecule), the maximum dimension (D_{max}) and the molecular weight of the sample. SAXS can also provide low resolution shape reconstructions and information on the flexibility and conformational changes of biomolecules.[150]

SANS, on the other hand, leverages the scattering of neutrons by nuclei. One of the main advantages of SANS over SAXS is its sensitivity to light elements such as hydrogen, which are abundant in biological samples. In addition, the neutron scattering lengths for different isotopes of the same element can vary significantly. This property forms the basis of contrast variation (or contrast matching), a powerful technique used in SANS to selectively highlight or mask specific components within complex biological systems.[151] The contrast variation technique in SANS takes advantage of the different scattering properties of hydrogen (H) and its isotope deuterium (D).[152] By replacing hydrogen atoms with deuterium in specific parts of a molecule, it is possible to adjust the scattering contrast between different regions of the sample. This ability to selectively vary the contrast is crucial for studying multi-component systems such as protein-lipid complexes, protein-nucleic acid assemblies or large protein complexes.[153] In practice, contrast variation is achieved by preparing samples in solvents with different ratios of H_2O (bulk water) and D_2O (deuterated water). By adjusting the D_2O concentration it is possible to match the scattering length density of the solvent to that of specific components of the sample. When the scattering length density of the solvent is matched to a particular part of the sample, that part becomes effectively ″invisible″ to neutrons, allowing other components to be studied in isolation. This selective focus provides unique insights into the organisation, conformation and interactions of complex biological structures.[154] For example, in the study of membrane proteins, contrast variation can be used to match the scattering signal of the lipid bilayer with that of the solvent. This matching allows the signals from the lipid bilayer and the solvent to be subtracted from the total solution signal, providing a direct measurement of the signal from the embedded protein. This technique isolates the scattering signals of individual molecules from the overall complex, facilitating detailed studies of protein-lipid or protein-protein interactions within membranes.

SAXS and SANS are particularly useful for studying large biomolecules and complexes that are difficult to crystallise, as these techniques do not require the sample to be in a crystalline state. Instead, they allow the analysis of biomolecules in their functional environment, often providing more physiologically relevant information than crystallography or other structural biology techniques. In addition, these methods are well suited to the study of dynamic processes such as protein folding, conformational changes upon ligand binding, and the assembly or disassembly of macromolecular complexes. The ability to study samples in solution allows SAXS and SANS to monitor structural changes in real time, providing insight into the kinetics and mechanisms underlying these biological processes.[155]

Integrating Techniques for Comprehensive Dynamics Studies

Each of these experimental techniques offers unique advantages for studying biomolecular dynamics, capturing a wide range of motions ranging from local atomic fluctuations to large-scale conformational changes. Together, these methods collectively provide a deeper and more comprehensive understanding of how biomolecules function within their native environment. By integrating data from multiple techniques, the inherent limitations of individual methods can be overcome to provide a comprehensive view of dynamic processes. For example, combining NMR relaxation with SAXS not only provides information about the local flexibility but also reveals global conformational changes, offering a detailed picture of molecular behaviour. Similarly, FRET measurements can complement time-resolved spectroscopy by providing real-time, distance-dependent observations of specific molecular interactions, enhancing the understanding of dynamic assemblies and conformational transitions.

2.5 Introduction to Computational Biology

2.5.1 Definition and Role of Computational Biology

Computational biology is an interdisciplinary field that combines principles from biology, computer science, mathematics and physics to analyse and model biological systems. It focuses on the development and application of *in silico* techniques to understand the complexity of biological data and to predict the behaviour of biological systems. In the context of biomolecular research, computational biology contributes to the study of the structures and dynamics of biomolecules, providing insights that complement experimental data.

Defining Computational Biology

At its core, computational biology is the use of algorithms, mathematical models and computer simulations to study biological phenomena. It comprises a wide range of approaches, including molecular dynamics simulations,[156] quantum mechanics,[157] machine learning[158] and bioinformatics.[159] These methods are used to model not only the structure but also the behaviour of biological molecules, to analyse large data sets and to predict the outcome of biological interactions. Computational biology provides a theoretical framework for understanding the principles of life at the molecular level, providing insights that are often difficult or impossible to obtain using experimental techniques alone.

Role in Studying Biomolecular Structures

One of the fundamental approaches is homology modelling, which predicts the three-dimensional structure of a biomolecule based on its similarity to known structures.[160] By aligning the sequence of the target protein with that of a homologous protein with a known structure, it is possible to build a model that approximates the structure of the target. This method is particularly useful for proteins that have significant sequence similarity to known structures, allowing reliable predictions to be made even in the absence of direct experimental data.[161] One of the most widely used homology modelling tools is SWISS-MODEL, which provides an automated online platform for generating high quality structural models.[162]

Ab initio or *de novo* methods, on the other hand, predict protein structures from scratch, relying exclusively on the physical and chemical principles that govern protein folding.[163]

These methods do not require a homologous template and can model proteins with novel folds. Although computationally intensive, *ab initio* techniques have advanced significantly with the development of sophisticated algorithms and increased computing power, allowing the prediction of complex structures.[164]

The advent of artificial intelligence (AI) has revolutionised structure prediction, as demonstrated by the success of AlphaFold[54] and RoseTTAFold.[165] AI-based structure prediction methods use deep learning algorithms trained on large datasets of known protein structures to predict the folding patterns of proteins. AlphaFold, for example, has achieved remarkable accuracy in predicting protein structures, approaching the level of precision achieved by experimental methods.[166] These AI-driven techniques are now widely used in structural biology, providing fast and reliable structural models to guide experimental design and functional analysis.[167]

Role in Understanding Biomolecular Dynamics

Beyond static structures, computational biology can shed light on the dynamic aspects of biomolecules. Dynamics is essential for understanding how biomolecules interact, change shape and achieve their biological functions. Computational approaches such as molecular dynamics (MD) simulations are powerful tools that model the movements of atoms within a biomolecule over time, providing a detailed view of the behaviour of the molecule in its, although limited, native environment. MD simulations can capture a wide range of motions, from fast, local fluctuations such as bond vibrations and side-chain rotations, to more substantial conformational changes that can occur on microsecond to millisecond timescales.[168] Simulations can provide insights into how local motions contribute to overall molecular function, stability and interactions with other molecules.

While MD simulations are increasingly able to approach the timescales on which significant conformational changes can be observed, they often excel at revealing smaller scales. MD simulations model the atomic-scale motions of biomolecules over time, revealing, for example, how proteins fold and how ligands bind to receptors.[169] MD can simulate the complex environments in which biomolecules operate, including solvent effects, ionic strength and temperature variations.[156] In addition to MD simulations, quantum mechanics/molecular mechanics (QM/MM) methods are used to study enzyme catalysis at a more detailed level.[170] These hybrid methods combine quantum mechanical calculations for the reactive site with molecular mechanics for the surrounding environment, providing a more accurate description of the catalytic process.[171] Understanding enzyme catalysis through computational methods not only reveals the mechanisms of action, but also aids the design of inhibitors and the engineering of enzymes for industrial applications.[172]

Exploring Molecular Interactions and Complex Formation

Computational biology can be applied to the study of biomolecular interactions and complex formation. Molecular docking simulations are widely used to predict how small molecules, such as drugs, interact with larger biomolecular targets, such as proteins or nucleic acids.[173] By modelling the binding affinity and orientation of ligands, computational approaches can identify potential drug candidates and predict their efficacy.

In recent years, AI-based methods have significantly advanced the prediction of protein-protein and protein-ligand interactions. Tools such as AlphaFold Multimer[174] extend the capabilities of single-chain structure prediction to multi-chain complexes, enabling accurate modelling of protein complexes. These AI approaches predict the interface residues and overall architecture of the complexes, providing insights into how proteins interact to form functional assemblies.

Understanding the Effects of Mutations

Another important application of computational biology is to understand the effects of genetic mutations on the structure and function of biomolecules. Mutations can alter the stability, folding and interactions of biomolecules, leading to changes in cellular behaviour. Computational models can predict the effects of specific mutations on the structure and dynamics of proteins, providing insights into the molecular basis of genetic disorders.[175]

Scoring functions are widely used in computational biology to predict the effects of mutations.[176] These functions evaluate the stability of protein structures and the effects of amino acid substitutions, leading to the identification of potentially deleterious mutations or, conversely, driving the design of desired properties. AI tools have further enhanced mutation analysis by using machine learning algorithms trained on experimental mutation data to predict the consequences of mutations. These tools can distinguish between benign and pathogenic mutations, predict changes in binding affinity, and identify potential drug resistance mutations.[177–179]

Generate Hypotheses and Drive Experiments

One of the key strengths of computational biology is its ability to generate hypotheses about molecular function. By simulating molecular processes and analysing large datasets, computational methods can identify patterns and predict outcomes that guide experimental research. This iterative process of testing and refining computational predictions with experimental data accelerates the discovery of new biological principles and therapeutic targets.

Computational biology also makes it possible to explore specific scenarios in advance of the experiment. For example, by simulating the effects of different environmental conditions, such as temperature or pH, it is possible to predict how these factors will affect function and stability, and then design the experiment according to the results obtained *in silico*.

2.5.2 Computational Biology Limitations

While computational biology provides powerful tools for the study of biomolecular structure and dynamics, it is important to recognise the inherent limitations and challenges associated with these approaches. Despite their potential to provide detailed insights into molecular behaviour, computational methods face several limitations that can affect the accuracy and reliability of their predictions.

Sampling Limitations in Molecular Dynamics

One of the main limitations of MD simulations is sampling. Capturing the full range of relevant conformational states and dynamic behaviours is often challenging due to the limited timescales that can be practically simulated. While dedicated hardware has extended the achievable simulation timescales to milliseconds,[180,181] many biologically important processes, such as large-scale conformational shifts, protein folding or rare binding events, can occur on timescales of milliseconds to seconds or longer. These events are often not captured in typical MD simulations, leading to incomplete sampling and potentially missing critical aspects of molecular function.[182]

In addition, MD simulations can become trapped in local energy minima, especially when exploring complex energy landscapes with multiple conformational states. This can lead to biased sampling, where certain states are overrepresented while others that may be biologically relevant are not adequately explored. Enhanced sampling methods, such as replica exchange[183] or metadynamics[184] have been developed to address these limitations, but they present their own challenges and may not always be applicable to all systems.[185]

Accuracy and Limitations of Force Fields

The accuracy of MD simulations is highly dependent on the force fields (FFs) used to model the interactions between atoms. FFs are mathematical functions that describe the potential energy of a system based on the positions of its atoms. While significant progress has been made in developing FFs that accurately represent a wide range of biomolecular interactions, they are still approximations that cannot capture all the details and nuances of real molecular interactions.[186,187] Limitations in FF accuracy can lead to errors in predicted structures, binding affinities and reaction mechanisms.[188]

For example, FFs may struggle to accurately model interactions involving metal ions, charged groups, or highly polarised environments, which are common in biological systems. Additionally, the treatment of solvent effects, hydrogen bonding, and dispersion forces may not be sufficiently precise, leading to deviations from experimental observations. These limitations underscore the need for continuous refinement of FFs and validation of simulation results against experimental data.[189]

See section 2.6.1 for more information on FFs.

Challenges in Quantum Mechanics Calculations

Quantum mechanics (QM) methods provide a more detailed description of molecular interactions by explicitly considering the electronic structure of atoms. These methods can be used to study chemical reactions, enzyme catalysis and interactions involving electronic transitions.[68] However, the high level of detail provided by QM calculations comes at a significant computational cost. QM methods are inherently slower than classical MD simulations, limiting their applicability to small systems or specific regions of interest within larger molecules.[171]

The computational cost of QM methods increases rapidly with the size of the system, making it difficult to apply these techniques to large biomolecules or complex environments. Hybrid methods, such as quantum mechanics/molecular mechanics (QM/MM), combine the accuracy of QM for the active site with the efficiency of classical mechanics for the surrounding environment.[190] While QM/MM methods have extended the applicability of QM calculations, they still face challenges in balancing accuracy and computational feasibility.[191]

Limitations in Molecular Docking and Scoring Functions

Molecular docking is a widely used computational technique for predicting how small molecules will bind to target proteins. While docking provides valuable insight into potential binding modes and affinities, it is not without limitations. Docking algorithms are often based on simplified representations of both the ligand and the target protein, which cannot fully capture the flexibility and conformational changes that occur upon binding.[192] As a result, docking predictions can sometimes be inaccurate or fail to identify the true binding mode.[193]

Scoring functions, which evaluate the binding affinity of a docked complex, are another source of limitations.[176] These functions are based on empirical or theoretical models that estimate the strength of the interactions between the ligand and the target. However, scoring functions may not take into account all relevant factors such as solvation effects, entropic contributions or the dynamic nature of the binding process.[194,195] This can lead to false positives or negatives in virtual screening campaigns, highlighting the need for more sophisticated scoring methods and the integration of additional experimental data.[196]

Limitations of AI-Based Structure Prediction

The advent of artificial intelligence (AI) applied to biology has significantly advanced structure prediction. However, AI-based structure prediction methods have their limitations. These algorithms are trained on large datasets of known protein structures, which means that their predictive power is highest for proteins that are similar to those already present in the training set. Novel protein folds or structures that do not have close homologs in the training data may be predicted with lower accuracy.[197]

In addition, AI-based methods face challenges in predicting the structures of intrinsically disordered regions. Because IDRs lack a stable three-dimensional structure, AI tools can identify these regions but cannot provide detailed structural information. This limitation reflects a broader challenge in computational biology: capturing the behaviour of flexible, dynamic regions that do not conform to well-defined, stable structures. As a result, while AI tools are excellent at predicting structured regions, they provide limited insight into the functional roles of disordered regions.[128]

Computational Power and Environmental Impact

Another significant limitation is the intensive computing power required for many computational biology approaches. High-performance computing (HPC) resources are often required to perform MD simulations, QM calculations and AI-based predictions. Access to such resources can be a barrier for researchers with limited computing infrastructure. In addition, the environmental impact of computational research is increasingly recognised, as the energy consumption of HPC facilities contributes to carbon emissions and climate change.[198] The demand for computing power is expected to grow as the complexity and size of simulations increase, highlighting the need for more efficient algorithms and sustainable computing practices.[199]

2.5.3 Historical Perspective

The field of computational biology has undergone a remarkable evolution, paralleling advances in computing power, algorithms and the growth of biological data. This journey from early computational analyses to the sophisticated simulations and models of today has been marked by a number of key developments.

Early Computational Biology: The Foundation

The roots of computational biology can be traced back to the mid-20th century when early computers were first applied to biological problems. One of the earliest and most influential works was the development of the mathematical model of enzyme kinetics by Leonor Michaelis and Maud Leonora Menten in 1913, which laid the foundations for quantitative biology.[200] However, it was not until the 1960s that computers began to be used extensively in biological research.

In the 1960s and 1970s, pioneers such as Margaret Oakley Dayhoff began using computers to compare protein sequences, leading to the creation of the first protein sequence databases and the Atlas of Protein Sequence and Structure.[201] Dayhoff's work laid the foundations for bioinformatics by introducing the concept of evolutionary trees and the first methods for sequence alignment, which are crucial for understanding molecular evolution and the relationships between different species.

The Advent of Sequence Analysis and Molecular Modelling

The explosion of molecular biology in the 1970s, marked by advances in DNA sequencing and the discovery of restriction enzymes, accelerated the need for computational tools. In the 1980s, the development of sequence alignment algorithms, such as the Needleman-Wunsch[202] (1970) and Smith-Waterman[203] (1981) algorithms, provided the basis for comparing DNA and protein sequences. These methods made it possible to identify homologous sequences and conserved motifs, which are essential for understanding functional domains in proteins.

In parallel with these developments, the field of molecular modelling began to take shape. Early molecular dynamics simulations, pioneered by Martin Karplus and colleagues in the late 1970s, demonstrated the potential of computational methods to study the structural dynamics of proteins.[204] This era also saw the development of the Ramachandran plot by Gopalasamudram Narayana Iyer Ramachandran in 1963, which provided insights into the allowable angles of peptide backbones, information used to predict protein structure.[205]

Introduction of Protein Structure Databases

The 1970s and 1980s saw the establishment of structural biology databases. The Protein Data Bank (PDB), founded in 1971, became a central repository for 3D structural data of biological macromolecules.[206] This resource enabled researchers worldwide to access and analyse protein structures and catalysed the development of structural bioinformatics and molecular visualisation tools such as Visual Molecular Dynamics[207] (VMD, 1995), developed by Klaus Shulten and colleagues, and PyMOL[208] (2000), developed by Warren Lyford DeLano and colleagues.

The Human Genome Project and the Rise of Bioinformatics

The launch of the Human Genome Project (HGP) in 1990 marked a turning point for computational biology.[209,210] The HGP drove the development of new computational methods for sequencing, assembling and annotating genomic data. The sheer volume of data generated required advances in data storage, management and analysis, leading to the establishment of bioinformatics as a discipline in its own right.[211]

During this period, the BLAST algorithm, developed by Stephen Frank Altschul and colleagues in 1990, revolutionised sequence analysis. BLAST allowed rapid comparison of nucleotide and protein sequences against large databases and became an indispensable tool in genomics.[212]

Computational Phylogenetics and Evolutionary Biology

The 1990s also saw significant advances in computational phylogenetics, with the development of maximum likelihood and Bayesian methods for reconstructing evolutionary trees. Tools such as PHYLIP (1980), PAUP (1996) and MrBayes (2001) became standard in the analysis of evolutionary relationships, providing insights into the origins and diversification of species based on genetic data.[213–216]

Structural Biology and Computational Advancements in the 21st Century

In the early 21st century, computational biology continued to expand, driven by exponential increases in computing power and the availability of high-throughput experimental data. The integration of computational methods with experimental techniques such as X-ray crystallography, NMR spectroscopy, SAS techniques, and cryo-electron microscopy has led to more accurate and detailed structural models of macromolecules.

Major milestones include the development of Rosetta[217] (1999), a software suite for protein structure prediction and design, and the success of the Critical Assessment of Structure Prediction (CASP) experiments initiated in 1994.[218] CASP provided a platform for assessing the accuracy of computational protein structure predictions and has guided improvements in modelling techniques over the years.[166]

The Genomics Revolution and Systems Biology

The sequencing of the human genome, completed in 2003, and subsequent advances in next-generation sequencing technologies have transformed computational biology. The ability to sequence whole genomes rapidly and cost-effectively has led to the emergence of comparative genomics, personalised medicine and metagenomics.[214] Databases such as GenBank and the European Nucleotide Archive (ENA) have grown exponentially, providing vast resources for comparative analyses.[219] The recognition that biological systems function as complex networks of interacting components gave rise to systems biology in the early 2000s.[220] Computational models began to incorporate data from genomics, transcriptomics, proteomics and metabolomics to provide a holistic view of cellular function. The development of tools for network analysis and pathway modelling, such as Cytoscape[221] (2002), facilitated the understanding of complex biological systems.

AI and Machine Learning Transform Computational Biology

In recent years, the advent of artificial intelligence and machine learning has transformed computational biology. Techniques such as AlphaFold,[54] developed by DeepMind, and RoseTTAFold,[222] developed by David Baker's lab, have achieved unprecedented accuracy in predicting protein structures directly from amino acid sequences, overcoming challenges that have persisted for decades. The success of AlphaFold in the CASP14 competition in 2020 marked a turning point, demonstrating that AI could predict protein structures with atomic-level accuracy.[166,223]

Machine learning approaches have also been applied to other areas of computational biology, including drug discovery, where AI models are used to predict drug-target interactions, optimise drug design and identify potential therapeutic targets.[224–227]

The Rise of CRISPR and Genome Editing Technologies

The discovery of CRISPR-Cas9 as a genome editing tool in 2012 had a profound impact on computational biology.[228,229] The ability to precisely edit genomes required the development of computational tools to design and evaluate guide RNAs, predict off-target effects, and analyse the results of genome editing experiments. Databases and tools such as CRISPRseek[230] and CRISPRdirect[231] have become essential resources for working with CRISPR technology.

2.6 Computational Techniques for Studying Biomolecules

In this section I describe the basics of the techniques I have used most in the projects I have contributed to.

2.6.1 Details of Molecular Dynamics

Molecular dynamics (MD) simulations are computational methods widely used to explore and predict the structure, behaviour, and thermodynamic properties of molecular systems that are too complex for traditional experimental approaches.[232–234] In MD, a system is represented as a model of interacting particles and its evolution over time is tracked through a dynamic trajectory derived by numerical integration of Newton's equations of motion.[233] Different theoretical frameworks can be employed to describe these models. For example, quantum mechanics (QM) approaches explicitly model electrons and calculate interaction energies by solving the electronic structures of molecules. However, due to the high computational requirements, QM methods are generally limited to relatively small systems of only a few hundred atoms.[235] For larger and more complex systems, molecular mechanics (MM) approaches are used, which treat molecules as collections of atoms or groups of atoms, thereby reducing the computational effort.[232]

However, MM approaches lack the ability to model chemical reactions because they do not allow for changes in the topology of the system, such as bond breaking or formation. In classical MD simulations, atoms are often modelled as rigid spheres connected by springs, representing the electron clouds involved in covalent bonding. The Born-Oppenheimer approximation is often used, which assumes that the electrons instantaneously adapt to the motion of the nuclei, effectively decoupling their motion.[234] This allows only the motion of the nuclei to be included in the calculation. The force acting on each atom in the system is calculated from the positions of all the other atoms, as defined by the following equation:

$$\overrightarrow{F_i}(t) = m_i \overrightarrow{a_i}(t) = m_i \frac{d\overrightarrow{v_i}(t)}{dt} = m_i \frac{d^2\overrightarrow{r_i}(t)}{dt^2}$$

In this equation, $\overrightarrow{F_i}(t)$ represents the force on atom i, m_i is the mass of atom i, $\overrightarrow{a_i}(t)$ is its acceleration, $\overrightarrow{v_i}(t)$ is its velocity, $\overrightarrow{r_i}(t)$ is its position, and t denotes time. An appropriate time step, typically around 2 femtoseconds to match the fastest vibrational frequencies (such as C-H bond vibrations), is chosen to integrate Newton's equations and update atomic positions. Initial positions are usually derived from crystallographic data.[236] The calculation of new atomic positions requires the definition of both velocities and forces: velocities are often sampled from a Boltzmann distribution, while forces are derived as the negative gradient of the potential energy described by the force field equations. A major challenge in MD simulations is the handling of the huge number of atoms in biological systems, which can reach hundreds of thousands or more, especially when solvent molecules are taken into account. Since the Newton's equations of motion cannot be solved analytically for such large systems, numerical integrators are used. These algorithms approximate solutions and update atomic positions (r_i) based on initial positions, velocities and the steric energy of the system. The approximation is often based on the Taylor series, which predicts the behaviour of a function around a given point:[237,238]

$$\overrightarrow{r_i}(t+\Delta t) = \overrightarrow{r_i}(t) + \frac{\Delta\overrightarrow{r_i}(t)}{\Delta t}\Delta t + \frac{1}{2}\frac{d^2\overrightarrow{r_i}(t)}{\Delta t^2}\Delta t^2 + ...$$

$$\overrightarrow{v_i}(t+\Delta t) = \overrightarrow{v_i}(t) + \frac{d\overrightarrow{v_i}(t)}{\Delta t}\Delta t + \frac{1}{2}\frac{d^2\overrightarrow{v_i}(t)}{\Delta t^2}\Delta t^2 + ...$$

$$\overrightarrow{a_i}(t+\Delta t) = \overrightarrow{a_i}(t) + \frac{d\overrightarrow{a_i}(t)}{\Delta t}\Delta t + ...$$

These series can be more concisely expressed as:

$$\overrightarrow{r_i}(t+\Delta t) = \overrightarrow{r_i}(t) + \overrightarrow{v_i}(t)\Delta t + \frac{1}{2}\overrightarrow{a_i}(t)\Delta t^2 + ...$$

The Verlet integrator is one of the most frequently used methods due to its accuracy and simplicity. It calculates new atomic positions by expanding $\overrightarrow{r_i}(t)$ at times $t+\Delta t$ and $t-\Delta t$, incorporating acceleration to provide:

$$\overrightarrow{r_i}(t+\Delta t) = 2\overrightarrow{r_i}(t) - \overrightarrow{r_i}(t-\Delta t) + \overrightarrow{a_i}(t)\Delta t^2 + ...$$

Except for the initial step, where velocities are required to calculate $\overrightarrow{r_i}(t)$ from $\overrightarrow{r_i}(t-\Delta t)$, velocities are not explicitly calculated but can be derived using:

$$\overrightarrow{v_i}(t) = \frac{\overrightarrow{r_i}(t+\Delta t) - \overrightarrow{r_i}(t-\Delta t)}{2\Delta t} + ...$$

MD simulations fundamentally rely on force fields to model how the energy of a system varies with the positions of its atoms. A force field (FF) is essentially a set of mathematical expressions and parameters designed to approximate the potential energy of a molecular system. These parameters can be derived from various experimental techniques, including X-ray diffraction, electron diffraction, NMR and infrared spectroscopy, as well as from *ab initio* or semi-empirical quantum mechanical calculations.[237] By replacing the real potential energy with a simplified model, force fields achieve a balance between computational efficiency and the ability to accurately reproduce the physical and chemical properties of the system. Numerous force fields have been developed, each calibrated against empirical data to ensure that its set of parameters reproduces specific physicochemical properties in agreement with experimental observations.[237] Despite differences in complexity between different force field models, a generalised expression for a force field can be expressed as:

$$\begin{aligned} \mathrm{U} = &\sum_{\text{bonds}} \frac{1}{2}k_b(r-r_0)^2 + \sum_{\text{angles}} \frac{1}{2}k_a(\theta-\theta_0)^2 + \sum_{\text{torsions}} \frac{V_n}{2}[1+\cos(n\theta-\delta)] \\ &+ \sum_{\text{improper}} V_{\text{imp}} + \sum_{\text{LJ}} 4\epsilon_{\text{ij}}\left(\frac{\sigma_{\text{ij}}^{12}}{r_{\text{ij}}^{12}} - \frac{\sigma_{\text{ij}}^{6}}{r_{\text{ij}}^{6}}\right) + \sum_{\text{elec}} \frac{q_i q_j}{r_{ij}} \end{aligned}$$

The first four terms of this expression capture intramolecular interactions, including bond stretching, angle bending, dihedral torsion, and improper torsions, which help to maintain molecular planarity. The last two terms represent intermolecular interactions: the 12-6 Lennard-Jones potential accounts for repulsive and van der Waals forces, while Coulombic interactions describe electrostatic forces between charged particles.[237] Bond stretching is typically modelled using a harmonic potential, where the bond length deviations from the equilibrium position follow a quadratic function. This harmonic approximation is usually effective for small displacements, but fails to accurately describe bond stretching beyond 10% of the equilibrium bond length, making it unsuitable for modelling bond breakage events that occur during chemical reactions. Similarly, angular deflection, which describes the deviation of bond angles from their equilibrium values, is often represented by a harmonic potential that captures the energy cost associated with these deviations. Torsional interactions play a crucial role in defining the conformational preferences of macromolecules. These torsional energies are usually modelled by a cosine function, which can account for multiple periodic minima corresponding to different stable conformations. The parameters for these torsional potentials are often derived from high-level *ab initio* quantum calculations and further refined on the basis of experimental data. Improper torsions are also considered to enforce the planarity of certain structural elements, such as aromatic rings and peptide bonds, where simple torsional potentials may not provide sufficient constraints.[237] In addition to these intramolecular forces, intermolecular interactions are essential for accurate modelling of molecular systems. The Lennard-Jones potential captures both the short-range repulsive interactions (arising from the Pauli exclusion principle) and the longer-range van der Waals attractions, which are required to model non-bonded interactions. Electrostatic interactions are governed by Coulomb's law, which calculates the force between charged particles scaled by their partial charges and the distance between them.

2.6.2 Details of Molecular Docking

Molecular docking is a computational technique used in structural biology and drug discovery to predict the putative binding sites of a small molecule, known as a ligand, on a target macromolecule. By simulating the specific interactions between the ligand and its target, docking helps to elucidate the binding mode, approximate the binding affinity and provide a molecular description, even at the atomistic level, of the ligand-receptor relationship, which is the basis for drug design and discovery.[239,240] This technique supports drug development at various stages, including the study of molecular activity, the optimisation of lead compounds and the virtual screening of large compound libraries to identify potential drug candidates. Identifying the most favourable binding pose of the ligand to the target protein requires a combination of search algorithms to explore the conformational space of the ligand and scoring functions to evaluate and rank the potential binding poses based on their predicted binding affinities.[241]

Scoring functions are required for the docking process as they quantify the quality of the interaction between the ligand and the target molecule and guide the identification of the best binding pose. These functions can be divided into three main types: empirical, force field based and knowledge based.[242] Empirical scoring functions use a weighted sum of interaction terms derived from experimental binding data to approximate binding affinity. Force field-based scoring functions calculate the interaction energy using classical force fields, taking into account van der Waals interactions, electrostatic forces and hydrogen bonding. Knowledge-based scoring functions, on the other hand, derive their parameters from statistical analysis of known protein-ligand complexes, using structural data to generate the scoring.[243]

Search algorithms used in molecular docking are designed to efficiently navigate conformational space given the large number of possible ligand poses. These algorithms are generally divided into systematic and stochastic methods.[241] Systematic search methods scan the conformational space in a predefined, deterministic manner, often using grid-based approaches to explore the possible poses. Although thorough, these methods can be computationally expensive and may not be practical for systems with a high degree of flexibility. Stochastic search methods, such as genetic algorithms, Monte Carlo simulations and simulated annealing, introduce random variations into the ligand conformation and iteratively refine these variations based on scoring function evaluations. These methods are particularly well suited to high-dimensional problems, such as flexible ligand-protein docking, because they can efficiently sample a wider range of conformations without the need for exhaustive searching.[243]

In most molecular docking simulations, the target system is treated as a rigid structure while the ligand is allowed to explore different conformations. This approach simplifies the computational requirements, but can overlook significant induced-fit effects where the protein undergoes conformational changes upon ligand binding. Advanced docking methods take into account protein flexibility, either by modelling it explicitly or by using ensemble docking techniques that consider multiple protein conformations.[244,245] Ligand flexibility is also a critical factor, as ligands with numerous rotatable bonds have a larger conformational space, which can exponentially increase the complexity and time required for docking simulations.

The evaluation of docking results involves more than just identifying the lowest energy binding site. A thorough evaluation requires consideration of the chemical and structural complementarity between the ligand and the protein binding site. Key factors include hydrogen bonding, hydrophobic interactions, electrostatic complementarity and steric fit. In addition, it is important to evaluate the predicted binding affinities against experimental data, if available, to validate the accuracy of the docking predictions.[243]

2.6.3 Deep Learning Models in Structure Prediction and Protein Design

The computational approaches based on deep learning models use large datasets and complex algorithms to predict the three-dimensional structures of proteins from their amino acid sequences and to design new proteins with specific functions or to modify existing proteins to enhance or change their properties. Models such as AlphaFold and RoseTTAFold have set new benchmarks in structure prediction accuracy, while denoising diffusion models such as RFDiffusion and other generative methods have opened up new avenues in protein design.

Protein Structure Prediction

Traditional methods of structure prediction, such as homology and *ab initio* modelling, often struggle with accuracy and scalability, especially for proteins without close homologs in existing structural databases.[246] Machine learning models, particularly those based on deep learning, have dramatically improved the accuracy of structure prediction. In particular, AlphaFold, developed by DeepMind, and RoseTTAFold, developed by David Baker's group, have demonstrated the ability to predict protein structures with near-experimental accuracy.[54,222]

AlphaFold uses a deep neural network architecture that combines multiple sequence alignments (MSAs), evolutionary couplings and structural templates to predict distances and angles between residues, which are then used to generate a three-dimensional structure. The model uses a combination of convolutional neural networks (CNNs) to capture local sequence features and attention mechanisms to model long-range interactions. The output of these networks is fed into a gradient descent-based structure refinement process, which iteratively improves the accuracy of the predicted structure.[54]

Similarly, RoseTTAFold[222] integrates a three-track neural network architecture that simultaneously processes sequence information, pairwise distances and predicted coordinates, allowing iterative refinement and direct prediction of atomic coordinates. This approach significantly reduces the computational complexity and time required for structure prediction, making it suitable for high throughput applications.[222]

***De Novo* Protein Design and Protein Modification**

Machine learning models are also driving innovation in *de novo* protein design and the modification of existing proteins.[247] In *de novo* design, the aim is to create entirely new proteins that do not exist in nature but have specific structural or functional properties. In protein modification, existing proteins are modified to increase their stability, alter their function, redesign regions to remove or include new domains, or improve their binding affinity to other molecules, such as drugs or substrates.

Generative models, including variational autoencoders (VAEs), generative adversarial networks (GANs) and, more recently, denoising diffusion models, are at the forefront of *de novo* protein design.[248,249] These models are trained on datasets of protein sequences and structures and learn to generate new sequences that fold into stable and functional three-dimensional structures. Denoising diffusion models,[226] for example, start with a random cloud of backbone atoms and iteratively refine it by removing ”noise” to generate a structure with the desired features. Another model, called Protein Message Passing Neural Network[225] (ProteinMPNN), predicts the sequence that encodes for that structure. This approach allows the generation of novel proteins that meet specific design criteria, such as binding to a particular target.[250] In the context of modifying existing proteins, these models can be adapt to redesign specific regions of a protein, such as the scaffold around an active site.[251]

2.7 Integrative Modelling: Combining Computational and Experimental Approaches

While computational methods provide powerful tools for simulating biomolecular behaviour and predicting structures, they are often limited by approximations, sampling issues and the inherent assumptions of their models. Similarly, experimental techniques provide direct observations and empirical data, but can be limited by resolution limitations, sample preparation challenges and the static nature of the information they provide. Integrative modelling seeks to exploit the strengths of both computational and experimental approaches, creating a synergistic framework that overcomes the limitations of each method and provides a more comprehensive understanding of biomolecular systems.

2.7.1 The Importance of Combining Approaches

Experimental data can provide validation for computational predictions, ensuring that the models accurately represent biological reality. Computational methods, such as molecular dynamics simulations, rely on force fields and parameters that approximate the interactions between atoms and molecules. While these models have been refined over time, they are still subject to inaccuracies and may not fully capture the complexity of real molecular interactions. By incorporating experimental data, such as NMR and SAS restraints or cryo-EM density maps, computational models can be corrected and refined to better match observed behaviour, improving the reliability and accuracy of predictions. Experimental data can helping to overcome sampling limitations and reduce the conformational space that needs to be explored. Indeed, molecules can adopt numerous conformations, and reaching the functionally relevant states within this space can be challenging. Experimental techniques, such as FRET, DEER, or SAS, can provide information about the likely conformations or distances within a molecule, which can be used to restrain or bias simulations towards these regions. This targeted approach allows focus on specific relevant states.

The integration of experimental data can help to bridge the limitations of each experimental technique by providing complementary information. For example, X-ray crystallography provides high-resolution structures of static conformations, but may not capture dynamic aspects or transient states. NMR spectroscopy provides information on molecular dynamics, but may lack the spatial resolution of crystallography. Cryo-EM can image large complexes under near-native conditions, but may struggle with smaller or more flexible proteins. By combining data from different experimental techniques within a computational framework, integrative modelling can provide a holistic view that captures both structural detail and dynamic behaviour, providing insights not possible with any single method.

Integrative modelling addresses the limitations of both computational and experimental approaches by creating a feedback loop where experimental data validate and refine computational models, and computational predictions guide and interpret experimental results. This iterative process enhances the reliability and accuracy of biomolecular studies. For instance, when studying protein-ligand interactions, experimental techniques such as X-ray crystallography or cryo-EM can provide structural snapshots of the binding site, while molecular dynamics simulations can explore the binding pathway and dynamics.

Integrative modelling also provides a framework for studying systems that are inaccessible to traditional methods alone. For example, intrinsically disordered regions, which lack stable structures and are highly dynamic, cannot be assessed using techniques that require well-defined structures, such as X-ray crystallography. By combining experimental techniques that capture dynamic behaviour, such as NMR and SAXS, with computational simulations, integrative modelling can provide insights that generate conformational ensembles in agreement with experimental data.

2.7.2 Techniques for Integrative Modelling

Integrative modelling combines computational simulations with experimental data to construct detailed and accurate representations of biomolecular systems. By integrating these two approaches, it is possible to leverage the strengths of each, using experimental data to validate and refine computational models, while computational techniques provide dynamic and structural insights that extend beyond experimental observations. Several methodologies have been developed to facilitate this integration, ensuring that computational models reflect biologically relevant behaviours and conformations.

Data Reweighting

A posteriori reweighting is used to adjust the results of computational techniques that generate molecular conformations to better match experimental data.[252] For example, after performing an MD simulation, the frames of the resulting conformational ensemble can be reweighted based on experimental observations. This process involves assigning different weights to the simulated conformations, favouring those that are in better agreement with experimental data such as NMR chemical shifts, FRET or DEER derived distances, SAXS radii of gyration. The reweighted ensemble provides a more accurate representation of the conformational space of the biomolecule, reflecting both the simulation and the experimental evidence. Reweighting techniques are particularly useful for interpreting experimental data averaged over multiple states or conformations. By reweighting the simulation data it is possible to identify the most likely conformational states and quantify the population of each state. This can lead to more reliable and informative models.

Simulation Restraints

In contrast to a posteriori reweighting, restraints for simulations incorporate experimental data directly into the calculation, driving the molecular dynamics to conform with experimental observations as the simulation progresses. This approach applies restraints, derived from experimental measurements, to specific degrees of freedom within the system. Restraints can be based on data from a variety of experimental techniques, including NMR, FRET, DEER, SAS, cryo-EM. For example, distance restraints derived from DEER experiments can be used to maintain specific distances between spin-labelled sites during an MD simulation. This ensures that the simulation remains consistent with experimental observations and explores conformations that are experimentally validated. The restraints, which can be thought of as energy penalties, can be applied using relatively simple approaches, such as the introduction of linear or harmonic potentials, or through more sophisticated methodologies, such as metainference.[253] Simple harmonic restraints penalise deviations from experimentally determined measurements, effectively biasing the simulation towards conformations that satisfy these constraints. Linear restraints, on the other hand, can provide a softer enforcement of the experimental data, allowing more flexibility in the conformational space while still favouring agreement with experimental observations. Metainference represents a more advanced approach to applying restraints in integrative modelling. It is a Bayesian replica-averaging framework that operates according to the maximum entropy principle, ensuring that the resulting ensemble of conformations is as unbiased as possible while still being consistent with experimental data. In metainference, multiple replicas of the system are simulated simultaneously, each preferably representing a different state of the conformational space. The experimental data are used to apply a global restraint across all replicas, penalising deviations from the average predicted data to the experimental observables. The advantage of metainference lies in its ability to account for both the uncertainty in the experimental data and the variability within the conformational ensemble. By treating experimental data as probabilistic constraints rather than fixed values, metainference allows for the generation of ensembles that reflect the inherent flexibility and heterogeneity of biomolecules.[254]

Data Driven Docking

Data-driven docking is another integrative modelling technique that uses experimental data to inform and enhance the docking process, improving the accuracy and reliability of predicted biomolecular interactions. Traditional docking approaches often assume that the binding site of a ligand to a protein is unknown, resulting in a blind search of the entire surface of the target protein. This can lead to numerous false positives and less reliable predictions, especially when the target protein has multiple potential binding sites. Incorporating experimental data into the docking process allows for more targeted and accurate predictions. If the binding site of the target protein or its homologue is known from experimental observations, the docking process can be focused on the relevant region, significantly reducing the search space and increasing the accuracy of the results. This targeted approach is less "blind" and allows more accurate predictions of how a ligand will interact with the protein. Experimental data can also be used to make docking more specific by applying constraints based on known interactions or properties. For example, if a particular residue is known to play a critical role in binding, this information can be used to bias the docking process towards conformations that satisfy this interaction. This method is useful for predicting the binding modes of ligands that interact with proteins through key residues or functional motifs. An example of a software tool that employs data-driven docking is HADDOCK[255,256] (High Ambiguity Driven biomolecular DOCKing), developed by Alexandre Bonvin and colleagues.

3 — Manuscripts

In this chapter, I present personal contributions that illustrate the use of computational techniques in combination with experimental data to study biomolecular systems. The work presented here highlights the power of integrative modelling, where the synergy between computational simulations and experimental observations provides deeper insights into the structure, dynamics and function of biomolecules. By using computational approaches alongside experimental data, these studies aim to overcome the limitations of each method and provide more comprehensive and accurate models of biological systems. The manuscripts included in this chapter serve as practical examples of how integrative modelling can be used to address complex biological questions and generate hypotheses that can guide future experimental work.

The first manuscript discusses the development and implementation of a small-angle scattering model designed to reconstruct *in silico* scattering intensities directly from the atomic coordinates of a target system during molecular dynamics simulations. The method uses a coarse-grained model to efficiently calculate scattering intensities, thereby reducing computational costs. This model was combined with metainference.[253] This Bayesian replica-averaging framework allows multi-replica MD simulations to be driven to generate an ensemble of conformations in agreement with experimental data. By dynamically applying experimental restraints, the conformational ensemble of the protein gelsolin was generated, providing a realistic and accurate representation that matches the measured scattering data. This integrative approach not only improves the accuracy of simulations, but also provides insights into structural dynamics and exemplifies how computational models can be validated or refined using experimental data.

The second manuscript describes the first *in silico* inactivation of a human olfactory receptor, OR51E2, highlighting the role of calcium ions in receptor state transitions. Due to the inherent challenges associated with determining the structures of G protein-coupled receptors (GPCRs), particularly olfactory receptors, computational techniques provide an alternative for studying their activation and inactivation mechanisms. This work integrates several computational methods, including AI-based structure prediction, homology modelling and MD simulations. Our results propose a novel molecular mechanism for olfactory receptor inactivation and provide a basis for experimental validation, illustrating how computational hypotheses can guide experimental studies.

The third manuscript explores the molecular mechanisms underlying a phenotype observed in a barley mutant, *TM2490*, which exhibits a pale green colouration while retaining wild-type growth and morphology. This phenotype is attributed to a missense mutation in the *Xan-h* gene encoding the magnesium chelatase subunit I (CHLI), an enzyme involved in chlorophyll synthesis. Although the mutation was experimentally identified and characterised, the precise molecular rationale connecting the phenotype to the mutation remained unclear. To address this, we used AI-based methods to reconstruct the structure of the CHLI protein, providing a model to study the effects of the mutation. Comparative structural analysis and molecular docking simulations were performed to understand how the mutation alters the structure and function of the protein. Our results provide a molecular explanation for the observed phenotype, linking the structural effects of the mutation to the physiological characteristics of the barley mutant. By combining computational modelling with experimental characterisation, this work provides insights that could drive breeding strategies for crops with improved photosynthetic efficiency.

3.1 Accurate and Efficient SAXS/SANS Implementation Including Solvation Layer Effects Suitable for Molecular Simulations

The combination of Small-Angle X-ray and Neutron Scattering (SAXS/SANS or SAS) experiments with molecular dynamics (MD) simulations is an effective strategy for the characterisation of biomolecules in solution.[257] On the one hand, the limited resolution of SAS benefits from the atomistic detail provided by MD; on the other hand, the integration of experimental data helps to reduce the inaccuracies of MD. Although promising, this approach remains hampered by high computational costs. In particular, the multiple scattering intensity calculations performed on-the-fly alongside the MD simulation make this method prohibitively expensive, even on the latest High-Performance Computing systems. One way to overcome this limitation is to calculate the intensity of the system of interest on a coarse-grained model, thus aggregating the scattering behaviour of groups of atoms into larger particles.[258] Previously, we presented a hybrid resolution method that allows atomistic SAXS-restricted MD simulation by using a Martini coarse-grained approach to efficiently back-calculate scattering intensities;[259] in our last work, we enhance this technique by developing a novel hybrid-SAS method that is faster, more accurate, extended to the SANS intensity calculation and that is compatible with both proteins and nucleic acids. Furthermore, an implicit and user-definable solvation layer contribution is included in the calculation to allow the reconstruction of a more realistic scattering behaviour in solution. This layer depends on solvent-solute interactions and, being typically more electron/neutron dense than the bulk solvent, actively contributes to the scattering signal.[260] To ensure a fast and simple use of our method and to broaden its application, we have included it in PLUMED-ISDB, a module part of PLUMED,[261] an open-source software designed to enhance and extend various MD engines or to be used as a stand-alone package to perform a wide range of advanced analyses of complex biomolecular systems.

3.1.1 Personal Contribution

With the exception of the gelsolin experiments, I was actively involved in all aspects of this work. My contributions included:

1. **Design and Development of the Hybrid SAS Model:** The hySAS model development involved the evaluation of different forward models across multiple systems (proteins and nucleic acids) to achieve an optimal balance between computational performance and accuracy. The aim was to reduce computational bottlenecks while maintaining the fidelity of the *in silico* reconstruction of the SAS intensity of biomolecules. I explored different modelling strategies, ranging from multi-resolution system-specific approaches to a transferable fixed-resolution model, leading to the development of single bead per amino acid and three bead per nucleic acid mapping. In addition, I assessed other methods and tools that take into account explicit and implicit solvent effects, which are critical for effective calibration of our SAS model.

2. **Computational Work:** I managed the preparation of protein and nucleic acid systems, performing both standard and enhanced molecular dynamics (MD) simulations. To ensure fast and easy use of our method and to broaden its application, I contributed to the implementation of the forward model in the PLUMED software, which allows hySAS to be coupled with different MD engines and to be used independently of the force field adopted. I carried out all the analyses, including evaluating the predictive capabilities of the model, interpreting the experimental SAS data and evaluating the MD simulation trajectories.

3. **Manuscript Writing:** I have contributed to the writing of the manuscript in every part of it.

JCTC Journal of Chemical Theory and Computation

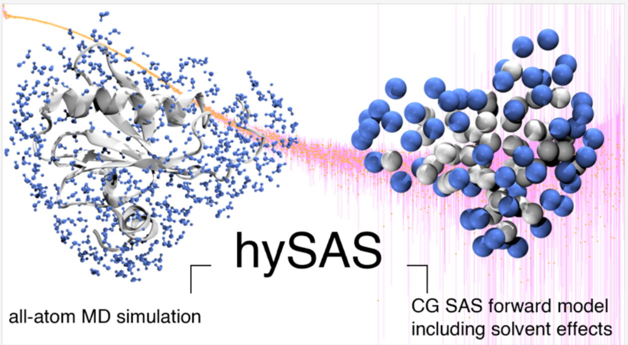
Open Access

Accurate and Efficient SAXS/SANS Implementation Including Solvation Layer Effects Suitable for Molecular Simulations

Federico Ballabio, Cristina Paissoni, Michela Bollati, Matteo de Rosa,* Riccardo Capelli,* and Carlo Camilloni*

Cite This: J. Chem. Theory Comput. 2023, 19, 8401−8413 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Small-angle X-ray and neutron scattering (SAXS/SANS) provide valuable insights into the structure and dynamics of biomolecules in solution, complementing a wide range of structural techniques, including molecular dynamics simulations. As contrast-based methods, they are sensitive not only to structural properties but also to solvent−solute interactions. Their use in molecular dynamics simulations requires a forward model that should be as fast and accurate as possible. In this work, we demonstrate the feasibility of calculating SAXS and SANS intensities using a coarse-grained representation consisting of one bead per amino acid and three beads per nucleic acid, with form factors that can be corrected on the fly to account for solvation effects at no additional computational cost. By coupling this forward model with molecular dynamics simulations restrained with SAS data, it is possible to determine conformational ensembles or refine the structure and dynamics of proteins and nucleic acids in agreement with the experimental results. To assess the robustness of this approach, we applied it to gelsolin, for which we acquired SAXS data on its closed state, and to a UP1-microRNA complex, for which we used previously collected measurements. Our hybrid-resolution small-angle scattering (hySAS) implementation, being distributed in PLUMED, can be used with atomistic and coarse-grained simulations using diverse restraining strategies.

1. INTRODUCTION

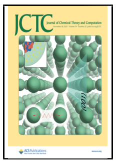
Small-angle scattering (SAS) techniques based on X-rays (SAXS) or neutrons (SANS) are established, valuable, and widely used tools in structural biology for the characterization of biomolecules in solution. These methods allow the size, shape, stoichiometry, and dynamics of biomolecules to be assessed under near-physiological conditions, using reasonable concentrations, and without the need of labeling agents.[1,2] Moreover, the size and the disorder level of the system are not a limitation, enabling the study of diverse biomolecular species.[3−5] Indeed, SAS techniques can efficiently complement nuclear magnetic resonance (NMR) spectroscopy and fluorescence resonance energy transfer (FRET) measurements to provide global features when studying multidomain proteins, intrinsically disordered proteins, and larger complexes.[6] Furthermore, SAS is particularly suitable for the analysis or the integration with molecular dynamics (MD) simulations, using either reweighing or restraining techniques.[7] This compatibility arises from the relative simplicity of calculating the forward model from the coordinates of an atomic resolution structure.

Briefly, the SAS intensity of a randomly oriented, N—atom, molecule in a vacuum can be calculated by the Debye equation

$$I(q) = \langle I(\mathbf{q})\rangle = \langle \sum_{i=1}^{N}\sum_{j=1}^{N} f_i(q) f_j(q) \exp(i\mathbf{q} \times \mathbf{r}_{ij})\rangle \cong \sum_{i=1}^{N}\sum_{j=1}^{N} f_i(q) f_j(q) \frac{\sin(qr_{ij})}{qr_{ij}} \quad (1)$$

The intensity is described as a function of the momentum transfer $q = |\mathbf{q}| = 4\pi\ (\sin\theta/\lambda)$, where 2θ is the scattering angle, λ the source wavelength, and $r_{ij} = |\mathbf{r}_{ij}| = |r_i - r_j|$ is the distance from the atom i to the atom j, which represents the relative position of atoms i and j in the sample. The notation $\langle\cdots\rangle$ refers to the spherical average, required to integrate the scattered intensity over all directions that have the same magnitude of q. In the case of SAXS, the radiation−matter interaction between the X-ray photon and the electron cloud of

Received: August 7, 2023
Revised: October 11, 2023
Accepted: October 24, 2023
Published: November 3, 2023



atom i is described by the atomic scattering factor $f_i(q)$, which can be approximated with the Cromer−Mann equation

$$f_i(q) = \sum_{k=1}^{4} a_k \exp[-b_k(q/4\pi)^2] + c = f_i^{\text{atomic}}(q) \tag{2}$$

The empirical and atom-type specific parameters a_k, b_k, and c are available in the International Tables for Crystallography.[8,9] To account for the solvent effects, each atomic scattering factor $f_i(q)$ is modified by subtracting a spherical Gaussian which depends on ρ_0, the electron density of the solvent (e.g., 0.334 e Å^{-3} for bulk water), and ν_i[10] the volume of the solvent displaced by the atom i, following the expression

$$\begin{aligned} f_i'(q) &= f_i(q) - \nu_i\rho_0 \exp\left(-\frac{q^2\nu_i^{2/3}}{4\pi}\right) \\ &= f_i^{\text{atomic}}(q) - \rho_0 f_i^{\text{solvent}}(q) \end{aligned} \tag{3}$$

Since the neutron wavelength is significantly larger than the nucleus dimension, in the case of SANS, the neutron scattering amplitude results to be isotropic, i.e., independent of the scattering angle. Therefore, eq 2 can be approximated to $f_i^{\text{atomic}}(q) = b_i$. The b_i constants, which are available in the literature,[11,12] depend on the number of neutrons and protons that constitute the nucleus. Consequently, isotopes of the same element with different numbers of neutrons, such as hydrogen and deuterium, have different neutron scattering lengths. This feature provides the basis of the contrast variation technique[13] a powerful advantage of neutron scattering over X-rays, which is usually achieved by using mixtures of hydrogenated and deuterated water in varying proportions. To account for this combination in solvent composition, eq 3 is modified to

$$f_i'(q) = f_i^{atomic}(q) - \eta\rho_0 f_i^{solvent}(q) \tag{4}$$

where

$$\eta = 0.1(b_{\text{O}} + 2(b_{\text{H}}(1-d) + db_{\text{D}})) \tag{5}$$

with b_{O}, b_{H}, and b_{D} as the neutron scattering amplitudes of oxygen, hydrogen, and deuterium, respectively, and d as the deuterium concentration (from 0 to 1, which corresponds to a percentage range of 0−100%). The coefficient 0.1 serves as a scaling factor to account for the 10 electrons per water molecule when converting from electron to molecule density. It should be noted that this modification does not consider the possible effects of hydrogen−deuterium exchange between the solvent and the solvent-exposed residues of the biomolecule.

Although eq 1 accounts for the solvent displaced by the solute in the calculation of the scattering signal, it does not consider the contribution of the solvation shell. This layer depends on solvent−solute interactions and is typically more electron-dense than that of the bulk solvent. For example, the hydration layer has been reported to be up to 20−25% more electron-dense than bulk water.[14−16] This phenomenon can result in an apparent increase in the radius of gyration of the solute.[17] The contribution of the solvation layer can be included in calculations through explicit solvent modeling, as implemented in software such as WAXSiS[18,19] and Capriqorn[20] or implicitly like in CRYSOL[21]/CRYSON,[14] FoXS,[22] and Pepsi-SAXS.[23] The explicit solvent methods consider the positions of the solvent atoms in the surrounding shell while calculating the scattering signal of the molecule in solution. However, this approach is computationally expensive due to the large number of solvent atoms that must be considered in addition to those of the molecule. Furthermore, it may still be inaccurate because of the limitations of the force field (FF) in the description of water−water and water-solute interactions.[24,25] Implicit solvent modeling methods, on the other hand, allow the calculation of the solvation layer contribution to the scattering signal without the need to model the solvent atoms explicitly. This reduces the computational cost but results in an approximate representation of the shell.[26] In general, implicit solvent methods require the introduction of a solvation layer term in eq 3

$$f_i'(q) = f_i^{\text{atomic}}(q) - \rho_0 f_i^{\text{solvent}}(q) + f_i^{\text{solvation layer}}(q) \tag{6}$$

The way the $f_i^{\text{solvation layer}}(q)$ is calculated is slightly different depending on the software used. For example, in CRYSOL 2.x,[27] this term depends on the contrast between the border layer, which is considered as an envelope of fixed width surrounding the particle, and the bulk solvent.

Equation 1 requires the evaluation of all of the pairwise interatomic distances within the molecule of interest, thus resulting in N^2 calculations, where N is the number of the atoms involved, making it highly demanding for large biomolecules. This problem is exacerbated when multiple evaluations of the scattering profile are required, as in the case of MD simulations restrained by SAS data, resulting in severe performance degradation. A successful strategy to mitigate this computational burden is to coarse-grain the representation of the molecule.[28−32] This simplification can be achieved by combining the scattering behavior of groups of atoms into larger beads while preserving the overall scattering properties of the molecule. This is made possible by the intrinsically low-resolution nature of SAS data, which are more sensitive to the overall shape and size of the molecule rather than its atomic level details. Depending on the specific coarse-graining method used, both the criteria for assigning atoms to beads and the placement of their center can vary significantly. In general, for a coarse-grained system, eq 1 becomes

$$I(q) \cong \sum_{i=1}^{M}\sum_{j=1}^{M} F_i(q)F_j(q)\frac{\sin(qR_{ij})}{qR_{ij}} \tag{7}$$

where M is the number of beads, R_{ij} is the relative distance between the center of the bead i and the center of the bead j, and $F(q)$ is the bead form factor, which mirrors the scattering intensities of individual beads. There are several approaches to calculating $F(q)$; among them, the single-bead approximation (SBA) proposed by Yang et al.[28] has proved to be one of the most computationally streamlined methods, which is fast but dependable. According to SBA

$$F_j(q) = \left[\sum_{i\in j}\sum_{k\in j} f_i'(q)f_k'(q)\frac{\sin(qr_{ik})}{qr_{ik}}\right]^{1/2} \tag{8}$$

where $f_i'(q)$ and $f_k'(q)$ are the atomic scattering factors of the atoms i and k belonging to bead j, which are modified to include only the solvent-excluded volume term as in eqs 3 and 4, for SAXS and SANS, respectively. Niebling et al.[32] have effectively applied the SBA with the Martini 2.2[33] coarse-grained scheme to derive SAXS bead form factors for proteins, and we have further extended it to nucleic acids.[34] With this forward model, we used SAXS data to restrain simulations based on the Martini force field[35] but also based on atomistic force fields, for both proteins and nucleic acids. In this latter hybrid scheme, called the hySAXS approach, the simulation is performed at atomic resolution, while the SAXS intensity of the respective frames is calculated on a coarse-grained model.[36−39]

Even when the Martini representation is used to determine the scattering intensity with hySAXS, the study of relatively large systems can be challenging. Furthermore, this approach was not designed to account for possible corrections due to the solvation layer effects. Here, we present a novel hySAS method for proteins and nucleic acids that is faster and more accurate with the inclusion of a solvent layer contribution and extended to the calculation of SANS intensity. As a case study, we applied the hySAS approach to determine the conformational ensemble of human gelsolin (GSN). This 83 kDa protein (reviewed by Nag et al.[40]) is composed of six homologous domains (named G1 to G6) connected by flexible linker regions and is considered a master regulator of actin dynamics, thanks to its severing and capping activities.[41] GSN and the other members of the superfamily play an important role in several physiological processes, such as cell division and mobility, trafficking, signal transduction, immunomodulation, and inflammation.[42,43] GSN is also responsible for a hereditary amyloidosis,[44] and it is involved in several other diseases, particularly cancer (reviewed by Li et al.[45]). Each GSN domain harbors a Ca^{2+} binding site, and binding to the ion triggers local changes and domain rearrangements that shift the protein from a closed to an open conformation.[46] In the absence of Ca^{2+}, the actin binding sites are buried, limiting the ability of the GSN to interact with actin filaments. In this inactive state, GSN can be crystallized,[47] but the resolution is relatively low and several stretches of the protein are too flexible to be modeled; such flexibility has been shown to be relevant for GSN physiopathology.[48−50] In this work, we have determined the ensemble of gelsolin structures in the closed and inactive state using SAXS data measured in the absence of calcium. Furthermore, as a second example of the applicability of hySAS, we refined a previously published protein−RNA complex. This newly introduced hySAS and our previous implementations are already available in the ISDB[51] module of PLUMED[52,53] software, an open-source software designed to enhance and extend various MD engines or to be used as a stand-alone package to perform a wide range of advanced analysis of complex biomolecular systems.

2. THEORY AND METHODS

2.1. SAS Form Factors with the Solvation Layer Contribution. Here, we introduce a novel hySAS method for proteins and nucleic acids where we use a single-bead (1B) representation to describe the scattering behavior of an amino acid and a three-bead (3B) mapping for a nucleotide, one for the phosphate group, one for the pentose sugar, and one for the nitrogenous base. This choice allow us to achieve better performance and to alleviate a source of inaccuracy in the Martini representation, specifically the need to extrapolate the bead form factor when it assumes negative values.[32] Importantly, to include the solvent layer contribution for small q values, we reformulate the SBA $F(q)$ as the sum of three terms

$$
\begin{aligned}
F_i(q) = \Bigg[& \sum_{k \in i} \sum_{l \in i} f_k^{\text{atomic}}(q) f_l^{\text{atomic}}(q) \frac{\sin(q r_{kl})}{q r_{kl}} \\
& + \rho^2 \sum_{k \in i} \sum_{l \in i} f_k^{\text{solvent}}(q) f_l^{\text{solvent}}(q) \frac{\sin(q r_{kl})}{q r_{kl}} \\
& - \rho \sum_{k \in i} \sum_{l \in i} [f_k^{\text{atomic}}(q) f_l^{\text{solvent}}(q) + f_k^{\text{solvent}}(q) \\
& f_l^{\text{atomic}}(q)] \frac{\sin(q r_{kl})}{q r_{kl}} \Bigg]^{1/2} \\
= {} & [F_i^{\text{atomic}} + \rho^2 F_i^{\text{solvent}} - \rho F_i^{\text{mixed}}]^{1/2} \qquad (9)
\end{aligned}
$$

This approximation allows the F_i^{atomic}, F_i^{solvent}, and F_i^{mixed} terms to be precalculated separately and for each bead type, regardless of the solvent-specific ρ_0 (and the deuteration fraction in the case of SANS). Therefore, in addition to the option of using a buffer other than bulk water, it is possible to assign modified solvation densities to different beads as a proxy for the effect of the solvation shell. More precisely, the ρ_0 value of the beads exposed to the solvent can be adjusted to implicitly include the solvation layer contribution (SLC) through a user-defined parameter. This correction can be described as $\rho = (\rho_0 - \text{SLC parameter})$. For this purpose, the solvent-accessible surface area (SASA) for each amino acid, nucleotide sugar, phosphate group or base is calculated on the fly during the MD simulation or only for a single frame using the efficient LCPO method.[54] Of note, in the case of SANS, the hydrogen−deuterium exchange is also considered. To achieve this result, we have precalculated the three terms of eq 9 for each bead type using both deuterium $f^{\text{atomic}}(q)$ and hydrogen $f^{\text{atomic}}(q)$. For the beads exposed to the solvent, each time eq 9 is solved, the terms obtained with deuterium $f^{\text{atomic}}(q)$ are used instead of those obtained with hydrogen $f^{\text{atomic}}(q)$, with a probability equal to the deuterium concentration in the buffer. For the same bead type, $f^{\text{solvent}}(q)$ is identical for SAXS and SANS, both for hydrogenated and deuterated beads, as it depends exclusively on the parameter ν. The ρ_0 value, the SLC parameter, the SASA threshold to consider a residue solvated, and the deuterium fraction in SANS can be defined by the user.

2.2. Bead Form Factor Parametrization and Validation. We computed the F_i^{atomic}, F_i^{solvent}, and F_i^{mixed} terms of eq 9 for all of the amino acids, as well as for nucleic acid bases, the pentose sugars, and the phosphate group, for both SAXS and SANS. Concerning proteins, the three terms per amino acid were calculated and averaged over 1000 frames extracted at equidistant intervals from a 2.7 μs MD trajectory of GSN. The heterogeneous structural composition of this 755-residue protein makes it an ideal model for a comprehensive conformational sampling. Indeed, in addition to encompassing all of the standard amino acids, GSN features an IDP-like N-terminal region of ∼25 residues and six structured domains rich in α-helices and β-sheets, connected by flexible linkers. To validate the transferability of the GSN terms to other systems, we generated two additional independent term sets, derived from a 270 ns MD simulation of the B1 immunoglobulin-binding domain of streptococcal protein G (B1), and from a 1.35 μs trajectory of the green fluorescent protein (GFP). The three components of the bead form factors of all of the standard amino acids have been calculated. Furthermore, we have also included the scattering behavior of the histidine with both the δ- and ε-nitrogen of the imidazole ring protonated. A different strategy was employed for nucleotides. Considering the lower accuracy of the FF for nucleic acids,[55] we preferred to calculate and average the terms from nonredundant molecular structures obtained from the Protein Data Bank (PDB). We used a set of 167 noncomplexed DNA structures[56] and a set of 75 RNA structures[57] that we had already prepared and used in our previous work.[34] To validate the parameters, 120 DNA and 43 RNA structures with no missing heavy atoms were selected from these repositories as the training subset, while the remaining structures were used as the validation subset. The final terms were computed on the two complete repositories (Table S1). We derived the form factor components of the five nucleobases (adenine, cytosine, thymine, uracil, and guanine), the phosphate group, and the DNA and RNA pentose sugars. In addition, we included two other DNA/RNA bead types for the 5′-end and the 3′-end pentose sugar with a hydroxyl moiety at carbon C5′ and C3′, respectively. Finally, each term belonging to either a protein or nucleic acid was fitted to a sixth-order polynomial. This means that F_i^{atomic}, F_i^{solvent}, and F_i^{mixed} are described by a total of 21 parameters.

2.3. Computational Details. Protein bead form factor parametrization was performed on mature human GSN. The initial model was determined from the PDB entry 3FFN, whose missing loops and N-terminus were reconstructed using AlphaFold2.[58] Regarding the 56 residues B1, and the 230 residues GFP, the structures are based on PDB entries 1PGB and 1GFL, respectively. All of the structures were prepared with the following procedure. The histidine orientation and protonation states were optimized using Schrödinger Maestro Suite, release 2021-4.[59] The topology was built using DES-Amber[59] FF and the system was solvated with the TIP4P-D[60] water model in a dodecahedron box with a NaCl concentration of 100 mM. After two preliminary minimization steps (steepest descent and conjugate gradient algorithms), a 2 ns long NPT simulation was performed with the protein atoms restrained to their minimized positions. For GSN, 675 ns of classical MD simulation was computed for each of the 4 replicas, collecting a total of 2.7 μs. For GFP, we ran a single replica MD simulation of 1.35 μs, while for B1, we ran 4 replicas for 67.5 ns each, reaching 270 ns.

The plain MD simulations of GSN, B1, and GFP were also used to evaluate the performance and accuracy of calculating SAS intensities at different resolutions. For nucleic acids, we followed the previous procedure to prepare and perform a 35 ns simulation of the 1,187 nucleotides large subunit ribosome fragment (PDB ID 1Z58) and a 14 ns simulation of single-stranded 12-mer RNA (AGUAGAUUAGCA). The former was used to assess the timing of the SAS intensity calculation and the latter to assess the accuracy.

For the GSN refinement, driven by SAXS-restrained MD simulation, the previous structure was modified. Since the experimental SAXS measurements were collected on a full-length GSN fused to a N-terminal His$_6$-tag, we modeled an additional 23 residues, corresponding to the sequence "MGSSHHHHHHSSGLVPRGSHMAS", resulting in a 778-residue protein that was prepared as described previously. We ran 2x 1 μs metainference[61] multireplica simulations (10 replicas, 100 ns each), one with and one without the solvation layer correction enabled. The representative SAXS intensities selected as restraints range between the q values of 0.01 Å^{-1} and 0.25 Å^{-1} with a stride of 0.015 Å^{-1}. The analysis was performed over the last 50 ns of each trajectory.

Regarding the protein−RNA complex refinement, we adopted the MD input files prepared in our previous work.[34] In summary, AMBER14SB[62] FF with parmbsc1[63] parameters and the TIP3P[64] water model were used to build the topology. To preserve the protein−RNA interface, we introduce harmonic biases on the distances between the phenylalanine residues and bases involved in nonbonded interactions; furthermore, we also added a restraining potential on the secondary structures of the protein, following the same procedure described by Kooshapur et al.[65] The metainference simulations were performed for 4.5 ns with and without the solvation layer correction activated, using 35 selected SAXS intensities with q values between 0.008 Å^{-1} and 0.3 Å^{-1} as restraints.

All of the simulations were performed using GROMACS 2021.6,[66] PLUMED2,[52,53] and the PLUMED-ISDB[51] module. Plots were generated using the matplotlib[67] 3.6.0 package, while the open-source software VMD[68] and PyMOL[69] were used for structural visualization of biomolecules. Relevant input files and trajectories are available on Zenodo[70] and the PLUMED-NEST as plumID:23.029.

2.4. Gelsolin Expression, Purification, and SAXS Data Collection. Recombinant full-length GSN protein, carrying an N-terminal His$_6$-tag, was produced as previously described.[50,71] Briefly, the human plasma isoform of GSN devoid of the signal peptide (mature form) was produced in *Escherichia coli* SHuffle cells (New England Biolabs) upon addition of 0.5 mM IPTG and incubation for 16 h at 18 °C. Cells were lysed in a Basic Z Bench top (Constant Systems Limited, U.K.) at 25 kPSI, and the clarified extract passed through a HisTrap HP column (all chromatographic media from GE-Healthcare). Further polishing was obtained by anion exchange (Resource Q), followed by size-exclusion chromatography (HiLoad 16/600 Superdex 200). For SAXS analysis, the protein was diluted to 2.01 mg/mL in 20 mM HEPES, pH 7.4, 100 mM NaCl, and 1 mM EDTA. GSN batch data were collected at the B21 BioSAXS beamline of the Diamond Synchrotron (Didcot, Oxfordshire, UK).[72] Data and model are deposited in the SASBDB[73] as SASDSN7.

3. RESULTS AND DISCUSSION

3.1. Single-Bead Mapping for Amino Acids and Three-Bead Mapping for Nucleotides Are Fast and Accurate for Small q Values. To assess the impact of the number of elements in a system on the speed of the SAS intensity calculation, we compared the time required by PLUMED to determine intensities from MD trajectory frames. We used different resolutions, including all-atom (AA), Martini scheme with transferable parameters (MT), and single-bead per amino acid (1B)/three beads per nucleotide (3B) mappings with the corresponding transferable parameters. For the analysis, we selected a GSN trajectory consisting of 6442 frames to evaluate the performance on proteins and a 500 frames trajectory of a large subunit ribosome fragment to evaluate nucleic acids. The intensities were calculated for 31 q values, in the range 1 × 10^{-10} to 0.3 Å^{-1}, every 0.01 Å^{-1}. As expected, the resolution had a dramatic effect on the calculation time. For proteins, it took approximately 5 days to determine the SAS intensity at AA details (11,558 atoms). The same calculation was achieved in about 143 min (48.4-fold speedup) using the MT mapping (1627 beads) and

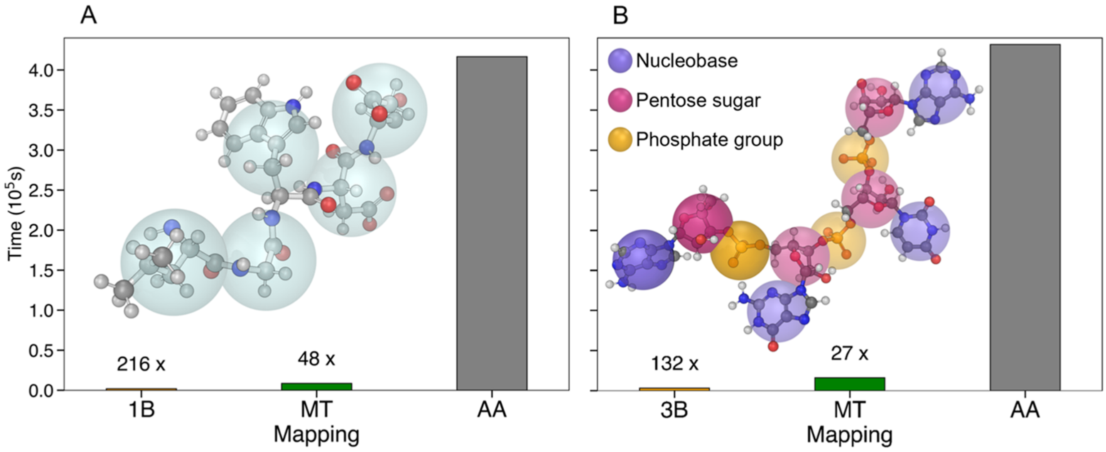


Figure 1. SAS intensity calculation timings. (A) The 6442 frame MD trajectory of 755 residues GSN was used as input for PLUMED-ISDB plugin to calculate the corresponding SAS intensities for 31 q values, using different mapping resolutions. The time required for completion at AA details (11,558 atoms) is 416,594 s and with MT mapping (1627 beads) is 8,615 s, while for 1B (755 beads) is 1925 s. 1B and MT are 216 times and 48 times faster than AA, respectively. As an example, five residues are shown at atomistic (ball and sticks visualization) and 1B resolution (light blue beads). (B) The 500 frames MD trajectory of 1187 nucleotide RNA strand was used to calculate the corresponding SAS intensities for 31 q values. The time required for completion at AA resolution (38,287 atoms) is 432,022 s and for MT mapping (7796 beads) is 15,912 s, while for 3B (3,560 beads) is 3263 s. 3B and MT are 132 times and 27 times faster than AA, respectively. As an example, four nucleotides are shown at atomistic (ball and sticks visualization) and 3B resolution (nucleobase in blue, pentose sugar in violet, phosphate group in orange). The timings were evaluated under the same conditions on a single core of a workstation equipped with an Intel Xeon E5-2660v3 CPU.

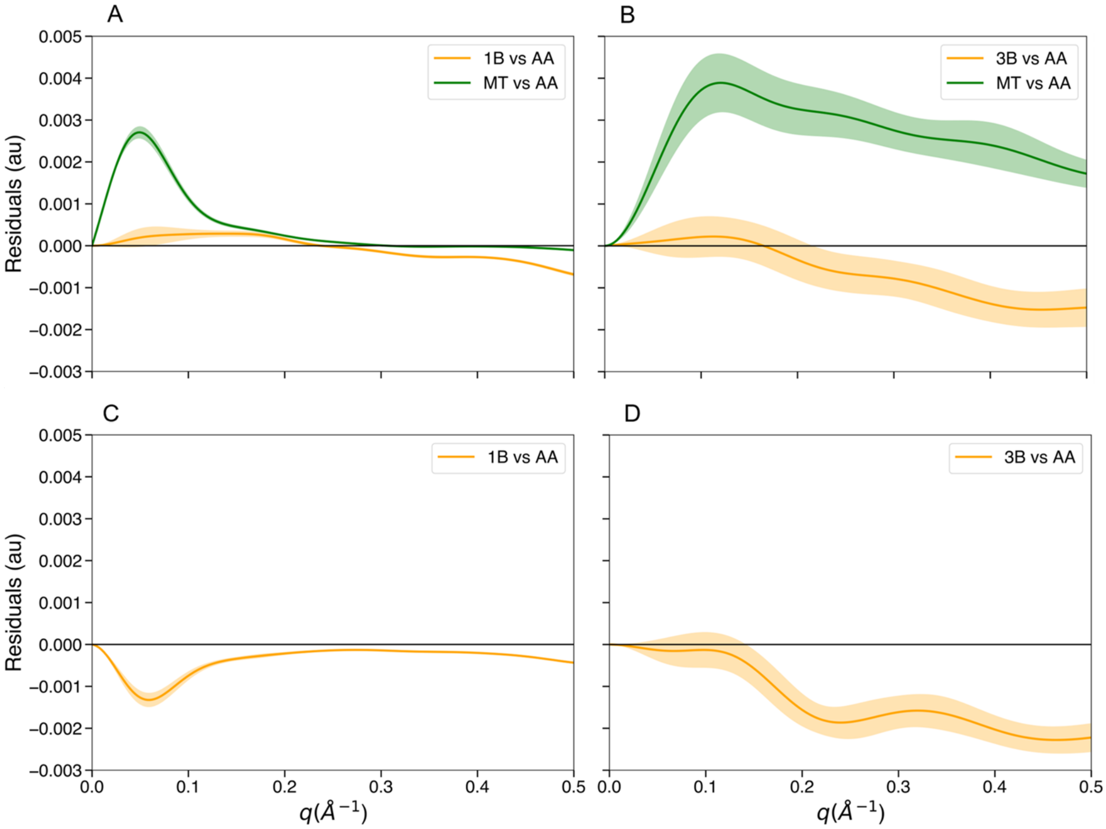


Figure 2. Validation of the 1B/3B mappings in the calculation of scattering intensities. The SAS profile of each frame from MD trajectories was calculated at atomic and coarse-grained resolution, for 201 q values ranging from 1×10^{-10} to 0.5 Å^{-1}. (A) Average and standard deviation on 6442 GSN frames of the SAXS residuals between MT and AA (green) and between 1B and AA (orange). (B) Average and standard deviation on 7256 12-mer RNA frames of the SAXS residuals between MT and AA (green) and between 3B and AA (orange). (C) Average and standard deviation on 6442 GSN frames of the SANS residuals between 1B and AA (orange). (D) Average and standard deviation on 7256 12-mer RNA frames of the SANS residuals between 3B and AA (orange).

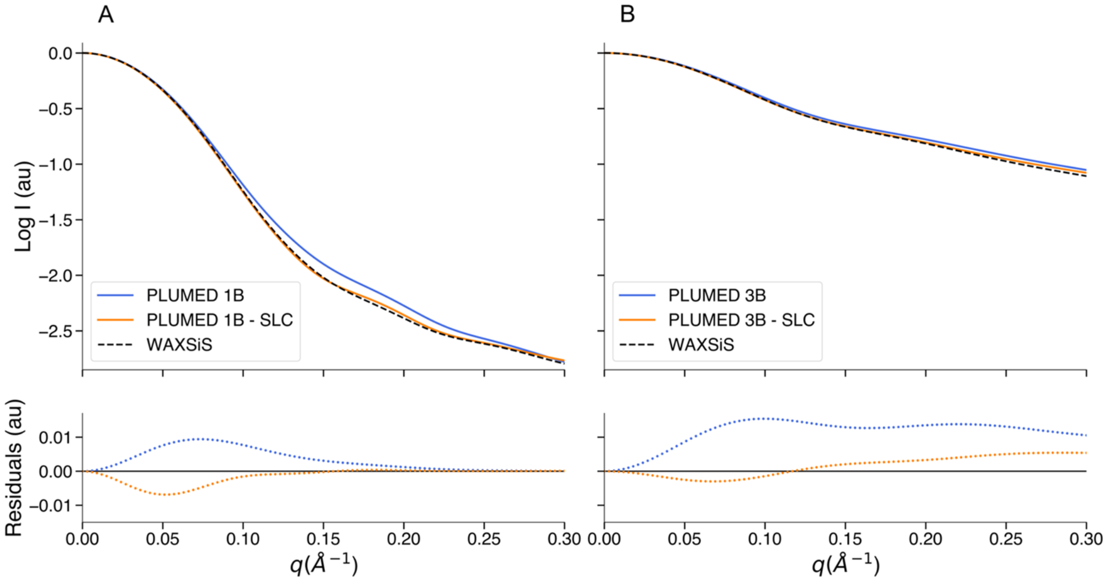


Figure 3. Solvation layer contribution in the 1B/3B SAXS intensity calculation. (A) Upper panel: logarithm of the SAXS profile of a representative, randomly selected, GSN frame calculated using 1B mapping (blue), 1B mapping with the best combination of SLC (0.08) and SC (0.6 nm^2) found for this frame (orange), and using WAXSiS (black). Bottom panel: residuals of 1B (blue) and 1B with SLC (orange) using the WAXSiS intensity as the reference. (B) Upper panel: logarithm of the SAXS profile of a representative, randomly selected, 12-mer RNA frame calculated using 3B mapping (blue), 3B mapping with the best combination of SLC (0.120) and SC (1.0 nm^2) found for this frame (orange), and using WAXSiS (black). Bottom panel: residuals of 3B (blue) and 3B with SLC (orange) using the WAXSiS intensity as the reference. All of the SAXS intensities were calculated for 101 q values, up to 0.3 Å^{-1}.

approximately 32 min (216.4-fold speedup) using the 1B representation (755 beads) (Figure 1A). Similarly, for nucleic acids, the calculation time reduced from around 5 days at AA resolution (38,287 atoms) to about 265 min (27.2-fold speedup) using MT mapping (7796 beads) and about 54 min (132.4-fold speedup) using the 3B representation (3560 beads) (Figure 1B).

In addition to performance evaluation, we also assessed the accuracy of 1*B*/3B mappings and parameters in the calculation of scattering intensities. For this benchmark, we selected 6442 equidistant frames from the GSN MD simulation, 6502 from B1, 9622 from GFP, and 7256 from the 12-mer RNA strand and calculated the SAXS and SANS intensity in both coarse-grained and atomistic details for 201 q values, in the range of 1 × 10^{-10} to 0.5 Å^{-1}, every 0.0025 Å^{-1}. For each frame and each q value, the intensity calculated with 1*B*/3B was compared with the corresponding intensity at atomistic resolution, which was taken as the reference. For SAXS we also included the comparison between MT and AA. The GSN SAXS intensities calculated with 1B mapping showed better agreement with those obtained with AA resolution than with MT up to 0.3 Å^{-1}, since the difference (residuals) between 1B and AA intensities is smaller than the difference between MT and AA for the same set of q values (Figure 2A). A similar behavior has been observed also for B1 (Figure S1A, left panel) and GFP (Figure S1B, left panel) SAXS intensities. This phenomenon, which is probably amplified by the approximation introduced in the calculation of the MT bead form factors, shows that for small q values, the atomic details are not critical in the determination of the intensity. Considering B1, which is the worst case scenario we observed, the SAXS intensity computed with 1B differs by less than 0.5% from that calculated at atomistic resolution in the range 0−0.3 Å^{-1}. Regarding the SANS intensity calculation with 1B mapping, the results obtained for GSN (Figure 2C) and GFP (Figure S1B, right panel) were comparable to those of SAXS, whereas for B1, the accuracy decreased, with a maximum difference between 1B and AA scattering profiles of about 1.5% (Figure S1A, right panel). As for the proteins, the calculation of the SAXS intensity on RNA with 3B mapping also proves to be accurate, with better agreement with AA resolution than with MT (Figure 2B). Finally, the difference between the RNA SANS intensity computed with 3B and that computed with AA shows a level of accuracy close to that observed for SAXS (Figure 2D). These results were obtained without considering the solvation layer contribution. To assess the transferability and validate the 1B parameters obtained from GSN, we generated additional independent sets of parameters from B1 and GFP. Using the 1B parameters obtained from B1, we calculated the SAXS intensities on the B1 trajectory frames and compared them with the corresponding AA intensities. The same B1 frames were employed to calculate the 1B intensity using the parameters obtained from GSN, and these intensities were also compared with the AA profiles. The two obtained residuals are nearly superimposable (Figure S2A), differing from each other by less than 0.1% at most. We followed the same procedure with GFP, and similarly, the residuals calculated with the 1B parameters from GFP are in strong agreement with the residuals calculated with the 1B parameters from GSN (Figure S2B). To validate the 3B parameters, the nucleic acid repositories previously described in Section 2 were divided into two subsets. We selected 120 DNA and 43 RNA PDB files as the training set to compute the 3B form factor parameters since all of the heavy atoms are solved in these structures. We calculated the SAXS intensity of each structure belonging to this set with 3B mapping and at AA resolution and evaluated the respective residuals. We performed the same analysis using the 3B parameters obtained from the training set on the remaining 47 DNA and 32 RNA structures, which we considered as the validation subset. The average of the

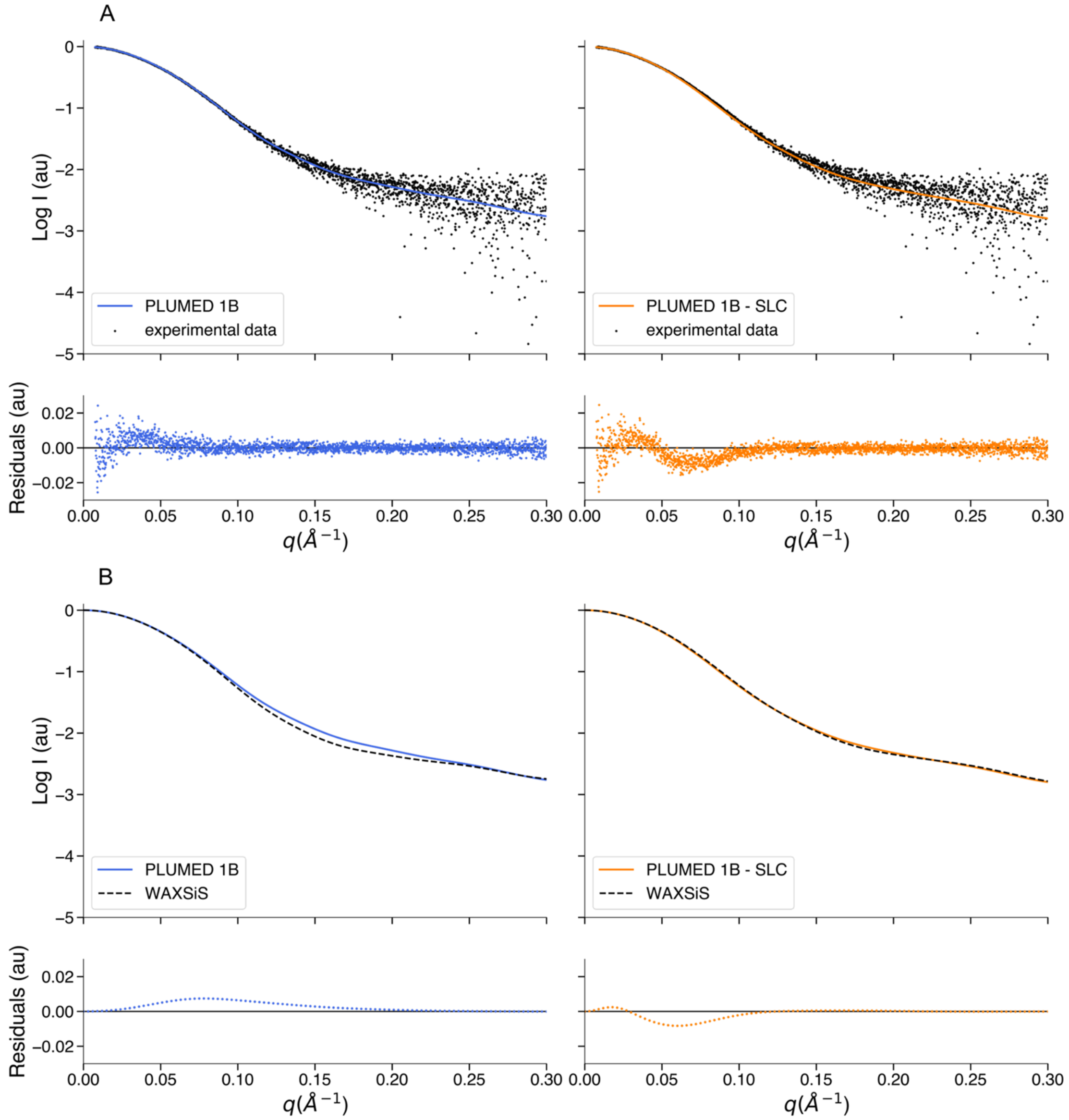


Figure 4. Agreement between hySAS, experimental SAXS data, and WAXSiS for the gelsolin ensembles. (A) Left panel: comparison between the logarithm of the average GSN SAXS profile calculated using 1B mapping without SLC (blue) and the logarithm of the experimental SAXS data (black dots). Right panel: comparison between the logarithm of the average GSN SAXS profile calculated using 1B mapping with SLC (orange) and the logarithm of the experimental SAXS data (black dots). (B) Left panel: comparison between the logarithm of the average GSN SAXS profile calculated using 1B mapping without SLC (blue) and the logarithm of the average WAXSiS profile (black dashed line). Right panel: comparison between the logarithm of the average GSN SAXS profile calculated using 1B mapping with SLC (orange) and the logarithm of the average WAXSiS profile (black dashed line). All of the residuals are calculated as the difference between the two intensities considered.

residuals from the training set and the average of the residuals from the validation set differed by a maximum of 0.12%. Furthermore, although the residuals from the validation set are more dispersed, the average is closer to the reference than the average of the residuals from the training set (Figure S3).

3.2. Inclusion of the Solvation Layer Contribution Allows Matching the SAXS Intensity Calculated by WAXSiS. The 1B and 3B form factors can be modified to include the solvation layer contribution in the SAS intensity calculation. Whether for a single PDB file or an MD trajectory, this process requires the calculation of the SASA to assess which beads are exposed to the solvent. This procedure is performed by the LCPO[54] algorithm implemented[74] in PLUMED. The reliability of the method was verified by comparing the results obtained with LCPO with the results obtained for the same frames with the sasa module[75] integrated in GROMACS (Figure S4). To evaluate the SLC, we used as a reference the intensities calculated by WAXSiS (Wide Angle X-ray Scattering in Solvent), a web server hosted at Saarland University, which allows the calculation of SAXS/WAXS profiles based on short MD simulations in an explicit solvent.[18,19] We extracted 10 equidistant frames from each of the previously described trajectories of GSN, B1, GFP, and 12-mer RNA. For all of these frames, we calculated the SAXS intensity using 1B/3B mapping with the SLC parameter set to 0.04, 0.06, 0.07, 0.08, 0.09, 0.095, 0.10, 0.11, and 0.12 and with the SASA cutoff (SC) of 0.4, 0.6, 0.7, 0.8, 1.0, and 1.2 nm^2, in all of the possible combinations. The same frames were used as

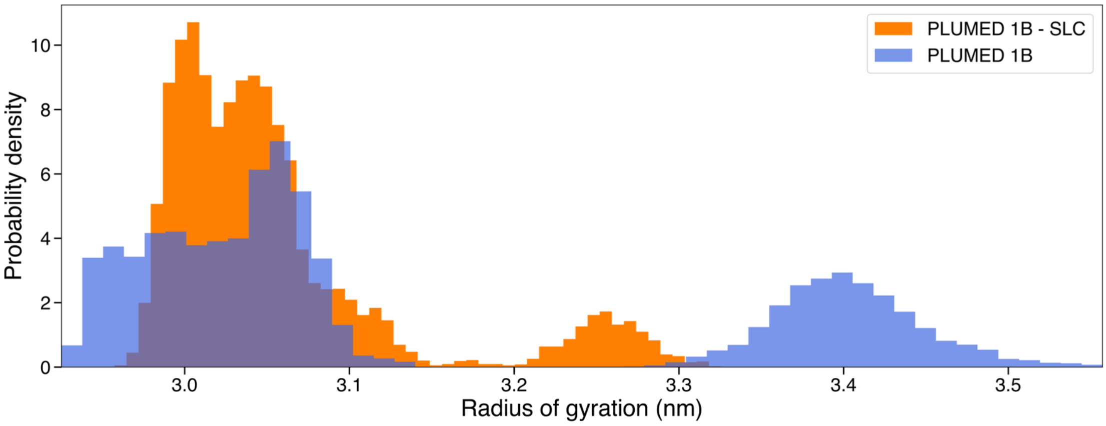


Figure 5. Radius of gyration and probability density histograms of GSN ensembles. The probability density distribution of the radius of gyration was calculated over 10,000 frames obtained using hySAS with SLC, colored in orange, while the distribution calculated over 10,000 frames obtained using hySAS without SLC is colored in blue. The area under each histogram integrates to 1.

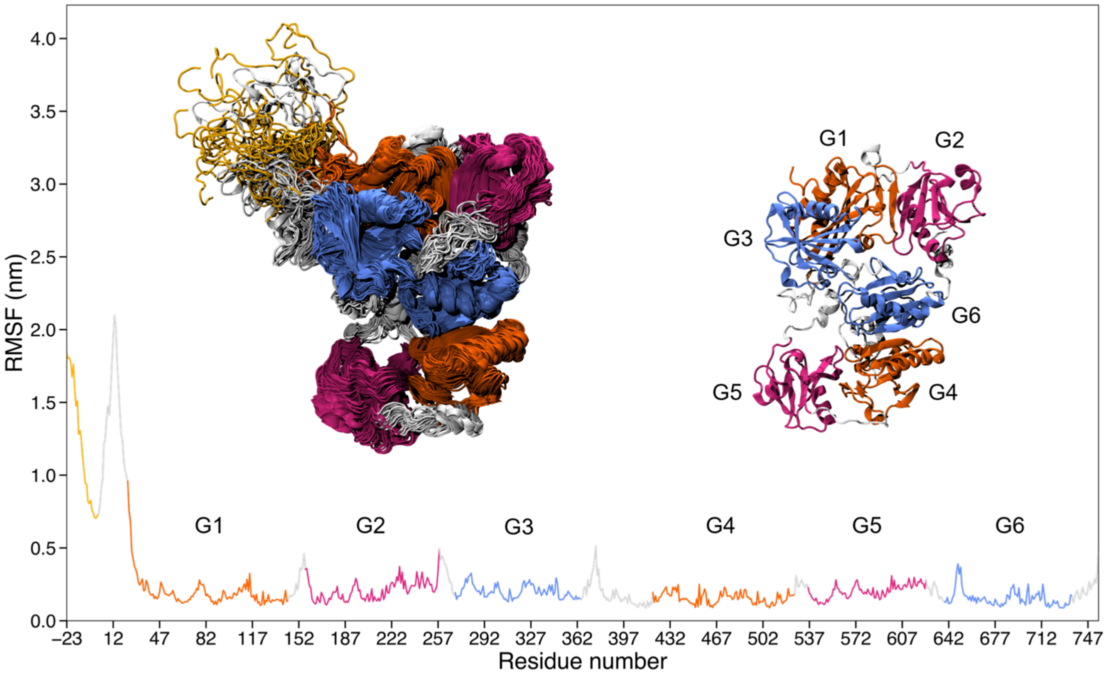


Figure 6. RMSF analysis of the GSN ensemble (with SLC). The flexibility of the protein was assessed by calculating the root-mean-square fluctuation of all residues. The residue numbering sequence on the *x*-axis includes the N-terminal His$_6$-tag (from −23 to −1) and the full-length human plasma isoform of GSN (1 to 755). The domains sharing the highest sequence and structural similarity are shown with the same color code: G1 and G4 are colored in orange, G2 and G5 in purple, and G3 and G6 in blue. The linkers and tails are colored in light gray, while the His$_6$-tag is colored in yellow. On the left, 50 equidistant frames from the analyzed trajectories are superimposed as a representative example of the conformational ensemble. The GSN structure on the right is that obtained by X-ray crystallography (PDB ID: 3FFN).

input to calculate the SAXS profiles with WAXSiS. We specified in the web server options an explicit solvent envelope of 7 Å from the surface of the biomolecule, and we selected the maximum available simulation length ($2 \times 10^6/N^{-0.77}$ frames, where N is the approximate number of atoms in the hydration layer). As in the previous analyses, we calculated the residuals between the intensity computed with 1B/3B and the intensity computed with WAXSiS, that we consider as the reference. Although some combinations of SLC and SC gave surprising results, leading to SAXS profiles practically identical to those calculated by WAXSiS (Figures 3 and S5), we found that using any of the indicated values of SLC and SC gave a better agreement with WAXSiS intensity than using 1B/3B without SLC. For a clearer overview, we computed the root-mean-square error (RMSE) between the logarithm (base 10) of the SAXS intensity calculated with AA, MT, and 1B/3B (with all of the SLC/SC combinations) and the logarithm of the SAXS intensity calculated with WAXSiS. The results obtained from all of the extracted frames were averaged for each system (Table S2). This analysis showed that the 1B/3B with SLC gave better results than 1B/3B without SLC but also compared to MT and AA resolution. For GSN, B1, and GFP, the SLC values that lead to the best results are generally between 0.08 and 0.1 with the SC of 0.7−0.8 nm^2. Instead, for the 12-mer RNA, an SLC greater than 0.1 with SC between 0.8 and 1 nm^2 is more in agreement with the reference.

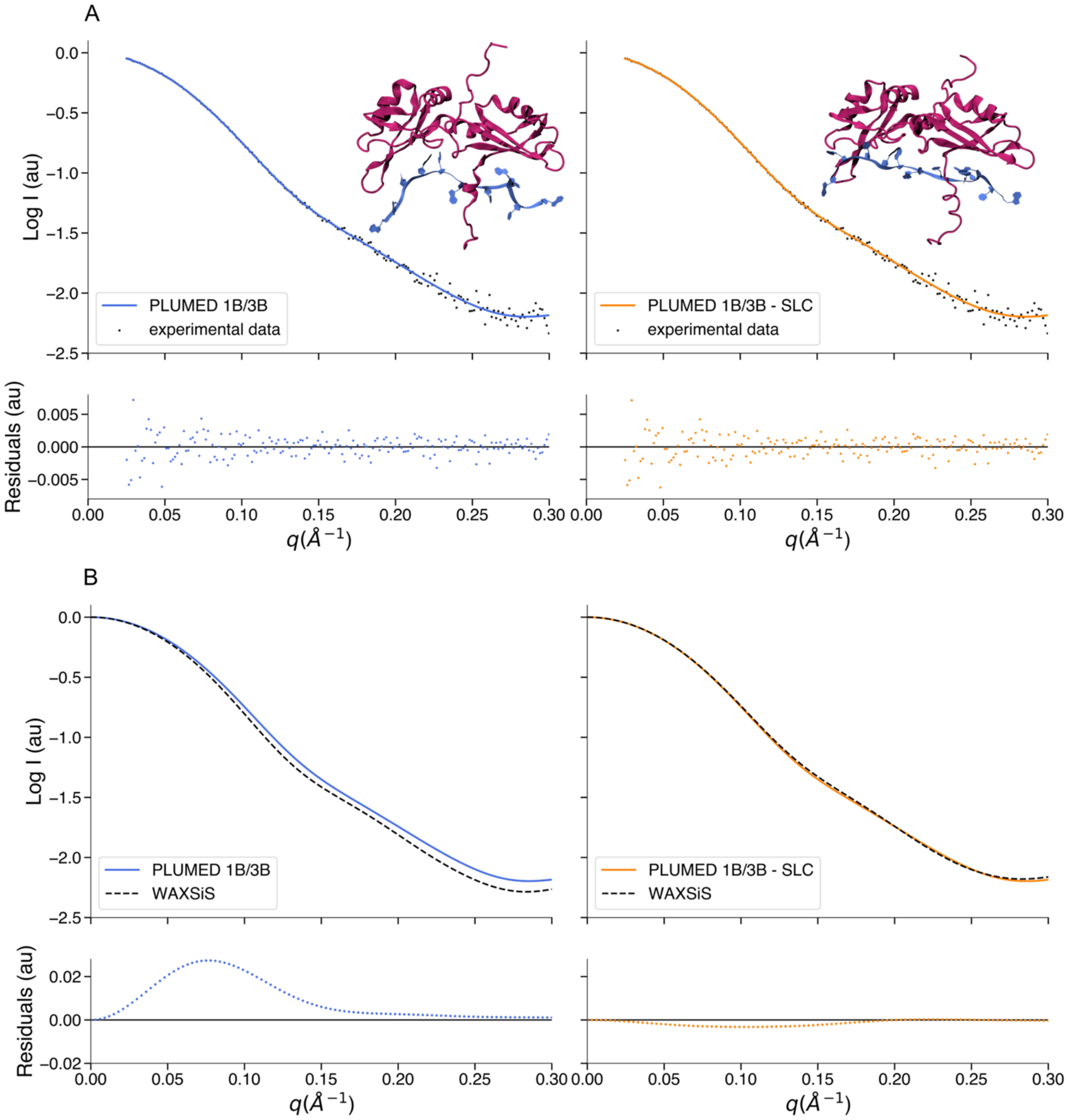


Figure 7. Comparison between hySAS, experimental SAXS data, and WAXSiS for UP1-RNA complex models. (A) Left panel: comparison between the logarithm of the protein−RNA complex SAXS intensity calculated using 1B/3B mapping without SLC (blue) and the logarithm of the experimental SAXS data (black dots). Right panel: comparison between the logarithm of the protein−RNA complex SAXS intensity calculated using 1B/3B mapping with the SLC (orange) and the logarithm of the experimental SAXS data (black dots). The upper right section of each panel shows the protein−RNA complex frame responsible for the relative intensity profile (purple/cartoon representation for UP1, blue/ribbon representation for 12-mer RNA). (B) Left panel: comparison between the logarithm of the protein−RNA complex SAXS intensity calculated using 1B/3B mapping without SLC (blue) and the logarithm of the WAXSiS profile (black dashed line). Right panel: comparison between the logarithm of the protein−RNA complex SAXS intensity calculated using 1B/3B mapping with the SLC (orange) and the logarithm of the WAXSiS profile (black dashed line). All of the residuals are calculated as the difference between the two intensities considered.

3.3. Solvation Layer Contribution Results in a Smaller Radius of Gyration and an Overall Decrease in the Fluctuations of Gelsolin.

The size and large structural variability of GSN made it an excellent candidate to provide a realistic evaluation of our method and to assess its applicability to practical scenarios. Specifically, we generated two independent GSN conformational ensembles through metainference multireplica simulations, using experimental SAXS data as a restraint and the 1B mapping and parameters to compute the forward models. One of the two ensembles was obtained by enabling SLC with a value of 0.08 and with the SC of 0.7 nm^2. We selected these settings based on the result of the analyses reported in the previous paragraph, as well as being particularly appropriate for GSN, this combination was also reasonable for B1 and GFP (Table S2). To define the ensembles, we considered only the second half of the trajectory of each replica, where the correlation between the forward model and the experimental intensity was stably close to 1, with a constant metainference score. From each ensemble, we extracted 1,000 equally distant frames, recalculated the SAXS profile using PLUMED with 1B mapping, and determined the average profile. For the ensemble frames with hydration layer correction, we used the same SLC and SC settings as those for the refinement. The two average profiles, representing the two ensembles, were directly compared with the experimental SAXS data (Figure 4A). We observed that both profiles show good agreement with the experimental SAXS data, with a chi-squared of 0.8 and 1.4 without and with the SLC term, respectively. To verify that our hydration layer correction is working properly, we also calculated the SAXS intensities using the WAXSiS web server and determined the corresponding average profiles. We compared the PLUMED profile with the WAXSiS profile of the ensemble obtained without SLC (Figure 4B, left panel) and the PLUMED profile with the WAXSiS profile of the ensemble obtained with SLC (Figure 4B, right panel). In this case, we found that the agreement between PLUMED and WAXSiS is higher when comparing the intensities calculated from the ensemble with SLC (RMSE: 1.7×10^{-2}) than when comparing the intensities calculated from the ensemble without SLC (RMSE: 6.4×10^{-2}). This indicates that although the hydration layer contribution in the SAXS intensity calculation is not critical to match the experimental data, hySAS can match WAXSiS with the appropriate SLC/SC settings. We analyzed both the ensembles in terms of radius of gyration and root-mean-square fluctuations (RMSFs) to gauge the effect of the SLC on the resulting conformations. Although both the ensembles showed a bimodal distribution of the radius of gyration, the one obtained with the inclusion of the SLC was, as possibly expected, more compact with an average radius of gyration of 3.05 nm, compared to the one generated without the SLC, which showed an average radius of gyration of 3.14 nm (Figure 5). Interestingly, a similar behavior was observed regarding the RMSF. The ensemble calculated applying the SLC shows systematically lower fluctuations, with an average RMSF of 0.26 nm, compared to the other ensemble, which has an RMSF of 0.38 nm (Figures 6, S6, and S7). Focusing on the SLC-corrected ensemble, the main contribution to the radius of gyration and the RMSF comes from the long N-terminal disordered region with significant fluctuations also found in the two main linkers connecting the G2 domain to the G3 domain and the G3 domain to the G4 domain (Figure 6). Referring to high-resolution data for some of the isolated domains (also in the presence of Ca^{2+} and/or actin),[76−78] GSN appears reasonably stable, suggesting that the model with smaller fluctuations is preferable.

3.4. Solvation Layer Contribution Results in a Lower Radius of Gyration in the Refinement of a Protein−RNA Complex. In addition to generating conformational ensembles, hySAS can also be used to refine single structures to enhance consistency with experimental SAS data. As an example of the latter application, we choose to improve a model of a previously published protein−RNA complex.[65] This system consists of the 199-residue unwinding protein 1 (UP1) interacting with a 12-mer single strand derived from the primary transcript of the 18*a* microRNA. The complex was originally refined using metainference, SAXS, and NMR data as restraints and successively tested with hySAXS and the Martini bead form factors.[34] Here, we repeated the latter test using the same input files and data but 1B and 3B mapping to compute the forward models. We generated a short trajectory with and without SLC with a value of 0.12 and an SC of 0.8 nm^2. From each trajectory, we obtained a refined structure with a chi-squared of 1 with respect to the SAXS data (Figure 7A). As for the GSN, to verify our method, we compared the intensities computed from the two selected conformations with the corresponding intensities recalculated with the WAXSiS web server. We obtained an RMSE of 6.4×10^{-2} between the PLUMED and the WAXSiS logarithm of the intensities (Figure 7B, left panel) when using the conformation generated without employing the hydration layer correction. However, when using the conformation calculated with the SLC, the RMSE drops to 1.2×10^{-2} (Figure 6B, right panel). Therefore, also in this case, the use of our SL allows us to obtain SAS profiles in agreement with WAXSiS.

Comparing the two resulting refined structures, it is possible to observe a difference in their radius of gyration, with the one obtained without using the SLC characterized by a radius of 2.26 nm as observed in our previous work,[65] and the one obtained using the SLC term by a radius of 2.20 nm. This difference is the result of more relaxed terminal regions of UP1.

4. CONCLUSIONS

The integration of experimental data in simulations is a powerful approach to increase the resolution of the former and the accuracy of the latter.[79−81] This integration is based on two elements: (i) a forward model for the calculation of an experimental observable, given a conformation and (ii) an integration strategy (e.g., restraints or reweighting based on either the maximum entropy principle or Bayesian inference[82,83]). The forward model should be accurate and computationally efficient when the goal is to apply a restraint in a simulation. In this work, we have presented an implementation of a SAXS and SANS forward model that efficiently exploits the limited resolution of these experimental techniques. In particular, it allows protein and nucleic acid scattering to be represented by a single-bead per amino acid and a three-bead per nucleic acid residue, and more importantly, it enables the effective on-the-fly inclusion of solute−solvent scattering corrections at no cost. We showed that the inclusion of this correction modifies the resulting conformations by mildly decreasing their radius of gyration, as expected, and matching WAXSiS, a more accurate but expensive forward model. The method presented here is already deployed in PLUMED, thus allowing its use in combination with different molecular dynamics engines, restraining strategies including metainference[61] and maximum entropy[84,85]/caliber[86] approaches or enhanced sampling techniques such as metadynamics[87] and umbrella sampling.[88]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c00864.

Extensive comparison of the performance of MT, 1B, and AA representation for proteins (considering also SLC), as well as the same comparison for 3B and AA for nucleic acids; table containing the list of PDB accession codes used to compute the 3B parameters; tables with the combinations of SASA threshold and the solvent density for SLC correction for both protein and nucleic acids; RMSF plot for the GSN without SLC; and ΔRMSF plot for the two simulations with and without SLC (PDF)

Accession Codes

All data are available via Zenodo with record 8192455, the PLUMED-NEST as plumID:23.029, and the SASBDB with accession SASDSN7.

AUTHOR INFORMATION

Corresponding Authors

Matteo de Rosa – *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy; Istituto di Biofisica, Consiglio Nazionale delle Ricerche (IBF-CNR), 20133 Milano, Italy*; Email: matteo.derosa@cnr.it

Riccardo Capelli – *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy*; orcid.org/0000-0001-9522-3132; Email: riccardo.capelli@unimi.it

Carlo Camilloni – *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy*; orcid.org/0000-0002-9923-8590; Email: carlo.camilloni@unimi.it

Authors

Federico Ballabio – *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy*

Cristina Paissoni – *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy*

Michela Bollati – *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy; Istituto di Biofisica, Consiglio Nazionale delle Ricerche (IBF-CNR), 20133 Milano, Italy*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c00864

Author Contributions

F.B. performed the computational work, M.B. and M.d.R performed the experiments on gelsolin, F.B., C.P, R.C., and C.C. designed the work and developed the hybrid SAS model, and F.B. and C.C. wrote the manuscript with contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank Diamond Light Source for beamtime (proposal MX20221) and the staff of beamlines B21 for assistance with data collection. They acknowledge CINECA for the availability of high-performance computing resources and support through the ISCRA award initiative.

ABBREVIATIONS

SAXS, small-angle X-ray scattering; SANS, small-angle neutron scattering; MD, molecular dynamics; SLC, solvation layer contribution; GSN, gelsolin; B1, B1 immunoglobulin-binding domain of streptococcal protein G; GFP, green fluorescent protein; FF, force field; PDB, protein data bank; RMSE, root-mean-square error; RMSF, root-mean-square fluctuation

REFERENCES

(1) Tuukkanen, A. T.; Spilotros, A.; Svergun, D. I. Progress in small-angle scattering from biological solutions at high-brilliance synchrotrons. *IUCrJ.* **2017**, *4* (5), 518–528.

(2) Koch, M. H. J.; Vachette, P.; Svergun, D. I. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q. Rev. Biophys.* **2003**, *36* (2), 147–227.

(3) Kikhney, A. G.; Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* **2015**, *589* (19 Pt A), 2570–2577 From NLM..

(4) Naudi-Fabra, S.; Tengo, M.; Jensen, M. R.; Blackledge, M.; Milles, S. Quantitative Description of Intrinsically Disordered Proteins Using Single-Molecule FRET, NMR, and SAXS. *J. Am. Chem. Soc.* **2021**, *143* (48), 20109–20121.

(5) Stelzl, L. S.; Pietrek, L. M.; Holla, A.; Oroz, J.; Sikora, M.; Köfinger, J.; Schuler, B.; Zweckstetter, M.; Hummer, G. Global Structure of the Intrinsically Disordered Protein Tau Emerges from Its Local Structure. *JACS Au* **2022**, *2* (3), 673–686.

(6) Aznauryan, M.; Delgado, L.; Soranno, A.; Nettels, D.; Huang, J.-r.; Labhardt, A. M.; Grzesiek, S.; Schuler, B. Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113* (37), E5389–E5398.

(7) Hub, J. S. Interpreting solution X-ray scattering data using molecular simulations. *Curr. Opin. Struct. Biol.* **2018**, *49*, 18–26.

(8) Cromer, D. T.; Waber, J. T. Scattering factors computed from relativistic Dirac-Slater wave functions. *Acta Crystallogr.* **1965**, *18* (1), 104–109.

(9) Brown, P. J.; Fox, A. G.; Maslen, E. N.; O'Keefe, M. A.; Willis, B. T. M. Intensity of Diffracted Intensities. In *International Tables for Crystallography*; International Union of Crystallography, 2006; pp 554–595.

(10) Fraser, R. D. B.; MacRae, T. P.; Suzuki, E. An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules. *J. Appl. Crystallogr.* **1978**, *11* (6), 693–694.

(11) Feigin, L. A.; Svergun, D. I. *Structure Analysis by Small-Angle X-ray and Neutron Scattering*; Springer Science & Business Media, 2013.

(12) Sears, V. F. Neutron scattering lengths and cross sections. *Neutron News* **1992**, *3* (3), 26–37.

(13) Heller, W. T. Small-angle neutron scattering and contrast variation: a powerful combination for studying biological structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66* (Pt 11), 1213–1217.

(14) Svergun, D. I.; Richard, S.; Koch, M. H.; Sayers, Z.; Kuprin, S.; Zaccai, G. Protein hydration in solution: experimental observation by x-ray and neutron scattering. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95* (5), 2267–2272.

(15) Perkins, S. J. X-ray and neutron scattering analyses of hydration shells: a molecular interpretation based on sequence predictions and modelling fits. *Biophys. Chem.* **2001**, *93* (2–3), 129–139.

(16) Linse, J.-B.; Jochen, S. H. Scrutinizing the protein hydration shell from molecular dynamics simulations against consensus small-angle scattering data. *bioRxiv* **2023**, DOI: 10.1101/2023.06.13.544709.

(17) Merzel, F.; Smith, J. C. Is the first hydration shell of lysozyme of higher density than bulk water? *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (8), 5378–5383.

(18) Chen, P.-C.; Hub, J. S. Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data. *Biophys. J.* **2014**, *107* (2), 435–447.

(19) Knight, C. J.; Hub, J. S. WAXSiS: a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics. *Nucleic Acids Res.* **2015**, *43* (W1), W225–230.

(20) Köfinger, J.; Hummer, G. Atomic-resolution structural information from scattering experiments on macromolecules in solution. *Phys. Rev. E* **2013**, *87* (5), No. 052712.

(21) Franke, D.; Petoukhov, M. V.; Konarev, P. V.; Panjkovich, A.; Tuukkanen, A.; Mertens, H. D. T.; Kikhney, A. G.; Hajizadeh, N. R.; Franklin, J. M.; Jeffries, C. M.; Svergun, D. I. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* **2017**, *50* (Pt 4), 1212–1225.

(22) Schneidman-Duhovny, D.; Hammel, M.; Sali, A. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* **2010**, *38* (Web Serverissue), W540–544.

(23) Grudinin, S.; Garkavenko, M.; Kazennov, A. Pepsi-SAXS: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Crystallogr., Sect. D: Struct. Biol.* **2017**, *73* (Pt 5), 449–464.

(24) Izadi, S.; Onufriev, A. V. Accuracy limit of rigid 3-point water models. *J. Chem. Phys.* **2016**, *145* (7), No. 074501.

(25) Conde, M. M.; Gonzalez, M. A.; Abascal, J. L. F.; Vega, C. Determining the phase diagram of water from direct coexistence simulations: the phase diagram of the TIP4P/2005 model revisited. *J. Chem. Phys.* **2013**, *139* (15), No. 154505.

(26) Bernetti, M.; Bussi, G. Comparing state-of-the-art approaches to back-calculate SAXS spectra from atomistic molecular dynamics simulations. *Eur. Phys. J. B* **2021**, *94* (9), 180.

(27) Svergun, D.; Barberato, C.; Koch, M. H. J. CRYSOL− a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **1995**, *28* (6), 768−773.

(28) Yang, S.; Park, S.; Makowski, L.; Roux, B. A Rapid Coarse Residue-Based Computational Method for X-Ray Solution Scattering Characterization of Protein Folds and Multiple Conformational States of Large Protein Complexes. *Biophys. J.* **2009**, *96* (11), 4449−4463.

(29) Ravikumar, K. M.; Huang, W.; Yang, S. Fast-SAXS-pro: A unified approach to computing SAXS profiles of DNA, RNA, protein, and their complexes. *J. Chem. Phys.* **2013**, *138* (2), No. 183a, DOI: 10.1063/1.4774148.

(30) Stovgaard, K.; Andreetta, C.; Ferkinghoff-Borg, J.; Hamelryck, T. Calculation of accurate small angle X-ray scattering curves from coarse-grained protein models. *BMC Bioinf.* **2010**, *11* (1), No. 429.

(31) Zheng, W.; Tekpinar, M. Accurate Flexible Fitting of High-Resolution Protein Structures to Small-Angle X-Ray Scattering Data Using a Coarse-Grained Model with Implicit Hydration Shell. *Biophys. J.* **2011**, *101* (12), 2981−2991.

(32) Niebling, S.; Bjorling, A.; Westenhoff, S. MARTINI bead form factors for the analysis of time-resolved X-ray scattering of proteins. *J. Appl. Crystallogr.* **2014**, *47* (4), 1190−1198.

(33) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini model. *Chem. Soc. Rev.* **2013**, *42* (16), 6801−6822.

(34) Paissoni, C.; Jussupow, A.; Camilloni, C. Martini bead form factors for nucleic acids and their application in the refinement of protein-nucleic acid complexes against SAXS data. *J. Appl. Crystallogr.* **2019**, *52* (2), 394−402.

(35) Jussupow, A.; Messias, A. C.; Stehle, R.; Geerlof, A.; Solbak, S. M. Ø.; Paissoni, C.; Bach, A.; Sattler, M.; Camilloni, C. The dynamics of linear polyubiquitin. *Sci. Adv.* **2020**, *6* (42), No. eabc3786.

(36) Paissoni, C.; Jussupow, A.; Camilloni, C. Determination of Protein Structural Ensembles by Hybrid-Resolution SAXS Restrained Molecular Dynamics. *J. Chem. Theory Comput.* **2020**, *16* (4), 2825−2834.

(37) Saad, D.; Paissoni, C.; Chaves-Sanjuan, A.; Nardini, M.; Mantovani, R.; Gnesutta, N.; Camilloni, C. High Conformational Flexibility of the E2F1/DP1/DNA Complex. *J. Mol. Biol.* **2021**, *433* (18), No. 167119.

(38) Paissoni, C.; Camilloni, C. How to Determine Accurate Conformational Ensembles by Metadynamics Metainference: A Chignolin Study Case. *Front. Mol. Biosci.* **2021**, *8*, No. 694130, DOI: 10.3389/fmolb.2021.694130.

(39) Ahmed, M. C.; Skaanning, L. K.; Jussupow, A.; Newcombe, E. A.; Kragelund, B. B.; Camilloni, C.; Langkilde, A. E.; Lindorff-Larsen, K. Refinement of α-Synuclein Ensembles Against SAXS Data: Comparison of Force Fields and Methods. *Front. Mol. Biosci.* **2021**, *8*, No. 654333, DOI: 10.3389/fmolb.2021.654333.

(40) Nag, S.; Larsson, M.; Robinson, R. C.; Burtnick, L. D. Gelsolin: The tail of a molecular gymnast. *Cytoskeleton* **2013**, *70* (7), 360−384.

(41) Yin, H. L.; Stossel, T. P. Control of cytoplasmic actin gel−sol transformation by gelsolin, a calcium-dependent regulatory protein. *Nature* **1979**, *281* (5732), 583−586.

(42) Sun, H. Q.; Yamamoto, M.; Mejillano, M.; Yin, H. L. Gelsolin, a multifunctional actin regulatory protein. *J. Biol. Chem.* **1999**, *274* (47), 33179−33182 From NLM..

(43) Piktel, E.; Levental, I.; Durnaś, B.; Janmey, P. A.; Bucki, R. Plasma Gelsolin: Indicator of Inflammation and Its Potential as a Diagnostic Tool and Therapeutic Target. *Int. J. Mol. Sci.* **2018**, *19* (9), 2516.

(44) Solomon, J. P.; Page, L. J.; Balch, W. E.; Kelly, J. W. Gelsolin amyloidosis: genetics, biochemistry, pathology and possible strategies for therapeutic intervention. *Crit. Rev. Biochem. Mol. Biol.* **2012**, *47* (3), 282−296.

(45) Li, G. H.; Arora, P. D.; Chen, Y.; McCulloch, C. A.; Liu, P. Multifunctional roles of gelsolin in health and diseases. *Med. Res. Rev.* **2012**, *32* (5), 999−1025.

(46) Ashish; Paine, M. S.; Perryman, P. B.; Yang, L.; Yin, H. L.; Krueger, J. K. Global Structure Changes Associated with Ca2+ Activation of Full-length Human Plasma Gelsolin*. *J. Biol. Chem.* **2007**, *282* (35), 25884−25892.

(47) Nag, S.; Ma, Q.; Wang, H.; Chumnarnsilpa, S.; Lee, W. L.; Larsson, M.; Kannan, B.; Hernandez-Valladares, M.; Burtnick, L. D.; Robinson, R. C. Ca^{2+} binding by domain 2 plays a critical role in the activation and stabilization of gelsolin. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106* (33), 13713−13718.

(48) Zorgati, H.; Larsson, M.; Ren, W.; Sim, A. Y. L.; Gettemans, J.; Grimes, J. M.; Li, W.; Robinson, R. C. The role of gelsolin domain 3 in familial amyloidosis (Finnish type). *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116* (28), 13958−13963.

(49) de Rosa, M.; Barbiroli, A.; Bonì, F.; Scalone, E.; Mattioni, D.; Vanoni, M. A.; Patrone, M.; Bollati, M.; Mastrangelo, E.; Giorgino, T.; Milani, M. The structure of N184K amyloidogenic variant of gelsolin highlights the role of the H-bond network for protein stability and aggregation properties. *Eur. Biophys. J.* **2020**, *49* (1), 11−19.

(50) Bollati, M.; Diomede, L.; Giorgino, T.; Natale, C.; Fagnani, E.; Boniardi, I.; Barbiroli, A.; Alemani, R.; Beeg, M.; Gobbi, M.; et al. A novel hotspot of gelsolin instability triggers an alternative mechanism of amyloid aggregation. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 6355−6365.

(51) Bonomi, M.; Camilloni, C. Integrative structural and dynamical biology with PLUMED-ISDB. *Bioinformatics* **2017**, *33* (24), 3999−4000.

(52) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604−613.

(53) Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G. A.; Banáš, P.; Barducci, A.; Bernetti, M.; Bolhuis, P. G.; Bottaro, S.; Branduardi, D.; et al. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **2019**, *16* (8), 670−673.

(54) Weiser, J.; Shenkin, P. S.; Still, W. C. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comput. Chem.* **1999**, *20* (2), 217−230.

(55) Šponer, J.; Bussi, G.; Krepl, M.; Banáš, P.; Bottaro, S.; Cunha, R. A.; Gil-Ley, A.; Pinamonti, G.; Poblete, S.; Jurečka, P.; et al. RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. *Chem. Rev.* **2018**, *118* (8), 4177−4338.

(56) Svozil, D.; Kalina, J.; Omelka, M.; Schneider, B. DNA conformations and their sequence preferences. *Nucleic Acids Res.* **2008**, *36* (11), 3690−3706.

(57) Bernauer, J.; Huang, X.; Sim, A. Y. L.; Levitt, M. Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA* **2011**, *17* (6), 1066−1075.

(58) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

(59) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27* (3), 221−234.

(60) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119* (16), 5113−5123.

(61) Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Sci.Adv.* **2016**, 2 (1), No. e1501177.

(62) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696−3713.

(63) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; et al. Parmbsc1: a refined force field for DNA simulations. *Nat. Methods* **2016**, *13* (1), 55−58.

(64) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79* (2), 926−935.

(65) Kooshapur, H.; Choudhury, N. R.; Simon, B.; Mühlbauer, M.; Jussupow, A.; Fernandez, N.; Jones, A. N.; Dallmann, A.; Gabel, F.; Camilloni, C.; et al. Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by hnRNP A1. *Nat. Commun.* **2018**, *9* (1), No. 2479.

(66) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.

(67) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90−95.

(68) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33−38.

(69) Schrödinger, L. L. C. The PyMOL Molecular Graphics System, Version 2.6.

(70) Ballabio, F.; Capelli, R.; Camilloni, C. Supporting data for: ″An accurate and efficient SAXS/SANS implementation including solvation layer effects suitable for restrained Molecular Dynamics simulations.". Zenodo: 2023.

(71) Giorgino, T.; Mattioni, D.; Hassan, A.; Milani, M.; Mastrangelo, E.; Barbiroli, A.; Verhelle, A.; Gettemans, J.; Barzago, M. M.; Diomede, L.; de Rosa, M. Nanobody interaction unveils structure, dynamics and proteotoxicity of the Finnish-type amyloidogenic gelsolin variant. *Biochim. Biophys. Acta, Mol. Basis Dis.* **2019**, *1865* (3), 648−660.

(72) Cowieson, N. P.; Edwards-Gayle, C. J. C.; Inoue, K.; Khunti, N. S.; Doutch, J.; Williams, E.; Daniels, S.; Preece, G.; Krumpa, N. A.; Sutter, J. P.; et al. Beamline B21: high-throughput small-angle X-ray scattering at Diamond Light Source. *J. Synchrotron Radiat.* **2020**, *27* (5), 1438−1446.

(73) Valentini, E.; Kikhney, A. G.; Previtali, G.; Jeffries, C. M.; Svergun, D. I. SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res.* **2015**, *43* (D1), D357−D363.

(74) Arsiccio, A.; Shea, J.-E. Protein Cold Denaturation in Implicit Solvent Simulations: A Transfer Free Energy Approach. *J. Phys. Chem. B* **2021**, *125* (20), 5222−5232.

(75) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *J. Comput. Chem.* **1995**, *16* (3), 273−284.

(76) Takeda, S.; Fujiwara, I.; Sugimoto, Y.; Oda, T.; Narita, A.; Maéda, Y. Novel inter-domain Ca2+-binding site in the gelsolin superfamily protein fragmin. *J. Muscle Res. Cell Motil.* **2020**, *41* (1), 153−162.

(77) Bollati, M.; Scalone, E.; Bonì, F.; Mastrangelo, E.; Giorgino, T.; Milani, M.; de Rosa, M. High-resolution crystal structure of gelsolin domain 2 in complex with the physiological calcium ion. *Biochem. Biophys. Res. Commun.* **2019**, *518* (1), 94−99.

(78) Vorobiev, S.; Strokopytov, B.; Drubin, D. G.; Frieden, C.; Ono, S.; Condeelis, J.; Rubenstein, P. A.; Almo, S. C. The structure of nonvertebrate actin: Implications for the ATP hydrolytic mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (10), 5760−5765.

(79) Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **2017**, *42*, 106−116 From NLM..

(80) Orioli, S.; Larsen, A. H.; Bottaro, S.; Lindorff-Larsen, K. How to learn from inconsistencies: Integrating molecular simulations with experimental data. *Prog. Mol. Biol. Transl. Sci.* **2020**, *170*, 123−176 From NLM..

(81) Habeck, M. Bayesian methods in integrative structure modeling. *Biol. Chem.* **2023**, *404*, 741.

(82) Hummer, G.; Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **2015**, *143* (24), No. 243150.

(83) Rangan, R.; Bonomi, M.; Heller, G. T.; Cesari, A.; Bussi, G.; Vendruscolo, M. Determination of Structural Ensembles of Proteins: Restraining vs Reweighting. *J. Chem. Theory Comput.* **2018**, *14* (12), 6632−6641.

(84) Cavalli, A.; Camilloni, C.; Vendruscolo, M. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* **2013**, *138* (9), No. 094112.

(85) Cesari, A.; Gil-Ley, A.; Bussi, G. Combining Simulations and Solution Experiments as a Paradigm for RNA Force Field Refinement. *J. Chem. Theory Comput.* **2016**, *12* (12), 6192−6200.

(86) Capelli, R.; Tiana, G.; Camilloni, C. An implementation of the maximum-caliber principle by replica-averaged time-resolved restrained simulations. *J. Chem. Phys.* **2018**, *148* (18), No. 184114.

(87) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (20), 12562−12566.

(88) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23* (2), 187−199.

Supporting Information for: An Accurate and Efficient SAXS/SANS Implementation Including Solvation Layer Effects Suitable for Molecular Simulations.

Federico Ballabio[1], Cristina Paissoni[1], Michela Bollati[1,2], Matteo de Rosa[1,2,*], Riccardo Capelli[1,*], and Carlo Camilloni[1*].

[1]Dipartimento di Bioscienze, Università degli Studi di Milano, via Celoria 26, 20133 Milano, Italy
[2]Istituto di Biofisica, Consiglio Nazionale delle Ricerche (IBF-CNR), via Alfonso Corti 12, 20133 Milano, Italy

Type	PDB Codes
RNA	157D, 1CSL, 1D4R, 1DQF, 1DUH, 1DUQ, 1F1T, 1F27, 1FIR, 1G2J, 1I9V, 1I9X, 1J9H, 1JZV, 1K9W, 1KD5, 1KFO, 1KH6, 1KXK, 1L2X, 1L3Z, 1MHK, 1MME, 1MSY, 1NBS, 1NUJ, 1P79, 1Q93, 1QBP, 1SA9, 1SDR, 1T0D, 1T0E, 1U9S, 1X8W, 1X9C, 1X9K, 1XJR, 1Y0Q, 1YFG, 1YKQ, 1YZD, 1Z43, 1Z58, 205D, 255D, 259D, 280D, 2A0P, 2A2E, 2A64, 2AO5, 2B8R, 2G3S, 2G91, 2H0S, 2NOK, 2OE6, 2TRA, 333D, 353D, 357D, 361D, 377D, 387D, 397D, 402D, 405D, 406D, 409D, 413D, 433D, 434D, 438D, 472D.
DNA (A-form)	118D, 137D, 138D, 160D, 1D78, 1D79, 1DNZ, 1KGK, 1M77, 1MA8, 1MLX, 1NZG, 1VJ4, 1VT5, 1VTB, 1XJX, 1Z7I, 1ZEX, 1ZEY, 1ZF1, 1ZF6, 1ZF8, 1ZF9, 1ZFA, 243D, 260D, 295D, 2D94, 317D, 338D, 344D, 345D, 348D, 349D, 368D, 369D, 370D, 371D, 395D, 396D, 399D, 414D, 440D, 9DNA.
DNA (B-form)	122D, 123D, 158D, 183D, 196D, 1BD1, 1BNA, 1CW9, 1D23, 1D3R, 1D49, 1D56, 1D8G, 1D8X, 1DOU, 1DPN, 1EDR, 1EHV, 1EN3, 1EN8, 1EN9, 1ENE, 1ENN, 1FQ2, 1G75, 1I3T, 1IKK, 1J8L, 1JGR, 1L4J, 1L6B, 1M6G, 1N1O, 1NVN, 1NVY, 1P4Y, 1P54, 1S23, 1S2R, 1SGS, 1SK5, 1UB8, 1VE8, 1ZF0, 1ZF3, 1ZF4, 1ZF5, 1ZF7, 1ZFB, 1ZFF, 1ZFG, 232D, 251D, 2D25, 307D, 355D, 3DNB, 403D, 423D, 428D, 431D, 436D, 454D, 455D, 456D, 460D,463D, 476D, 477D, 5DNB, 9BNA.
DNA (Z-form)	131D, 145D, 181D, 1D48, 1D53, 1DA2, 1DCG, 1DJ6, 1DNF, 1I0T, 1ICK, 1JES, 1LJX, 1OMK, 1VTT, 1VTW, 1XA2, 1XAM, 1ZNA, 210D, 211D, 242D, 292D, 293D, 2DCG, 313D, 314D, 331D, 336D, 351D, 362D, 400D,417D.
DNA (Quadruplexes	184D, 190D, 191D, 1BQJ, 1CN0, 1JPQ, 1L1H, 1MF5, 1O0K, 1QYK, 1QYL, 1V3N, 1V3O, 1V3P, 200D, 241D, 244D, 284D, 352D.

Table S1. List of PDB files used to compute the 3B parameters. The underlined 43 codes for RNA and 120 for DNA indicate the structures belonging to the initial training set. The remaining 32 RNA structures and 47 DNA structures were used as the validation set. The final parameters were calculated from the full set of 242 PDB structures.

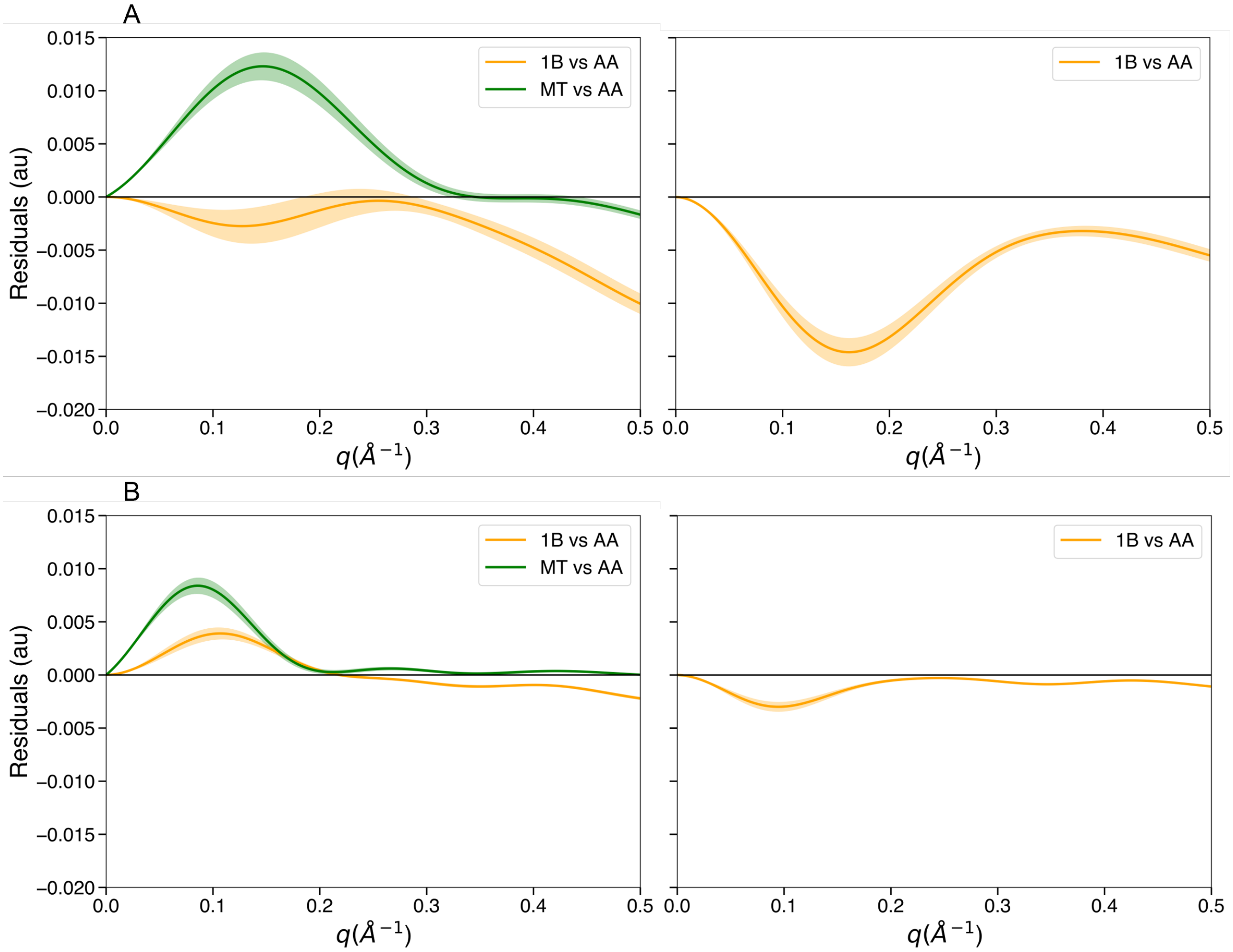


Figure S1. Accuracy evaluation of coarse-grained mappings in the calculation of scattering intensities. The SAS profile of each frame from MD trajectories was calculated with coarse-grained mappings and at AA resolution, for 201 q values ranging from $1{\cdot}10^{-10}$ Å^{-1} to 0.5 Å^{-1}. A) Left panel: average and standard deviation on 6,502 B1 frames of the SAXS residuals between MT and AA (green), and between 1B and AA (orange). Right panel: average and standard deviation of SANS residual between 1B and AA (orange). B) Left panel: average and standard deviation on 9,622 GFP frames of the SAXS residuals between MT and AA (green), and between 1B and AA (orange). Right panel: average and standard deviation of SANS residual between 1B and AA (orange).

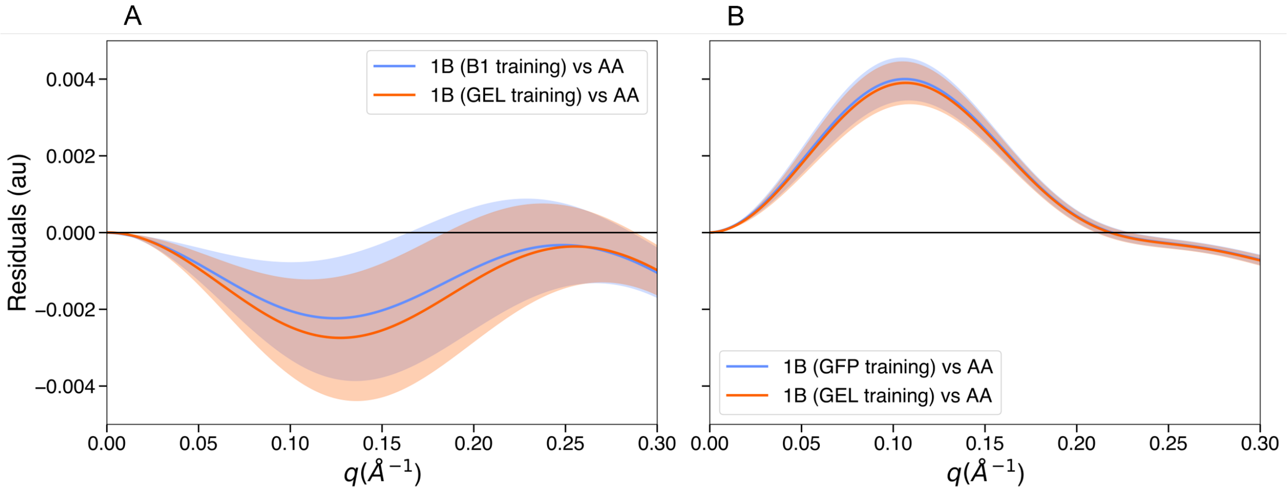


Figure S2. Transferability assessment of 1B parameters. The SAXS profile of each frame from B1 and GFP MD trajectories was calculated with 1B mapping employing B1 and GFP parameters, respectively, with 1B mapping and GSN parameters, and at AA resolution, for 201 *q* values ranging from $1 \cdot 10^{-10}$ Å^{-1} to 0.3 Å^{-1} A) Average and standard deviation on 6,502 B1 frames of the SAXS residuals between 1B with B1 parameters and AA (blue), and between 1B with GSN parameters and AA (red). B) Average and standard deviation on 9,622 GFP frames of the SAXS residuals between 1B with GFP parameters and AA (blue), and between 1B with GSN parameters and AA (red).

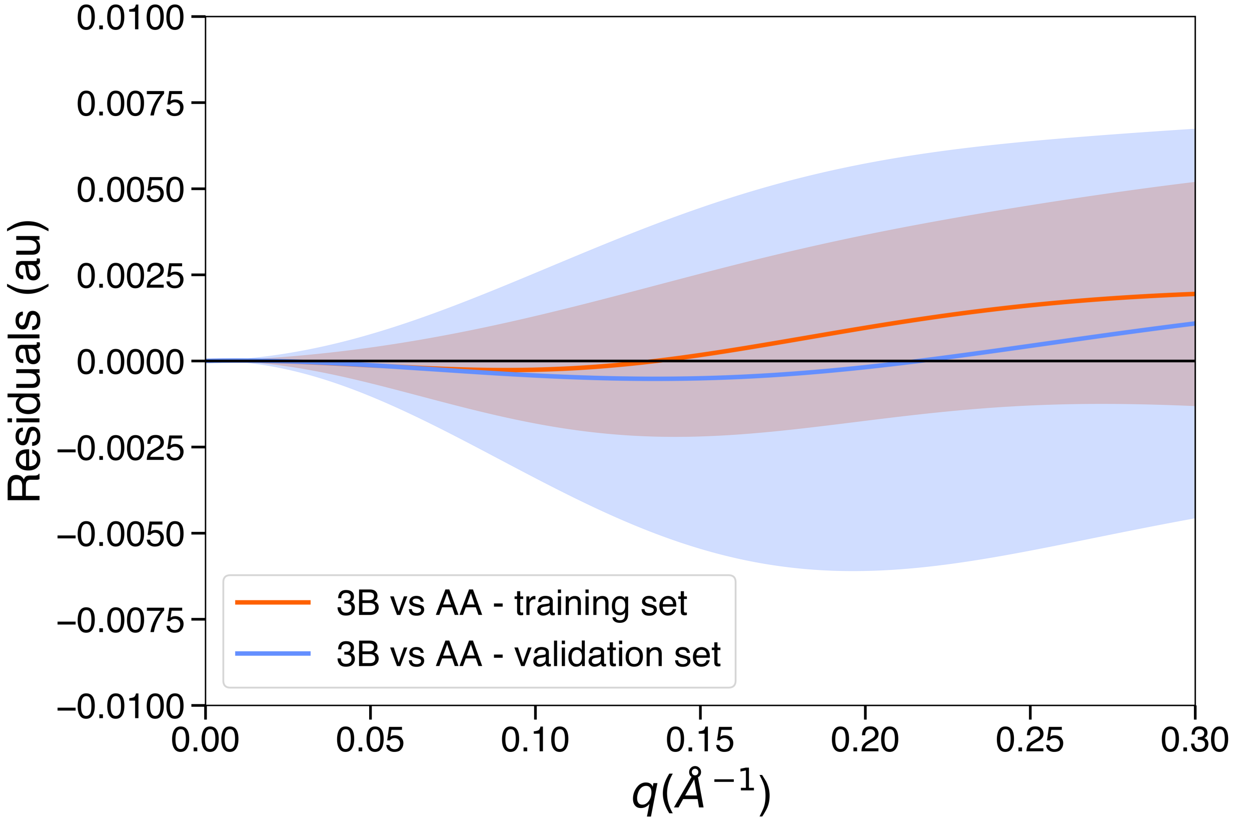


Figure S3. Transferability assessment of 3B parameters. The SAXS profile of each structure belonging to the training and to the validation sets was calculated with 3B mapping and the parameters computed from the PDB training set, and at AA resolution. The intensity was calculated for 201 *q* values ranging from $1 \cdot 10^{-10}$ Å^{-1} to 0.3 Å^{-1}. In red the average and standard deviation on 163 PDB structures (training set) of the SAXS residuals between 3B mapping and AA resolution. In blue the average and standard deviation on 79 PDB structures (validation set) of the SAXS residuals between 3B mapping and AA resolution.

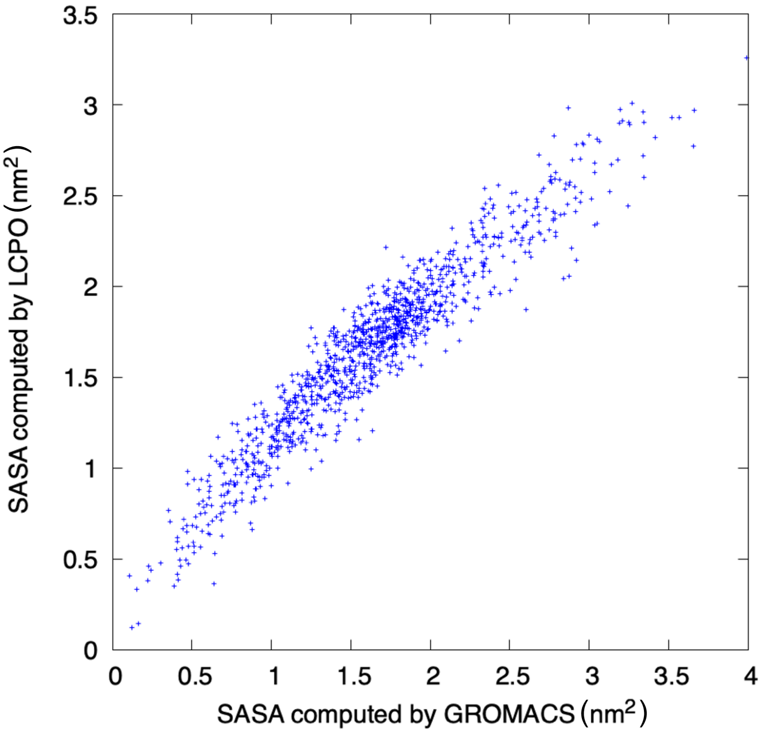


Figure S4. Comparison of SASA calculation between PLUMED LCPO and GROMACS. The SASA of each residue of a GSN frame randomly extracted from a MD trajectory was computed using the LCPO algorithm implemented in PLUMED (y-axis) and with the sasa module of GROMACS (x-axis). Each blue cross represents one residue.

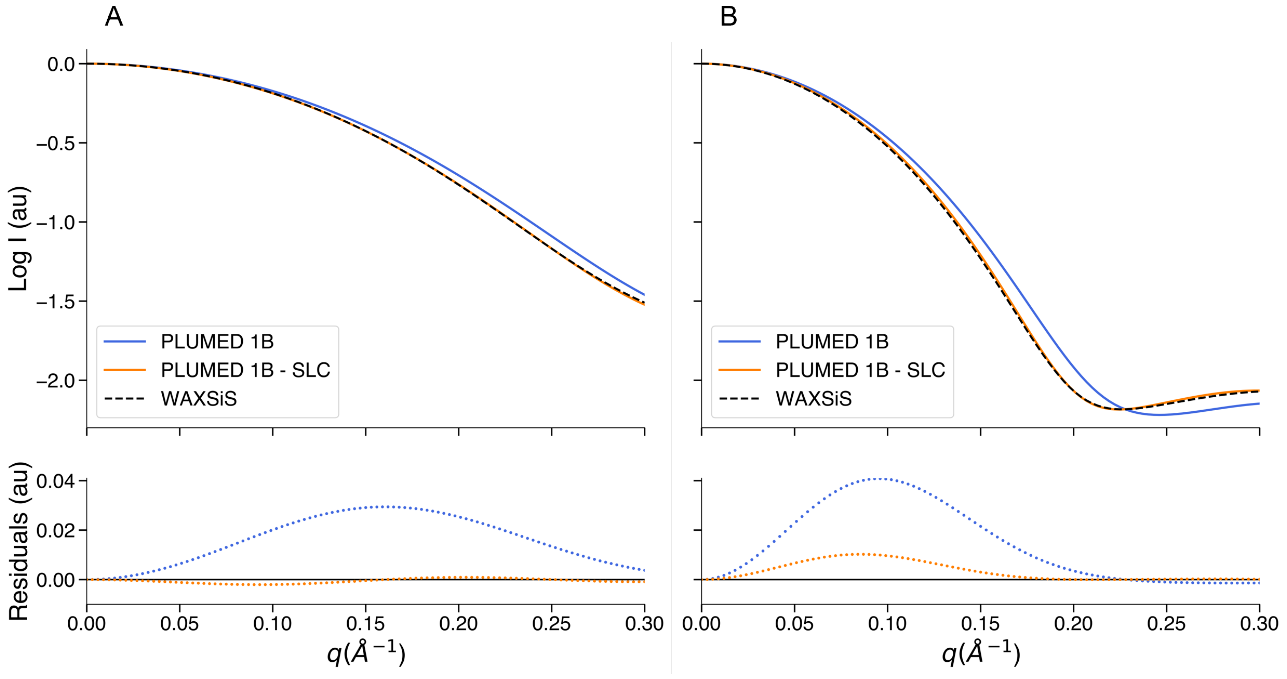


Figure S5. The solvation layer contribution in the 1B SAXS intensity calculation. A) Upper panel: base 10 logarithm of the SAXS profile of a representative, randomly selected, B1 frame calculated using 1B mapping (blue), 1B mapping with the best combination of SLC (0.09) and SC (1.2 nm^2) found for this frame (orange) and using WAXSiS (black dashed line). Bottom panel: residuals of 1B (blue) and 1B with SLC (orange) using the WAXSiS intensity as reference. B) Upper panel: base 10 logarithm of the SAXS profile of a representative, randomly selected, GFP frame calculated using 1B mapping (blue), 1B mapping with the best combination of SLC (0.08) and SC (0.7 nm^2) found for this frame (orange) and using WAXSiS (black dashed line). Bottom panel: residuals of 1B (blue) and 1B with SLC (orange) using the WAXSiS intensity as reference. All the SAXS intensities were calculated for 101 *q* values, up to 0.3 Å^{-1}.

A

RMSE (e-02)

MAPPING	**SLC**	**SC: 1.2**	**SC: 1.0**	**SC: 0.8**	**SC: 0.7**	**SC: 0.6**	**SC: 0.4**	//
AA	//	//	//	//	//	//	//	6.5
MT	//	//	//	//	//	//	//	7.8
1B	//	//	//	//	//	//	//	7.1
1B	0.040	5.2	4.4	3.7	3.5	3.4	3.6	//
1B	0.060	4.5	3.5	2.7	2.5	2.4	2.8	//
1B	0.070	4.2	3.2	2.4	2.2	2.1	2.6	//
1B	0.080	4.0	2.9	2.2	2.1	2.0	2.4	//
1B	0.090	3.8	2.7	2.2	2.2	2.1	2.4	//
1B	0.095	3.7	2.7	2.3	2.2	2.2	2.3	//
1B	0.100	3.6	2.6	2.3	2.3	2.3	2.4	//
1B	0.110	3.5	2.6	2.5	2.6	2.6	2.4	//
1B	0.120	3.4	2.7	2.8	2.8	2.9	2.5	//

B

RMSE (e-02)

MAPPING	**SLC**	**SC: 1.2**	**SC: 1.0**	**SC: 0.8**	**SC: 0.7**	**SC: 0.6**	**SC: 0.4**	//
AA	//	//	//	//	//	//	//	7.5
MT	//	//	//	//	//	//	//	9.0
1B	//	//	//	//	//	//	//	7.2
1B	0.040	4.7	4.3	4.0	3.9	4.0	4.2	//
1B	0.060	3.7	3.3	3.0	2.9	3.1	3.4	//
1B	0.070	3.2	2.9	2.6	2.5	2.7	3.1	//
1B	0.080	2.8	2.5	2.3	2.2	2.4	2.8	//
1B	0.090	2.5	2.3	2.0	1.9	2.2	2.6	//
1B	0.095	2.4	2.2	1.9	1.8	2.1	2.5	//
1B	0.100	2.2	2.1	1.8	1.6	2.0	2.4	//
1B	0.110	2.0	2.0	1.6	1.5	1.8	2.3	//
1B	0.120	1.9	1.9	1.5	1.4	1.8	2.2	//

C

		RMSE (e-02)						
MAPPING	**SLC**	**SC: 1.2**	**SC: 1.0**	**SC: 0.8**	**SC: 0.7**	**SC: 0.6**	**SC: 0.4**	//
AA	//	//	//	//	//	//	//	7.0
MT	//	//	//	//	//	//	//	8.0
1B	//	//	//	//	//	//	//	8.1
1B	0.040	6.2	5.5	4.6	4.3	4.2	4.2	//
1B	0.060	5.4	4.5	3.6	3.3	3.3	3.6	//
1B	0.070	5.1	4.1	3.3	3.1	3.1	3.7	//
1B	0.080	4.8	3.9	3.2	3.1	3.2	4.0	//
1B	0.090	4.5	3.7	3.2	3.2	3.5	4.4	//
1B	0.095	4.4	3.6	3.3	3.3	3.7	4.6	//
1B	0.100	4.4	3.6	3.4	3.5	3.9	4.8	//
1B	0.110	4.3	3.5	3.6	3.9	4.4	5.4	//
1B	0.120	4.2	3.6	4.0	4.4	4.9	5.9	//

D

		RMSE (e-02)						
MAPPING	**SLC**	**SC: 1.2**	**SC: 1.0**	**SC: 0.8**	**SC: 0.7**	**SC: 0.6**	**SC: 0.4**	//
AA	//	//	//	//	//	//	//	3.9
MT	//	//	//	//	//	//	//	4.7
3B	//	//	//	//	//	//	//	3.8
3B	0.040	3.1	2.9	2.9	2.9	3.1	3.4	//
3B	0.060	2.8	2.5	2.5	2.6	2.9	3.3	//
3B	0.070	2.7	2.4	2.3	2.5	2.8	3.1	//
3B	0.080	2.5	2.2	2.2	2.3	2.7	3.1	//
3B	0.090	2.4	2.1	2.0	2.2	2.6	3.1	//
3B	0.095	2.4	2.0	2.0	2.2	2.5	3.0	//
3B	0.100	2.3	2.0	1.9	2.1	2.6	3.0	//
3B	0.110	2.2	1.8	1.8	2.0	2.4	3.0	//
3B	0.120	2.1	1.7	1.7	1.9	2.3	2.9	//

Table S2. SLC and SC evaluation in SAXS intensity calculation. Each table shows the RMSE between the logarithm (base 10) of the SAXS intensity calculated with AA, MT, 1B/3B (with different values of SLC/SC) and the logarithm of the SAXS intensity calculated with WAXSiS, averaged for 10 equidistant frames extracted from A) GSN, B) B1, C) GFP and D) 12-*mer* RNA MD trajectories. The SC is expressed in nm^2.

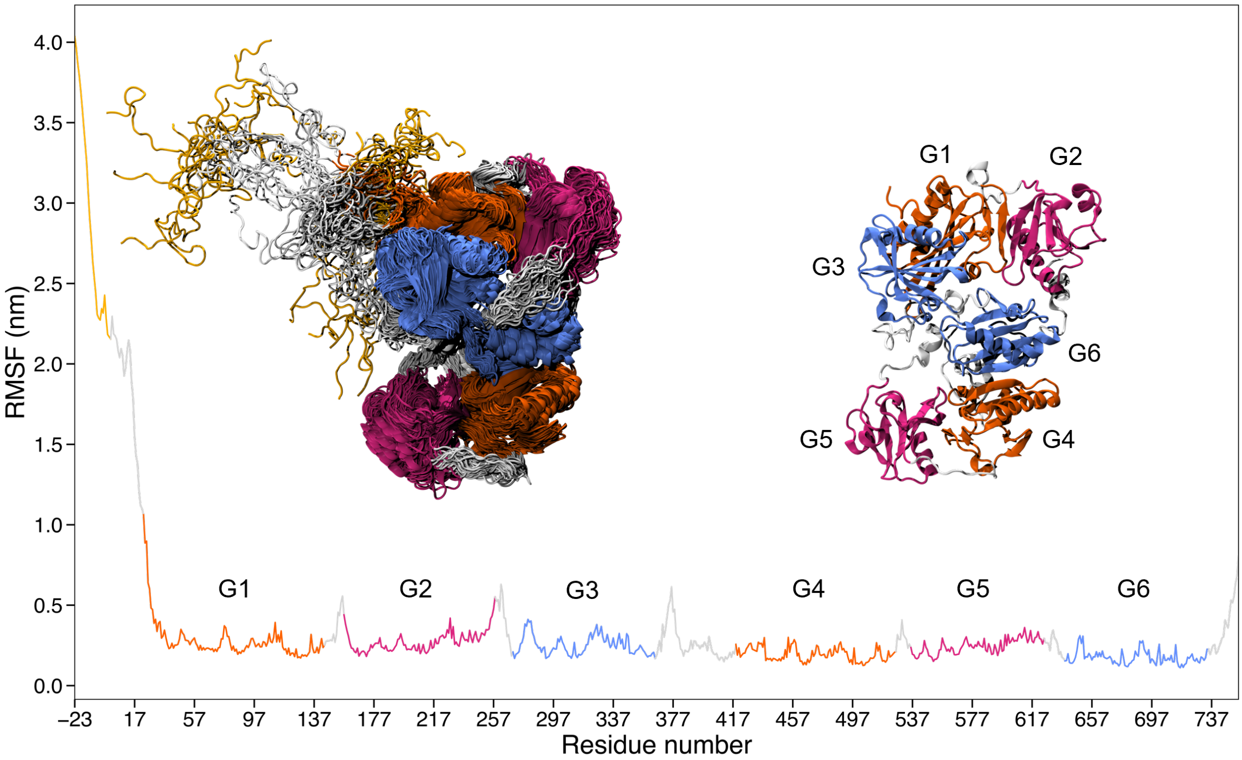


Figure S6. RMSF analysis of the GSN ensemble (without SLC). The flexibility of the protein was assessed by calculating the root-mean-square-fluctuation of all residues. The residue numbering sequence on the x-axis includes the N-terminal His$_6$-tag (from -23 to -1) and the full-length human plasma isoform of GSN (1 to 755). The domains sharing the highest sequence and structural similarity are shown with the same colour code: G1 and G4 in orange, G2 and G5 in purple, G3 and G6 in blue. The linkers and tails are coloured in light grey while the His$_6$-tag is coloured in yellow. On the left, 50 equidistant frames from the analysed trajectories are superimposed as a representative example of the conformational ensemble. The GSN structure on the right is that obtained by X-ray crystallography (PDB ID: 3FFN).

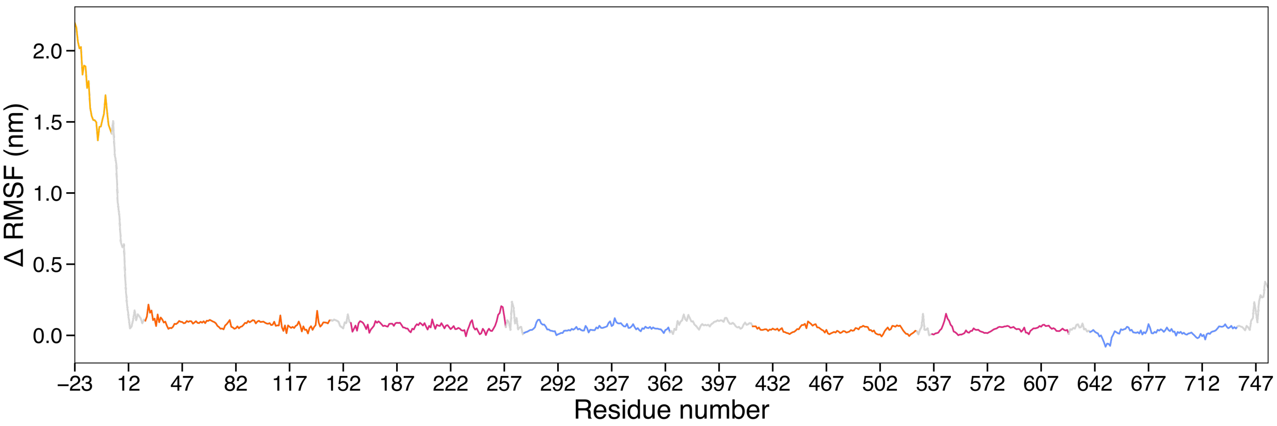


Figure S7. Difference in RMSF between the residues of the GSN ensemble obtained with SLC and the residues of the GSN ensemble obtained without SLC. The residue numbering sequence on the x-axis includes the N-terminal His6-tag (from -23 to -1) and the full-length human plasma isoform of GSN (1 to 755). The domains sharing the highest sequence and structural similarity are shown with the same colour code: G1 and G4 in orange, G2 and G5 in purple, G3 and G6 in blue. The linkers and tails are coloured in light grey while the His$_6$-tag is coloured in yellow.

3.1.3 Future Perspectives

Following the publication of the manuscript, although the forward model has remained unchanged, the corresponding code implemented in PLUMED has been continuously improved in terms of both performance and robustness. Numerous checks have been added to ensure the quality of the input PDB file required to define the beads, particularly in the case of the 3B mapping for nucleic acids. These controls are designed to mitigate, where possible, the generation of incorrect results when the initial structure contains errors. In addition, two new bead types have been defined: a 5′-phosphorylated end with an additional hydroxyl moiety at the phosphorus atom for nucleic acids, and the oxidised cysteine residue involved in a disulfide bridge.

To facilitate the use of hySAS, I have also written a practical manual, which can be found in the next section, 3.1.4, and on the PLUMED tutorials web page.[262] This manual not only provides tutorials demonstrating the use of hySAS but also includes instructions for installing and running the tool on high-performance computing (HPC) systems.

Finally, the characterisation of gelsolin continues. This protein undergoes significant domain rearrangements with increasing concentrations of calcium ions in solution, transitioning from a closed to an open conformation. To characterise this transition, hySAS will be used with SAXS data obtained at different calcium concentrations, complemented by additional techniques:

1. **Double Electron-Electron Resonance (DEER) Spectroscopy.**[96] This technique allows the measurement of distances between paramagnetic tags, in particular nitroxide spin labels such as methanethiosulfonate (MTSL), up to 8 nm. By taking measurements at increasing calcium concentrations and placing the labels at different positions on the protein surface, it is possible to determine the relative distances between domains as the calcium ion concentration in solution increases.

2. **Multi-*e*GO.**[263,264] This hybrid multistate structure-based model combines classical molecular dynamics with a structure-based approach. Specifically, electrostatic and van der Waals interactions are replaced by structure-based contacts, eliminating the need for electrostatics, explicit water, and hydrogen atoms. This method will enable efficient sampling of the gelsolin open state.

3.1.4 Manual

The following manual is available on the PLUMED tutorials website[262] under the accession ID 24.001.

hySAS intro & overview

The combination of Small-Angle X-ray and Neutron Scattering (SAXS/SANS or SAS) experiments with molecular dynamics (MD) simulations is an effective strategy for the characterisation of biomolecules in solution. On the one hand, the limited resolution of SAS can benefit from the MD contribution, and on the other hand, the inaccuracy of MD can be mitigated by using the experimental data to drive the simulations and generate conformational ensembles in agreement with the SAS data.

To achieve this result, an energy penalty is introduced to the system potential. This bias depends on the difference between the SAS experimental data and the SAS profile calculated in real-time from the system coordinates retrieved from the ongoing simulation. In this way, all the conformations that are not in agreement with the experimental data are discouraged.

The predictor used to calculate the SAS intensity from the coordinates of the molecule is called the “forward model”.

Although very promising, this approach is hampered by its high computational cost. One way to overcome this limitation is to calculate the intensity of the system of interest using a coarse-grained model, thus aggregating the scattering behaviour of groups of atoms into single particles. This does not mean that the simulation is also performed with a coarse-grained model, but that the SAS signal is calculated at a lower resolution: the simulation retains all the atomistic details!

Previously, we presented a hybrid resolution method that combines atomistic MD simulations with a Martini coarse-grained SAXS forward model.

We further enhance this technique by developing a novel hybrid-SAS method that is faster, more accurate, extended to the SANS intensity calculation and that is compatible with both proteins and nucleic acids. In the new hySAS forward model, an amino acid is represented by a single bead, while a nucleic acid is represented by three beads: one for the base, one for the sugar and one for the phosphate group. The centre of each particle is placed at the centre of mass of the atoms belonging to the bead itself. Protein-nucleic acid complexes are also compatible with the method.

In addition, an implicit and user-definable solvation layer contribution is included in the calculation to allow the reconstruction of a more realistic scattering behaviour in solution. This layer depends on solvent-solute interactions and, being typically more electron/neutron dense than the bulk solvent, actively contributes to the scattering signal. To account for this phenomenon, just for the beads that are exposed to the solvent is aplyied an electron density correction. For this reason, the solvent-accessible surface area of a bead is calculated from the coordinates of the heavy atoms belonging to that bead using the LCPO algorithm.

Some additional features are introduced specifically for SANS. It is possible to handle the percentage of deuterium in solution and implicitly account for hydrogen-deuterium exchange for the bead that are exposed to the solvent.

As well as generating a conformational ensemble or refining a structure, it is possible to use the forward model to calculate the SAS profile from a PDB or a trajectory using the PLUMED driver.

Installing PLUMED2 with ARRAYFIRE support

The purpose of this tutorial is to guide the user step-by-step through the process of compiling and installing PLUMED with ARRAYFIRE, an open-source library that supports a wide range of hardware accelerators for parallel computing.

In this specific scenario, the library is used to take advantage of CUDA-based GPUs to speed up the SAS calculation (ATOMISTIC / MARTINI / PARAMETERS / ONEBEAD representations). Furthermore, in order to show a real-life limit case, this installation example is performed on Leonardo, an HPC hosted by CINECA and managed by the SLURM scheduler.

To keep the working environment clean and compatible with other setups, in this guide the modules are not automatically loaded via .bashrc, .bash_aliases, or other configuration files, but are invoked during installation and then managed via the SLURM batch script used to submit jobs to the HPC.

1. Loading essential modules

To install ArrayFire, certain dependencies are required. Often, these dependencies are already available on HPC systems and can be loaded through modules. It is important to note that dependencies may also have their own dependencies, such as compilers or essential libraries. In general, to view the available modules on a SLURM-based HPC, you can use the command:

```
module avail
```

For a comprehensive list of requirements and additional installation suggestions, please visit the official ARRAYFIRE GitHub page. Continuing with the Leonardo example, here is the list of basic modules that can be loaded directly from the shell:

```
module load profile/lifesc
module load gmp/6.2.1
module load mpfr/4.1.0
module load mpc/1.2.1
module load gcc/11.3.0
module load zlib/1.2.13--gcc--11.3.0
module load openmpi/4.1.4--gcc--11.3.0-cuda-11.8
module load openblas/0.3.21--gcc--11.3.0
module load cblas/2015-06-06--gcc--11.3.0
module load gsl/2.7.1--gcc--11.3.0
module load cuda/11.8
```

Other required modules are available from Leonardo, but for general purposes and to provide useful examples, the other dependencies will be installed manually in the next steps. It is assumed that the downloads are stored in the home folder.

2. Building dependencies from source

Let's start by creating a folder where all the codes will be installed:

```
mkdir $HOME/build
```

FFTW

Download the “Fastest Fourier Transform in the West” package version 3.3.10:

```
wget https://www.fftw.org/fftw-3.3.10.tar.gz
```

Extract the archive:

```
tar -xzvf fftw-3.3.10.tar.gz
```

Copy the extracted folder to fftw-3.3.10_f and move into it:

```
cp -r fftw-3.3.10 fftw-3.3.10_f && cd fftw-3.3.10_f
```

The FFTW libraries must be built separately with single and double precision support. The next command configures the installation with 32-bit floating point support:

```
./configure --prefix=$HOME/build/fftw/ --enable-float --enable-shared
 CFLAGS="-march=native" --enable-avx2 --enable-sse2
```

Configuration details:

```
--prefix
```

Specifies the installation directory.

```
--enable-float
```

Enables single precision.

```
--enable-shared
```

Enables the compilation of shared libraries.

```
--enable-avx2 --enable-sse2
```

Enable the use of AVX2 and SSE2 instructions respectively, which can improve the performance of operations on CPUs that support these instructions.

```
CFLAGS="-march=native"
```

Sets compiler options to optimize the code for the specific architecture of the machine on which it is being compiled.

Compile and install the source code:

```
make && make install
```

In case of slowdowns, specify the number of processors to use, in this case 8:

```
make -j 8
```

Move to the fftw-3.3.10 folder and configure new instructions with the 64-bit floating point support:

```
cd ../fftw-3.3.10

./configure --prefix=$HOME/build/fftw/ --enable-shared

CFLAGS="-march=native" --enable-avx2 --enable-sse2
```

Compile and install the source code:

```
make && make install
```

Set the LD_LIBRARY_PATH environment variable directly in the shell:

```
export LD_LIBRARY_PATH="$HOME/build/fftw/lib/:$LD_LIBRARY_PATH"
```

Boost

Return to the download folder. Download the “Boost C++” libraries version 1.85.0:

```
wget https://boostorg.jfrog.io/artifactory/main/release/1.85.0/
source/boost_1_85_0.tar.gz
```

Extract the archive:

```
tar -xzvf boost_1_85_0.tar.gz
```

Move into the boost folder:

```
cd boost_1_85_0
```

Initialise the Boost build system:

```
./bootstrap.sh
```

Install Boost libraries specifying the directory where Boost will be installed:

```
./b2 install --prefix=$HOME/build/boost
```

Add the Boost library path to the LD_LIBRARY_PATH and the include path to the PATH:

```
export PATH="$HOME/boost/include/:$PATH"

export LD_LIBRARY_PATH="$HOME/build/boost/lib/:$LD_LIBRARY_PATH"
```

Spdlog

Return to the download folder. Clone and move into the spdlog repository:

```
git clone https://github.com/gabime/spdlog.git

cd spdlog
```

Checkout the specific version 1.9.2:

```
git checkout v1.9.2
```

Create build directory and navigate into it:

```
mkdir build && cd build
```

Run CMake to configure the build:

```
cmake .. -DCMAKE_INSTALL_PREFIX=$HOME/build/spdlog

-DSPDLOG_BUILD_SHARED=ON
```

Compile and install the source code:

```
make && make install
```

Set the environment variables:

```
export PATH="$HOME/build/spdlog/include/:$PATH"
export LD_LIBRARY_PATH="$HOME/build/spdlog/lib64/:$LD_LIBRARY_PATH"
```

Link to libcuda library

The next step can be tricky as it involves identifying a specific CUDA library on the cluster and linking it to your home directory. As the installation takes place on the login node, the CUDA drivers may not be directly available during code compilation. As an alternative, the stub library `libcuda.so` must be found on the HPC system. If the location is not clear, use the following command to find it:

```
find / -name libcuda.so 2>/dev/null
```

For Leonardo, the `libcuda.so` compatible with toolkit version 11.8 is located at:

```
/leonardo/prod/opt/compilers/cuda/11.8/none/lib64/stubs/libcuda.so
```

Create a folder in your home directory to link the library:

```
mkdir $HOME/libs && cd $HOME/libs
```

Link the library:

```
ln -s /leonardo/prod/opt/compilers/cuda/11.8/none/lib64/stubs/
libcuda.so libcuda.so.1
```

Update the LD_LIBRARY_PATH environment variable:

```
export LD_LIBRARY_PATH="$HOME/libs/:$LD_LIBRARY_PATH"
```

3. Building and installing ArrayFire from source

Return to the download folder. Clone the ArrayFire repository:

```
git clone --recursive https://github.com/arrayfire/arrayfire.git
```

Move into the arrayfire directory and checkout the specific version (v3.9.0):

```
cd arrayfire
git checkout v3.9.0
```

Create a build directory and navigate into it:

```
mkdir build && cd build
```

Configure the build system with CMake:

```
cmake .. -DCMAKE_INSTALL_PREFIX=$HOME/build/arrayfire
-DAF_BUILD_OPENCL=OFF -DAF_BUILD_CPU=OFF -DAF_BUILD_CUDA=ON
```

```
-DAF_BUILD_FORGE=OFF -DFFTW_INCLUDE_DIR=$HOME/build/fftw/include
-DFFTWF_LIBRARY=$HOME/build/fftw/lib/libfftw3f.so
-DFFTW_LIBRARY=$HOME/build/fftw/lib/libfftw3.so
-DBoost_DIR=$HOME/build/boost
-DBoost_INCLUDE_DIR=$HOME/build/boost/include
-DCUDA_TOOLKIT_ROOT_DIR=/leonardo/prod/opt/compilers/cuda/11.8/none
-DNVPRUNE=/leonardo/prod/opt/compilers/cuda/11.8/none/bin/nvprune
-Dspdlog_DIR=$HOME/build/spdlog/lib64/cmake/spdlog
```

Configuration details:

```
-DCMAKE_INSTALL_PREFIX=$HOME/build/arrayfire
```

Specifies the installation directory for ArrayFire.

```
-DAF_BUILD_OPENCL=OFF
```

Disables the OpenCL backend.

```
-DAF_BUILD_CPU=OFF
```

Disables the CPU backend.

```
-DAF_BUILD_CUDA=ON
```

Enables the CUDA backend.

```
-DAF_BUILD_FORGE=OFF
```

Disables the Forge library build.

```
-DFFTW_INCLUDE_DIR=$HOME/build/fftw/include
```

Specifies the directory containing the FFTW include files.

```
-DFFTWF_LIBRARY=$HOME/build/fftw/lib/libfftw3f.so
```

Specifies the single-precision FFTW library.

```
-DFFTW_LIBRARY=$HOME/build/fftw/lib/libfftw3.so
```

Specifies the double-precision FFTW library.

```
-DBoost_DIR=$HOME/build/boost:
```

Specifies the directory containing Boost.

```
-DBoost_INCLUDE_DIR=$HOME/build/boost/include
```

Specifies the Boost include directory.

```
-DCUDA_TOOLKIT_ROOT_DIR=/leonardo/prod/opt/compilers/cuda/11.8/none
```

Specifies the CUDA toolkit root directory.

```
-DNVPRUNE=/leonardo/prod/opt/compilers/cuda/11.8/none/bin/nvprune
```

Specifies the path to the nvprune tool.

```
-Dspdlog_DIR=$HOME/build/spdlog/lib64/cmake/spdlog
```

Specifies the directory containing the spdlog CMake configuration files.

Build and install ArrayFire:

```
make && make install
```

Check that `libafcuda.so` is correctly built, all the required shared libraries must be resolved:

```
ldd $HOME/build/arrayfire/lib64/libafcuda.so
```

Update the LD_LIBRARY_PATH environment variable:

```
export LD_LIBRARY_PATH="$HOME/build/arrayfire/lib64/:$LD_LIBRARY_PATH"
```

4. Installing PLUMED2 with ArrayFire support

Clone the PLUMED repository:

```
git clone --recursive https://github.com/plumed/plumed2.git
```

Navigate to the PLUMED directory:

```
cd plumed2
```

Configure the build system:

```
./configure --prefix=$HOME/build/plumed CC=mpicc CXX=mpicxx
--enable-modules=all --enable-asmjit --enable-fftw
LDFLAGS="-L$HOME/build/arrayfire/lib64/
-Wl,-rpath,$HOME/build/arrayfire/lib64/
-L$HOME/build/fftw/lib/ -Wl,-rpath,$HOME/build/fftw/lib/"
CPPFLAGS="-I$HOME/build/arrayfire/include
-I$HOME/build/fftw/include" --enable-af_cuda --verbose
```

Configuration details:

```
CC=mpicc CXX=mpicxx
```

Specifies the MPI compilers to use.

```
--enable-modules=all
```

Enables all PLUMED modules.

```
--enable-asmjit
```

Enables the AsmJit library.

```
--enable-fftw
```

Enables FFTW support.

```
LDFLAGS
```

Specifies linker flags:

```
-L$HOME/build/arrayfire/lib64/
```

Adds the ArrayFire library directory to the linker search path.

```
-Wl,-rpath,$HOME/build/arrayfire/lib64/
```

Adds the ArrayFire library directory to the runtime library search path.

```
-L$HOME/build/fftw/lib/
```

Adds the FFTW library directory to the linker search path.

```
-Wl,-rpath,$HOME/build/fftw/lib/
```

Adds the FFTW library directory to the runtime library search path.

```
CPPFLAGS
```

Specifies preprocessor flags:

```
-I$HOME/build/arrayfire/include
```

Adds the ArrayFire include directory to the compiler search path.

```
-I$HOME/build/fftw/include
```

Adds the FFTW include directory to the compiler search path.

```
--enable-af_cuda
```

Enables ArrayFire CUDA support.

```
--verbose
```

Enables verbose output during the configuration process.

Build and install PLUMED:

```
make && make install
```

5. Logout & login

To prevent compatibility issues caused by exporting the stub library, completely logout of the HPC shell and perform a clean login. Avoiding this step could result in “CUDA device native identification” errors.

6. Write & run the SLURM batch script

Create a SLURM batch file, e.g. `RUN.sh`, using a text editor:

```
vim RUN.sh
```

Set the job configuration, the modules, the environment variables and run the PLUMED driver command:

```
#!/bin/bash


#SBATCH -A _PROJECT_
#SBATCH -p _PARTITION_
#SBATCH --time HH:MM:SS


#SBATCH --job-name=PLUMED_AF
#SBATCH -N 1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=8
#SBATCH --gres=gpu:1
#SBATCH -o OUT.log


### MODULES ###
module load profile/lifesc
module load gmp/6.2.1
module load mpfr/4.1.0
module load mpc/1.2.1
module load gcc/11.3.0
module load zlib/1.2.13--gcc--11.3.0
module load openmpi/4.1.4--gcc--11.3.0-cuda-11.8
module load openblas/0.3.21--gcc--11.3.0
module load cblas/2015-06-06--gcc--11.3.0
module load gsl/2.7.1--gcc--11.3.0
module load cuda/11.8
```

```
### DEPENDENCIES ###
export LD_LIBRARY_PATH="$HOME/build/fftw/lib/:$LD_LIBRARY_PATH"
export PATH="$HOME/build/boost/include/:$PATH"
export LD_LIBRARY_PATH="$HOME/boost/lib/:$LD_LIBRARY_PATH"
export PATH="$HOME/build/spdlog/include/:$PATH"
export LD_LIBRARY_PATH="$HOME/build/spdlog/lib64/:$LD_LIBRARY_PATH"
export LD_LIBRARY_PATH="$HOME/build/arrayfire/lib64/:$LD_LIBRARY_PATH"

### PLUMED ###
export PATH="$HOME/build/plumed/bin:$PATH"
export LD_LIBRARY_PATH="$HOME/build/plumed/lib/:$LD_LIBRARY_PATH"
export PKG_CONFIG_PATH=
"$HOME/build/plumed/lib/pkgconfig/:$PKG_CONFIG_PATH"
export PLUMED_KERNEL="$HOME/build/plumed/lib/libplumedKernel.so"

export PLUMED_NUM_THREADS=$SLURM_CPUS_PER_TASK

### CMD ###
plumed driver --plumed plumed.dat --mf_xtc trj.xtc
```

In this example, the plumed command launches the driver to analyse the molecular dynamics trajectory `trj.xtc` according to the `plumed.dat` file. For the details regarding the `plumed.dat`instructions, refer to the Tutorial 1 of this guide.
Run the SLURM script file:

```
sbatch RUN.sh
```

7. Running GROMACS with PLUMED support

In addition to analysing a PDB or MD trajectory with the driver, it is possible to generate a conformational ensemble that agrees with SAS data using an MD engine. To enable this feature in GROMACS, it must be patched with the plumed-patch module and compiled with MPI support. Here is an example of a SLURM script that, after loading dependencies and environment variables, launches GROMACS with PLUMED support:

```
#!/bin/bash

#SBATCH -A _PROJECT_
#SBATCH -p _PARTITION_
#SBATCH --time HH:MM:SS

#SBATCH --job-name="hysas_metainference"
#SBATCH -o OUT.log

#SBATCH -N 1
#SBATCH --ntasks-per-node=4
#SBATCH --cpus-per-task=8
#SBATCH --gres=gpu:4
#SBATCH -o OUT.log

### MODULES ###
module load profile/lifesc
module load gmp/6.2.1
module load mpfr/4.1.0
module load mpc/1.2.1
module load gcc/11.3.0
module load zlib/1.2.13--gcc--11.3.0
module load openmpi/4.1.4--gcc--11.3.0-cuda-11.8
module load openblas/0.3.21--gcc--11.3.0
module load cblas/2015-06-06--gcc--11.3.0
module load gsl/2.7.1--gcc--11.3.0
module load cuda/11.8
```

```
### DEPENDENCIES ###
export LD_LIBRARY_PATH="$HOME/build/fftw/lib/:$LD_LIBRARY_PATH"
export PATH="$HOME/build/boost/include/:$PATH"
export LD_LIBRARY_PATH="$HOME/boost/lib/:$LD_LIBRARY_PATH"
export PATH="$HOME/build/spdlog/include/:$PATH"
export LD_LIBRARY_PATH="$HOME/build/spdlog/lib64/:$LD_LIBRARY_PATH"
export LD_LIBRARY_PATH="$HOME/build/arrayfire/lib64/:$LD_LIBRARY_PATH"

### PLUMED ###
export PATH="$HOME/build/plumed/bin:$PATH"
export LD_LIBRARY_PATH="$HOME/build/plumed/lib/:$LD_LIBRARY_PATH"
export PKG_CONFIG_PATH=
"$HOME/build/plumed/lib/pkgconfig/:$PKG_CONFIG_PATH"
export PLUMED_KERNEL="$HOME/build/plumed/lib/libplumedKernel.so"

### GROMACS ###
source $HOME/build/gmx24_mpi/bin/GMXRC

### EXPORT ###
export OMP_PROC_BIND=true
export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK

### CMD ###
MPIRUN=$(which mpirun)
MDRUN=$(which gmx_mpi)
NP=‘expr $SLURM_JOB_NUM_NODES \* $SLURM_NTASKS_PER_NODE‘

$MPIRUN -np $NP --map-by socket $MDRUN mdrun -v -deffnm replica -pin on
-ntomp $SLURM_CPUS_PER_TASK -nsteps -1 -dlb yes -nb gpu -bonded gpu
-pme gpu -plumed ../plumed.dat -multidir 0 1 2 3 -cpi replica.cpt
-maxh 23.9
```

In this example a single node is dedicated to perform a multidir simulation of 4 replicas, one for each GPU.

8. Final notes

This guide was written and verified in June 2024, using the latest version of PLUMED available at that time. The names, paths, and versions of the modules can vary between HPC systems. Even for Leonardo, a software stack update might make this guide partially incompatible. However, all provided examples should offer a broad logic to customize and adapt the installation on local machines or other HPC systems.

Preparing the input files

The hySAS ONEBEAD mapping (1 bead for 1 amino acid, 3 beads for 1 nucleodide) requires two PDBs, one for the MOLINFO and one for the TEMPLATE actions. This applies to both SAXS and SAS modules. The MOLINFO PDB may contain atoms, residues, chains, that are not included for the SAS signal calculation, while the TEMPLATE PDB must contain only the residues that are used for the ONEBEAD conversion. Some general rules:

1. Consistency between MOLINFO and TEMPLATE PDBs

The numbering of atoms and residues must be consistent between the MOLINFO and TEMPLATE PDBs. For example, if MOLINFO is provided with a PDB containing 100 residues, but the SAS calculation is performed on residues in the range 20-80, it is necessary to provide TEMPLATE with a PDB containing only these residues, while maintaining their numbering: if the 20th residue starts with atom number 319 and the 80th residue ends with atom number 1325, the TEMPLATE PDB must start with ATOM 319 and residue number 20, and end with ATOM 1325 and residue number 80. Warning: If a residue is included in the SAS calculation, all corresponding atoms must be present in the PDB. Missing atoms from a residue will cause the bead mapping step to fail.

2. Consistency between MOLINFO and the molfile

When using the plumed driver, the supplied molfile (PDB, XTC, . . .) to be analysed must be consistent with the PDB provided in MOLINFO. Specifically, there must be the same number of atoms in the same order.

3. MOLINFO PDB numbering

The MOLINFO PDB must start with ATOM number 1. The residue number can start with a positive number other than 1.

4. Providing the same file to MOLINFO and TEMPLATE

It is possible to provide MOLINFO and TEMPLATE with the same PDB, taking into account all three points above.

5. PDB format

The naming conventions for atoms and residues can vary depending on the force field or software used. This variability, particularly in the case of nucleic acids, is a major limitation in the hySAS mapping process. Detecting the residue type and analysing the atoms belonging to a bead is essential to:

- calculate the Solvent-Accessible Surface Area, required to introduce the solvation layer contribution or the hydrogen-deuterium exchange (SANS only);
- calculate the bead centre of mass, used to define the centre of the bead, a feature which has a major influence on the final SAS profile calculation.

For these reasons, as a preliminary step, each bead type and atom composition is verified. In order to perform this process, it is not feasible to consider all possible atom and residue names.

- Amino acids

Several histidine residue names in different protonation states are accepted: HIS, HIE/HID/HIP (AMBER), HSE/HSD/HSP (CHARMM).

Besides CYS, the cysteine involved in disulfide bridge is also allowed. In the latter case, the residue name must be CYX.

- Nucleic acids

Only the AMBER OL3 nomenclature for RNA residue and atom names, and the AMBER OL15 nomenclature for DNA residue and atom names, are accepted. The easiest way to convert a PDB to these formats is to use the pdb4amber script which is part of the free AmberTools suite.

This is the command:

```
pdb4amber -i original_DNA.pdb -o DNA_amber.pdb -d -v --add-missing-atoms
```

Warning: The `-d` `--add-missing-atoms` flags combination adds the hydrogen atoms to the PDB even if they already exist with a different nomenclature. This means that this step could double the number of atoms in the PDB. A solution could be to remove the original hydrogens from the PDB allowing the script to add them again with the correct nomenclature. Note that the order of the atoms in the MOLINFO and TEMPLATE PDBs, as well as in a previously generated trajectory to be analysed with the driver, must be consistent.
Alternatively, the LEaP script, also part of the AmberTools suite, can convert the PDB. Here is a quick step-by-step guide:
- RNA
create a `leapRNA.in` file:

```
vim leapRNA.in
```

copy the following instructions and save the file:

```
source leaprc.RNA.OL3

rna_molecule = loadpdb original_RNA.pdb

savepdb rna_molecule RNA_amber.pdb

quit
```

Command details:

```
source leaprc.RNA.OL3
```

Load the parameter set for RNA molecules, specifically the OL3 force field parameters.

```
rna_molecule = loadpdb original_RNA.pdb
```

Load the RNA molecule from the specified PDB file (original_RNA.pdb) into the LEaP program and assign it to the variable rna_molecule. Replace `original_RNA.pdb` with the PDB name you wish to convert.

```
savepdb rna molecule RNA amber.pdb
```

Save the RNA molecule that was loaded into the LEaP environment into a new PDB file named `RNA_amber.pdb`. This new PDB file will be formatted according to the conventions used by the AMBER software suite. Replace `RNA_amber.pdb` with your preferred PDB name.

Run LEaP:

```
tleap -s -f leapRNA.in
```

- DNA

create a `leapDNA.in` file:

```
vim leapDNA.in
```

copy the following instructions and save the file:

```
source leaprc.DNA.OL15

dna_molecule = loadpdb original_DNA.pdb

savepdb dna_molecule DNA_amber.pdb

quit
```

Run LEaP:

```
tleap -s -f leapDNA.in
```

6. Special beads

DNA and RNA

Three additional bead types are available for DNA and RNA besides phosphate group, pentose sugar, and nucleobase:

- 5′-end pentose sugar capped with a hydroxyl moiety at C5′. To enable this, the residue name in the PDB must be followed by “5”. For example, in the case of cytosine, the corresponding residue must be edited to DC5 or to C5 in DNA and RNA, respectively.
- 3′-end pentose sugar capped with an hydroxyl moiety at C3′. To enable this, the residue name in the PDB must be followed by “3”. For example, in the case of cytosine, the corresponding residue must be edited to DC3 or to C3 in DNA and RNA, respectively. - 5′-phosphorylated end with an additional hydroxyl moiety at P. To enable this, the residue name in the PDB must be followed by “T”. For example, in the case of cytosine, the corresponding residue must be edited to DCT or to CT in DNA and RNA, respectively. The additional O atom can be named in the PDB as OP3 or O3P, while the additional H can be named HOP3.

Proteins

An additional bead type is available for proteins:

- Cysteine residue involved in disulfide bridge (the residue in the PDB must be named CYX).

7. Atom names

For practical purposes, the table below lists all DNA/RNA atom names accepted for each bead category. Note that an additional bead check is performed to ensure that the correct atom is associated with the correct nucleotidee bead type. For example, if an O2′ atom name is detected in a thymine bead, an error message is triggered.

Pentose Bead Atoms	**Nucleobase Bead Atoms**	**PO Bead Atoms**
O5′, C5′, O4′, C4′, O3′, C3′, O2′, C2′, C1′, H5′, H5″, H4′, H3′, H2′, H2″, H2′2, H1′, HO5′, HO3′, HO2′, H5′1, H5′2, HO′2, H2′1, H5T, H3T.	N1, N2, N3, N4, N6, N7, N9, C2, C4, C5, C6, C7, C8, O2, O4, O6, H1, H2, H3, H5, H6, H8, H21, H22, H41, H42, H61, H62, H71, H72, H73.	P, OP1, OP2, OP3, O1P, O2P, O3P, HP, HOP3.

Tutorial 1 - Determining a scattering profile with hySAS

In the next example, we calculate the scattering profile of a structure consisting of atoms 1-11558 using the ONEBEAD approach. This structure could be made of protein(s), DNA, RNA, or complexes. As a first step, let's create and open the `plumed.dat` file, which will contain all the instructions:

```
vim plumed.dat
```

The first requirement is to provide the PDB of the structure via MOLINFO, in this case `template_AA.pdb`. Detailed information about the input file can be found in the technical support section.

```
MOLINFO STRUCTURE=template_AA.pdb
```

The SAS instructions need the corresponding action option, which can be either SAXS or SANS. In the following example, we proceed by opening the SAXS action:

```
MOLINFO STRUCTURE=template_AA.pdb


SAXS ...
```

The SAXS calculation options must be listed with the appropriate flags, in any order. Here, we include atoms 1-11878 and specify the ONEBEAD mapping (1 bead for 1 amino acid, 3 beads for 1 nucleotide). The same `template_AA.pdb` is used for both MOLINFO and TEMPLATE. The GPU flag allows the calculation of the scattering profile(s) on the GPU; this feature requires PLUMED to be compiled with support for the ArrayFire library. See the corresponding technical support section in this manual for more information. The electron density of the solvent is set to 0.334 electrons per cubic angstrom (bulk water) with the SOLVDENS flag. In the example, the solvation layer contribution is ignored by setting SOLVATION_CORRECTION=0.00. For this reason, other flags such as SASA_CUTOFF and SOLVATION_STRIDE are not required. Further examples are given at the end of this page.

```
MOLINFO STRUCTURE=template_AA.pdb


SAXS ...
        ATOMS=1-11878
        ONEBEAD
        LABEL=saxsdata
        GPU
        TEMPLATE=template_AA.pdb
        SOLVDENS=0.334
        SOLVATION_CORRECTION=0.00
```

After the option flags, the user has to provide the list of q-values for which the corresponding SAS intensity must be calculated. In this example, we provide 121 q-values, from almost zero to 0.3 Å^{-1} (which corresponds to the maximum momentum transfer available for the ONEBEAD mapping), every 0.0025 Å^{-1}. The more the q-values, the more intensive the calculation. After listing the q-values, the SAXS action must be closed and the intensities printed. In the case of a trajectory, it is possible to select the printing frequency. In the example, the intensities are printed at every simulation step, specifying STRIDE=1. Additional {} brackets have been added to prevent a readability issue, they are optional to run the code. Save and close the file.

```
MOLINFO STRUCTURE=template_AA.pdb


SAXS ...
        ATOMS=1-11878
        ONEBEAD
        LABEL=saxsdata
        #GPU            #This flag requires Arrayfire
        TEMPLATE=template_AA.pdb
        SOLVDENS=0.334
        SOLVATION_CORRECTION=0.00

QVALUE1=0.0000000001
QVALUE2=0.0025
QVALUE3=0.0050
QVALUE4=0.0075
QVALUE5=0.0100
QVALUE6=0.0125
QVALUE7=0.0150
```

```
QVALUE8=0.0175

QVALUE9=0.0200

QVALUE10=0.0225

QVALUE11=0.0250

QVALUE12=0.0275

QVALUE13=0.0300

QVALUE14=0.0325

QVALUE15=0.0350

QVALUE16=0.0375

QVALUE17=0.0400

QVALUE18=0.0425

QVALUE19=0.0450

QVALUE20=0.0475

QVALUE21=0.0500

QVALUE22=0.0525

QVALUE23=0.0550

QVALUE24=0.0575

QVALUE25=0.0600

QVALUE26=0.0625

QVALUE27=0.0650

QVALUE28=0.0675

QVALUE29=0.0700

QVALUE30=0.0725

QVALUE31=0.0750

QVALUE32=0.0775

QVALUE33=0.0800

QVALUE34=0.0825

QVALUE35=0.0850

QVALUE36=0.0875

QVALUE37=0.0900

QVALUE38=0.0925
```

```
QVALUE39=0.0950

QVALUE40=0.0975

QVALUE41=0.1000

QVALUE42=0.1025

QVALUE43=0.1050

QVALUE44=0.1075

QVALUE45=0.1100

QVALUE46=0.1125

QVALUE47=0.1150

QVALUE48=0.1175

QVALUE49=0.1200

QVALUE50=0.1225

QVALUE51=0.1250

QVALUE52=0.1275

QVALUE53=0.1300

QVALUE54=0.1325

QVALUE55=0.1350

QVALUE56=0.1375

QVALUE57=0.1400

QVALUE58=0.1425

QVALUE59=0.1450

QVALUE60=0.1475

QVALUE61=0.1500

QVALUE62=0.1525

QVALUE63=0.1550

QVALUE64=0.1575

QVALUE65=0.1600

QVALUE66=0.1625

QVALUE67=0.1650

QVALUE68=0.1675

QVALUE69=0.1700
```

```
QVALUE70=0.1725
QVALUE71=0.1750
QVALUE72=0.1775
QVALUE73=0.1800
QVALUE74=0.1825
QVALUE75=0.1850
QVALUE76=0.1875
QVALUE77=0.1900
QVALUE78=0.1925
QVALUE79=0.1950
QVALUE80=0.1975
QVALUE81=0.2000
QVALUE82=0.2025
QVALUE83=0.2050
QVALUE84=0.2075
QVALUE85=0.2100
QVALUE86=0.2125
QVALUE87=0.2150
QVALUE88=0.2175
QVALUE89=0.2200
QVALUE90=0.2225
QVALUE91=0.2250
QVALUE92=0.2275
QVALUE93=0.2300
QVALUE94=0.2325
QVALUE95=0.2350
QVALUE96=0.2375
QVALUE97=0.2400
QVALUE98=0.2425
QVALUE99=0.2450
QVALUE100=0.2475
```

```
QVALUE101=0.2500
QVALUE102=0.2525
QVALUE103=0.2550
QVALUE104=0.2575
QVALUE105=0.2600
QVALUE106=0.2625
QVALUE107=0.2650
QVALUE108=0.2675
QVALUE109=0.2700
QVALUE110=0.2725
QVALUE111=0.2750
QVALUE112=0.2775
QVALUE113=0.2800
QVALUE114=0.2825
QVALUE115=0.2850
QVALUE116=0.2875
QVALUE117=0.2900
QVALUE118=0.2925
QVALUE119=0.2950
QVALUE120=0.2975
QVALUE121=0.3000

... SAXS

PRINT ARG={(saxsdata\.q-.*)} STRIDE=1 FILE=SAXSINT
```

As an additional example, we provide a SANS action with a different setup. The solvent layer contribution is set to 80 electrons/nm^3. Since we want to calculate the SANS intensity, there will be an automatic conversion from electron density to water molecule density. Unlike the previous example, the solvation correction is now activated. We specified the solvent exposed area threshold for a bead to 1 nm^2 using the flag SASA_CUTOFF.

The bead exposure is evaluated by the LCPO algorithm, and since calculating it for each step of the simulation could be computationally expensive, it is possible to set the frequency for the solvent-accessible surface area estimation with the SOLVATION_STRIDE flag. By default, it is set to 10 steps. Finally, the fraction of deuterated solvent is specified with DEUTER_CONC. This information is also used to configure the implicit hydrogen-deuterium exchange: with a probability equal to the deuterium concentration in the solvent (60% in this case), an exposed bead is considered deuterated.

```
SANS ...
        ATOMS=1-11878
        ONEBEAD
        LABEL=saxsdata
        GPU
        TEMPLATE=template_AA.pdb
        SOLVDENS=0.334
        DEUTER_CONC=0.6
        SOLVATION_CORRECTION=0.80
        SASA_CUTOFF=1.0
        SOLVATION_STRIDE=10

QVALUE1=
QVALUE2=

... SANS
```

Before running the driver, it is possible to set the number of CPU cores to be used for the analysis as an environment variable in the terminal where the PLUMED driver will be launched. In the following example 8 cores are used:

```
export PLUMED_NUM_THREADS=8
```

Finally, run the driver to analyse a:

PDB

```
plumed driver --plumed plumed.dat --mf_pdb template_AA.pdb
```

In this case the PDB is the same that is provided to MOLINFO, TEMPLATE and analysed.

Trajectory

```
plumed driver --plumed plumed.dat --mf_xtc trj.xtc
```

Flag	Default value	Description
NOPBC	false	Ignore the periodic boundary conditions when calculating distances
SERIAL	false	Perform the calculation in serial - for debug purpose
DEVICEID	-1	Identifier of the GPU to be used
GPU	false	Calculate SAXS/SANS using ARRAYFIRE on a GPU
ABSOLUTE	false	Absolute intensity: the intensities for each q-value are not normalized for the intensity at q=0.
ATOMISTIC	false	Calculate SAXS/SANS using the atomistic model
MARTINI	false	Calculate SAXS using the Martini model
ONEBEAD	false	Calculate SAXS/SANS for a single bead model
TEMPLATE	template.pdb	A PDB file is required for ONEBEAD mapping
ATOMS		The atoms to be included in the calculation
QVALUE		Selected scattering lengths in inverse angstroms are given as QVALUE1, QVALUE2, ...
PARAMETERS		Used parameter Keywords like PARAMETERS1, PARAMETERS2. These are used to calculate the form factor for the *i*th atom/bead
PARAMETERSFILE		Read the PARAMETERS from a file
DEUTER_CONC	0.	Fraction of deuterated solvent. For SANS only.
SOLVDENS	0.334	Density of the solvent
SOLVATION_CORRECTION	0.0	Solvation layer electron density correction (ONEBEAD only)
SASA_CUTOFF	1.0	SASA value to consider a residue as exposed to the solvent (ONEBEAD only)
EXPINT		Experimental intensity for a specific q value
SOLVATION_STRIDE	10	Number of steps between every new residues solvation estimation via LCPO (ONEBEAD only)
SCALE_EXPINT	1.0	Scaling value for experimental data normalization. Cannot be used with ABSOLUTE.

Table 1. Description of flags and their default values used in the SAXS/SANS calculation.

Tutorial 2 - Generating a conformational ensemble with GMX and hySAS

The main purpose of hySAS is to allow an MD engine, independently of the force field, to generate a conformational ensemble that is in agreement with the experimental data. To achieve this, a potential is added to the system to discourage conformations that do not match the measured signal. Here, we propose two examples of coupling methods to introduce this energy penalty during the MD simulation.

Most of the SAS options have been discussed in Tutorial 1, here we present two additional flags: EXPINT and SCALE_EXPINT. For each QVALUE is required an EXPINT value, which consists in the experimental intensity measured at that specific momentum transfer. All the EXPINT values must be rescaled to the SAXS intensity at q = 0. To facilitate this operation, it is possible to automatically rescale all the experimental intensity entries with SCALE_EXPINT, which allows to provide the intensity at q = 0. In this way, each EXPINT is divided by SCALE_EXPINT.

In general, the more the q-values and the corresponding experimental intensities, the more intensive the calculation.

1. Harmonic-linear restraint

In this example, the SAXS profile is calculated using the ONEBEAD mapping, specifying a solvent electron density of 334 electrons/nm^3 and a correction of 80 electrons/nm^3 for the beads exposed to the solvent. The solvent-accessible surface calculation, which is required to assess the beads' exposure to the solvent, is performed every 10 steps of the simulation. The theoretical intensity at q = 0 is defined by SCALE_EXPINT, while the experimentally determined intensity for a given QVALUE is reported by EXPINT. All the EXPINT values are rescaled by the SCALE_EXPINT value.

The STATS action is then used to calculate statistical properties between the intensities computed from the coordinates of the simulated system for each QVALUE and the corresponding experimental references defined with EXPINT. Specifically, the correlation between the *in silico* intensities and the experimental intensities is considered as an argument for RESTRAINT, an action from the BIAS module. As additional information useful to follow the evolution of the simulation, the radius of gyration of the system is printed every 100 steps. Additional {} brackets have been added to prevent a readability issue, they are optional to run the code. When curly brackets are used, ensure that the entire list of arguments is enclosed within one set of curly brackets.

Detailed information on STATS, RESTRAINT and GYRATION can be found in the PLUMED manual.

```
MOLINFO STRUCTURE=template_AA.pdb


SAXS ...
        ATOMS=1-11878
        ONEBEAD
        TEMPLATE=template_AA.pdb
        #GPU            #This flag requires Arrayfire
        SOLVDENS=0.334
        SOLVATION_CORRECTION=0.080
        SOLVATION_STRIDE=10
        SASA_CUTOFF=0.7
        SCALE_EXPINT=0.281543E+00


QVALUE1=0.00444189      EXPINT1=0.279832E+00
QVALUE2=0.0133257       EXPINT2=0.266507E+00
QVALUE3=0.0222823       EXPINT3=0.241456E+00
QVALUE4=0.0312965       EXPINT4=0.207871E+00
QVALUE5=0.0402661       EXPINT5=0.170297E+00
QVALUE6=0.0534267       EXPINT6=0.116142E+00
QVALUE7=0.0786047       EXPINT7=0.427377E-01
QVALUE8=0.103746        EXPINT8=0.143592E-01
QVALUE9=0.128888        EXPINT9=0.775681E-02
QVALUE10=0.153991       EXPINT10=0.564444E-02
QVALUE11=0.179113       EXPINT11=0.395988E-02
QVALUE12=0.20428        EXPINT12=0.324523E-02
QVALUE13=0.229423       EXPINT13=0.321736E-02
... SAXS


#### INFO  ####
```

```
st_saxs: STATS ARG={(SAXS\.q-.*)} PARARG={(SAXS\.exp-.*)}
rg_saxs: GYRATION TYPE=RADIUS ATOMS=1-11878

#### RESTRAINTS ####
RESTRAINT ARG=st_saxs.corr AT=1 KAPPA=0 SLOPE=-10000

#### OUT ####
PRINT ARG=rg_saxs STRIDE=100 FILE=GYRATION
PRINT ARG={st_saxs.*,(SAXS\.q-.*),
(SAXS\.exp-.*)} STRIDE=100 FILE=STAT_SAXS
PRINT ARG={(SAXS\.q-.*)} STRIDE=100 FILE=SAXSINT
PRINT ARG=st_saxs.corr STRIDE=100 FILE=CORRELATION
```

Metainference

Changing the coupling method does not affect the main SAXS/SANS actions, which remain unchanged. In this example we use metainference, a Bayesian replica-averaging framework that restrain the average predicted data, in this case the SAXS signal calculated from the system coordinates, to be close to the experimental observable, generating an ensemble in accordance to the maximum entropy principle.

Unlike the previous example, we do not use the SCALE_EXPINT flag because the EXPINT values have already been rescaled. In both the examples the maximum momentum transfer accepted for the ONEBEAD mapping is 0.3 Å^{-1}.

```
MOLINFO STRUCTURE=template_AA.pdb

SAXS ...

        LABEL=saxsdata
        ATOMS=1-11878
        ONEBEAD
        TEMPLATE=template_AA.pdb
```

```
#GPU            #This flag requires Arrayfire
SOLVDENS=0.334
SOLVATION_CORRECTION=0.080
SOLVATION_STRIDE=10
SASA_CUTOFF=0.7

#QVALUE RANGE 0.01-0.25 (stride: 0.0150014)

QVALUE1=0.0101007       EXPINT1=0.9655268182607186
QVALUE2=0.0251021       EXPINT2=0.8074336052768617
QVALUE3=0.0401035       EXPINT3=0.5867844124283979
QVALUE4=0.0551049       EXPINT4=0.3778068043742405
QVALUE5=0.0701063       EXPINT5=0.2199053983683388
QVALUE6=0.0851077       EXPINT6=0.1176236764450616
QVALUE7=0.1001091       EXPINT7=0.0598961117861482
QVALUE8=0.1151105       EXPINT8=0.0320698663426488
QVALUE9=0.1301119       EXPINT9=0.0197961291442457
QVALUE10=0.1451133      EXPINT10=0.0131910258635653
QVALUE11=0.1601147      EXPINT11=0.0090075507724353
QVALUE12=0.1751161      EXPINT12=0.0070054764797778
QVALUE13=0.1901175      EXPINT13=0.0060413990626627
QVALUE14=0.2051189      EXPINT14=0.0049136955389689
QVALUE15=0.2201203      EXPINT15=0.0039878840479083
QVALUE16=0.2351217      EXPINT16=0.0037107186252386
QVALUE17=0.2501231      EXPINT17=0.0033106578718972

OPTSIGMAMEAN=SEM_MAX
SIGMA_MAX_STEPS=500000
AVERAGING=500
DOSCORE SIGMA_MEAN0=0.5
SCALEDATA SCALE0=1.0
```

```
        DSCALE=0.01 SCALE_PRIOR=GAUSSIAN
        SIGMA0=0.5 SIGMA_MIN=0.0001 SIGMA_MAX=0.5
        NOISETYPE=MGAUSS
... SAXS

#### METAINFERENCE ####
saxsbias: BIASVALUE ARG={(saxsdata\.score)} STRIDE=2
ens: ENSEMBLE ARG={(saxsdata\.q-.*)}

#### STATISTICS ####
statcg: STATS ARG={(ens.*)} PARARG={(saxsdata\.exp-.*)}

#### PRINT ####
PRINT ARG={(ens.*)} STRIDE=1000 FILE=ENS.SAXSINT
PRINT ARG={(saxsdata\.score),(saxsdata\.scale),
(saxsdata\.acceptSigma),
(saxsdata\.sigma.*)} STRIDE=1000 FILE=BAYES.SAXS
PRINT ARG={(saxsdata\.q-.*)} STRIDE=1000 FILE=SAXSINT
PRINT ARG=statcg.* STRIDE=1000 FILE=ST.SAXSCG
```

3.2 Calcium-Driven *In Silico* Inactivation of a Human Olfactory Receptor

The conformational changes and molecular determinants involved in the activation and inactivation of olfactory receptors remain poorly understood, largely due to the inherent challenges in determining the structure of this GPCR family.[265] In this study, we perform the first *in silico* inactivation of the human olfactory receptor OR51E2, revealing a potential role for calcium ion in receptor state transitions. Using molecular dynamics simulations, we show that the presence of a divalent ion at the ion binding site, coordinated by two conserved acidic residues, stabilises the receptor in its inactive state. Conversely, protonation of these residues alone does not induce inactivation within the microsecond timescale of our simulations. These results propose a novel molecular mechanism for olfactory receptor inactivation, provide insights for experimental validation, and suggest a broader role for divalent ions in GPCR signalling.

3.2.1 Personal Contribution

I was involved in various aspects of this work, including:

1. **Preparation of Structural Models.** Special attention was paid to the assignment of protonation states, in particular for residues D69 and E110, belonging to the ion binding site, under different conditions (no ions, Na^+, Ca^{2+}), using the Schrödinger Maestro suite.[266] The preparation also involved using the CHARMM-GUI online platform[267,268] to construct the 3:1 POPC:cholesterol membrane in which the receptor is embedded.

2. **Molecular Dynamics Simulation Setup.** The simulations were configured using GROMACS[269] and performed on high-performance computing (HPC) systems, totaling 81 µs of simulation time. This included:

   - 5 µs × 5 replicates for each of the three charged models (with no ions in the ion binding pocket, with Na^+ bound, and with Ca^{2+} bound);
   - 1 µs × 6 replicates for the neutral model with no ions in the ion binding pocket.

3. **Molecular Dynamics Trajectory Analysis.** I have developed custom Python3 scripts[270] to analyse molecular dynamics trajectories, with a focus on hydrogen bond analysis. These scripts allow the identification of unique interactions by comparing a selected trajectory with the others, filtering for hydrogen bonds that meet a specified occupancy threshold.

4. **Manuscript Writing.** I have contributed to every part of the manuscript.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Letter

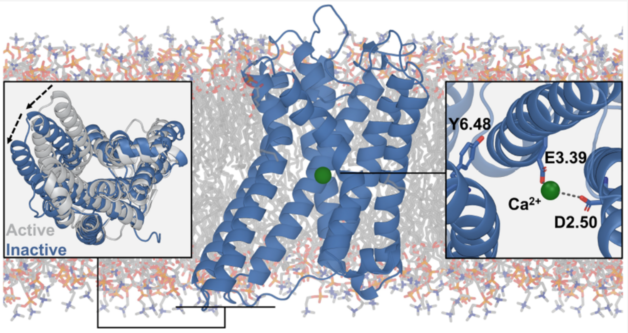
Calcium-Driven In Silico Inactivation of a Human Olfactory Receptor

Lorenza Pirona,[§] Federico Ballabio,[§] Mercedes Alfonso-Prieto, and Riccardo Capelli*

Cite This: *J. Chem. Inf. Model.* 2024, 64, 2971−2978 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Conformational changes as well as molecular determinants related to the activation and inactivation of olfactory receptors are still poorly understood due to the intrinsic difficulties in the structural determination of this GPCR family. Here, we perform, for the first time, the *in silico* inactivation of human olfactory receptor OR51E2, highlighting the possible role of calcium in this receptor state transition. Using molecular dynamics simulations, we show that a divalent ion in the ion binding site, coordinated by two acidic residues at positions 2.50 and 3.39 conserved across most ORs, stabilizes the receptor in its inactive state. In contrast, protonation of the same two acidic residues is not sufficient to drive inactivation within the microsecond timescale of our simulations. Our findings suggest a novel molecular mechanism for OR inactivation, potentially guiding experimental validation and offering insights into the possible broader role of divalent ions in GPCR signaling.



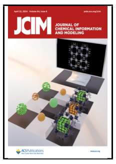
Olfactory receptors (ORs) are a family of G protein-coupled receptors (GPCRs) implicated in odor perception.[1] This group of class A GPCRs comprises approximately 400 members, representing half of the GPCRs encoded in the human genome.[2] The function of ORs is not limited to olfaction, as they have been identified in various extranasal tissues and have been shown to play a role in a plethora of physiological and pathological processes.[3,4] This important clinical perspective, combined with the fact that non-olfactory GPCRs are the target of 34% of the total number of FDA-approved drugs,[5] makes the study of the OR family one of the most promising fields for biochemistry and pharmacology.

Despite their potential as drug targets,[6] the study of ORs has been hindered by the lack of structural data: the first experimental structure was published in March 2023,[7] consisting of human OR51E2 in its active state, bound to an agonist (propionate) and a mini-G protein. In light of this limitation, the modeling community has approached the OR family with different techniques, from homology modeling[8−11] to *de novo* structure prediction.[12,13] Such computational models have provided useful insights into the mechanism of OR-ligand recognition at the atomistic level as well as pinpointed crucial residues for OR function, in combination with experimental mutagenesis data.

Recent cryoEM work by Choi et al.[14] unveiled, for the first time, both the active and inactive conformations of a chimeric OR, OR52$_{cs}$, which was constructed using a consensus sequence strategy based on 26 members of the OR52 family. The availability of experimental structural data for both conformational states of the same OR receptor opens the opportunity for cross-validation of *in silico* models. Furthermore, this enables a computational exploration of the activation or inactivation process, aligning with the approach taken in similar computational studies on other class A GPCRs.[15,16]

In this study, we focus on the inactivation of human OR51E2, leveraging the cryoEM structure of this receptor in its active state.[7] Our primary objective is to simulate the inactivation of the receptor *in silico* over a microsecond timescale, drawing upon the pioneering work of Dror et al. on the β_2-adrenergic receptor.[15] Moreover, OR51E2, along with 90% of the OR family members, bears two acidic residues in the ion binding site[13] (2.50 and 3.39 using the Ballesteros−Weinstein class A GPCR generic numbering[17]), indicating the potential involvement of these charged residues in the ion binding mechanism and conformational dynamics of ORs. This hypothesis is supported by our previous observations of sodium-bound OR51E2 in the inactive state.[13] Accordingly, we aim at elucidating the role of the ion binding site for OR conformational plasticity, also building on previous work on sodium binding to other non-olfactory class A GPCRs.[18,19]

Received: February 12, 2024
Revised: March 20, 2024
Accepted: March 20, 2024
Published: March 25, 2024



ACS Publications

https://doi.org/10.1021/acs.jcim.4c00249

SIMULATIONS WITH NO IONS IN THE BINDING SITE

Initially, the receptor structure was built from its experimentally determined active conformation (PDB code: 8F76[7]), with the agonist and mini-G protein removed. We employed the CHARMM-GUI web server[20,21] for receptor setup and its embedding into a 3:1 POPC/cholesterol bilayer (see Methods section and Supporting Information for further details). Following system preparation, we adopted a multistep equilibration protocol, applying restraints to preserve the initial fold while relaxing the system (refer to Methods), as done in our previous work.[13] This was followed by five independent unrestrained MD simulations, each lasting 5 μs.

In all five apo OR51E2 replicas, a partial loss of the receptor fold was consistently observed (see Figure 1a), which is in line with the results obtained when starting from *de novo* structures of OR51E2.[13] The region most impacted by this partial unfolding was the interface between transmembrane helices 6 and 7 (TM6 and TM7, respectively), which widens significantly during the simulation (from 8 to 15 Å).

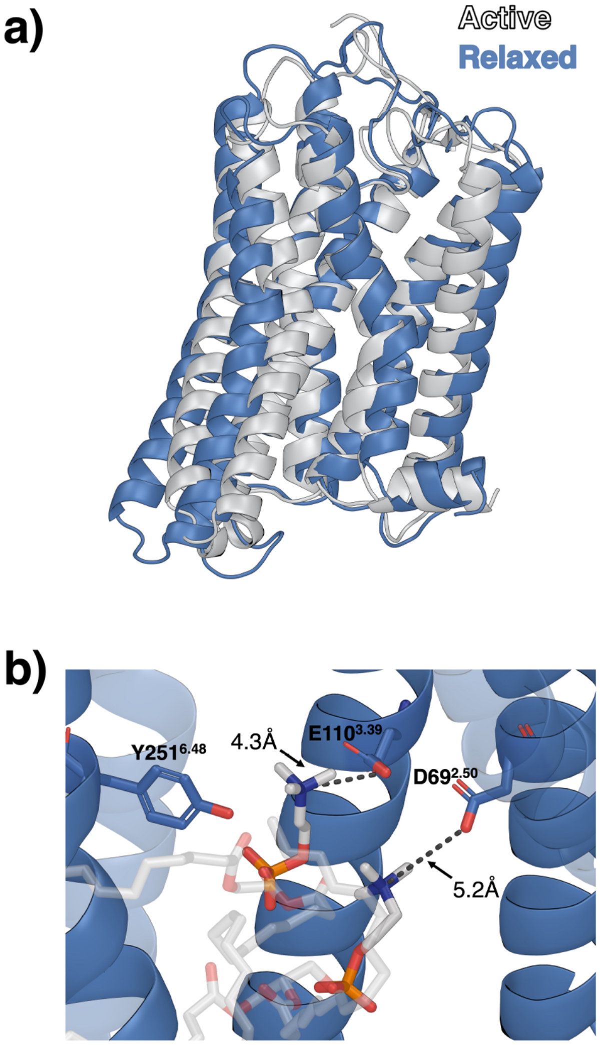


Figure 1. OR51E2 from simulations without ions in the binding pocket. (a) Superposition of the cluster centroid obtained from the five trajectories (blue) started from the cryoEM structure in its active state (white), evidencing the widening of the TM6-TM7 interface. (b) Detail of the ion binding pocket, highlighting the interaction between the charged residues D69$^{2.50}$ and E110$^{3.39}$ and the two POPC headgroups entering the receptor bundle. Additionally, Y251$^{6.48}$ is oriented toward the ion binding pocket. Hydrogen atoms have been omitted for clarity, and the distances refer to the frame shown.

Additionally, the expansion of the TM6-TM7 interface coincides with POPC molecules “diving” into the ion binding pocket, as their positively charged choline groups orient toward the charged residues D69$^{2.50}$ and E110$^{3.39}$ (see Figure 1b). These events clearly indicate a significant charge imbalance near the ion binding pocket, leading to lipid snorkeling through the apolar membrane. Within this structural rearrangement, we observed several phenomena related to the system conformational dynamics: (i) Water molecules freely traversed from the intracellular to the extracellular side of the receptor. (ii) The toggle switch Y251$^{6.48}$ established multiple transient interactions with the charged headgroups of POPC lipids (both choline and phosphate moieties), E110$^{3.39}$, and S111$^{3.40}$. Lastly, (iii) the negatively charged side chains of D69$^{2.50}$ and E110$^{3.39}$, whose charge is locally unbalanced, interacted with nearby polar and neutral amino acid side chains, and also electrostatically attracted the charged groups that entered the ion binding site (see Table S2 in the Supporting Information). Remarkably, we observed spontaneous binding of a Na$^+$ ion in two out of the five simulations (see Table S2 and Figure S1). The ion entered from either the extracellular (replica 4) or intracellular (replica 1) side of the membrane; however, it is important to note that due to the use of periodic boundary conditions (PBC), the ionic concentration is nearly identical on both sides of the membrane.

As an independent confirmation of our observation of an electrostatic interaction between the ion binding pocket and charged groups, we considered the final frame of 1 μs simulation started from the apo structure of OR52$_{cs}$ (retrieved at https://github.com/sek24/natcomm2023); such simulation was performed under conditions similar to ours (except for the use of KCl instead of NaCl). Also in the case of OR52$_{cs}$, two POPC molecules snorkel into the ion binding pocket. However, in our simulations, the two lipid molecules wedge themselves between TM6 and TM7, while in the simulations of Choi et al. they do so between TM5 and TM6. In addition, the final OR52$_{cs}$ frame shows a positively charged ion (K$^+$) located close to D69$^{2.50}$ and E110$^{3.39}$, similar to the spontaneous sodium binding event observed in two of our replica simulations.

SIMULATIONS WITH SODIUM IN THE ION BINDING SITE

After observing the spontaneous binding of Na$^+$ in the ion binding site, we carried out a new set of five replica simulations. In these simulations, Na$^+$ was manually positioned in the ion binding pocket of the initial structure while retaining the same procedure and parameters as in the previous section. The simulation length was 5 μs per replica. This approach aligns with the observations from our previous work on the *de novo* structures of the same receptor in the inactive state,[13] which maintained their fold only when sodium was present in the ion binding pocket. However, here we began with the cryoEM structure of OR51E2 in the active state[7] and conducted more and longer replicas.

Overall, the simulations involving Na$^+$ showed slightly enhanced receptor stability (see Figure 2), which is reflected

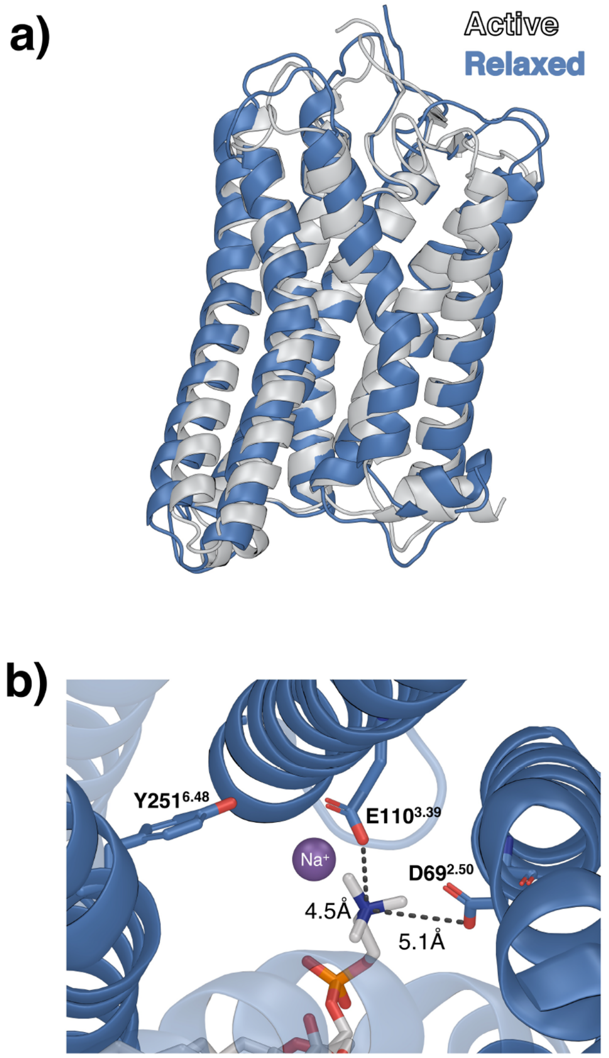


Figure 2. OR51E2 from simulations with Na$^+$ ion in the binding pocket. (a) Superposition of the cluster centroid obtained from the five trajectories (blue) on the cryoEM structure in its active state (white). (b) Detail of the ion binding pocket, highlighting the interactions between Na$^+$ and D69$^{2.50}$ and E110$^{3.39}$, as well as between POPC and the same two acidic residues. Hydrogen atoms have been omitted for clarity, and the distances refer to the frame shown.

in less POPC snorkeling into the ion binding pocket (Figure S4), reduced water permeability (Table S2), and lower RMSD values (Figure S3) of the C$_\alpha$ atoms of the transmembrane domains with respect to the cryoEM structure. Differently from the simulations with no ions, two trajectories (replicas 1 and 3) did not reveal any POPC snorkeling. In the other three trajectories, we observed the POPC headgroup moving toward the ion binding site, passing either through the interface between TM5 and TM6 and interacting mainly with E110$^{3.39}$ (replica 5 and the last 50 ns of replica 2) or between TM6 and TM7, engaging residues D69$^{2.50}$ and E110$^{3.39}$ (replicas 4 and 5). Interestingly, Na$^+$ unbinding was detected in replica 5 in the last microsecond, where the ion moved to the intracellular space. Shortly after this event, a second POPC entered the ion binding site. Finally, a new Na$^+$ ion spontaneously translocated from the extracellular space to the ion binding site within 1 μs from the initial dissociation event (Figure S3 and Table S2). This phenomenon supports, once again, the need for a charged group at this location, counteracting the negative charge of the two acidic residues at positions 2.50 and 3.39.

In terms of water permeability, we observed reduced water passing through the protein compared with the simulations with no ions (Table S2). This trend was evident in the first three replicas, particularly in the second trajectory. Considering that the formation of a hydrated pathway spanning the receptor is connected to the reorganization of the H-bonds around the ion binding site,[19] we focused on replica 2 to identify changes in nonbonded interactions at the atomistic level. Namely, we performed a comparative H-bond analysis across all 10 replicas described thus far (five with Na$^+$ and five without) to identify H-bonds that, in terms of occupancy, exhibited the most significant differences between the second replica and the others. The interaction between D69$^{2.50}$ and Y291$^{7.53}$ was particularly notable. The total occupancy of this H-bond (i.e., considering both oxygen atoms of the side chain of D69$^{2.50}$) in the second replica of the Na$^+$-bound system persisted for 76% of the trajectory. In contrast, replicas 1, 3, 4, and 5 showed occupancies of 19%, 52%, 3%, and 4%, respectively. Extending this analysis to the replicas without the ion in the binding site, this occupancy reached a maximum of 6% (replica 4), while in all the others it was around 1%. Y291$^{7.53}$ has been reported as a conserved residue in class A GPCRs, implicated in the diffusion of Na$^+$ ions,[18,19] in the formation of a water pathway inside the receptor, and in the receptor activation mechanism.[22] Another interesting H-bond pattern was observed between S111$^{3.40}$ and Y251$^{6.48}$, which is significantly more frequent in the presence of sodium. This increased persistence is particularly evident in three out of five trajectories with sodium (occupancies of 41%, 78%, 92%, 78%, and 4% in replicas 1−5, respectively). In contrast, only one replica from the trajectories lacking ions exhibits a stable H-bond between these two residues (occupancy of 90% in replica 1, 0% in all the others).

To assess the stability of the receptor fold, we performed an RMSD calculation using the cryoEM structure of OR51E2 in its active state[7] as the reference. Overall, the simulations with Na$^+$ bound maintained the fold better (i.e., lower RMSD values, see Figure S3) than the simulations without ions (Figure S1), with some minor variability among them. Interestingly, peaks in the RMSD time evolution are linked to lipid snorkeling and increased water permeability in the receptor (Figure S4).

SIMULATIONS WITH CALCIUM IN THE ION BINDING SITE

From the previous simulations, we can conclude that the charge imbalance in the ion binding site is still present, even if attenuated, when Na$^+$ is bound. Hence, we decided to perform another round of five independent simulations, positioning a Ca^{2+} ion in the binding site while maintaining all of the simulation parameters previously used (see Methods and Supporting Information). From a technical point of view, to minimize the limitations of the point charge representation of divalent ions, we manually placed calcium in the ion binding site, rather than trying to simulate its diffusion from the bulk solvent. Based on sequence analyses, class A GPCRs containing two acidic residues in the ion binding pocket have been hypothesized to bind Ca^{2+};[23] however, to the best of our knowledge, this possibility has not yet been investigated in the context of ORs.

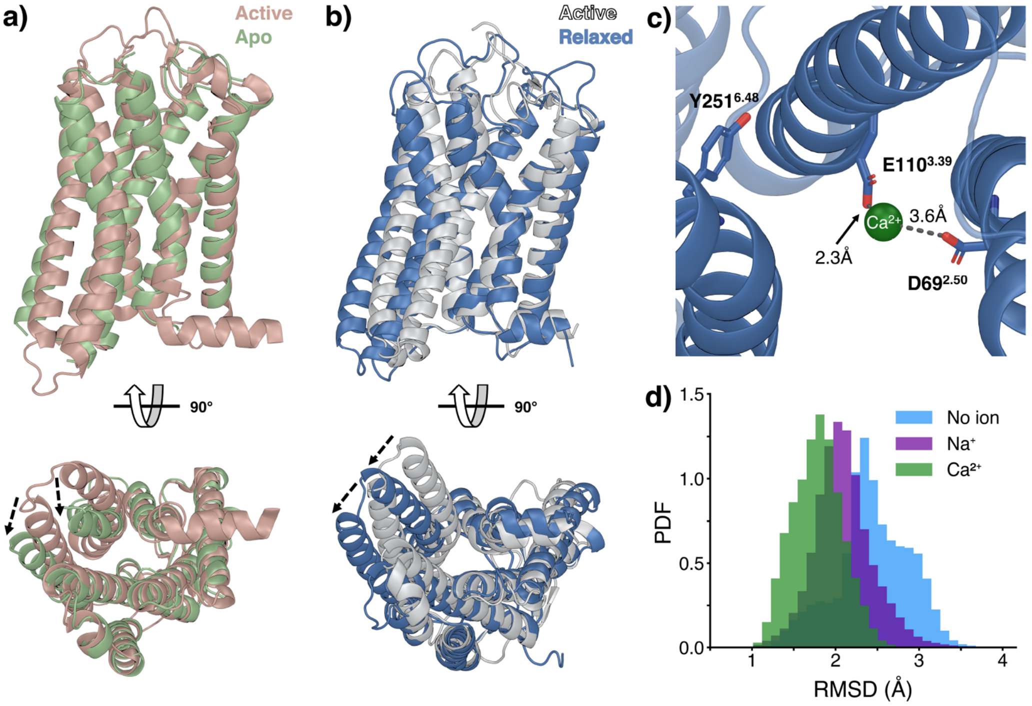


Figure 3. Results of the *in silico* Ca^{2+}-bound OR51E2 simulations compared to the available experimental structures. (a) Superposition of the active (orange) and apo (green) structures of $OR52_{cs}$ solved by Choi et al.[14] (b) Comparison of the OR51E2 experimental active structure[7] (white) and the cluster centroid obtained from the five trajectories with Ca^{2+} in the ion binding site (blue). (c) Detail of the ion binding pocket, highlighting the interactions between Ca^{2+} and the two acidic residues, $D69^{2.50}$ and $E110^{3.39}$. Hydrogen atoms have been omitted for clarity, and the distances refer to the frame shown. (d) Histogram of the RMSD of the simulations performed in these conditions: without ions (light blue), with Na^+ (purple), and with Ca^{2+} (green) in the ion binding site) with respect to TM3−TM5−TM6 C_α of the apo form of $OR52_{cs}$.

The simulations with Ca^{2+} bound further support the stabilization of the receptor fold in the presence of an ion in the binding site, as observed in some of the Na^+-bound simulations. In addition to the lower RMSD values (Figure S5), comparison of the RMSF of the three simulation data sets (Figure S7) further confirms increased stability and reduced flexibility of the receptor with calcium bound, particularly in the TM5−TM6−TM7 region. Analysis of the H-bonds across the five Ca^{2+}-bound replicas did not reveal any exclusive interactions that are absent in the trajectories with sodium. Furthermore, in contrast to the simulations with Na^+, $Y291^{7.53}$ is able to form a hydrogen bond with $D69^{2.50}$ in replicas 2 and 4, but only for about 25% of the trajectory. In replicas 1 and 5, $Y291^{7.53}$ establishes stable hydrophobic interactions with $V44^{1.53}$, $L62^{2.43}$, and $I298^{8.50}$. Generally, in the simulations with Ca^{2+}, we observed that the side chain of $Y291^{7.53}$ preferentially orients toward TM1 and TM2. Moreover, the improved electrostatic complementarity between the +2 charge of calcium and the −2 charge of the ion binding site appears to preclude lipid snorkeling, as we did not observe any POPC head groups reaching the ion binding site over the total accumulated 25 μs time. In particular, we noticed tight coordination of the acidic residues $D69^{2.50}$ and $E110^{3.39}$ with the Ca^{2+} ion (Figure 3c). In this context, visual inspection during the simulation reveals a reduced (almost negligible) amount of water passing between the two sides of the membrane through the receptor (Table S2).

The main global dynamic effect observed during the Ca^{2+}-bound simulations is characterized by the rotational movement of TM5−TM6 around TM3 acting as a pivot (Figure 3b). Such conformational change happens early for all of the replicas, during the first 0.3−0.7 μs, and appears in agreement with the transition observed for the only experimentally determined active−inactive pair of an OR to date (Figure 3a). Specifically, we performed the calculation of RMSD for all 75 μs of the simulations, taking as reference the C_α atoms of TM3, TM5, and TM6 of the $OR52_{cs}$ structure, as labeled with the structure alignment tool in the GPCRdb.[24] We can observe that the RMSD distributions (Figure 3d) are closer to the $OR52_{cs}$ inactive state in the presence of higher-charge ions in the binding site (no ion > Na^+ > Ca^{2+}). Thus, we can consider the simulations presented in this work as the first observation of calcium-driven *in silico* inactivation of a human OR. Moreover, the TM5−TM6 rotation observed during our simulations for OR51E2, and by comparison of static structures of $OR52_{cs}$, could be considered as the hallmark of inactivation for the OR family.

■ SIMULATIONS WITH PROTONATED ACIDIC RESIDUES AT POSITIONS 2.50 AND 3.39

To discriminate if receptor inactivation was due simply to neutralization of the doubly charged ion binding pocket by Ca^{2+}, we repeated our calculations with a different, neutral protonation state of the two acidic residues $D69^{2.50}$ and $E110^{3.39}$. Operatively, we equilibrated the system with this new protonation state and ran six 1 μs-long independent replicas (that is, longer than the observed timescale for inactivation in the charged system). From the analysis of the trajectories of the neutral system, we can observe three main effects. (i) The receptor undergoes a partial active-to-intermediate transition, with a rotation of the TM5-TM6 block, but with a smaller angle with respect to the Ca^{2+}-bound charged system (see

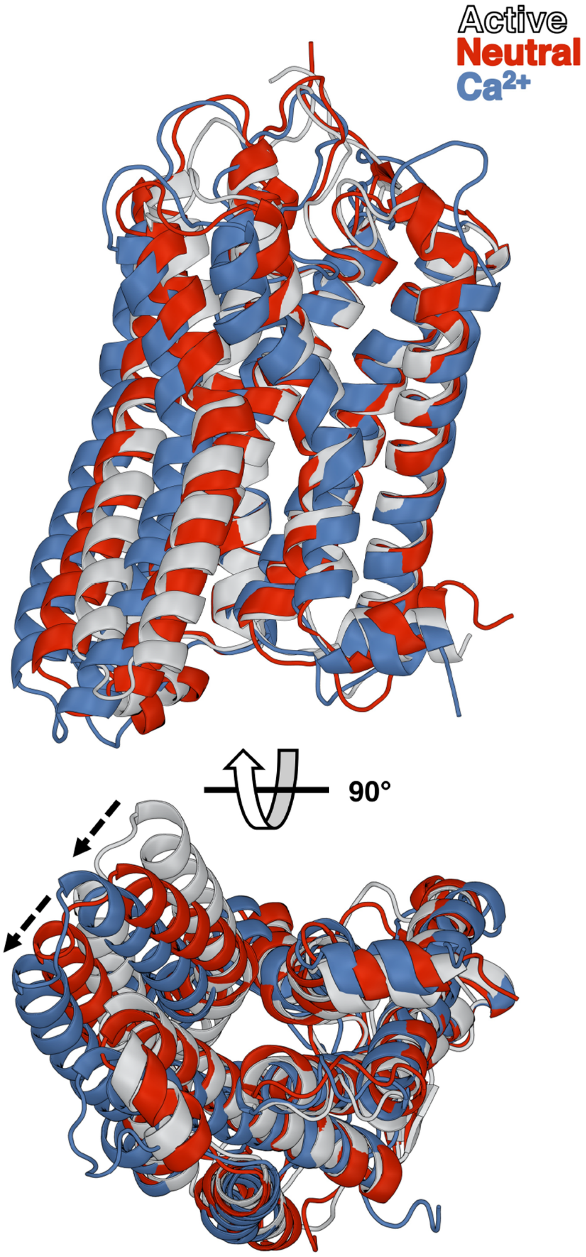


Figure 4. Results of the simulations with protonated residues at positions 2.50 and 3.39. Superposition of the cryoEM structure in its active state (white), the cluster centroid obtained from the five trajectories with Ca^{2+} bound (blue), and the cluster centroid obtained from the six trajectories with neutral $D^{2.50}$ and $E^{3.39}$ (red). The progression of TM5−TM6 rotation in the three structures is indicated with dashed arrows.

Figure 4 and Figure S8). (ii) No positively charged ions enter the ion binding site, and (iii) we do not observe any lipid snorkeling by the POPC molecules of the membrane. These last two observations were not unexpected, given the absence of charge imbalance in the ion binding pocket upon protonation of the acidic residues. The absence of ion binding is in line with a previous work on non-olfactory class A GPCRs, where the protonation of the sole acidic residue in the ion pocket facilitates the unbinding of Na^+.[19] Taken together, these data indicate that it is not only charge neutralization of the two acidic residues in the ion binding site that drives the transition to the inactive state, but also the presence of a divalent ion. This observation is within the limitations of the μs timescale of our simulations and the force field representation of divalent ions and fixed protonation states.

■ DISCUSSION

The conservation of two acidic residues in the ion binding site of ∼98% of human ORs,[13] at positions 2.50 and 3.39 (Figure 5a), raises the question of whether not only monovalent, but also divalent ions, can occupy this site and modulate receptor function. The simulations carried out in this work show that indeed calcium is more efficient than sodium at stabilizing the inactive state of a prototypical OR, OR51E2. Neutralization of the ion binding site by protonation of the two acidic residues is not sufficient to drive receptor inactivation, further supporting the strict requirement for a divalent ion bound in this pocket to stabilize the inactive receptor state.

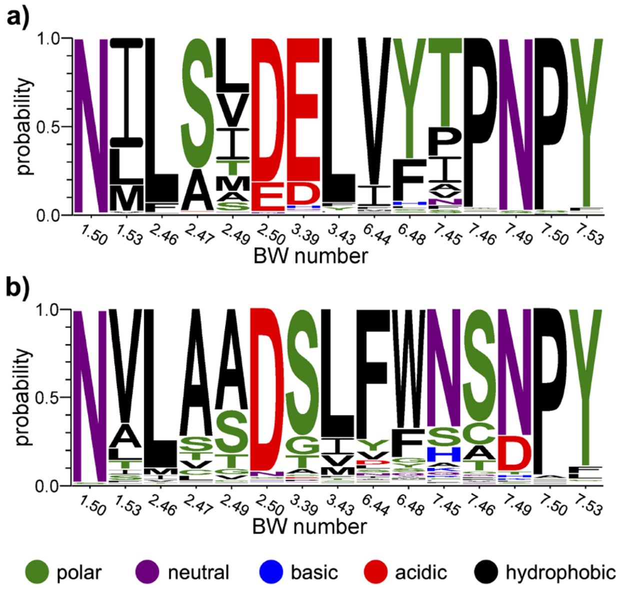


Figure 5. Sequence logos depicting conservation of residues lining the ion binding site in class A GPCRs. (a) Human olfactory receptors, based on the multiple sequence alignment of 412 sequences in Fierro et al.[36] (b) Human non-olfactory class A GPCRs, using the 286 sequences available in the GPCRdb sequence alignment tool[37] as of January 2024. Sequence logos were generated with WebLogo version 3.7.12,[38] and amino acids are colored according to their chemical properties.

The possibility of divalent ions binding to class A GPCRs containing two acidic residues in the ion binding site was already highlighted in the review by Katritch et al. as an "outstanding question".[23] Besides the conserved $D^{2.50}$, previous sequence analyses[23,25] have shown that a subset of non-olfactory class A GPCRs bear a second acidic residue in the ion binding site. Here, we extended such analyses and estimated the number to be ∼21%. The second acidic residue is typically located at position 7.49 (∼18%), but we also observed a non-negligible number of receptors where it is located at positions 6.44 (∼2%) and 3.39 (∼1%; see Figure 5b and Table S3). However, as of January 2024, only the experimental structures of melanocortin receptors[26−32] have shown Ca^{2+} bound to two aspartates at positions 2.50 and 7.49, acting as a cofactor for agonist binding. Instead, experimental structures of proteinase-activated receptors[33,34] and cysteinyl leukotriene receptors,[35] also bearing two aspartates at positions 2.50 and 7.49, have been solved with Na^+ and an antagonist bound.

We thus believe that the investigation of calcium ions effect is an extremely promising venue in the context of GPCRs,[39] in particular for ORs, which also needs an experimental counterpart. Despite the great advancements in the experimental determination of GPCR structures, the state of the art in cryoEM is still unable to discriminate between water and ions.[40] A possible help may come from the use of buffers with higher calcium concentrations, focusing our attention on possible ion−protein interactions. Moreover, while most non-olfactory class A GPCRs with a second acidic residue (see Table S3) bear an aspartate at either position 7.49 (~86%) or 6.44 (~8%), the majority of ORs (~90%) have a longer side chain acidic residue (glutamate) and at a different position (3.39).[13] This is expected to shift the ion location within the binding site, thus resulting in different effects of ion binding on the receptor function. Interestingly, two chemokine receptors (CCR1 and CCR3) and one peptide receptor (QRFP) also display glutamate at position 3.39, suggesting the possibility that calcium might also stabilize the inactive state of these receptors, similarly to what we proposed here for the ORs.

In summary, in this work we suggest a molecular mechanism for ORs inactivation, which relies on the presence of calcium in the ion binding site. In this regard, calcium-mediated OR inactivation could constitute a negative feedback mechanism to stop the olfactory signal triggered by odorant binding and subsequent calcium entry through cyclic nucleotide gated channels. The work shown here suffers from the technical limitations of classical modeling approaches in the presence of divalent ions and fixed protonation states. We envisage the possibility to use more advanced techniques such as QM/MM, which proved their efficacy in explaining classical model shortcomings in polarization effects in GPCRs,[41] exploiting the next generation exascale HPC machines. We hope that these computationally driven hypotheses could be validated by means of experimental techniques, highlighting the predictive power of molecular modeling and simulations to suggest biologically relevant structure−function relationships. Such experimental testing could include *in vitro* assays with varying concentrations of calcium in the extracellular medium or structural determination efforts with higher calcium (or divalent metal analogs) concentrations in the buffers used for X-ray, cryoEM, or native mass spectrometry.[42]

■ METHODS

System Preparation. The initial configuration of OR51E2 in its active state was obtained from the Protein Data Bank (PDB code: 8F76[7]). This structure was preprocessed using the Protein Preparation Wizard implemented in Schrödinger Maestro 2023−3,[43] which automatically assigns the amino acid protonation states. In particular, both $D69^{2.50}$ and $E110^{3.39}$ were predicted to be negatively charged at pH 7.4 and in the presence of a positively charged ion (either Na^+ or Ca^{2+}) in the ion binding site. Instead, in the absence of ions, both $D69^{2.50}$ and $E110^{3.39}$ were predicted to be protonated. We refer to these two alternative protonation states as "charged" and "neutral" forms of the receptor, respectively. The complete system was then assembled using the CHARMM-GUI[20,21] Web server. First, a disulfide bond was established between $C96^{3.25}$ and $C178^{45.50}$; then, a cubic box of dimensions 100 × 100 × 120 Å^3 was defined with the receptor embedded in a mixed lipid bilayer composed of POPC and cholesterol (3:1 ratio). The receptor-membrane system was then solvated in water with a NaCl concentration of 0.15 M. In the case of the calcium-bound simulations, we added a single Ca^{2+} ion and neutralized its charge by adding two additional Cl^- ions in the solution. The protein, lipids, and ions were parametrized using the CHARMM36m force field,[44] while water was described using the TIP3P[45] model.

Molecular Dynamics Simulations. In this work, the equilibration phase of the simulations was conducted following a protocol presented in our previous work;[13] further details are provided in the Supporting Information. For the subsequent production phase, we performed unrestrained molecular dynamics (MD), with a 2 fs time step. Simulations were 5 μs-long for the three charged systems and 1 μs-long for the neutral apo form. To maintain constant temperature and pressure conditions at 310 K and 1 bar, respectively, we employed the velocity rescaling thermostat[46] alongside the semi-isotropic cell rescaling barostat.[47] A total of 21 independent runs were executed, comprising (i) five replicates for each of the three charged models, i.e., without any ions in the ion binding pocket, with Na^+ bound, and with Ca^{2+} bound and (ii) six replicates of the neutral model without ions in the ion binding pocket. Each of these runs was initiated with distinct initial velocities. The simulations were all carried out using GROMACS,[48] version 2021.2.

■ ASSOCIATED CONTENT

Data Availability Statement

Data needed to reproduce the results shown in this paper (structures, topology, GROMACS input files, etc.) and the resulting trajectories are available at Zenodo (https://zenodo.org/doi/10.5281/zenodo.10589509). All the trajectories computed here for OR51E2 with charged $D^{2.50}$ and $E^{3.39}$ have been also uploaded on the GPCRmd,[49] with the accession codes 1976 (no ions), 1977 (Na^+-bound), and 1978 (Ca^{2+}-bound).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00249.

> Extended details of the MD simulations and tables with the Ballesteros−Weinstein numbering of OR51E2, with the main events observed in every MD replica of the charged systems and with the list of non-olfactory class A GPCRs with two acidic residues in the ion binding site, as well as figures showing the RMSD of all the replicas of the charged systems with respect to the experimental structure of OR51E2 in the active state, the minimum distance analysis between POPC lipid molecules and the ion binding site, the RMSF for all the simulations performed for the three charged systems, and RMSD histograms with respect to $OR52_{cs}$ for all systems, either charged or neutral (PDF).

■ AUTHOR INFORMATION

Corresponding Author

Riccardo Capelli − *Department of Biosciences, Università degli Studi di Milano, I-20133 Milano, Italy;* orcid.org/0000-0001-9522-3132; Email: riccardo.capelli@unimi.it

Authors

Lorenza Pirona − *Department of Biosciences, Università degli Studi di Milano, I-20133 Milano, Italy;* orcid.org/0009-0002-3716-1775

Federico Ballabio − *Department of Biosciences, Università degli Studi di Milano, I-20133 Milano, Italy;* orcid.org/0000-0001-5702-3674

Mercedes Alfonso-Prieto − *Computational Biomedicine, Institute for Neuroscience and Medicine INM-9, Forschungszentrum Jülich GmbH, D-54248 Jülich, Germany;* orcid.org/0000-0003-4509-4517

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00249

Author Contributions

[§]Equal contribution

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Prof. Carlo Camilloni and Dr. Antonella Paladino for fruitful discussion and Prof. Hee-Jung Choi for providing us, before its official deposit in the Protein Data Bank, the apo structure of $OR52_{cs}$. We also thank the anonymous reviewers for helping us in strengthening and improving the work presented here. We acknowledge EuroHPC Joint Undertaking for awarding access to the Karolina CPU cluster at IT4Innovations, Czech Republic (EHPC-REG-2022R03-049). We also acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This research was supported by Università degli Studi di Milano (Piano di supporto di ateneo Linea 2 2023-DBS Capelli) to R.C. and in part by the DFG Research Unit FOR2518 "Functional Dynamics of Ion Channels and Transporters − DynIon", project P6, to M.A.-P.

■ REFERENCES

(1) Buck, L.; Axel, R. A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* **1991**, *65*, 175−187.

(2) Alexander, S. P H; Christopoulos, A.; Davenport, A. P; Kelly, E.; Mathie, A.; Peters, J. A; Veale, E. L; Armstrong, J. F; Faccenda, E.; Harding, S. D; Pawson, A. J; Sharman, J. L; Southan, C.; Davies, J. A The concise guide to Pharmacology 2019/20: G protein-coupled receptors. *Br. J. Pharmacol.* **2019**, *176*, S21.

(3) Maßberg, D.; Hatt, H. Human olfactory receptors: novel cellular functions outside of the nose. *Physiol. Rev.* **2018**, *98*, 1739−1763.

(4) Lee, S.-J.; Depoortere, I.; Hatt, H. Therapeutic potential of ectopic olfactory and taste receptors. *Nat. Rev. Drug Discovery* **2019**, *18*, 116−138.

(5) Hauser, A. S.; Attwood, M. M.; Rask-Andersen, M.; Schiöth, H. B.; Gloriam, D. E. Trends in GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discovery* **2017**, *16*, 829−842.

(6) Naressi, R. G.; Schechtman, D.; Malnic, B. Odorant receptors as potential drug targets. *Trends Pharmacol. Sci.* **2023**, *44*, 11−14.

(7) Billesbølle, C. B.; de March, C. A.; van der Velden, W. J. C.; Ma, N.; Tewari, J.; del Torrent, C. L.; Li, L.; Faust, B.; Vaidehi, N.; Matsunami, H.; et al. Structural basis of odorant recognition by a human odorant receptor. *Nature* **2023**, *615*, 742−749.

(8) de March, C. A.; Yu, Y.; Ni, M. J.; Adipietro, K. A.; Matsunami, H.; Ma, M.; Golebiowski, J. Conserved residues control activation of mammalian G protein-coupled odorant receptors. *J. Am. Chem. Soc.* **2015**, *137*, 8611−8616.

(9) Wolf, S.; Jovancevic, N.; Gelis, L.; Pietsch, S.; Hatt, H.; Gerwert, K. Dynamical binding modes determine agonistic and antagonistic ligand effects in the prostate-specific G-protein coupled receptor (PSGR). *Sci. Rep.* **2017**, *7*, 16007.

(10) Yu, Y.; Ma, Z.; Pacalon, J.; Xu, L.; Li, W.; Belloir, C.; Topin, J.; Briand, L.; Golebiowski, J.; Cong, X. Extracellular loop 2 of G protein−coupled olfactory receptors is critical for odorant recognition. *J. Biol. Chem.* **2022**, *298*, 102331.

(11) Shim, T.; Pacalon, J.; Kim, W.-C.; Cong, X.; Topin, J.; Golebiowski, J.; Moon, C. The Third Extracellular Loop of Mammalian Odorant Receptors Is Involved in Ligand Binding. *International Journal of Molecular Sciences* **2022**, *23*, 12501.

(12) Nicoli, A.; Haag, F.; Marcinek, P.; He, R.; Kreißl, J.; Stein, J.; Marchetto, A.; Dunkel, A.; Hofmann, T.; Krautwurst, D.; et al. Modeling the orthosteric binding site of the G protein-coupled odorant receptor OR5K1. *J. Chem. Inf. Model.* **2023**, *63*, 2014−2029.

(13) Alfonso-Prieto, M.; Capelli, R. Machine Learning-Based Modeling of Olfactory Receptors in Their Inactive State: Human OR51E2 as a Case Study. *J. Chem. Inf. Model.* **2023**, *63*, 2911−2917.

(14) Choi, C.; Bae, J.; Kim, S.; Lee, S.; Kang, H.; Kim, J.; Bang, I.; Kim, K.; Huh, W.-K.; Seok, C.; Park, H.; Im, W.; Choi, H.-J. Understanding the molecular mechanisms of odorant binding and activation of the human OR52 family. *Nat. Commun.* **2023**, *14*, 8105.

(15) Dror, R. O.; Arlow, D. H.; Maragakis, P.; Mildorf, T. J.; Pan, A. C.; Xu, H.; Borhani, D. W.; Shaw, D. E. Activation mechanism of the β_2-adrenergic receptor. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 18684−18689.

(16) Dong, S. S.; Goddard, W. A.; Abrol, R. Conformational and Thermodynamic Landscape of GPCR Activation from Theory and Computation. *Biophys. J.* **2016**, *110*, 2618−2629.

(17) Ballesteros, J. A.; Weinstein, H. *Methods in Neurosciences*; Elsevier, 1995; Vol. *25*; pp 366−428.

(18) Miao, Y.; Caliman, A. D.; McCammon, J. A. Allosteric effects of sodium ion binding on activation of the M3 muscarinic G-protein-coupled receptor. *Biophys. J.* **2015**, *108*, 1796−1806.

(19) Vickery, O. N.; Carvalheda, C. A.; Zaidi, S. A.; Pisliakov, A. V.; Katritch, V.; Zachariae, U. Intracellular Transfer of Na+ in an Active-State G-Protein-Coupled Receptor. *Structure* **2018**, *26*, 171−180.e2.

(20) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859−1865.

(21) Lee, J.; Cheng, X.; Swails, J. M.; Yeom, M. S.; Eastman, P. K.; Lemkul, J. A.; Wei, S.; Buckner, J.; Jeong, J. C.; Qi, Y.; et al. CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J. Chem. Theory Comput.* **2016**, *12*, 405−413.

(22) Yuan, S.; Filipek, S.; Palczewski, K.; Vogel, H. Activation of G-protein-coupled receptors correlates with the formation of a continuous internal water pathway. *Nat. Commun.* **2014**, *5*, 4733.

(23) Katritch, V.; Fenalti, G.; Abola, E. E.; Roth, B. L.; Cherezov, V.; Stevens, R. C. Allosteric sodium in class A GPCR signaling. *Trends Biochem. Sci.* **2014**, *39*, 233−244.

(24) Pándy-Szekeres, G.; Caroli, J.; Mamyrbekov, A.; Kermani, A. A.; Keserű, G. M.; Kooistra, A. J.; Gloriam, D. E. GPCRdb in 2023: state-specific structure models using AlphaFold2 and new ligand resources. *Nucleic Acids Res.* **2023**, *51*, D395−D402.

(25) Zarzycka, B.; Zaidi, S. A.; Roth, B. L.; Katritch, V. Harnessing ion-binding sites for GPCR pharmacology. *Pharmacol. Rev.* **2019**, *71*, 571−595.

(26) Yu, J.; Gimenez, L. E.; Hernandez, C. C.; Wu, Y.; Wein, A. H.; Han, G. W.; McClary, K.; Mittal, S. R.; Burdsall, K.; Stauch, B.; et al. Determination of the melanocortin-4 receptor structure identifies Ca^{2+} as a cofactor for ligand binding. *Science* **2020**, *368*, 428−433.

(27) Heyder, N. A.; Kleinau, G.; Speck, D.; Schmidt, A.; Paisdzior, S.; Szczepek, M.; Bauer, B.; Koch, A.; Gallandi, M.; Kwiatkowski, D.; et al. Structures of active melanocortin-4 receptor−Gs-protein complexes with NDP-α-MSH and setmelanotide. *Cell Research* **2021**, *31*, 1176−1189.

(28) Israeli, H.; Degtjarik, O.; Fierro, F.; Chunilal, V.; Gill, A. K.; Roth, N. J.; Botta, J.; Prabahar, V.; Peleg, Y.; Chan, L. F.; et al.

Structure reveals the activation mechanism of the MC4 receptor to initiate satiation signaling. *Science* **2021**, *372*, 808−814.

(29) Ma, S.; Chen, Y.; Dai, A.; Yin, W.; Guo, J.; Yang, D.; Zhou, F.; Jiang, Y.; Wang, M.-W.; Xu, H. E. Structural mechanism of calcium-mediated hormone recognition and Gβ interaction by the human melanocortin-1 receptor. *Cell Research* **2021**, *31*, 1061−1071.

(30) Zhang, H.; Chen, L.-N.; Yang, D.; Mao, C.; Shen, Q.; Feng, W.; Shen, D.-D.; Dai, A.; Xie, S.; Zhou, Y.; et al. Structural insights into ligand recognition and activation of the melanocortin-4 receptor. *Cell Research* **2021**, *31*, 1163−1175.

(31) Feng, W.; Zhou, Q.; Chen, X.; Dai, A.; Cai, X.; Liu, X.; Zhao, F.; Chen, Y.; Ye, C.; Xu, Y.; Cong, Z.; Li, H.; Lin, S.; Yang, D.; Wang, M.-W. Structural insights into ligand recognition and subtype selectivity of the human melanocortin-3 and melanocortin-5 receptors. *Cell Discovery* **2023**, *9*, 81.

(32) Luo, P.; Feng, W.; Ma, S.; Dai, A.; Wu, K.; Chen, X.; Yuan, Q.; Cai, X.; Yang, D.; Wang, M.-W.; et al. Structural basis of signaling regulation of the human melanocortin-2 receptor by MRAP1. *Cell Research* **2023**, *33*, 46−54.

(33) Zhang, C.; Srinivasan, Y.; Arlow, D. H.; Fung, J. J.; Palmer, D.; Zheng, Y.; Green, H. F.; Pandey, A.; Dror, R. O.; Shaw, D. E.; et al. High-resolution crystal structure of human protease-activated receptor 1. *Nature* **2012**, *492*, 387−392.

(34) Cheng, R. K. Y.; Fiez-Vandal, C.; Schlenker, O.; Edman, K.; Aggeler, B.; Brown, D. G.; Brown, G. A.; Cooke, R. M.; Dumelin, C. E.; Doré, A. S.; et al. Structural insight into allosteric modulation of protease-activated receptor 2. *Nature* **2017**, *545*, 112−115.

(35) Luginina, A.; Gusach, A.; Marin, E.; Mishin, A.; Brouillette, R.; Popov, P.; Shiriaeva, A.; Besserer-Offroy, E.; Longpre, J.-M.; Lyapina, E.; Ishchenko, A.; Patel, N.; Polovinkin, V.; Safronova, N.; Bogorodskiy, A.; Edelweiss, E.; Hu, H.; Weierstall, U.; Liu, W.; Batyuk, A.; Gordeliy, V.; Han, G. W.; Sarret, P.; Katritch, V.; Borshchevskiy, V.; Cherezov, V. Structure-based mechanism of cysteinyl leukotriene receptor inhibition by antiasthmatic drugs. *Science Advances* **2019**, *5*, DOI: 10.1126/sciadv.aax2518.

(36) Fierro, F.; Suku, E.; Alfonso-Prieto, M.; Giorgetti, A.; Cichon, S.; Carloni, P. Agonist binding to chemosensory receptors: a systematic bioinformatics analysis. *Frontiers in Molecular Biosciences* **2017**, *4*, 63.

(37) Isberg, V.; de Graaf, C.; Bortolato, A.; Cherezov, V.; Katritch, V.; Marshall, F. H.; Mordalski, S.; Pin, J.-P.; Stevens, R. C.; Vriend, G.; et al. Generic GPCR residue numbers − aligning topology maps while minding the gaps. *Trends Pharmacol. Sci.* **2015**, *36*, 22−31.

(38) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: A Sequence Logo Generator: Figure 1. *Genome Res.* **2004**, *14*, 1188−1190.

(39) Zou, R.; Wang, X.; Li, S.; Chan, H. C. S.; Vogel, H.; Yuan, S. The role of metal ions in G protein-coupled receptor signalling and drug discovery. *WIREs Computational Molecular Science* **2022**, *12*, e1565.

(40) Pintilie, G.; Chiu, W. Validation, analysis and annotation of cryo-EM structures. *Acta Crystallographica Section D Structural Biology* **2021**, *77*, 1142−1152.

(41) Capelli, R.; Lyu, W.; Bolnykh, V.; Meloni, S.; Olsen, J. M. H.; Rothlisberger, U.; Parrinello, M.; Carloni, P. Accuracy of Molecular Simulation-Based Predictions of koff Values: A Metadynamics Study. *J. Phys. Chem. Lett.* **2020**, *11*, 6373−6381.

(42) Agasid, M. T.; Sørensen, L.; Urner, L. H.; Yan, J.; Robinson, C. V. The Effects of Sodium Ions on Ligand Binding and Conformational States of G Protein-Coupled Receptors—Insights from Mass Spectrometry. *J. Am. Chem. Soc.* **2021**, *143*, 4085−4089.

(43) *Maestro 2023−3*; Schrödinger, LLC: New York, 2023.

(44) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71−73.

(45) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(46) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(47) Bernetti, M.; Bussi, G. Pressure control using stochastic cell rescaling. *J. Chem. Phys.* **2020**, *153*, 114107.

(48) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19−25.

(49) Rodríguez-Espigares, I.; Torrens-Fontanals, M.; Tiemann, J. K. S.; Aranda-García, D.; Ramírez-Anguita, J. M.; Stepniewski, T. M.; Worp, N.; Varela-Rial, A.; Morales-Pastor, A.; Medel-Lacruz, B.; et al. GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods* **2020**, *17*, 777−787.

Supporting Information for: "Calcium-driven In Silico Inactivation of a Human Olfactory Receptor"

Lorenza Pirona,[†,¶] Federico Ballabio,[†,¶] Mercedes Alfonso-Prieto,[‡] and Riccardo Capelli[*,†]

[†]*Department of Biosciences, Università degli Studi di Milano, Via Celoria 26, I-20133 Milano, Italy*

[‡]*Computational Biomedicine, Institute for Neuroscience and Medicine INM-9, Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Straße, D-54248 Jülich, Germany*

[¶]*Equal contribution*

E-mail: riccardo.capelli@unimi.it

Preparation of the Structural Models

In the initial phase, the cryoEM structure of OR51E2 in its active state (PDB code: 8F76[1]) was prepared using the Protein Preparation Wizard in Schrödinger Maestro version 2023-3.[2] This process involved automated assignment of protonation states for amino acids based on their microenvironment. D69$^{2.50}$ and E110$^{3.39}$ were automatically identified as negatively charged in the presence of a Na^{+} or Ca^{2+} ion. In absence of any ions, both D69$^{2.50}$ and E110$^{3.39}$ were instead predicted to be protonated. This dependency of the protonation state of the ion binding site on the presence of ions has also been reported for non-olfactory class A receptors with a single acidic residue at position 2.50.[3] The thus-prepared models (neutral apo receptor and charged receptor in either apo, Na^{+}-bound and Ca^{2+}-bound forms) were then subjected to further refinement using CHARMM-GUI online platform[4,5] (https://charmm-gui.org/). This included establishing a disulfide bond between C96$^{3.25}$ and C178$^{45.50}$ and setting up a cubic simulation box of dimensions (100 Å) × (100 Å) × (120 Å). Within this box, the receptor was located at the center, embedded in a membrane comprising a 3:1 POPC:cholesterol mix. The system, including the lipid bilayer and receptor, was immersed in water with 150 mM NaCl, reflecting standard experimental conditions for GPCRs. In the replicas with the calcium ion bound, we also added two chloride ions to neutralize the extra 2+ charge. The CHARMM36m force field[6] was used to parameterize proteins, lipids, and ions, while water molecules were modeled using the TIP3P[7] model. When necessary, we manually positioned the sodium or calcium ion in the ion binding pocket of the olfactory receptor.

Molecular Dynamics Simulation Protocol

The molecular dynamics simulation protocol adopted was the same as the one presented in our previous work on inactive models of olfactory receptors.[8] In practice, it is an adaptation and extension of the standard procedure recommended by CHARMM-GUI for transmembrane proteins. It includes nine distinct steps (according to the CHARMM-GUI numbering system, from 0 to 6, plus two additional phases, 7 and 8):

- step 0: 5,000 steps of steepest descent minimization, constraining the protein backbone ($k = 4,000$ kJ/mol/nm^2) and side chains ($k = 2,000$ kJ/mol/nm^2), lipid phosphate groups ($k = 1,000$ kJ/mol/nm^2), and dihedrals ($k = 1,000$ kJ/mol/rad^2).

- Step 1: 125 ps of MD with a time step of 1 fs, restraining the protein backbone ($k = 4,000$ kJ/mol/nm^2) and side chains ($k = 2,000$ kJ/mol/nm^2), phosphate groups for POPC and hydroxyl group for cholesterol ($k = 1,000$ kJ/mol/nm^2), and dihedrals ($k = 1,000$ kJ/mol/rad^2).

- Step 2: 125 ps of MD with a time step of 1 fs, restraining the protein backbone ($k = 2,000$ kJ/mol/nm^2) and side chains ($k = 1,000$ kJ/mol/nm^2), phosphate groups for POPC and hydroxyl group for cholesterol ($k = 400$ kJ/mol/nm^2), and dihedrals ($k = 400$ kJ/mol/rad^2).

- Step 3: 125 ps of MD with a time step of 1 fs, restraining the protein backbone ($k = 1,000$ kJ/mol/nm^2) and side chains ($k = 500$ kJ/mol/nm^2), phosphate groups for POPC and hydroxyl group for cholesterol ($k = 400$ kJ/mol/nm^2) and dihedrals ($k = 200$ kJ/mol/rad^2).

- Step 4: 500 ps MD with a time step of 2 fs, restraining the protein backbone ($k = 500$ kJ/mol/nm^2) and side chains ($k = 200$ kJ/mol/nm^2), phosphate groups for POPC and hydroxyl group for cholesterol ($k = 200$ kJ/mol/nm^2) and dihedrals ($k = 200$ kJ/mol/rad^2).

- Step 5: 500 ps of MD with a time step of 2 fs, restraining the protein backbone ($k = 200$ kJ/mol/nm^2) and side chains ($k = 50$ kJ/mol/nm^2), phosphate groups for POPC and hydroxyl group for cholesterol ($k = 40$ kJ/mol/nm^2), and dihedrals ($k = 100$ kJ/mol/rad^2).

- Step 6: 100 ns MD with a time step of 2 fs, restraining the protein backbone ($k = 50$ kJ/mol/nm^2); this step is 10 times longer than the standard CHARMM-GUI protocol.

- Step 7: 100 ns of MD with a time step of 2 fs, restraining the protein backbone ($k = 5$ kJ/mol/nm^2); this is a completely new step that increases the length of the restrained equilibration.

The final production phase (step 8) consisted of MD without any restraint, 5 μs long for the charged systems and 1 μs for the neutral apo receptor. All simulations were performed with a 2 fs time step. The cutoff for van der Waals and short-range interactions was set to 10 Å , and long-range electrostatic interactions were calculated using the Ewald smooth particle mesh method.[9] Temperature control was achieved using the velocity rescale thermostat[10] at 310 K, and pressure was maintained at 1 bar using the semi-isotropic cell rescale barostat.[11] All simulations were performed with GROMACS[12] version 2021.2.

In total, we ran fifteen different production simulations: five replicas starting with no ions in the binding pocket, five starting with sodium in the ion binding pocket, and five starting with calcium in the binding pocket. Each simulation with charged D69$^{2.50}$ and E110$^{3.39}$ was 5 μs-long, while the replicas with protonated acidic residues were 1 μs-long.

Hydrogen Bonds Analysis

The 15 trajectories were analyzed using the 'Hydrogen Bonds' plugin from the Visual Molecular Dynamics[13] (VMD) suite, version 1.9.3. The analysis was configured to identify hydrogen bonds involving only polar atoms of the receptor, considering a donor-acceptor distance within 3.5 Å, and a tolerance of 30° deviation from the linear donor-hydrogen-acceptor angle. For comparison among replicas, a custom Python3[14] script was developed aimed at identifying hydrogen bonds in a selected trajectory that were not present in all the other trajectories, given a specified threshold for the hydrogen bond occupancy percentage.

Cluster Analysis

Cluster analyses were performed on the concatenated trajectories of the five replicas for each simulation condition (no ions, Na^+, and Ca^{2+}, respectively).

The clustering was performed using the gromos method[15] implemented in the GROMACS cluster tool, setting an RMSD cutoff of 4 Å. The RMSD was calculated on the C_α atoms of the transmembrane helices (TM1 to TM7), as defined in Table S1.

RMSD and RMSF analyses

All the RMSD and RMSF analyses shown here are performed on the C_α of the transmembrane helices of the OR51E2 (see Table S1). For the RMSD, we consider the cryoEM structure of the receptor in its active state[1] as the reference conformation. The RMSF calculations (Figure S8) have been performed on the concatenated trajectories for each the three simulation conditions, encompassing 25 μs of dynamics each.

Table S1: Ballesteros-Weinstein generic numbering for human OR51E2, as listed in the GPCRdb[16] (https://gpcrdb.org/residue/residuetabledisplay, accessed on December 2023).

TM1		**TM2**		**TM3**		**TM4**		**TM5**		**TM6**		**TM7**	
1x32	H23	2x37	A56	3x21	S92	4x38	N136	5x32	T191	6x26	S229	7x29	H268
1x33	F24	2x38	P57	3x22	F93	4x39	N137	5x33	L192	6x27	K230	7x30	P269
1x34	W25	2x39	M58	3x23	E94	4x40	T138	5x34	P193	6x28	S231	7x31	I270
1x35	V26	2x40	Y59	3x24	A95	4x41	V139	5x35	N194	6x29	E232	7x32	V271
1x36	G27	2x41	L60	3x25	C96	4x42	T140	5x36	V195	6x30	R233	7x33	R272
1x37	F28	2x42	F61	3x26	L97	4x43	A141	5x37	V196	6x31	A234	7x34	V273
1x38	P29	2x43	L62	3x27	T98	4x44	Q142	5x38	Y197	6x32	K235	7x35	V274
1x39	L30	2x44	C63	3x28	Q99	4x45	I143	5x39	G198	6x33	A236	7x36	M275
1x40	L31	2x45	M64	3x29	M100	4x46	G144	5x40	L199	6x34	F237	7x37	G276
1x41	S32	2x46	L65	3x30	F101	4x47	I145	5x41	T200	6x35	G238	7x38	D277
1x42	M33	2x47	A66	3x31	F102	4x48	V146	5x42	A201	6x36	T239	7x39	I278
1x43	Y34	2x48	A67	3x32	I103	4x49	A147	5x43	I202	6x37	C240	7x40	Y279
1x44	V35	2x49	I68	3x33	H104	4x50	V148	5x44	L203	6x38	V241	7x41	L280
1x45	V36	2x50	D69	3x34	A105	4x51	V149	5x45	L204	6x39	S242	7x42	L281
1x46	A37	2x51	L70	3x35	L106	4x52	R150	5x46	V205	6x40	H243	7x43	L282
1x47	M38	2x52	A71	3x36	S107	4x53	G151	5x47	M206	6x41	I244	7x45	P283
1x48	F39	2x53	L72	3x37	A108	4x54	S152	5x48	G207	6x42	G245	7x46	P284
1x49	G40	2x54	S73	3x38	I109	4x55	L153	5x49	V208	6x43	V246	7x47	V285
1x50	N41	2x55	T74	3x39	E110	4x56	F154	5x50	D209	6x44	V247	7x48	I286
1x51	C42	2x551	S75	3x40	S111	4x57	F155	5x51	V210	6x45	L248	7x49	N287
1x52	I43	2x56	T76	3x41	T112	4x58	F156	5x52	M211	6x46	A249	7x50	P288
1x53	V44	2x57	M77	3x42	I113	4x59	P157	5x53	F212	6x47	F250	7x51	I289
1x54	V45	2x58	P78	3x43	L114	4x60	L158	5x54	I213	6x48	Y251	7x52	I290
1x55	F46	2x59	K79	3x44	L115	4x61	P159	5x55	S214	6x49	V252	7x53	Y291
1x56	I47	2x60	I80	3x45	A116	4x62	L160	5x56	L215	6x50	P253	7x54	G292
1x57	V48	2x61	L81	3x46	M117	4x63	L161	5x57	S216	6x51	L254	7x55	A293
1x58	R49	2x62	A82	3x47	A118	4x64	I162	5x58	Y217	6x52	I255	7x56	K294
1x59	T50	2x63	L83	3x48	F119	4x65	K163	5x59	F218	6x53	G256		
1x60	E51	2x64	F84	3x49	D120	4x66	R164	5x60	L219	6x54	L257		
		2x65	W85	3x50	R121	4x67	L165	5x61	I220	6x55	S258		
		2x66	F86	3x51	Y122			5x62	I221	6x56	V259		
		2x67	D87	3x52	V123			5x63	R222	6x57	V260		
				3x53	A124			5x64	T223	6x58	H261		
				3x54	I125			5x65	V224	6x59	R262		
				3x55	C126			5x66	L225	6x60	F263		
				3x56	H127			5x67	Q226	6x61	G264		
								5x68	L227				
ICL1				**ICL2**		**ECL2**						**H8**	
12x48	R52			34x50	P128	45x50	C178					8x47	T295
12x49	S53			34x51	L129	45x51	V179					8x48	K296
12x50	L54			34x52	R130	45x52	H180					8x49	Q297
12x51	H55			34x53	H131							8x50	I298
				34x54	A132							8x51	R299
				34x55	A133							8x52	T300
				34x56	V134							8x53	R301
				34x57	L135							8x54	V302
												8x55	L303
												8x56	A304
												8x57	M305
												8x58	F306
												8x59	K307
												8x60	I308

Table S2: Relevant events observed in each replica.

Initial condition	Replica ID	Ion un/binding	POPC snorkeling	Role of $Y^{6.48}$	Role of $Y^{7.53}$	Water passing
No ion	1	Binding at 0.44 μs	1 POPC from 0.44 μs	H-bond with $S^{3.40}$	Interacts with $V^{1.53}$, $L^{2.63}$, $I^{8.50}$ from 0.17 to 1 μs and from 2 to 2.2 μs	Free passage
	2	No	2 POPC: 1st from 0.26 μs and 2nd from 1.49 μs	Interacts with POPC	Interacts with $V^{1.53}$, $L^{2.63}$, $I^{8.50}$ from 1.34 to 2.17 μs; then with POPC	Free passage
	3	No	1 POPC from 0.09 μs	Interacts with POPC	Interacts with POPC	Free passage
	4	Binding at 4.66 μs	1 POPC from 0.03 μs to 4.64 μs	H-bond with $E^{3.39}$ from 2.35 μs and $S^{3.40}$ from 4.49 μs	Interacts with POPC	Free passage
	5	No	1 POPC from 0.03 μs	Interacts with $E^{3.39}$, and with POPC from 0.03 μs	Interacts with POPC	Free passage
Na^{+}	1	No	No	H-bond with $S^{3.40}$ until 2.10 μs and with $E^{3.39}$ from 2.10 μs	Interacts with $M^{3.46}$, $I^{8.50}$ and from 1.7 μs with $D^{2.50}$	Free passage
	2	No	1 POPC from 4.14 μs	H-bond with $S^{3.40}$	Interacts with $D^{2.50}$	Few molecules
	3	No	No	H-bond with $S^{3.40}$	Interacts with $D^{2.50}$	Free passage
	4	No	1 POPC from 0.15 μs to 2.15 μs	H-bond with $S^{3.40}$	Points towards TM3-TM5	Free passage
	5	Unbinding at 4.2 μs, and binding at 4.95 μs	1 POPC from 0.76 μs to 3.15 μs 2 POPC: 1st from 3.83 μs and 2nd from 4.23 μs	H-bond with $S^{3.40}$ until 1.85 μs	Interacts with POPC	Free passage
Ca^{2+}	1	No	No	H-bond with $S^{3.40}$	Interacts with $V^{1.53}$, $L^{2.63}$, $I^{8.50}$ from 1.25 μs	Few molecules
	2	No	No	H-bond with $S^{3.40}$	Interacts with $D^{2.50}$ from 0.2 to 2.38 μs	Almost water-free
	3	No	No	H-bond with $S^{3.40}$	Interacts with $D^{2.50}$ from 2.9 to 3.5 μs	Almost water-free
	4	No	No	H-bond with $S^{3.40}$	Interacts with $D^{2.50}$ multiple times	Few molecules
	5	No	No	H-bond with $S^{3.40}$	Interacts with $V^{1.53}$, $L^{2.63}$, $I^{8.50}$ from 0.37 μs	Almost water-free

Table S3: Non-olfactory class A GPCRs with a second acidic residue in the ion binding site, besides the first acidic residue at position 2.50. Out of the 286 human sequences available in the sequence alignment tool of GPCRdb[17] as of January 2024, 60 bear a second acidic residue in the ion binding site, specifically, E$^{3.39}$ in three receptors, D$^{6.44}$ in five and D$^{7.49}$ in 55.

Uniprot ID	First	Second	Uniprot ID	First	Second
acthr_human	D$^{2.50}$	D$^{7.49}$	mshr_human	D$^{2.50}$	D$^{7.49}$
ccr1_human	D$^{2.50}$	E$^{3.39}$	ogr1_human	D$^{2.50}$	D$^{7.49}$
ccr3_human	D$^{2.50}$	E$^{3.39}$	oxer1_human	D$^{2.50}$	D$^{7.49}$
cltr1_human	D$^{2.50}$	D$^{7.49}$	p2ry1_human	D$^{2.50}$	D$^{7.49}$
ffar2_human	D$^{2.50}$	D$^{7.49}$	p2ry2_human	D$^{2.50}$	D$^{7.49}$
ffar3_human	D$^{2.50}$	D$^{7.49}$	p2ry4_human	D$^{2.50}$	D$^{7.49}$
fshr_human	D$^{2.50}$	D$^{6.44}$	p2ry6_human	D$^{2.50}$	D$^{7.49}$
gp132_human	E$^{2.50}$	D$^{7.49}$	p2ry8_human	D$^{2.50}$	D$^{7.49}$
gp171_human	D$^{2.50}$	D$^{7.49}$	p2y10_human	D$^{2.50}$	D$^{7.49}$
gp174_human	D$^{2.50}$	D$^{7.49}$	p2y12_human	D$^{2.50}$	D$^{7.49}$
gp183_human	D$^{2.50}$	D$^{7.49}$	p2y13_human	D$^{2.50}$	D$^{7.49}$
gpr17_human	D$^{2.50}$	D$^{7.49}$	p2y14_human	D$^{2.50}$	D$^{7.49}$
gpr18_human	D$^{2.50}$	D$^{7.49}$	par1_human	D$^{2.50}$	D$^{7.49}$
gpr20_human	D$^{2.50}$	D$^{7.49}$	par2_human	D$^{2.50}$	D$^{7.49}$
gpr34_human	D$^{2.50}$	D$^{7.49}$	par3_human	D$^{2.50}$	D$^{7.49}$
gpr35_human	D$^{2.50}$	D$^{7.49}$	par4_human	D$^{2.50}$	D$^{7.49}$
gpr42_human	D$^{2.50}$	D$^{7.49}$	pd2r_human	D$^{2.50}$	D$^{7.49}$
gpr4_human	D$^{2.50}$	D$^{7.49}$	pe2r1_human	D$^{2.50}$	D$^{7.49}$
gpr55_human	D$^{2.50}$	D$^{7.49}$	pe2r2_human	D$^{2.50}$	D$^{7.49}$
gpr87_human	D$^{2.50}$	D$^{7.49}$	pe2r3_human	D$^{2.50}$	D$^{7.49}$
hcar1_human	D$^{2.50}$	D$^{7.49}$	pe2r4_human	D$^{2.50}$	D$^{7.49}$
hcar2_human	D$^{2.50}$	D$^{7.49}$	pf2r_human	D$^{2.50}$	D$^{7.49}$
hcar3_human	D$^{2.50}$	D$^{7.49}$	pi2r_human	D$^{2.50}$	D$^{7.49}$
lpar4_human	D$^{2.50}$	D$^{7.49}$	psyr_human	D$^{2.50}$	D$^{7.49}$
lpar5_human	D$^{2.50}$	D$^{7.49}$	ptafr_human	D$^{2.50}$	D$^{7.49}$
lpar6_human	D$^{2.50}$	D$^{7.49}$	qrfpr_human	D$^{2.50}$	E$^{3.39}$
lshr_human	D$^{2.50}$	D$^{6.44}$	rxfp1_human	D$^{2.50}$	D$^{6.44}$
mc3r_human	D$^{2.50}$	D$^{7.49}$	rxfp2_human	D$^{2.50}$	D$^{6.44}$
mc4r_human	D$^{2.50}$	D$^{7.49}$	ta2r_human	D$^{2.50}$	D$^{7.49}$
mc5r_human	D$^{2.50}$	D$^{7.49}$	tshr_human	D$^{2.50}$	D$^{6.44}$

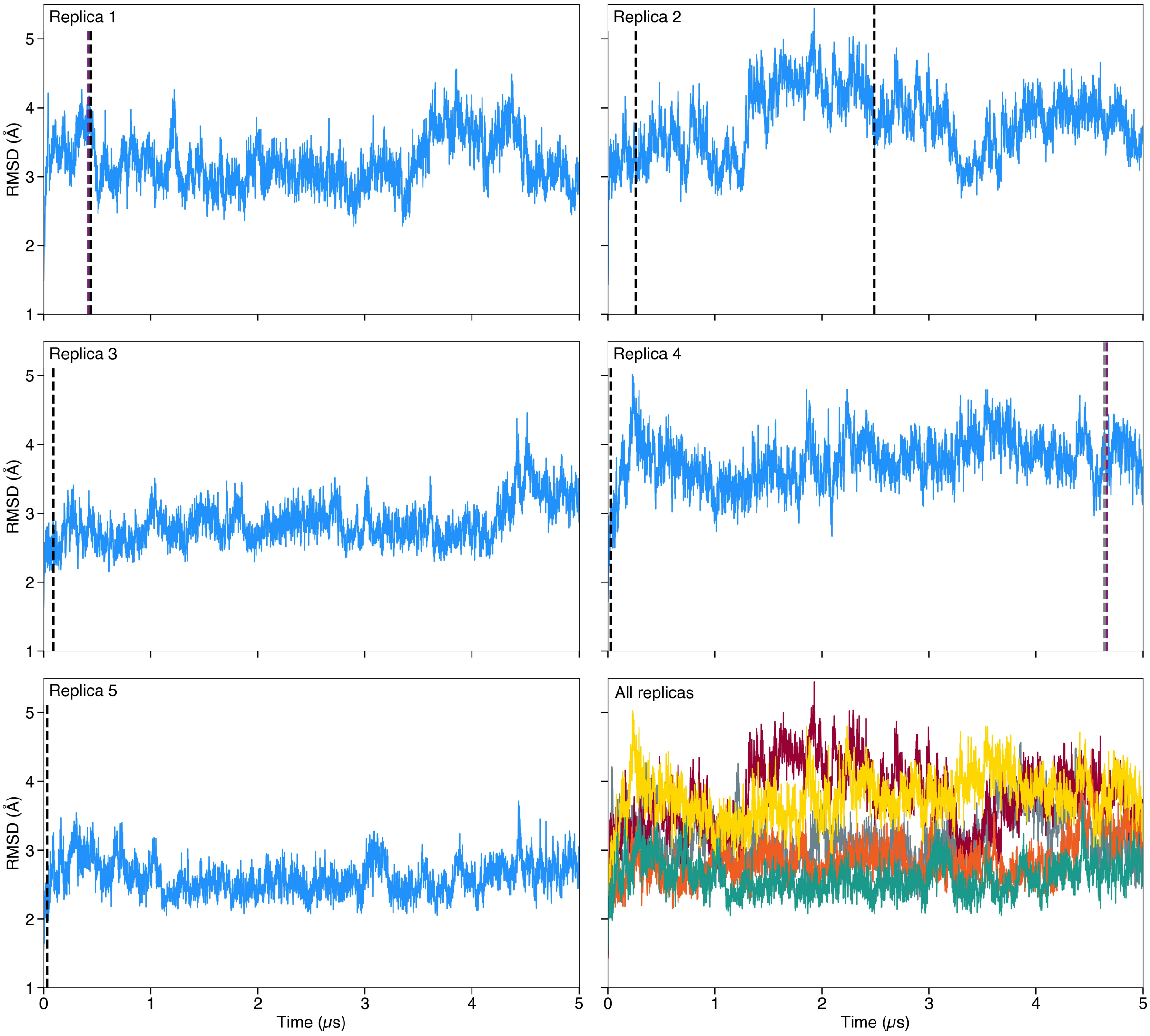


Figure S1: Time evolution of the root mean square deviation (RMSD) for the five MD simulations without ions in the ion binding site. The vertical dashed lines indicate the time at which a given event was observed. Sodium binding to the ion binding site is depicted by purple dashed lines, while POPC snorkeling into and out of the ion binding site is represented by black and grey dashed lines, respectively. The bottom right graph displays the combined data from all five replicas.

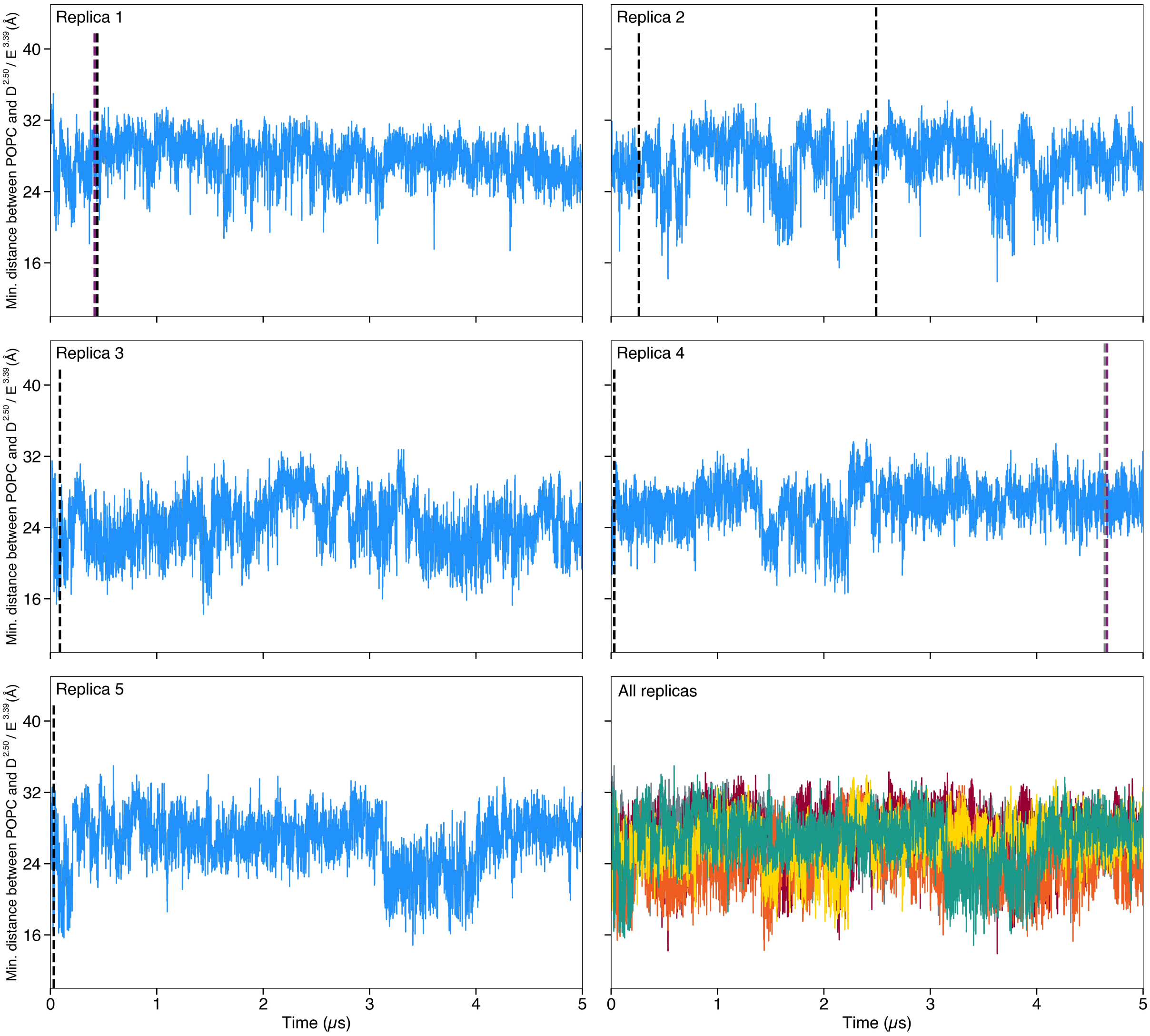


Figure S2: Time series of the sum of the minimum distance between the choline groups of POPC and residue D69$^{2.50}$ and the minimum distance between the choline groups of POPC and residue E110$^{3.39}$ in simulations without ions in the ion binding site. The vertical dashed lines indicate the time at which a given event was observed. Sodium binding to the ion binding site is depicted by purple dashed lines, while POPC snorkeling into and out of the ion binding site is represented by black and grey dashed lines, respectively. The bottom right graph displays the combined data from all five replicas.

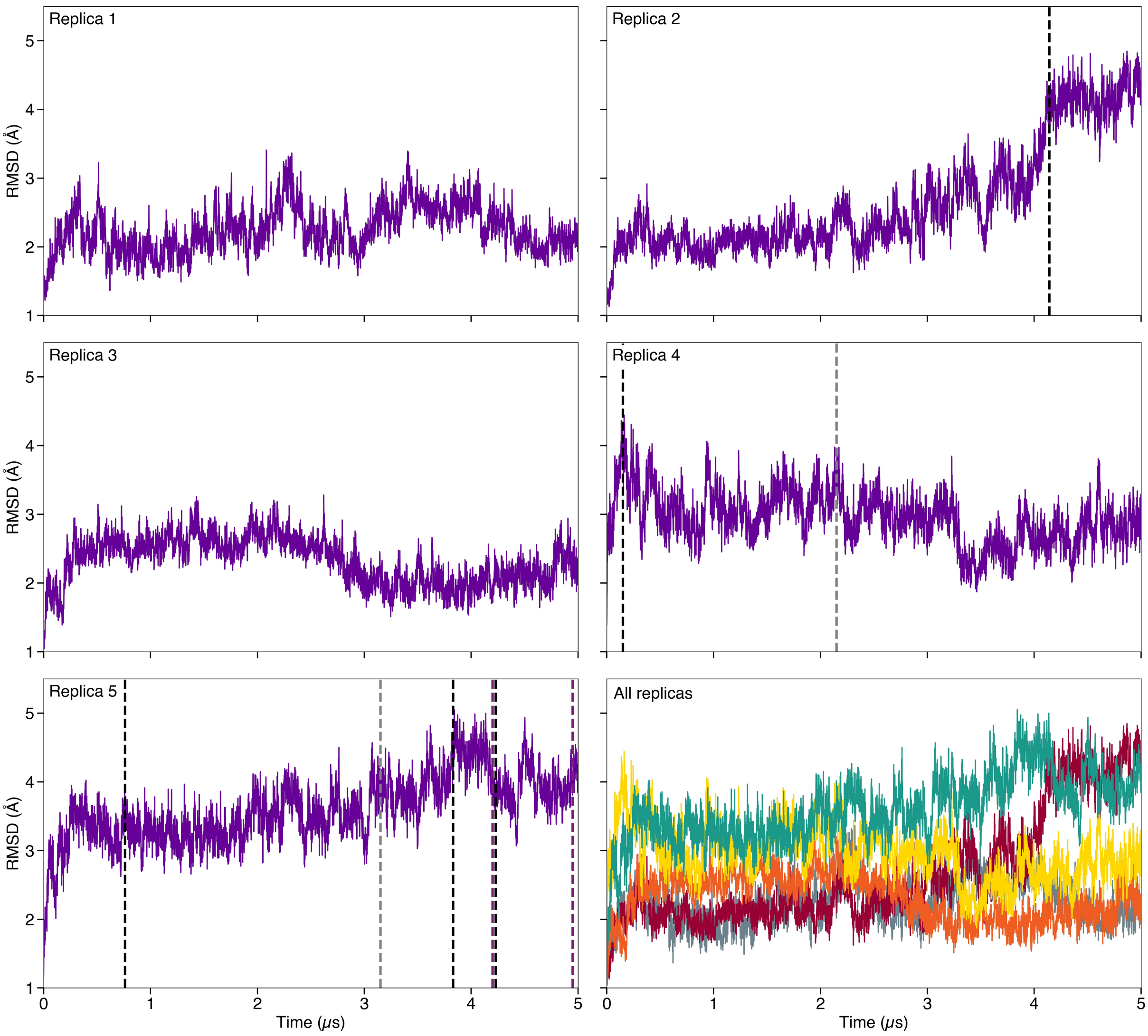


Figure S3: Time evolution of the root mean square deviation (RMSD) for the five MD simulations with Na$^+$ in the ion binding site. The vertical dashed lines indicate the time at which a given event was observed. Sodium unbinding to the ion binding site is depicted by purple dashed lines while POPC snorkeling into and out of the ion binding site is represented by black and grey dashed lines, respectively. The bottom right graph displays the combined data from all five replicas.

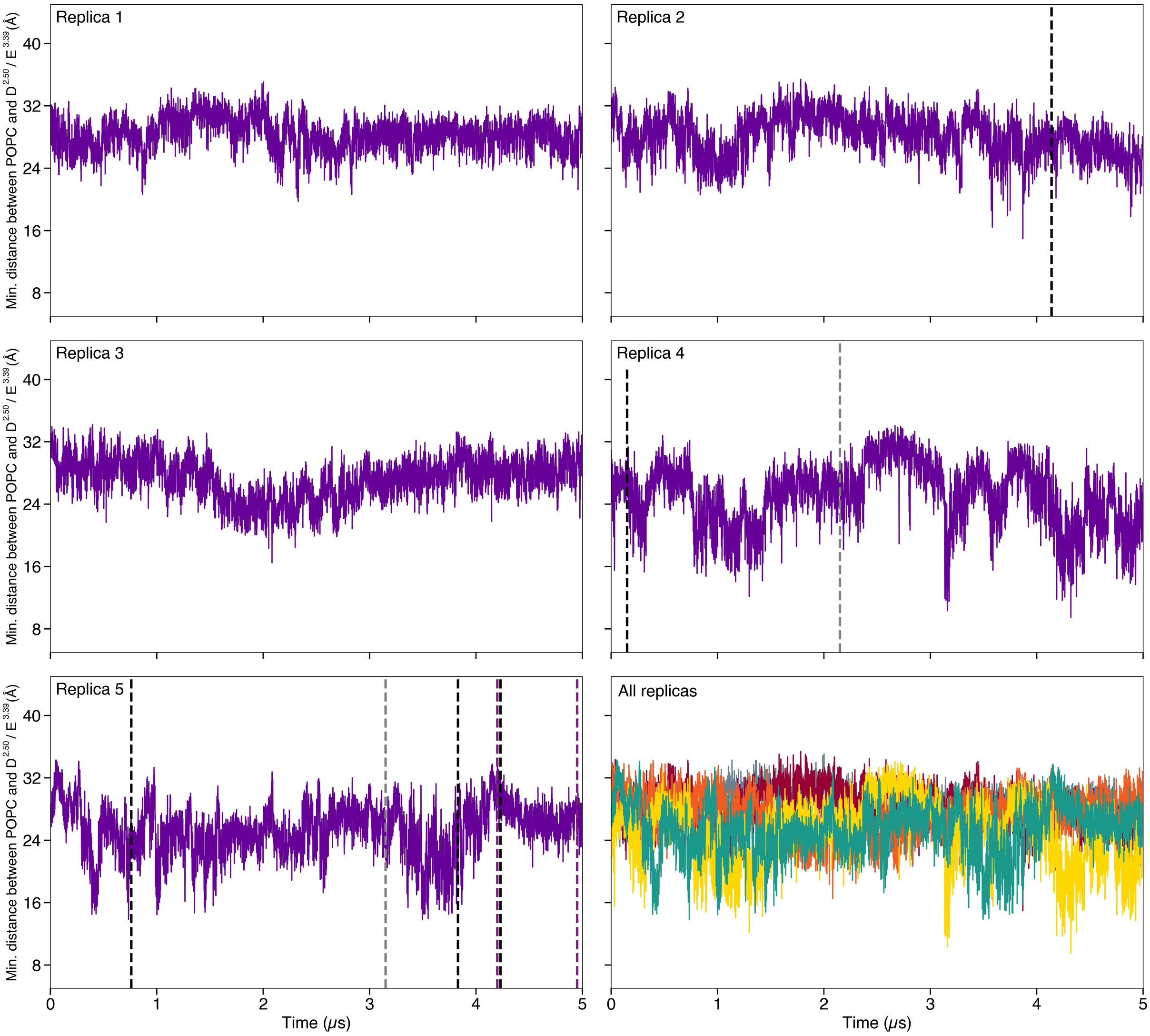


Figure S4: Time series of the sum of the minimum distance between the choline groups of POPC and residue D69$^{2.50}$ and the minimum distance between the choline groups of POPC and residue E110$^{3.39}$ in simulations with Na$^+$ in the ion binding site. The vertical dashed lines indicate the time at which a given event was observed. Sodium binding and unbinding in the ion binding site is depicted by purple dashed lines, while POPC snorkeling into and out of the ion binding site is represented by black and grey dashed lines, respectively. The bottom right graph displays the combined data from all five replicas.

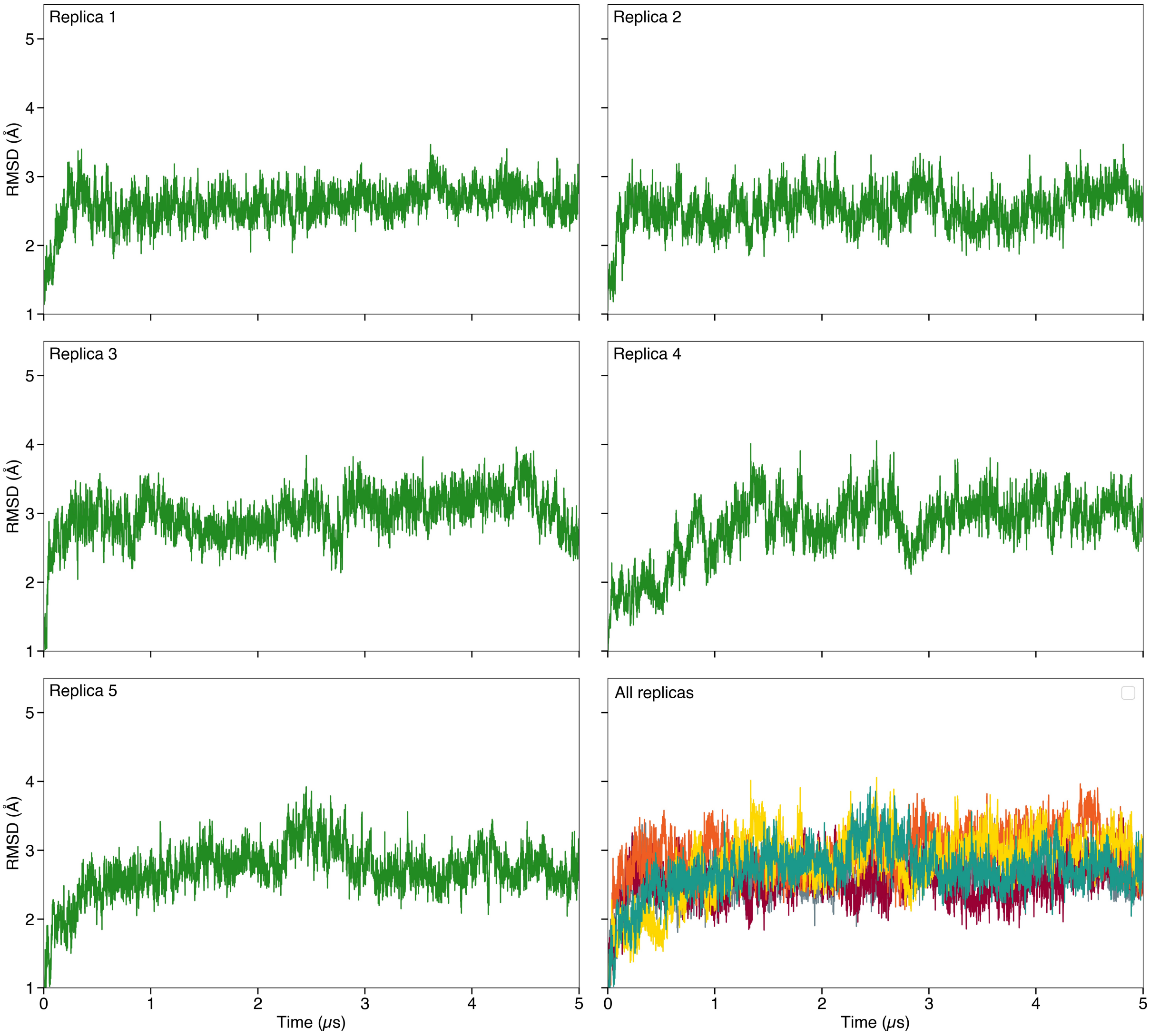


Figure S5: Time evolution of the root mean square deviation (RMSD) for the five MD simulations with Ca^{2+} in the ion binding site. The bottom right graph displays the combined data from all five replicas.

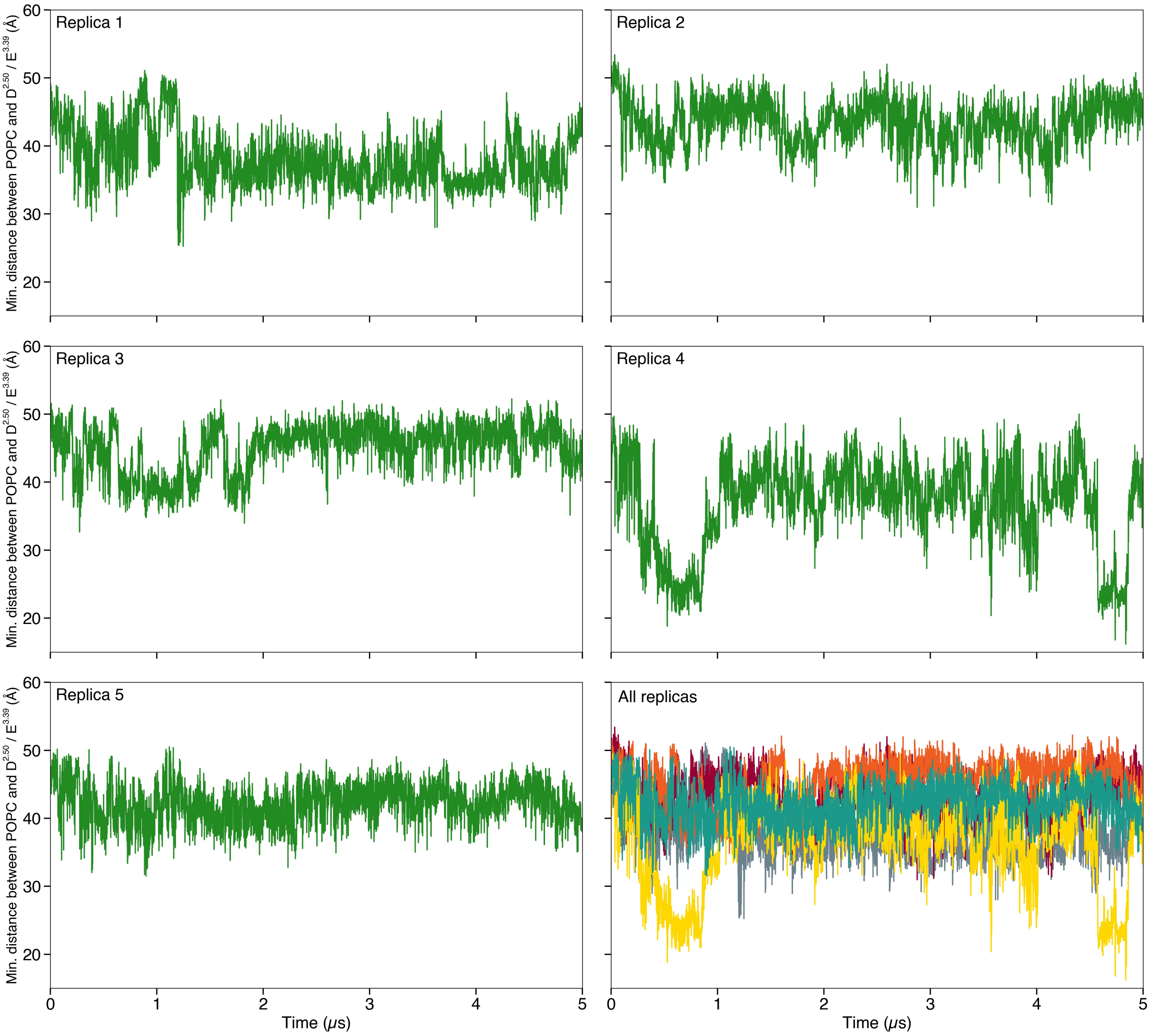


Figure S6: Time series of the sum of the minimum distance between the choline groups of POPC and residue D69$^{2.50}$ and the minimum distance between the choline groups of POPC and residue E110$^{3.39}$ in simulations with Ca^{2+} in the ion binding site. The bottom right graph displays the combined data from all five replicas.

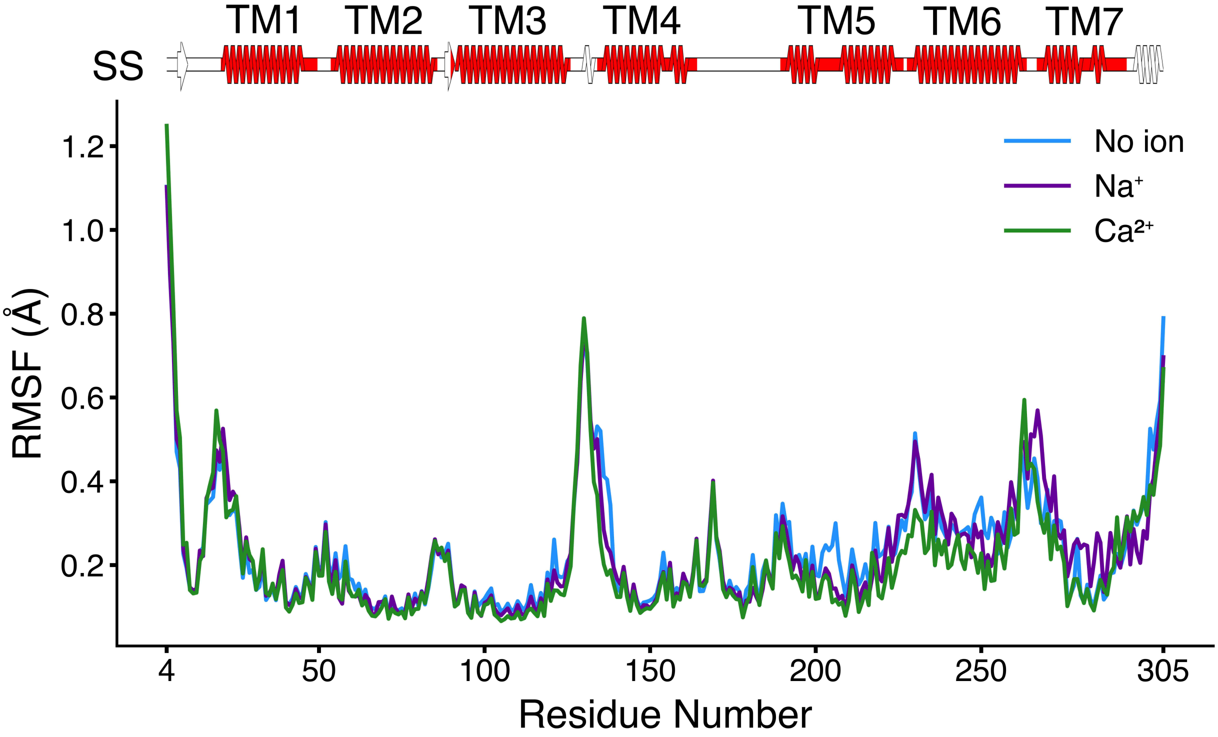


Figure S7: Root mean square fluctuation (RMSF) analysis with secondary structure representation. Top: secondary structure representation calculated and generated with SSDraw,[18] using as input the chain A of the OR51E2 cryo-EM structure (PDB ID: 8F76[1]). The transmembrane (TM) regions, as identified in table S1, are highlighted in red. Bottom: RMSF of the three sets of simulations.

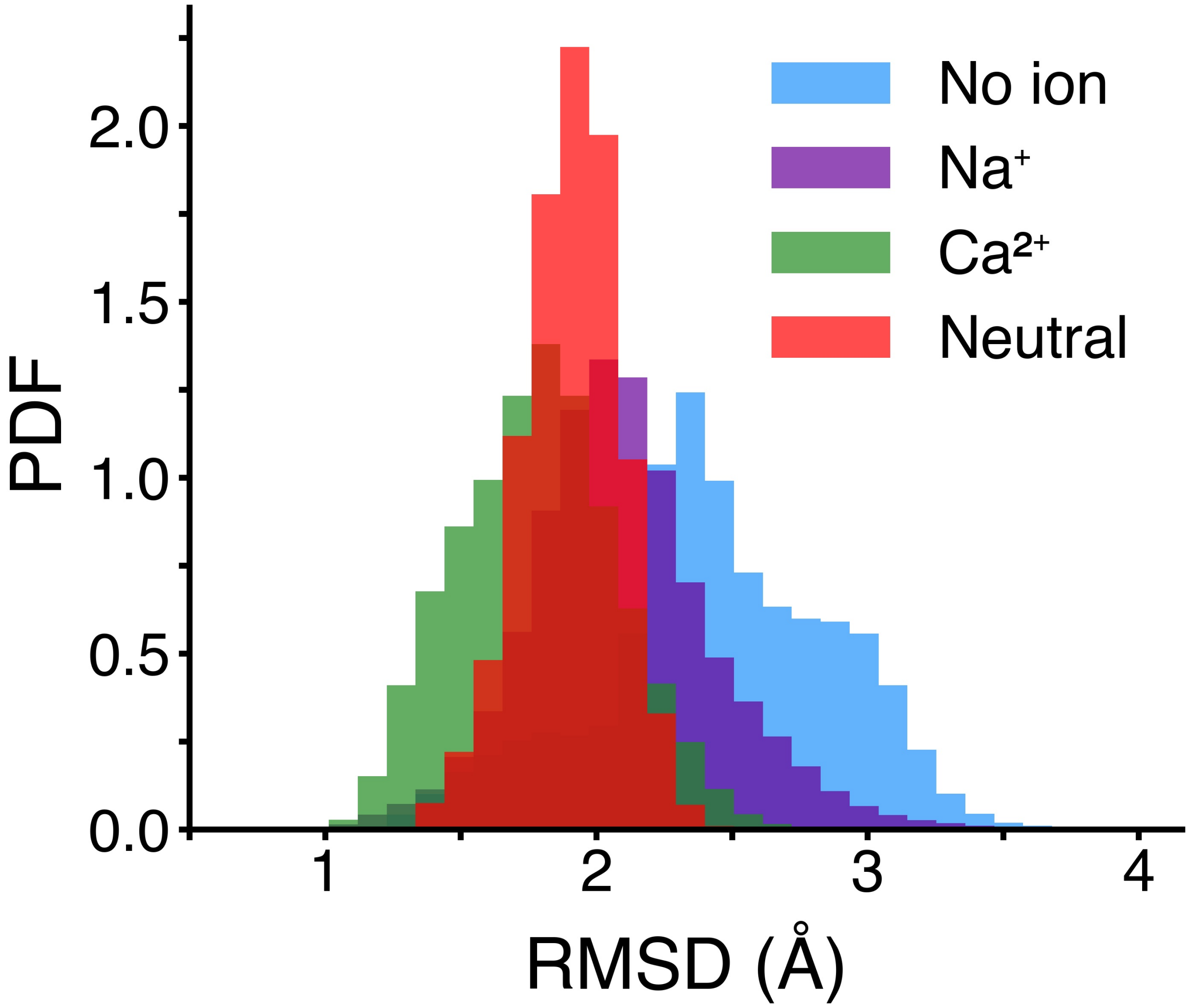


Figure S8: Histogram of the RMSD of the simulations with charged D69$^{2.50}$ and E110$^{3.39}$ performed in this work (without ions, with Na$^+$, and with Ca^{2+} in the ion binding site) and of the simulations with neutral D69$^{2.50}$ and E110$^{3.39}$ with respect to TM3-TM5-TM6 C$_\alpha$ of the apo form of OR52$_{cs}$. We can observe that also quantitatively the simulations with neutral binding pocket sampled receptor conformations in between those of the Na$^+$ and Ca^{2+} simulations.

References

(1) Billesbølle, C. B.; de March, C. A.; van der Velden, W. J. C.; Ma, N.; Tewari, J.; del Torrent, C. L.; Li, L.; Faust, B.; Vaidehi, N.; Matsunami, H.; Manglik, A. Structural basis of odorant recognition by a human odorant receptor. *Nature* **2023**, *615*, 742–749.

(2) Schrödinger, LLC Maestro 2023-3, 2023; New York, NY.

(3) Vickery, O. N.; Carvalheda, C. A.; Zaidi, S. A.; Pisliakov, A. V.; Katritch, V.; Zachariae, U. Intracellular Transfer of Na+ in an Active-State G-Protein-Coupled Receptor. *Structure* **2018**, *26*, 171–180.e2.

(4) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **2008**, *29*, 1859–1865.

(5) Lee, J.; Cheng, X.; Swails, J. M.; Yeom, M. S.; Eastman, P. K.; Lemkul, J. A.; Wei, S.; Buckner, J.; Jeong, J. C.; Qi, Y.; Jo, S.; Pande, V. S.; Case, D. A.; Brooks, C. L. I.; MacKerell, A. D. J.; Klauda, J. B.; Im, W. CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *Journal of Chemical Theory and Computation* **2016**, *12*, 405–413.

(6) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **2017**, *14*, 71–73.

(7) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

(8) Alfonso-Prieto, M.; Capelli, R. Machine Learning-Based Modeling of Olfactory Receptors in Their Inactive State: Human OR51E2 as a Case Study. *Journal of Chemical Information and Modeling* **2023**, *63*, 2911–2917.

(9) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **1995**, *103*, 8577–8593.

(10) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **2007**, *126*, 014101.

(11) Bernetti, M.; Bussi, G. Pressure control using stochastic cell rescaling. *The Journal of Chemical Physics* **2020**, *153*, 114107.

(12) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25.

(13) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(14) Van Rossum, G.; Drake, F. L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.

(15) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Van Gunsteren, W. F.; Mark, A. E. Peptide folding: when simulation meets experiment. *Angewandte Chemie International Edition* **1999**, *38*, 236–240.

(16) Pándy-Szekeres, G.; Caroli, J.; Mamyrbekov, A.; Kermani, A. A.; Keserű, G. M.; Kooistra, A. J.; Gloriam, D. E. GPCRdb in 2023: state-specific structure models using AlphaFold2 and new ligand resources. *Nucleic Acids Research* **2022**, *51*, D395–D402.

(17) Isberg, V.; de Graaf, C.; Bortolato, A.; Cherezov, V.; Katritch, V.; Marshall, F. H.; Mordalski, S.; Pin, J.-P.; Stevens, R. C.; Vriend, G.; Gloriam, D. E. Generic GPCR residue numbers – aligning topology maps while minding the gaps. *Trends in Pharmacological Sciences* **2015**, *36*, 22–31.

(18) Chen, E. A.; Porter, L. L. SSDraw: Software for generating comparative protein secondary structure diagrams. *Protein Science* **2023**, *32*, e4836.

3.2.3 Future Perspectives

Work on OR51E2 continues in two directions:

1. **Investigation of Ligand Binding to Inactive OR51E2.** Molecular docking and molecular dynamics simulations are used to explore ligand binding to the inactive state of OR51E2. Specifically, these methods are used to understand the orientation and location of different ligands within the receptor and to gain insight into the key interactions required for ligand binding. Three different ligands have been selected for this study: propionate, β-ionone, and 13-cis-retinoic acid. The rationale behind this choice is to explore molecules with diverse functions and structures. Propionate, an odorant that acts as an agonist, has a known binding site in OR51E2.[271] In contrast, β-ionone, which acts as an ectopic agonist, binds to a separate site close to the propionate binding site.[272] Finally, 13-cis-retinoic acid is a potential antagonist of OR51E2, although its binding mechanism and location within OR51E2 are currently unknown.[273]

2. **Simulation of OR51E2 with G-Protein Subunits in Membrane.** Future work will involve the modelling and simulating OR51E2, embedded in the membrane, in complex with $G\alpha_s$, $G\beta$, and $G\gamma$ subunits , under different ion binding conditions (no ions, Na^+, and Ca^{2+}). The aim is to study how these different ion binding states influence the conformational changes of the receptor and how these changes are propagated to the associated G-protein subunits.

3.3 A Missense Mutation in the Barley *Xan-h* Gene Encoding the Mg-Chelatase Subunit I Leads to a Viable Pale Green Line with Reduced Daily Transpiration Rate

The group of Professor Paolo Pesaresi (Department of Biosciences, University of Milan) is interested in studying photosynthetic barley mutants from mutagenised collections with the aim of identifying and characterising traits that can improve crop yield or enhance agricultural sustainability.[274,275] In particular, the barley *TM2490* mutant from the TILL-More[276] population shows altered photosynthetic parameters and a pale green colouration, but retains wild-type-like growth and morphology. This specific trait increases the availability of photons in the lower leaf layers and reduces the energy dissipated as heat by leaves exposed to direct sunlight. The group has identified and mapped the mutation responsible for TM2490 phenotype in *CHLI* ATPase subunit of magnesium chelatase, an enzyme that catalyses the insertion of magnesium into protoporphyrin IX and leads to chlorophyll synthesis.[277] This enzyme is highly conserved in all photosynthetic organisms.[278] To date, the structure of CHLI has been solved for several organisms but not for barley. This ATPase subunit assembles as a homo-hexameric ring.

In this context, my aim was to determine a barley CHLI model (*Hv*CHLI) and analyse the structural context of the TM2490 mutation.

3.3.1 Personal Contribution

My involvement in this work was focused exclusively on the computational aspects, which included the following tasks:

1. **Structure Prediction and Comparative Analysis.** I predicted the structure of the *Hv*CHLI subunit using DeepMind AlphaFold2.[54] This was followed by a comparative structural analysis with homologous proteins from different species. The goal was to identify the key residues that define the ATP binding cleft and to locate and analyse the specific residue whose mutation leads to the desired phenotype.

2. **Molecular Docking.** I performed molecular docking of the ATP molecule into the *Hv*CHLI binding cleft using the Schrödinger Maestro suite.[266] The docking process

was driven by the data from the comparative structural analysis, which was used as a constraint to accurately place the ATP molecule.

3. **Mutation Modelling and Analysis.** I modelled the mutation in the context of the ATP-bound state to identify differences in terms of interaction compared to the wild-type structure.

I contributed to the manuscript writing of the ″Materials and Methods″ and ″Results″ sections related to computational structural biology.

Plant Cell Reports (2024) 43:246
https://doi.org/10.1007/s00299-024-03328-2

ORIGINAL ARTICLE

A missense mutation in the barley *Xan-h* gene encoding the Mg-chelatase subunit I leads to a viable pale green line with reduced daily transpiration rate

Andrea Persello[1,8] · Luca Tadini[1] · Lisa Rotasperti[1] · Federico Ballabio[1] · Andrea Tagliani[1] · Viola Torricella[1] · Peter Jahns[2] · Ahan Dalal[3] · Menachem Moshelion[3] · Carlo Camilloni[1] · Serena Rosignoli[4] · Mats Hansson[5] · Luigi Cattivelli[6] · David S. Horner[1] · Laura Rossini[7] · Alessandro Tondelli[6] · Silvio Salvi[4] · Paolo Pesaresi[1]

Received: 4 May 2024 / Accepted: 12 September 2024

Abstract

Key message **The barley mutant *xan-h.chli-1* shows phenotypic features, such as reduced leaf chlorophyll content and daily transpiration rate, typical of wild barley accessions and landraces adapted to arid climatic conditions.**

Abstract The pale green trait, i.e. reduced chlorophyll content, has been shown to increase the efficiency of photosynthesis and biomass accumulation when photosynthetic microorganisms and tobacco plants are cultivated at high densities. Here, we assess the effects of reducing leaf chlorophyll content in barley by altering the chlorophyll biosynthesis pathway (CBP). To this end, we have isolated and characterised the pale green barley mutant *xan-h.chli-1*, which carries a missense mutation in the *Xan-h* gene for subunit I of Mg-chelatase (*Hv*CHLI), the first enzyme in the CBP. Intriguingly, *xan-h.chli-1* is the only known viable homozygous mutant at the *Xan-h* locus in barley. The Arg298Lys amino-acid substitution in the ATP-binding cleft causes a slight decrease in *Hv*CHLI protein abundance and a marked reduction in Mg-chelatase activity. Under controlled growth conditions, mutant plants display reduced accumulation of antenna and photosystem core subunits, together with reduced photosystem II yield relative to wild-type under moderate illumination, and consistently higher than wild-type levels at high light intensities. Moreover, the reduced content of leaf chlorophyll is associated with a stable reduction in daily transpiration rate, and slight decreases in total biomass accumulation and water-use efficiency, reminiscent of phenotypic features of wild barley accessions and landraces that thrive under arid climatic conditions.

Keywords Barley · Canopy photosynthesis · Pale green leaves · Chlorophyll biosynthesis · Mg-chelatase · Drought stress

Communicated by Xiaodong Yang.

Andrea Persello, Luca Tadini, Lisa Rotasperti have contributed equally to the manuscript.

✉ Paolo Pesaresi
paolo.pesaresi@unimi.it

[1] Department of Biosciences, University of Milan, 20133 Milan, Italy

[2] Plant Biochemistry, Heinrich Heine University, 40225 Düsseldorf, Germany

[3] The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, The Hebrew University of Jerusalem, 76100 Rehovot, Israel

[4] Department of Agricultural and Food Sciences, University of Bologna, 40127 Bologna, Italy

[5] Department of Biology, Lund University, 22362 Lund, Sweden

[6] Council for Agricultural Research and Economics (CREA) – Research Centre for Genomics and Bioinformatics, 29017 Fiorenzuola d'Arda, Italy

[7] Department of Agricultural and Environmental Sciences–Production, Landscape, Agroenergy (DiSAA), University of Milan, 20133 Milan, Italy

[8] Department of Industrial Engineering, University of Padua, 35100 Padua, Italy

Published online: 29 September 2024

Introduction

Climate change, increasing population growth and scarcity of land undermine the current paradigm of modern agriculture, and our approach to crop production must become more sustainable. While plant architecture and grain yield have been widely explored in modern breeding programs, photosynthetic traits have generally been neglected, and still offer great potential for further crop improvement and adaptation to cope with emerging climatic parameters (Long et al. 2015). The solar energy conversion efficiency (ECE) index, which is defined as the proportion of absorbed radiation that is converted into biomass, relies on whole-canopy photosynthetic efficiency, and overall crop biomass largely depends upon the ECE (Slattery and Ort 2021). This factor is especially relevant because ECE often falls below half of its theoretical maximum levels in crops (Slattery and Ort 2021). Due to competition for light and nutrients, which are crucial for reproductive success under natural conditions, plants accumulate chlorophylls and thylakoid antenna proteins in large excess with respect to the optimal required for autotrophic growth (Canham et al. 2011). In fact, the photosynthetic machinery saturates at approximately 25% of the maximum solar flux in C3 plant canopies, and this represents a major constraint on productivity in these species (Jansson et al. 2010). On the other hand, in anthropic environments, such as cultivated fields characterized by monocultures, competition among individual plants is disadvantageous, and new cultivars with reduced chlorophyll accumulation might become valuable resources. To this end, reduction of leaf chlorophyll content has been suggested to be highly effective in improving light penetration under high-density mass cultivation, and in mitigating high-light-related photo-oxidative damages, with great benefits for biomass yield (Melis 2009). In addition, the reduction of leaf chlorophyll content in crops, *i.e.* the use of pale green phenotypes, enhances light reflectance, which helps to alleviate the effects of heat waves triggered by global climate change (Genesio et al. 2021), and improves the efficiency of water use by reducing canopy temperature (Drewry et al. 2014; Galkin et al. 2018). Furthermore, independent studies have predicted that reductions in chlorophyll content should increase the efficiency of nitrogen use (Walker et al. 2018; Sakowska et al. 2018). Pale green crops can be created by manipulating a plethora of processes, such as the biogenesis and/or accumulation of antenna proteins—also known as the truncated light-harvesting antenna (TLA) strategy—and pigment biosynthesis (for a review, see Cutolo et al. 2023). For instance, increased photosynthetic performance and enhanced plant biomass accumulation were observed upon cultivation at high density under greenhouse conditions of a pale green tobacco line with downregulated expression of *cpSRP43* (Kirst et al. 2018). This nuclear gene codes for the 43-kDa chloroplast-localised signal recognition particle, which is responsible for the delivery of antenna proteins to the thylakoid membranes (Klimyuk et al. 1999), More recently, the barley mutant *happy under the sun 1* (*hus1*), which carries a premature stop codon in the corresponding *HvcpSRP43* gene and is characterised by a 50% reduction in the chlorophyll content of leaves, was shown to accumulate biomass and grains at levels comparable to those observed for the control cultivar Sebastian, when grown under field conditions at standard density. These findings demonstrate that crops can indeed decrease their investment in antenna proteins and chlorophyll biosynthesis significantly, without detrimental effects on productivity (Rotasperti et al. 2022). Conversely, a decrease of about 26% in biomass production was observed in the case of the pale green soybean mutant *MinnGold* under field conditions (Sakowska et al. 2018). Owing to a missense mutation in the nuclear gene encoding the CHLI subunit of the enzyme Mg-protoporphyrin IX chelatase (Mg-chelatase), this mutant synthesizes approximately 80% less chlorophyll than the green control plants. As the first enzyme specific for the chlorophyll biosynthetic pathway, Mg-chelatase is a multimeric complex that is responsible for the insertion of Mg^{2+} into the protoporphyrin IX tetrapyrrole ring. In plants, the enzyme complex consists of three subunits (Masuda 2008), designated CHLI (36–46 kDa), CHLD (60–87 kDa) and CHLH (120–155 kDa). In the presence of Mg^{2+} and ATP, the CHLI and CHLD subunits form a double homohexameric ring complex typical of members of the AAA + [ATPase associated with various cellular activities] protein superfamily (Elmlund et al. 2008; Lundqvist et al. 2013), which then interacts with the CHLH subunit responsible for binding protoporphyrin IX and inserting Mg^{2+} to form Mg-protoporphyrin IX (Farmer et al. 2019; Adams et al. 2020; Willows and Beale 1998). The ATP needed for this reaction is hydrolysed by the CHLI subunit (Lundqvist et al. 2010), and the ATP-binding pocket is formed by two neighbouring CHLI subunits via five key interaction motifs (Gao et al. 2020). While most photosynthetic species, including barley, have only one *Hv*CHLI isoform, *Arabidopsis thaliana* has two *CHLI* genes, *AtCHLI1* and *AtCHLI2*, with *AtCHLI1* being more highly expressed than *AtCHLI2* (Huang and Li 2009). Extensive characterization of *chli* mutants has been conducted in various land plants, including Arabidopsis, barley, maize, rice, pea, strawberry and tea (Zhang et al. 2023; Ma et al. 2023; H. Zhang et al. 2006; Wu et al. 2022; Huang and Li 2009; Braumann et al. 2014). Intriguingly, many forward genetic screens in barley mutant populations have identified chlorophyll-deficient lines with seedling-lethal phenotypes, designated as *Xantha* and *Chlorina* mutants, including *xan-h.38*, *xan-h.56*, *xan-h.57*, *xan-h.clo125*, *xan-h.clo157*, and *xan-h.clo161* (Braumann et al. 2014; Hansson et al. 1999). These *Xantha* and *Chlorina* mutants carry nonsense and missense mutations, respectively, in the *Xan-h* coding region. Interestingly, heterozygous missense mutations display stronger semidominance than nonsense mutations (Hansson et al. 2002; Braumann et al. 2014).

Alongside its key role in chlorophyll biosynthesis, Mg-chelatase has been reported to have a role in chloroplast-to-nucleus retrograde communication. Thus, the inactivation of Mg-chelatase due to a mutation in *CHLH* resulted in the Arabidopsis *gun5* mutant (*genomes uncoupled 5*), which deregulates the expression of the *Light Harvesting Complex B2* gene (*LHCB2*) (Mochizuki et al. 2001) upon inhibition of chloroplast biogenesis. The GUN4 protein (*genomes uncoupled 4*), a regulatory subunit found in oxygenic photosynthetic organisms, which binds to CHLH and stimulates its magnesium chelatase activity, has also been reported to participate in retrograde signalling (Larkin et al. 2003). Similarly, *chld* mutants that are deficient in Mg-chelatase activity show plastid-mediated deregulation of selected nuclear genes (Brzezowski et al. 2016; Huang and Li 2009). With regard to *chli* mutants, it was reported that *A. thaliana cs* and *ch42* and rice *chlorina-9* mutants do not show the *genomes uncoupled* phenotype (Mochizuki et al. 2001; Zhang et al. 2006), whereas both the semi-dominant Arabidopsis mutant *cs215/cs215* and the *Atchli1/Atchli1 Atchli2/Atchli2* double mutant do since they accumulate higher levels of *Light Harvesting Complex B1* (*LHCB1*) transcripts than the wild type upon impairment of chloroplast activity by norflurazon (NF) treatment (Huang and Li 2009). In barley, lethal mutations in any of the three Mg-chelatase genes cause the *genomes uncoupled* phenotype (Gadjieva et al. 2005).

In this study, we describe the pale green barley mutant line *TM2490*, which was isolated from the TILLMore mutagenized population (Talamè et al. 2008), and is characterised by a single point mutation in the *Xan-h* (*HORVU.MOREX.r3.7HG0738240*) gene, responsible for the Arg-to-Lys substitution at position 298 (R298K) in the *Hv*CHLI subunit. The homozygous mutant plants show reduced leaf chlorophyll content and increased photosynthetic efficiency at high light intensities and represent the only known viable homozygous *xan-h.chli-1* mutant in barley. In the following, we provide further insights into *Hv*CHLI function and chlorophyll accumulation in barley and explore the behaviour of the pale green leaf phenotype under drought stress conditions.

Results

The pale green phenotype of the *TM2490* barley mutant is caused by a missense mutation in *Xan-h*, the single-copy nuclear gene encoding the *Hv*CHLI subunit of Mg-chelatase

The chemically mutagenized TILLMore population (Talamè et al. 2008) was screened for pale green mutants with improved photosynthetic performance under field conditions (see Materials and Methods). Among M_4 mutant lines, the *TM2490* line was selected based on its reduced chlorophyll content, i.e. pale green leaf phenotype and enhanced photosynthetic performance with respect to the control (*cv.* Morex). Under controlled greenhouse conditions, the growth rate and plant architecture of the *TM2490* line were similar to those of the wild-type control. However, amounts of chlorophylls *a* and *b* (Chl*a* + Chl*b*) ranged from 50% of WT levels in the first and second leaves to only 25% in the sixth, i.e. penultimate, leaf, while no major differences in chlorophyll abundance were observed between mutant and control flag leaves (Fig. 1A, B). Similarly, in mutant plants, younger leaves showed a generally increased photosynthetic efficiency of photosystem II [Y(II)] under light conditions and optimal functionality of photosystem II (PSII) under dark conditions (Fv/Fm) relative to the control plants (Fig. 1C), while at later stages no major differences could be observed between mutant and control leaves.

To identify the mutation responsible for the pale green phenotype, a segregating F_2 population of 565 plants was generated by crossing the *TM2490* line (background *cv.* Morex) with *cv.* Barke. About one-quarter of the total population (131/565) showed the *TM2490*-like phenotype with reduced chlorophyll content and increased Fv/Fm values (WT-like 0.74 ± 0.02 vs *TM2490-like* 0.81 ± 0.02, Student's *t*-test < 0.001), typical of monogenic recessive inheritance (χ^2 test 3:1 WT:mut, not significant). Total RNA was then isolated from 100 F2 *TM2490*-like and 100 F2 WT-like plants and bulked in an equal ratio to generate two distinct RNA pools. Both RNA pools were subjected to polyA capture and paired-end sequencing, producing approximately 100 million 2 × 150-bp read pairs per pool. Reads were mapped on the reference genome sequence assembly of barley *cv.* Morex (Morex V3; Monat et al. 2019) to identify the allelic variants in each of the two pools. Plotting of the allele frequencies over SNP positions along the barley genome revealed a sharp peak along chromosome 7H, corresponding to a 20-Mb region (from 586.396.977 bp to 606.525.807 bp) in which allelic variants with frequencies higher than 0.5 and peaking at 1.0 in the *TM2490*-like pool were coupled with frequencies lower

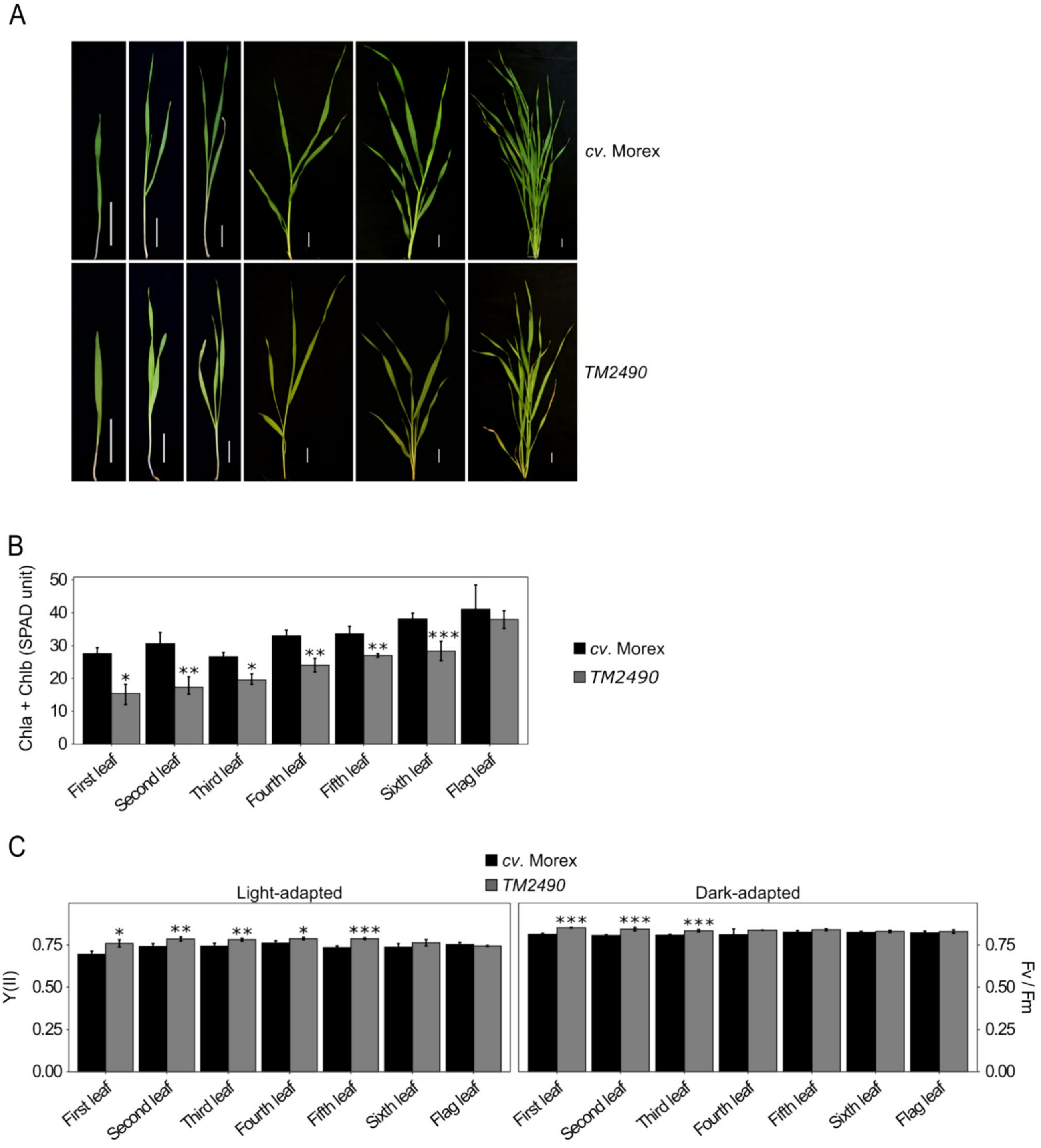


Fig. 1 Visible phenotypes of *cv.* Morex (control) and *TM2490* mutant plants grown under greenhouse conditions. **A** Images of *cv.* Morex control plant and the pale green mutant *TM2490* from coleoptile to flag-leaf stage. Scale bar = 2 cm. **B** Measurements of apparent chlorophyll content in *cv.* Morex and *TM2490* leaves (expressed as SPAD units) carried out on eight independent plants at different developmental stages. **C** Leaf photosynthetic performance of dark-adapted and light-adapted plants measured with the Handy PEA fluorometer in eight independent plants. Error bars on the histograms indicate standard deviations and the significance of the observed differences was assessed using Student's t-test (*** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$)

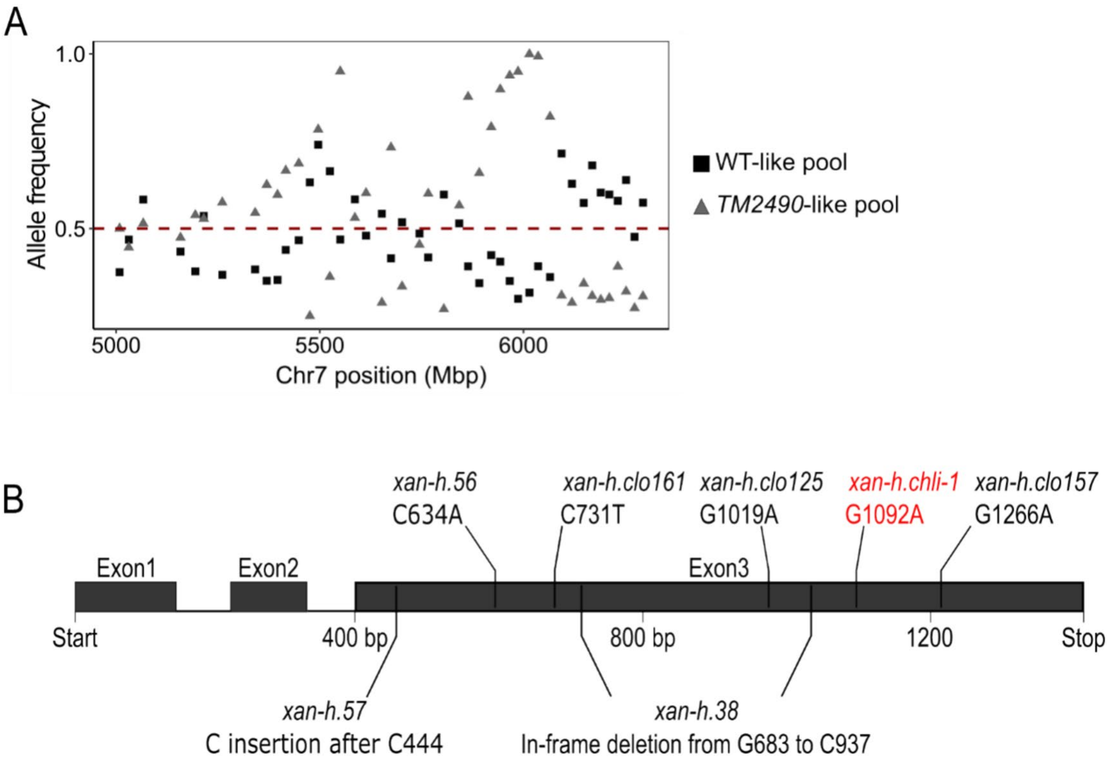


Fig. 2 Identification of the *TM2490* locus. **A** Comparison of allele frequency distributions in RNAseq pools obtained from WT-like and *TM2490*-like F2 individuals. Allele frequencies are indicated on the Y axis, genomic coordinates along Chr7 on the X axis. The peak of homozygous alleles in the *TM2490*-like pool corresponds to the 20-Mb candidate region around 6000 Mbp. The red line indicates the threshold allele frequency of 0.5. **B** Schematic representation of the *Xan-h* (*HORVU.MOREX.r3.7HG0738240*) locus, i.e. the single-copy gene chosen as the best candidate for the *TM2490* phenotype. Bars indicate the positions of known lethal mutations within the gene, together with the *TM2490* mutation, here indicated as *xan-h.chli-1*, with the respective SNPs. Boxes represent exons and lines indicate introns

than 0.5 in the WT-like pool (Fig. 2A). Within this region, the single-copy gene *HORVU.MOREX.r3.7HG0738240*, known as *Xan-h* locus, carried a G-to-A transition at position + 1092 from the translation start codon in *TM2490*-like plants (referred to as the *xan-h.chli-1* allele in the following; Fig. 2B), causing the R298K substitution (Fig. S1). The gene is annotated in the Barlex database as Mg-protoporphyrin IX chelatase subunit I (*Hv*CHLI), a 417-a.a. protein, with a predicted 56-a.a. chloroplast transit peptide (cTP) at the N-terminus, which is essential for the insertion of Mg^{2+} into protoporphyrin IX, the first chlorophyll-specific step of tetrapyrrole biosynthesis in photosynthetic organisms (Kobayashi et al. 2008; Huang and Li 2009). The protein is highly conserved from photosynthetic bacteria to higher plants as is the Arg residue at position 298 (Fig. S1).

To validate the association between the missense mutation in the *Xan-h* locus and the *TM2490* phenotype, allelism tests were performed, with the aid of the two other known mutant alleles at this locus, *xan-h.clo161* (Hansson et al. 1999) and *xan-h.56* (Braumann et al. 2014), both of which are recessive chlorotic lethals (see Fig. 2B). To this end, homozygous *TM2490* (*xan-h.chli-1*/*xan-h.chli-1*) plants were crossed with heterozygous *xan-h.clo161* (*Xan-h*/*xan-h.clo161*) and *xan-h.56* (*Xan-h*/*xan-h.56*) plants and F_1 seedlings were phenotypically and genetically analysed at the cotyledon stage. Approximately 50% of F_1 plants, carrying both *xan-h.chli-1* and either of the *Xan-h* alleles, showed a WT-like photosynthetic and dark-green leaf phenotype, while the biallelic *xan-h.clo161*/*xan-h.chli-1* seedlings were characterised by a dramatic reduction in chlorophyll content, impaired PSII activity (Fv/Fm) and seedling lethality, similar to those of homozygous *xan-h.clo161* mutant seedlings (Fig. S2A, C, D). The pale green phenotype with a significant reduction in leaf chlorophyll content was also observed in *xan-h.56*/*xan-h.chli-1* biallelic seedlings, despite showing WT-like PSII activity and the capability to complete the life cycle (Fig. S2B, E, F).

The functional status of the *xan-h.chli-1* mutant allele was further investigated by cloning its coding sequence into a binary vector under the control of the *CaMV35S* promoter and introducing it into the Arabidopsis *Atchli1*/*Atchli1* knock-out genetic background by Agrobacterium-mediated transformation (Huang and Li 2009). BLAST

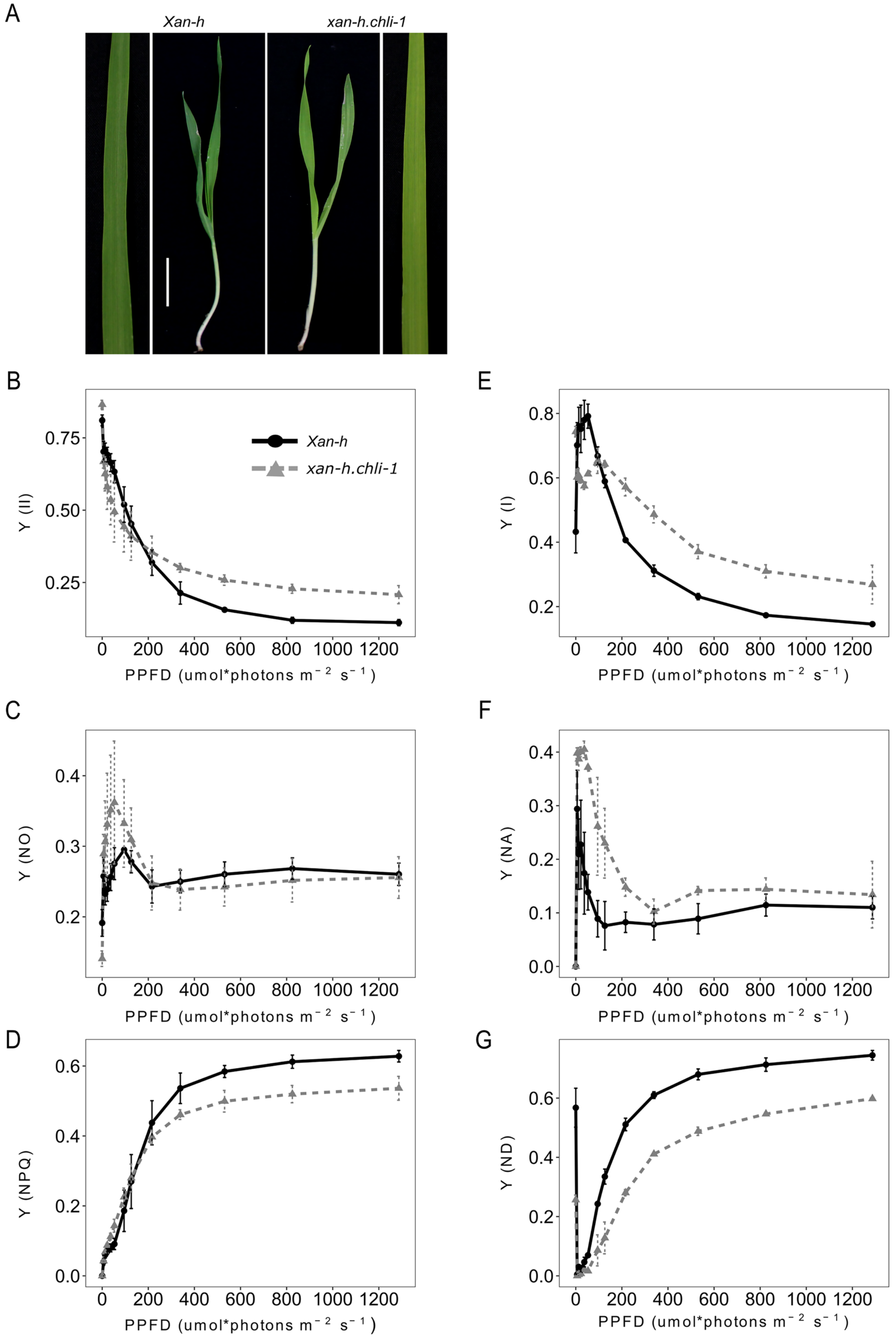

A
Xan-h
xan-h.chli-1
B
Y (II)
0.75
0.50
0.25
Xan-h
xan-h.chli-1
0 200 400 600 800 1000 1200
PPFD (umol*photons m^{-2} s^{-1})
E
Y (I)
0.8
0.6
0.4
0.2
C
Y (NO)
0.4
0.3
0.2
F
Y (NA)
0.4
0.3
0.2
0.1
0.0
D
Y (NPQ)
0.6
0.4
0.2
0.0
G
Y (ND)
0.6
0.4
0.2
0.0

Fig. 3 Representative phenotypes of *Xan-h* and *xan-h.chli-1* plants at the second-leaf stage following growth under greenhouse conditions. **A** *Xan-h* and *xan-h.chli-1* barley leaves were harvested 14 days after germination. Note that, in terms of leaf pigment content and photosynthetic performance, *xan-h.chli-1* plants (BC_2F_2 generation) were identical to *TM2490* plants at the M4 generation. Scale bar = 2 cm. Analyses of photosynthetic parameters were performed using the Dual-PAM 100 fluorometer. **B** The effective quantum yield of PSII [Y(II)], and quantum yields of non-regulated energy dissipation [Y(NO)] (**C**) and regulated energy dissipation of PSII [Y(NPQ)] (**D**). Measurements used to monitor PSII performance were carried out at increasing light intensities (from dark to 1287 µmol photons m^{-2} s^{-1}; 3-min exposure to each light intensity). Concomitantly, the effective quantum yield of PSI [Y(I)] (**E**), and the quantum yields of non-photochemical energy dissipation in PSI owing to acceptor-side limitation [Y(NA)] (**F**), and donor-side-limited heat dissipation [Y(ND)] (**G**), were determined. Curves show average values of three biological replicates, while bars indicate standard deviations. *PPFD* photosynthetic photon flux density

analyses indeed revealed that the *Hv*CHLI subunit from *cv.* Morex, encoded by a single-copy gene, shares high homology with the *A. thaliana* proteins *At*CHLI1 (78% identity) and *At*CHLI2 (81% identity) (Fig. S1). The *Atchli1/Atchli1* + *35S::Xan-h* line, carrying the WT *HvCHLI* coding sequence from *cv.* Morex and the endogenous Arabidopsis *At*CHLI2 protein, used here as the control, showed a fully complemented phenotype in terms of photosynthetic performance and chlorophyll accumulation in T1 lines and progenies. In contrast, the *35S::xan-h.chli-1* construct only partially complemented the *Atchli1/Atchli1* lethal phenotype, generating viable plant lines that were similar to both barley *TM2490* and Arabidopsis *cs/cs* mutants with respect to photosynthetic performance and total chlorophyll content (Fig. S3; Kobayashi et al. 2008). Overall, these data corroborated the hypothesis that the *xan-h.chli-1* mutant allele is responsible for the pale green phenotype and the altered *Hv*CHLI subunit of the homozygous *TM2490* line bearing the R298K amino acid substitution that hampers chlorophyll biosynthesis.

The *xan-h.chli-1* barley mutant shows a reduced chlorophyll content and increased photosynthetic efficiency under high light intensities

To extend the characterization of *TM2490*-related phenotype and minimize any possible influence of other chemically induced mutations in the *TM2490* genome, the mutant was backcrossed with the barley *cv.* Morex, and BC_2F_2 plants showing the *TM2490* phenotype (referred to as the *xan-h.chli-1* line in the following) were selected for detailed biochemical and physiological characterization, together with their wild-type-like siblings (referred to as *Xan-h* in the following). In particular, the second leaves of *Xan-h* and *xan-h.chli-1* plants were used for all the analyses reported from here on (Fig. 3A). Quantification of leaf pigments by high-performance liquid chromatography (HPLC) revealed the total chlorophyll content (Chl*a* + Chl*b*) in *xan-h.chli-1* amounted to about 57% of that in *Xan-h*, while the ratio of Chl*a* to Chl*b* in the mutant (3.97 ± 0.1) was higher than that in the WT (*Xan-h* 3.29 ± 0.1; see also Table 1). This difference was due to the reduced accumulation of Chl*a* in *xan-h.chli-1* line (59% of the *Xan-h* level) and an even more marked decrease in Chl*b* (49% of the *Xan-h* level). In addition, the pool of carotenoids associated with photosystem antenna complexes, such as lutein (Lut) and neoxanthin (Nx), showed a marked reduction in the mutant (to around 54% of *Xan-h* levels), while the β-carotene (β-Car) content, found mainly in photosystem cores, and in part also in antenna proteins of photosystem I, was decreased to 65% of *Xan-h* control (Table 1), indicating a general alteration of photosystems, albeit more pronounced at the level of antenna proteins. To investigate this aspect further, the second leaves of *Xan-h* and *xan-h.chli-1* plants were exposed to increasing actinic light intensities (0–1287 µmol photons m^{-2} s^{-1}) and the photosynthetic efficiency was assessed by monitoring the performance of PSII. In dark-adapted leaves, *xan-h.chli-1* showed a higher PSII quantum yield (Fv/Fm), which declined more rapidly than in *Xan-h* upon moderate light illumination [Y(II) less than 200 µmol photons m^{-2} s^{-1}; Fig. 3B]. Conversely, the PSII quantum yield of non-regulated energy dissipation [Y(NO)] was markedly higher in *xan-h.chli-1* at low-to-moderate light intensities – implying rather inefficient photochemical energy conversion overall compared to *Xan-h* leaves (Fig. 3C). In addition, upon exposure to 200–1287 µmol photons m^{-2} s^{-1} of actinic light, Y(II) values remained consistently higher in *xan-h.chli-1* than in *Xan-h*, possibly because the values for PSII quantum yield attributable to regulated energy dissipation [Y(NPQ)] were consistently lower in *xan-h.chli-1* leaves (Fig. 3D), while Y(NO) levels were identical in mutant and *Xan-h* samples. To investigate further the photosynthetic properties of *xan-h.chli-1* leaves, an identical experimental set-up was used to assess photosystem I (PSI) activity. The quantum yield of PSI [Y(I)] was higher in *xan-h.chli-1* under dark-adapted conditions and dropped to values lower than those seen in *Xan-h*, between 6 and 95 µmol photons m^{-2} s^{-1} (Fig. 3E), similarly to the Y(II) trend, and most probably because of less efficient energy transfer from the antenna to the PSI reaction centre. Furthermore, the Y(NA) parameter, i.e. the quantum yield of non-photochemical energy dissipation in PSI due to acceptor-side limitation (Fig. 3F), was much higher in *xan-h.chli-1* under low to moderate light conditions, while it decreased to *Xan-h* values at higher light levels, as soon as the photosynthesis control was engaged (Colombo et al. 2016). Similarly, the lower values of Y(ND), *i.e.* the non-photochemical PSI quantum yield due to donor-side-limited heat dissipation (Fig. 3G), at higher light intensities confirmed the greater efficiency of electron transport

Table 1 HPLC analysis of second-leaf pigment content in *Xan-h* and *xan-h.chli-1*. The pigment content was normalized to leaf fresh weight (FW) and is reported as pmol per mg of FW

	Nx	Lut	Chl*b*	Chl*a*	β-Car	VAZ	Chl*a* + Chl*b*	Chl*a*/*Chlb*
Xan-h	62 ± 11	189 ± 29	478 ± 101	1563 ± 302	182 ± 44	120 ± 20	2041 ± 402	3.29 ± 0.1
xan-h.chli-1	34 ± 9	103 ± 30	235 ± 52	930 ± 203	117 ± 30	104 ± 25	1164 ± 255	3.97 ± 0.1
T-test	**	**	***	**	*	ns	**	***

Average values ± standard deviation of three biological replicates are shown. The significance of the observed differences was evaluated with Student's *t*-test (*** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, ns = not significant)

Nx neoxanthin, *Lut* lutein, *Chl* chlorophyll, *β-Car* β-carotene, *VAZ* violaxanthin + antheraxanthin + zeaxanthin

from PSII to PSI in *xan-h.chli-1* leaves. Overall, our findings highlight the low photosynthetic efficiency of *xan-h.chli-1* under low-to-moderate actinic light intensities, although this parameter rises at higher intensities, most probably as a consequence of the reduced chlorophyll content and light absorption capacity of the pale green *xan-h.chli-1* leaves.

The functional status of the photosynthetic machinery was also analysed at the biochemical level by monitoring the protein composition of the thylakoid electron transport machinery by means of immunoblot analysis. In agreement with the pigment accumulation profile (Table 1), immunoblot analyses with antibodies specific for Lhca3, Lhcb1, Lhcb2 and Lhcb3 confirmed a general reduction (of at least 60%) in antenna proteins in *xan-h.chli-1* thylakoids, while lesser declines (of 20–40%) were observed for Lhca1, Lhca2 and Lhcb4. Only in the case of Lhcb5 were the levels attained identical between mutant and *Xan-h* samples (Fig. 4A). Moreover, in *xan-h.chli-1* thylakoid samples the levels of PSII core subunits (D1 and CP43) were reduced by around 30% and 50%, respectively, similar to what was observed for PsbS and two subunits of the Oxygen-Evolving Complex (OEC), PsbO and PsbR. Conversely, the PsbQ subunit of OEC showed a much more drastic reduction, accumulating to only around 14% of its level in *Xan-h*. Similarly to PSII, the PSI core subunits, PsaA and PsaD, accumulated in *xan-h.chli-1* thylakoids to lower levels than in *Xan-h*, while no major differences were observed in the accumulation of the cytochrome *f* subunit (PetA) or the plastocyanin electron carrier (PetE).

In order to test the effect of high-light exposure, *Xan-h* and *xan-h.chli-1* plants were grown under control conditions and then adapted to high-light for 0.5 and 8 h. Immunoblots analysis on the photosynthetic machinery and Mg-chelatase subunits revealed no major differences between control and high-light conditions on most of the subunits analysed (Fig. S4). Nevertheless, both genotypes showed a reduced accumulation of D1 and D2 PSII core subunits after 30 min of high-light exposure, particularly marked in *Xan-h* thylakoids, indicating a higher capability of *xan-h.chli-1* leaves to better adapt to highlight conditions, in agreement with the increased photosynthetic performance under high-light regimes (see Fig. 3). The accumulation of D1 and D2 subunits increased in both genotypes after 8 h of high-light exposure, due to their adaptation to the light environment.

Finally, the impact of the decreased chlorophyll and thylakoid protein contents on the chloroplast ultrastructure in the mutant was investigated by Transmission Electron Microscopy (TEM; Fig. 4B). TEM analyses, performed on growth-light-adapted plants, showed reduced accumulation of starch granules in *xan-h.chli-1* chloroplasts when compared to *Xan-h*, while no major alteration in the organization of grana and stroma lamellae was observed.

The *xan-h.chli-1* mutation impairs Mg-chelatase activity

Since the *xan-h.chli-1* mutation results in the R298K amino acid exchange (Fig. S1), the accumulation and activity of Mg-chelatase enzyme was quantified. Immunoblot analyses revealed that the *Hv*CHLH subunit accumulated to WT levels, while *Hv*CHLI and *Hv*CHLD were slightly reduced in *xan-h.chli-1* leaves (Fig. 5A). However, the regulatory subunit *Hv*GUN4 was almost twice as abundant in *xan-h.chli-1* as it was in *Xan-h*, possibly as a compensatory response to the decline in Mg-chelatase activity owing to partial impairment of *Hv*CHLI. To test whether the point mutation identified in the *xan-h.chli-1* allele affects its homodimerization, its coding sequence (devoid of the cTP-coding region) was tested for homodimer formation in a yeast two-hybrid assay, together with the variants *xan-h.clo125*, *xan-h.clo157* and *xan-h.clo161* and compared with the ability of *Xan-h* from *cv.* Morex to homodimerize as a control. As shown in Fig. 5B, colonies expressing *Xan-h* and *xan-h.chli-1* were able to grow on selective media, suggesting that the R298K missense mutation does not impair homodimer formation. In contrast to this result, colonies expressing *xan-h.clo125*, *xan-h.clo157* and *xan-h.clo161* variants were unable to grow on selective media, indicating that lethal mutations hamper the ability to form CHLI homodimers (Fig. 5B, Fig. S5). Moreover, stromal protein extracts from etiolated *Xan-h* and *xan-h.chli-1* seedlings were used to test the activity of the Mg-chelatase oligo-enzyme by measuring its ability to

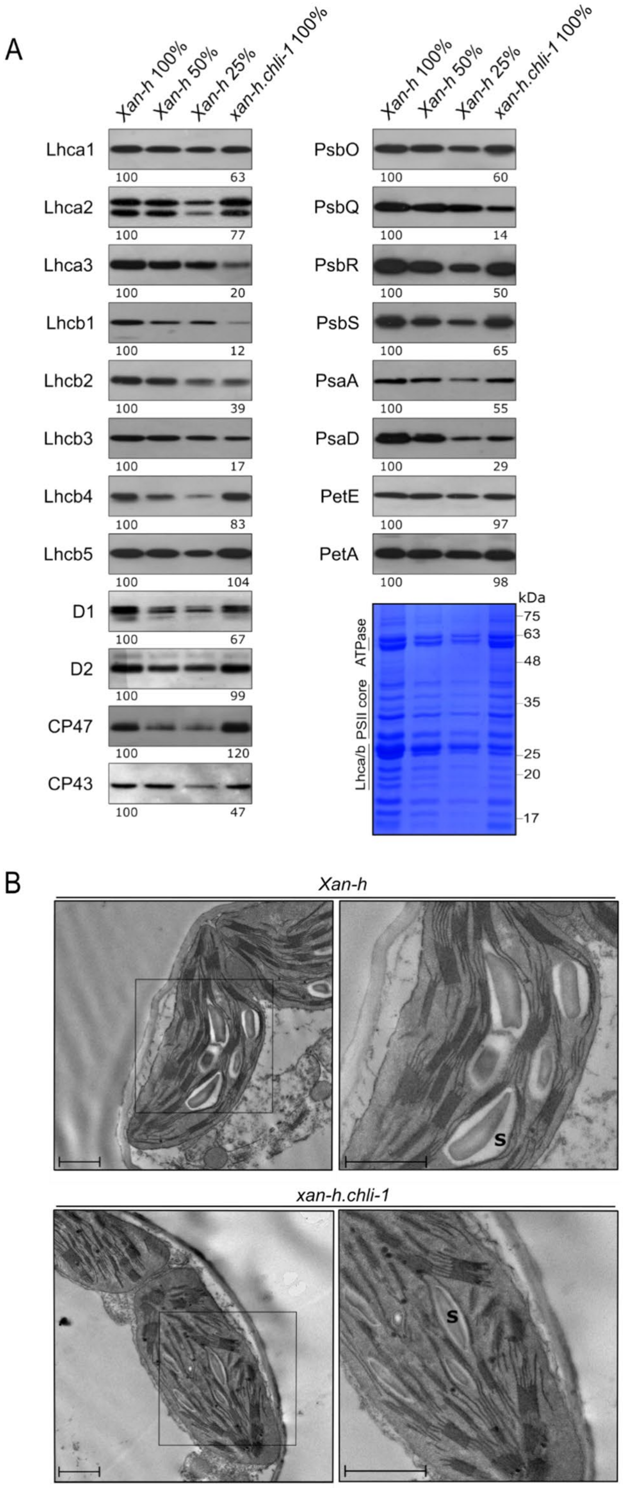


Fig. 4 Biochemical and ultrastructural characterization of thylakoid membranes from *Xan-h* and *xan-h.chli-1*. **A** Immunoblot analyses of thylakoid protein extracts from *Xan-h* and *xan-h.chli-1* leaf material, normalized with respect to fresh weight and probed with antibodies specific for subunits of thylakoid protein complexes. For relative quantification, 50% and 25% dilutions of *Xan-h* protein extracts were also loaded. One filter (representative of three biological replicates) is shown for each immunoblot. An SDS-PA gel stained with Coomassie Brilliant Blue (CBB) is shown as loading control. **B** TEM micrographs depict chloroplast ultrastructure in *Xan-h* (upper panels) and *xan-h.chli-1* (lower panels) samples. *S* starch granule; Scale bar = 1 µm

convert deuteroporphyrin IX into Mg-deuteroporphyrin IX in vitro. As expected, a marked reduction in Mg-chelatase activity was observed in the *xan-h.chli-1* protein relative to *Xan-h* samples and the mock control (Fig. 5C).

The R298K substitution in *Hv*CHLI may hamper its interaction with ATP

To analyse the consequences of the R298K substitution in a structural context, we modelled the configuration of the *Hv*CHLI subunit, as described in the Materials and Methods section. In general, the *Hv*CHLI ATPase subunit assembles as a closed ring and its quaternary structure results from the association of six identical monomers (Hansson et al. 2002; Lundqvist et al. 2010; Fig. 6A). In the reconstructed model, R298 protrudes towards the ATP-binding cleft at the interface between two monomers (Fig. 6B). This position is consistent with the X-ray structure of the *Synechocystis sp. PCC 6803 substr. Kazusa* CHLI subunit (PDB ID 6L8D; Gao et al. 2020), in which R298 corresponds to R233. A similar situation was also observed for other relevant, highly conserved residues, such as R356 (R291 in *Synechocystis*) and D274 (D209 in *Synechocystis*) (Fig. 6B). These two residues, whose replacements lead to lethal mutations in *xan-h.clo125* (D274N) and *xan-h.clo157* (R356K), are also located at the ATP-binding pocket. In particular, D274 and R356 belong to different alpha-helices of the same chain and interact with each other (Fig. 6C and Fig. S6A), establishing an intra-monomer hydrogen bond network that includes R393. The disruption of this interaction could affect the folding of the monomer and thus the formation of CHLI dimer (Fig. 5B and Fig. 6C). On the other hand, the R298K substitution does not lead to either the loss or formation of significant intra- or inter-monomer contacts (Fig. 6C). To further investigate the possible role of R298 in the context of the ATP binding site, the structural analysis was extended to AAA + proteins that are not related to photosynthesis. In particular, the hexameric structure of the chaperone Heat Shock Locus U [HSLU, PDB ID: 1DO0; (Bochtler et al. 2000)] from *Escherichia coli*, which has been resolved by X-ray analysis with its magnesium ion and ATP in the binding cleft, was utilised to this purpose. The high degree of homology between the ATP-binding domain of HSLU and the barley *Hv*CHLI model enabled us to infer distances between the magnesium ion and the conserved residues in both structures (Fig. S6B). With this information, the magnesium ion was positioned in the binding cleft of the model, and the distances were used to constrain the ATP docking mode (Fig. 6B, C). The results show that R356, whose homologue in *Synechocystis* has been described as part of the S-2 motif, can establish direct hydrogen-bond interactions with ATP (Fig. 6C, left panel), as does the Arg finger R275 (R210 in *Synechocystis*). Hydrogen bonds could also be established

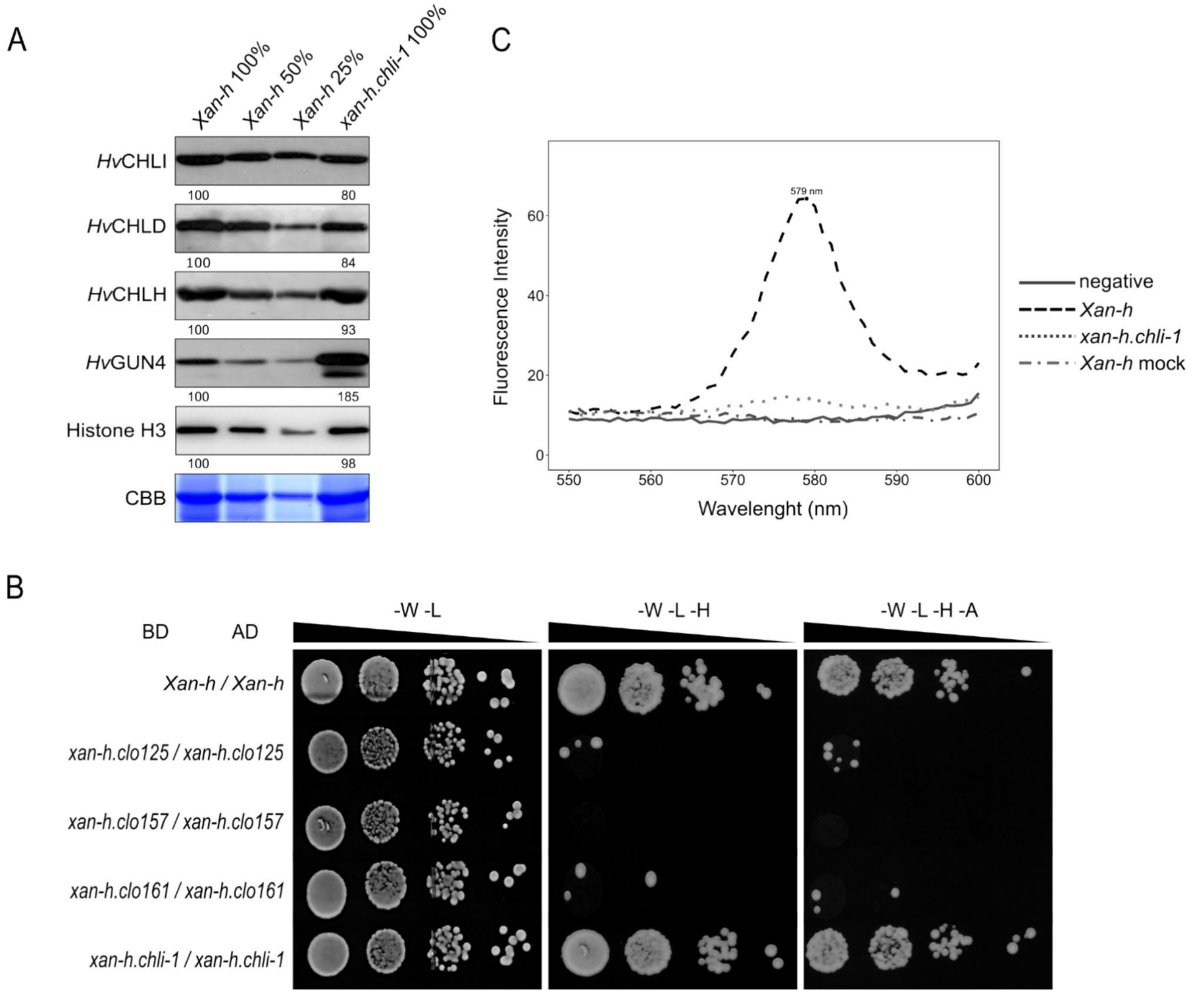


Fig. 5 Effects of the *xan-h.chli-1* mutation on the accumulation, assembly and activity of the Mg-Chelatase complex. **A** Immunoblot analyses of total protein extracts (normalized to leaf fresh weight) from *Xan-h* and *xan-h.chli-1* plants with antibodies specific for *Hv*CHLI, *Hv*CHLD, *Hv*CHLH and *Hv*GUN4, respectively. A CBB-stained gel corresponding to the RbcL region, and an immunoblot showing the histone H3 protein are shown as controls for equal loading. For protein quantification, 50% and 25% dilutions of *Xan-h* protein extracts were also loaded. One representative of three biological replicates is shown for each immunoblot. **B** Yeast two-hybrid interaction assays were performed on *Xan-h* and the mutant allelic variants *xan-h.chli-1*, *xan-h.clo125*, *xan-h.clo157* and *xan-h.clo161* in order to test their ability to self-interact (homodimerization). As highlighted by their growth on selective media (-W-L-H and -W-L-H-A), only the colonies expressing the wild-type *Xan-h* and its mutant *xan-h.chli-1* alleles were able to self-associate. BD, GAL4 DNA-binding domain, AD, GAL4 activation domain, -W –L, dropout medium devoid of Trp and Leu (permissive medium); -W -L -H, lacking Trp, Leu and His (selective medium), and -W -L -H -A, lacking Trp, Leu, His and Ade (selective medium). Serial dilutions were prepared for each strain. **C** In vitro assay of Mg-chelatase activity in etiolated leaf extracts from *Xan-h* and *xan-h.chli-1*. The fluorescence emission of the Mg-chelatase product Mg-deuteroporphyrin was recorded from 550 to 600 nm using an excitation wavelength of 408 nm. Protein extracts were normalized to total protein content. One representative chart (of three biological replicates) is shown

with the ATP molecule by the side-chain of R298 (Fig. 6C, left panel). This interaction could be perturbed or prevented by the R298K missense mutation, owing to the shorter side-chain and the presence of only one amino group in lysine (Fig. 6C, right panel). Overall, these observations suggest that R298K missense mutation might compromise the interaction of *Hv*CHLI dimers with ATP.

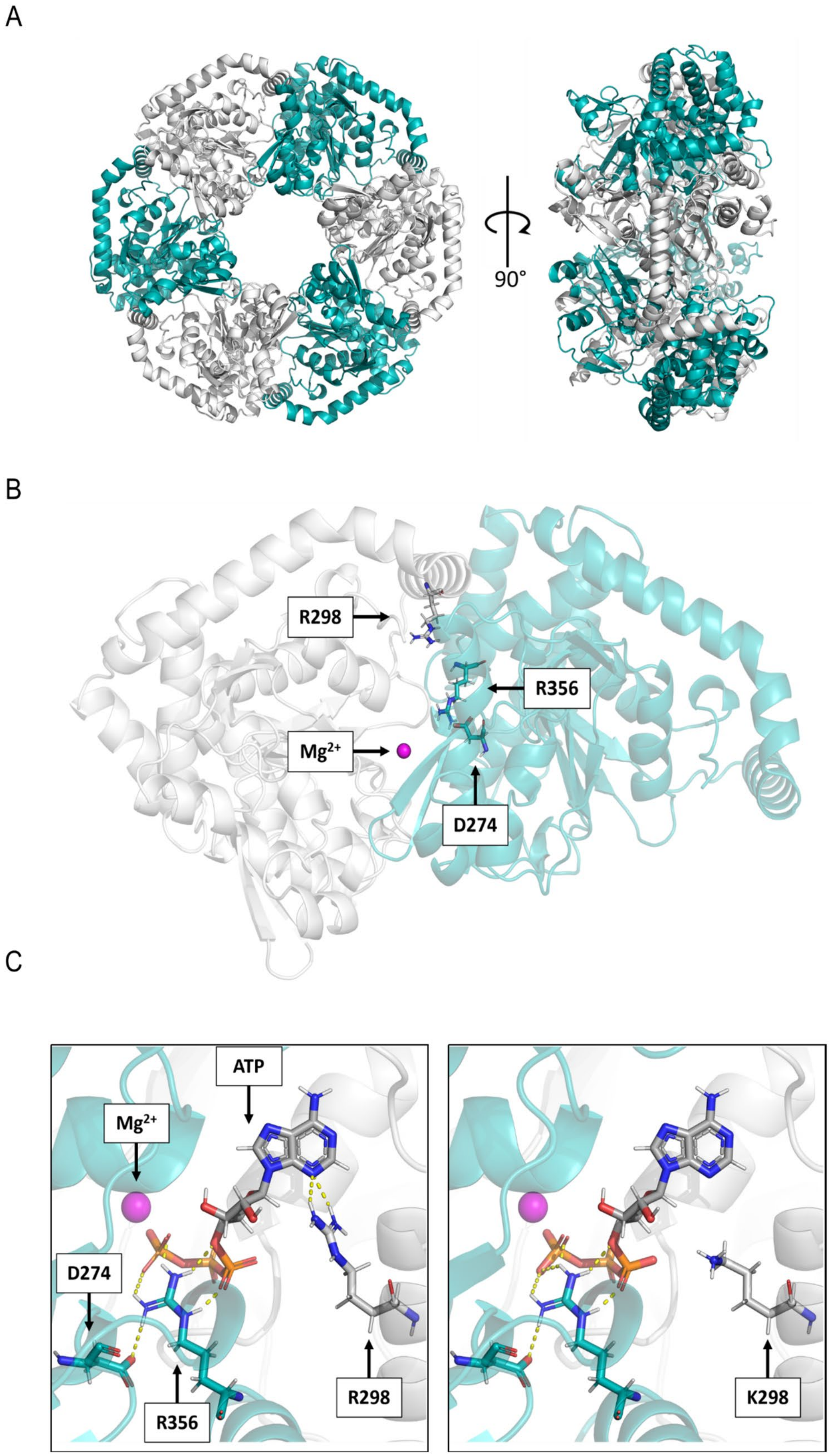


Fig. 6 Effect of the *xan-h.chli-1* mutation on the *Hv*CHLI hexamer structure. **A** Model of the AAA + ATPase subunit of the barley Mg-chelatase enzyme. The adjacent monomers are coloured in white and cyan. Left panel: frontal view of the homo-hexameric ring shown in cartoon representation; Right panel: side view of the ring in cartoon representation. **B** Frontal view of a single dimer. The two monomers are represented in transparent cartoon and coloured in white (chain A) and cyan (chain B), respectively. The constituent atoms of R298 in chain A are depicted in light grey (C atoms), blue (N), red (O), and white (H). D274 and R356 (S-2) of chain B are shown in cyan (C), blue (N), red (O), white (H) and represented as solid sticks. R298 in chain B, D274 and R356 from chain A are not highlighted. Mg^{2+} is shown as a magenta sphere. **C** ATP binding cleft with Mg^{2+} and docked ATP. The same colour scheme is used for D274 and R356 of chain B, with ATP shown in dark grey and P atoms in orange H-bonds are represented as yellow dashed lines

The reduced Mg-chelatase activity in *xan-h.chli-1* plants does not affect plastid-to-nucleus retrograde signaling or the expression of photosynthesis-associated nuclear genes

Since Mg-chelatase activity is markedly reduced in *xan-h.chli-1* chloroplast, we investigated the possibility that Arabidopsis and barley plants carrying the *xan-h.chli-1* allele might show the *genomes uncoupled* (*gun*) phenotype. To do so, barley *Xan-h*, *xan-h.chli-1* and *xan-h.56* seedlings were grown on MS medium in the presence or absence of norflurazon (NF), and levels of *Lhcb3* and *Rbcs* transcripts were determined. As shown in Fig. S7A, *Xan-h* and *xan-h.chli-1* seedlings were able to down-regulate the expression of *Lhcb3* and *Rbcs* genes in the presence of NF, indicating that *xan-h.chli-1* mutant does not display the *gun* phenotype, unlike the lethal *xan-h.56* allele used here as a positive control (Gadjieva et al. 2005). Similarly, Arabidopsis lines carry either the *Xan-h* or the mutant *xan-h.chli-1* allele from barley under the control of *35SCaMV* promoter markedly reduced the expression of *Lhcb3* and *Rbcs* genes in the presence of NF, like Col-0 and the *cs* mutant. As expected, the *gun5* mutant failed to repress the expression of *Lhcb3* and *Rbcs* genes, supporting the notion that the *xan-h.chli-1* mutation does not affect plastid-to-nucleus communication (Fig. S7B).

To further investigate the impact of the *xan-h.chli-1* mutant allele on leaf gene expression, a transcriptomic analysis was performed on *Xan-h* and *xan-h.chli-1* leaves obtained from plants grown under greenhouse conditions. Principal component analysis (PCA) revealed that the four transcriptome replicates of each genotype clustered together in two clearly separated groups (Fig. 7A). Moreover, differentially expressed genes (DEGs) were identified by filtering for the log-fold-change (logFC) and the adjusted p-value (padj), which resulted in the identification of 432 up-regulated and 335 down-regulated genes in *xan-h.chli-1* relative to *Xan-h* (Fig. 7B; Supplementary data). The relatively small number of DEGs agrees with the moderate distance between the two PCA clusters. This observation, together with the fact that Biological Process Gene Ontology (GO) term analysis resulted in no significant GO term enrichment, indicates that the *xan-h.chli-1* mutant allele does not cause major changes at the transcriptional level with respect to its *Xan-h* counterpart. SUBA5 location prediction was applied to the Arabidopsis homologs of the up and down-regulated genes. Among the up-regulated DEGs, 298 were found in the SUBA5 database and most of the encoded proteins were predicted to be active in the plasma membrane (23%), nucleus (21%) and cytosol (19%), while only 8% were targeted to plastids (Fig. 7C).

The majority of the 177 down-regulated genes found in the SUBA5 database were also predicted to be localized to the plasma membrane (27%), the nucleus (24%) or the cytosol (16%), while only 7% of the genes encoded plastid proteins, further confirming the limited impact of the *xan-h.chli-1* mutation on chloroplast functionality.

In light of the localization of the Mg-chelatase enzyme and the pale green phenotype of mutant plants, the chloroplast-related DEGs were analysed in detail. Twenty-three up-regulated nuclear genes were predicted or reported to encode proteins active in the chloroplast (Table S2). These included the *ATNTH1* gene encoding a DNA glycosylase-lyase involved in base excision repair of oxidative DNA damages and an M-type 4 thioredoxin with a role in the oxidative stress response. Genes coding for proteins with a role in jasmonic-acid-mediated stress responses, such as lipoxygenae 2 (*LOX2*), the lipase *DALL4*, and the allene oxidase synthase (*AOS*), and in drought and heat-stress responses, as in the case of *Heat Shock Protein 21* (*HSP21*) and *TRR14,* were also upregulated (see Table S2). In addition, several upregulated genes are reported to play a role during the early stages of chloroplast and seedling development, including early light-inducible protein 1 (*ELIP1*), plastid transcriptionally active chromosome 18 (*pTAC18*), raspberry 3 (*RSY3*), and arogenate dehydratase 6 (*ADT6*). Furthermore, the plastid type I signal peptidase 1 (*PLSP1*) and the plastid type I signal peptidase 2B (*PLSP2B*), which remove signal sequences from proteins translocated into the thylakoid lumen, were also upregulated.

The thirteen chloroplast-located down-regulated genes are mainly involved in protein folding and assembly, such as FK506-binding protein 13, a peptidyl-prolyl isomerase located in chloroplast thylakoid lumen which is considered to act as a protein folding catalyst, and *RAF2*, Rubisco Assembly Factor 2, in fatty acid and lipid biosynthesis, such as Acyl Activating Enzyme 16 (*AAE16*) and UDP-sulfoquinovose:DAG sulfoquinovosyltransferase 2 (*SQD2*), respectively, and during early developmental stages, as in the case of *BE1*, a putative glycoside hydrolase that plays a vital role during embryogenesis and in carbohydrate metabolism, and Late Embryogenesis Abundant (*LEA*) hydroxyproline-rich glycoprotein, which is thought to function in plant development and growth. Strikingly, none of the nuclear genes encoding subunits of the thylakoid photosynthetic apparatus were among those differentially regulated by the *xan-h.chli-1* mutant allele.

Moreover, neither the Mg-chelatase subunits nor enzymes involved in chlorophyll biosynthesis were found to be differentially expressed. On the other hand, the genes Early light-induced protein 1 (*HORVU.MOREX.r3.5HG0482040*), which prevents excess accumulation of free chlorophyll by inhibiting the entire chlorophyll biosynthesis pathway, and the Chlorophyllase-2 (*HORVU.MOREX.r3.3HG0235790*), involved in chlorophyll degradation, were found to be

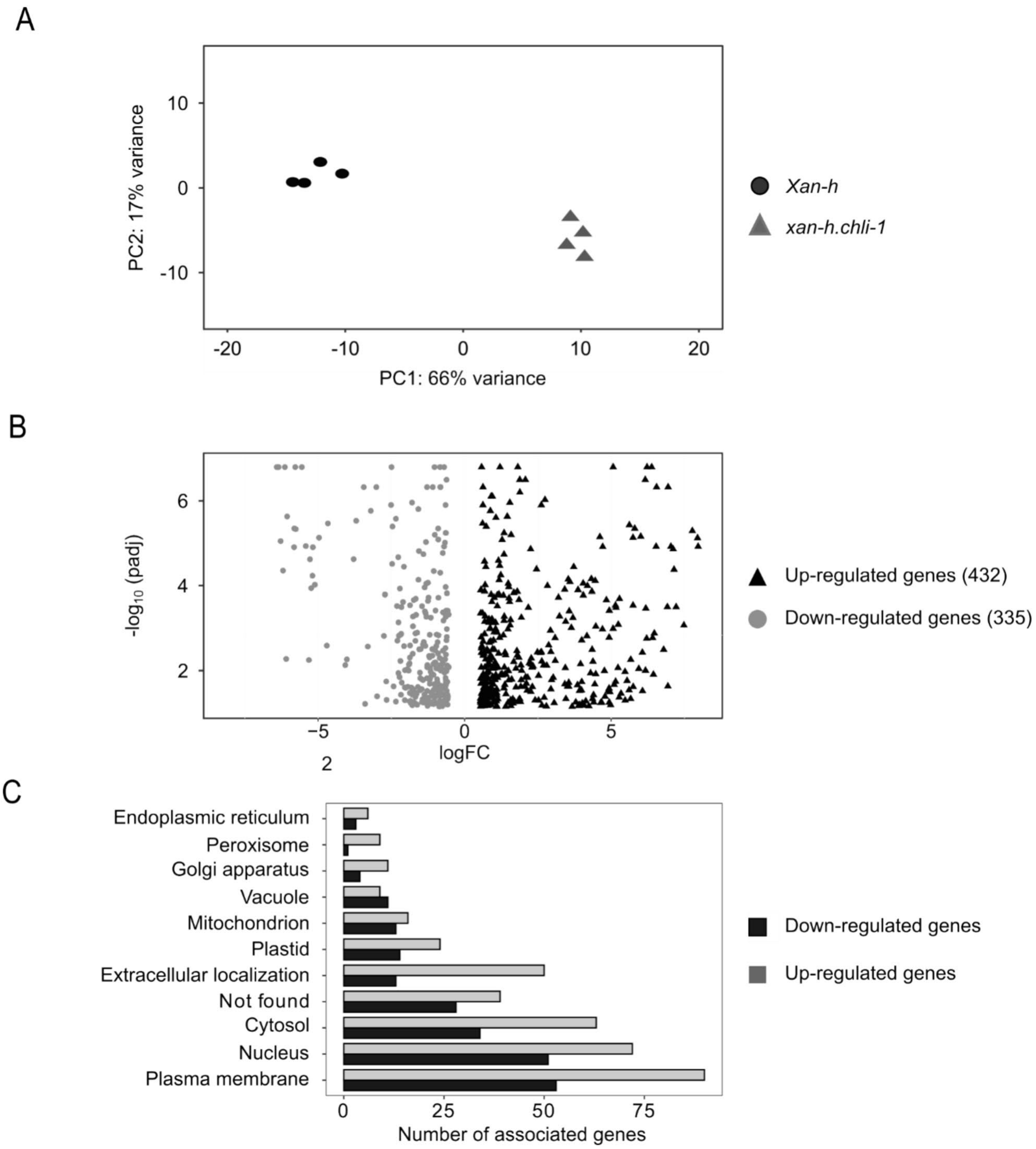


Fig. 7 Comparative transcriptomic analyses of *xan-h.chli-1* and *Xan-h* leaves grown under greenhouse conditions. **A** Principal component analysis (PCA) of the four biological replicates for each genotype. **B** Volcano plot of the differentially expressed genes (DEGs) filtered by the log of fold change (logFC) and the adjusted p-value (padj). **C** Subcellular localization of DEGs based on information available in the SUBA5 database (https://suba.live/)

up-regulated in the mutant. Three ABA-related genes (*HORVU.MOREX.r3.6HG0616980, HORVU.MOREX.r3.6HG0622710, HORVU.MOREX.r3.3HG0288800*) were also up-regulated in *xan-h.chli-1* samples, together with the nitrate transporter NPF6.3 (*HORVU.MOREX.r3.7HG0700030*) that functions in the stomatal opening (Guo et al. 2023).

The *xan-h.chli-1* mutant line is characterised by reduced daily transpiration rate

To investigate the growth advantages associated with the pale green leaf phenotype, *xan-h.chli-1* and *Xan-h* plants were grown in pots in Plantarray, a functional phenotyping platform (FPP), in a semi-controlled environmental greenhouse, with the aim of detecting small changes in specific physiological processes under both optimal and limiting watering regimes (Lupo and Moshelion 2024). Plant biomass and water flux measurements performed throughout the entire plant life cycle, from January 19^{th} to March 2^{nd}, 2023, allowed for calculations of transpiration and biomass gain of each plant (Appiah et al. 2023). Plants were initially grown under well-watered conditions for 23 days, followed by an 18-day period of drought-stress, till the late stem elongation stage, and then returned to standard conditions until harvesting, at early inflorescence emergence. Under well-watered conditions, the daily transpiration rate normalized to plant fresh weight was significantly lower in *xan-h.chli-1* plants (from 10 to 55% reduction) than in the *Xan-h* control, as shown in Fig. 8A, where data collected during 14 days are shown. Similar differences were observed during the drought stress period (Fig. 8B), where *xan-h.chli-1* plants showed a stable reduction in the daily transpiration rate, i.e. around 40–50% less than the *Xan-h* control. Strikingly, towards the end of the drought-stress period, when plants were under severe water deficiency, *Xan-h* daily transpiration rate decreased severely reaching values significantly lower than those recorded for *xan-h.chli-1*, most probably as a consequence of the fact that *xan-h.chli-1* plants were better able to tolerate the drought stress (Fig. 8B). However, this general reduction in transpiration rate came at the expense of total biomass accumulation (*xan-h.chli-1* 16.52 ± 4.42 gr *vs. Xan-h* 25.34 ± 2.57 gr; Fig. 8C) and water use efficiency (WUE), i.e. biomass gain per ml of transpired water (*xan-h.chli-1* 0.00364 ± 0.00064 gr ml^{-1} *vs Xan-h* 0.00446 ± 0.00036 gr ml^{-1}; Fig. 8D).

Discussion

The manipulation of leaf pigment content has been reported to enhance light use efficiency in high-density monocultures (Kirst et al. 2017). Many genetic targets are available for the alteration of leaf chlorophyll levels, as reviewed in Cutolo et al. 2023. Recently, we reported on the barley mutant *happy under the sun 1* (*hus1*), which is characterised by a 50% reduction in the chlorophyll content of leaves, owing to a premature stop codon in the *HvcpSRP43* gene that codes for the 43-kDa chloroplast Signal Recognition Particle (Rotasperti et al. 2022). However, when sown at standard density under field conditions, the yield of *hus1* plants was comparable to that of the wild type, implying that the reduction of leaf chlorophyll content is well tolerated in crops.

The *xan-h.chli-1* barley mutant phenotype is due to the reduced activity of Mg-chelatase enzyme

In the present work, we have characterized a novel chlorophyll-deficient mutant in barley. This pale-green mutant, *xan-h.chli-1,* is due to a missense mutation (R298K) in a highly conserved residue of the *Hv*CHLI protein—the smallest subunit of the Mg-chelatase enzyme, which catalyses the first unique step in chlorophyll biosynthesis (Lundqvist et al. 2010, 2013). Interestingly, while all of the mutants previously described at the *Xan-h* locus in barley (*xan-h.38*, *xan-h.56*, *xan-h.57* and *xan-h.clo125*, *xan-h.clo157*, *xan-h-clo161*) show a seedling-lethal phenotype (Hansson et al. 1999; Braumann et al. 2014), the *xan-h.chli-1* line is viable. This unique phenotype is due to the fact that the R298K missense mutation does not dramatically affect the accumulation of *Hv*CHLI, nor its ability to form homodimers, but rather results in a drastic reduction of the Mg-chelatase activity, as experimentally verified in vitro. The fact that the *xan-h.clo125*, *xan-h.clo157* and *xan-h.clo161* variants do not form homodimers in yeast two-hybrid assays, while the corresponding variants of the *Rhodobacter capsulatus bchI* gene oligomerize on a gel-filtration column in the presence of ATP (Hansson et al. 2002), may be ascribed to the relatively low homology (49% identity) between their amino-acid sequences.

The introgression of the *xan-h.chli-1* mutant allele into the lethal *Atchli1/Atchli1* mutant background (Huang and Li 2009) fully restored plant viability and reverted the albino *Atchli1/Atchli1* phenotype to a milder pale-green leaf colour, similar to those of the Arabidopsis *cs/cs* (Kobayashi et al. 2008) and barley *xan-h.chli-1* mutant phenotypes, confirming further that this ATPase motor, found in all kingdoms of living organisms, shares a common core structure and function (Ogura and Wilkinson 2001; Gao et al. 2020; Cha et al. 2010; Miller et al. 2014). Furthermore, the *xan-h.chli-1* allele also attenuated the lethal *xan-h.56* phenotype, as the biallelic mutant *xan-h.56/xan-h.chli-1* shows a pale green leaf phenotype and PSII functionality is comparable to that of *xan-h.chli-1* leaves. Since the homozygous *xan-h.56* mutant does not accumulate the *Hv*CHLI protein (Braumann et al. 2014), the pale-green phenotype of the heterozygote is attributable to the *xan-h.chli-1* allele alone. Conversely, the *xan-h.clo161* barley mutant showed a reduction in *Hv*CHLI accumulation relative to the wild-type, and the semi-dominant nature of this mutation indicates that the protein encoded by the *xan-h.clo161* allele has detrimental effects on the assembly and activity of the *Hv*CHLI hexamer (Hansson et al. 2002).

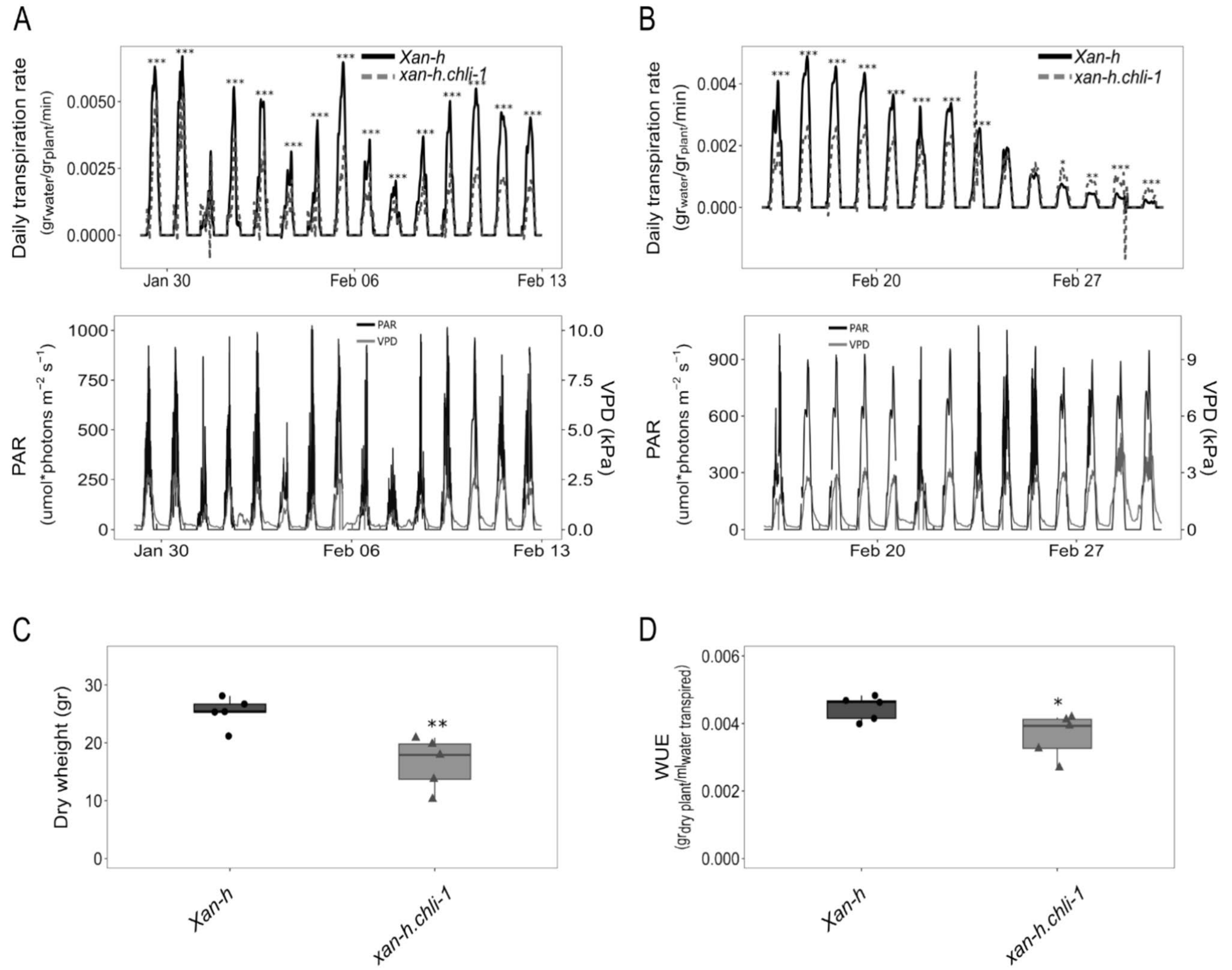


Fig. 8 Relative performance of *Xan-h* and *xan-h.chli-1* plants grown under optimal and drought-stress conditions, as estimated by the FPP phenotyping platform. **A** Upper panel: Daily transpiration rate normalized to plant fresh weight (g water/g plant/min) as evaluated for 14 days under well-watered conditions during daylight exposure from 6.00 am to 18.00 pm. To avoid overloading the Figure, data obtained during the night period are not shown. Lower panel: Photosynthetic active radiation intensities (PAR) and vapour pressure deficit (VPD) measured by a weather station for the 14 representative days under well-watered conditions. **B** Upper panel: daily transpiration rate normalized to plant fresh weight (g water/g plant/min) as evaluated for 14 days under drought-stress conditions, automatically maintained through the feedback-controlled irrigation system, during daylight exposure from 6.00 am to 18.00 pm. Lower panel: photosynthetic active radiation intensities (PAR) and vapour pressure deficit (VPD) measured by a weather station for the 14 days under drought-stress conditions. In all cases, the significance of the data was estimated using Student's *t*-test (***$P < 0.001$, **$P < 0.01$, *$P < 0.05$). **C** Plant dry weight (g) at the end of the experiment, i.e. upon completion of plant life cycle. The plant material was dried at 60 °C for 72 h. The significance of the observed differences was evaluated with Students *t*-test (** $P < 0.01$). **D** Water use efficiency (WUE) (g dry plant/ml water transpired) was measured by the total weight of dry plants at the end of the life cycle, normalized to total water transpired. Student's *t*-test was performed to estimate the significance of the observed differences (* $P < 0.05$). Average data of five biological replicates are shown

This accounts for the accumulation of both protein variants in the barley biallelic mutant *xan-h.clo161*/*xan-h.chli-1*, in which the interaction of the two variants within the *Hv*CHLI hexamer most probably results in a non-functional Mg-chelatase and a lethal *Chlorina*-like phenotype.

Our findings are also in agreement with the localization of the R298 residue, which is predicted to reside in the ATP-binding pocket and to interact directly with the ATP molecule. The model displays conformational similarity to the recently published CHLI hexamer structure from *Synechocystis* sp. PCC 6803 (PDB ID 6L8D, Gao et al. 2020), which exhibits 73% sequence identity with WT *Hv*CHLI. In particular, the altered interactions caused by the R298K substitution in the ATP-binding pocket suggest that the R298 residue is involved in either ATP binding and/or ATP hydrolysis.

The reduced chlorophyll content in *xan-h.chli-1* leaves increases photosynthetic efficiency under high light conditions

As expected, pale green *xan-h.chli-1* leaves showed a reduced content of both Chls *a* and *b* and an increase in the Chl*a*/Chl*b* ratio when compared to control plants. Furthermore, comparable reductions were observed in the accumulation of carotenoids, including β-carotene which is preferentially associated with the PSI and PSII cores (Caffarri et al. 2014). This indicates that—unlike the alteration of antenna protein biogenesis in *hus1* mutant (Rotasperti et al. 2022)—the impairment of chlorophyll biosynthesis leads to a general destabilization of the entire photosynthetic apparatus, as is confirmed by the reduced accumulation of antenna proteins and photosystem core subunits observed by immunoblots. In addition, the reorganization of electron transport in the thylakoids of *xan-h.chli-1* leaves appears to take place at the post-transcriptional level, as transcriptomic analysis revealed that none of the nuclear genes encoding photosynthesis-associated proteins were affected in the mutant.

Intriguingly, the reduction of antenna size in *xan-h.chli-1* leaves, together with the decline in photosystem core proteins, decreased the efficiency of photosynthesis only under low light intensities, whereas photosynthetic performance was enhanced relative to WT under high light levels, as already reported in the case of *hus1* plants (Rotasperti et al. 2022). This is most probably due to the reduction in thylakoid excitation pressure in *xan-h.chli-1* leaves exposed to high-light intensities, as indicated by the lower values of Y(NPQ) and Y(ND) parameters and the lower reduction of the abundance of D1 and D2 PSII core subunits observed upon exposure to highlight conditions (see Fig. S4). In this context, the pale green phenotype associated with the *xan-h.chli-1* mutant allele deserves to be investigated for its performance under field conditions since this trait has been reported to favour a more equal distribution of light under high-density field conditions with potential benefits for net photosynthetic efficiency across the entire canopy and grain yield (Kirst et al. 2018), as well as for the efficiency of nitrogen use (Walker et al. 2018; Sakowska et al. 2018). Unlike the soybean mutant *MinnGold* (Sakowska et al. 2018), characterised by a marked decrease in leaf chlorophyll content and biomass production under field conditions, *xan-h.chli-1* shows, indeed, a milder reduction of leaf chlorophyll content during the vegetative phase, while the chlorophyll content of the flag leaf, which contributes largely to grain yield (Niu et al. 2022), is almost identical to control plants.

The *xan-h.chli-1* allelic variant does not alter the chloroplast-to-nucleus retrograde communication and reduces the daily transpiration rate

Independent studies have described the Mg-chelatase to have a role in chloroplast-to-nucleus retrograde communication (Mochizuki et al. 2001; Larkin et al. 2003). However, *xan-h.chli-1* seedlings do not show the *genomes uncoupled* phenotype in the presence of Norflurazon, unlike seedlings that carry lethal allelic variants (Gadjieva et al. 2005). This is advantageous, since retrograde signalling plays a crucial role in the adaptation of plants to changing environments, and several *gun* mutants show impaired responses and heightened sensitivity to abiotic challenges (Song et al. 2018; Marino et al. 2019).

Moreover, *xan-h.chli-1* plants are characterised by a significantly lower transpiration rate, at the expense of total biomass accumulation and WUE, possibly due to the reduction of leaf temperature, predicted to be associated with the reduced leaf chlorophyll content (Drewry et al. 2014). Alternatively, the reduced daily transpiration rate might be the consequence of a marked decrease of Mg-chelatase activity together with the slight reduction of Mg-chelatase abundance, including CHLH subunit reported to bind abscisic acid (ABA) and to function in ABA signalling and stomatal movement (Shen et al. 2006; Wu et al. 2009). Furthermore, a specific role in the modulation of ABA signalling in guard cells was attributed to the CHLI subunit (Du et al. 2012).

This behaviour, combined with the reduction in Chl*b* leaf content, resembles the high-risk drought escape strategy adopted by certain wild barley accessions (*Hordeum vulgare* spp. *spontaneum*) that are adapted to stable and very dry environments where fitness (i.e. reproductive output and the quality of offspring) prevails over the achievement of the full production potential (Galkin et al. 2018). This finding supports the notion that pale-green leaves may have beneficial effects in harsh environments, as in the case of certain Syrian barley landraces and a few accessions of wild barley (*Hordeum vulgare* spp. *spontaneum*) in Israel, that grow under arid climatic conditions and are characterized by a pale green phenotype (Watanabe and Nakada 1999; Tardy et al. 1998; Galkin et al. 2018).

Overall, while crop breeding has led to the development of high-yielding cultivars, progress toward the development of crops that tolerate abiotic stresses has been very slow. Thus, the need to reduce the 'yield gap' and improve yields under a variety of stress conditions is of strategic importance for future food security (Sadras and Richards 2014; Cattivelli et al. 2008; Araus et al. 2002). In this context, the *xan-h.chli-1* mutant allele and its pale green phenotype have a potential for application in breeding programs that deserves to be investigated. To this end, the introgression of the *xan-h.chli-1* allele into elite barley cultivars, and collaboration with plant breeders, agronomists and crop physiologists to select the most appropriate yield-testing protocols, including te definition of growing plant densities and standard parameters to define yields are needed. Finally, in the medium term, the knowledge gained could be transferred to other cereals, including wheat, given the high degree of conservation of the chlorophyll biosynthetic pathway and the photosynthetic machinery in higher plants.

Materials and methods

Nucleotide and amino acid sequence analysis

Amino-acid and genome sequences of *Xan-h* (*HORVU.MOREX.r3.7HG0738240*), *AtCHLI1* (*At4g18480*) and *AtCHLI2* (*At5g45930*) were obtained from the ENSAMBLE-Plant database (plants.ensembl.org/index.html). Multiple sequence alignments were obtained locally with Muscle v5 (drive5.com/muscle5/) (Edgar 2022). Subcellular localization and chloroplast transit peptide (cTP) predictions were identified by TargetP (services.healthtech.dtu.dk/services/TargetP-2.0/).

Plant material and growth conditions

Barley (*Hordeum vulgare*) plants were cultivated on acid soil (Vigor plant-growth medium, based on Irish and Baltic peats, pH 6.0; pot volume 2.5 L; 2 plants per pot) supplemented with Osmocote fertilizer under controlled greenhouse conditions (around 500–600 µmol photons m^{-2} s^{-1} for 16 h and 8 h dark; GreenPower LED toplighting linear—Phillips). The greenhouse is located at the Botanical Garden "Città Studi" of the Univeristy of Milano (45°28′32.2″N—9°14′05.0″E). Temperatures were set to 20 °C during the day and 16 °C at night, with a relative humidity of 60%. High-light exposure was conducted by growing plants under control conditions (around 400–500 µmol photons m^{-2} s^{-1}) and transferring them to high light (around 1200–1400 µmol photons m $^{-2}$ s $^{-1}$) for 0.5 and 8 h.

Only in the case of Arabidopsis, Columbia-0 (Col-0) and mutant lines were grown on soil (acid sphagnum peat, Atami Bio-Gromix; pot volume 0.5 L; 5 plants per pot) in a climate chamber (Percival CLF AR-66L; 150 µmol photons m^{-2} s^{-1} for 16 h, and 8 h dark, 22 °C and a relative humidity of 60%), placed at the Department of Biosciences of University of Milano (45°28′35.6″N 9°14′02.0″E).

The barley *TM2490* line was identified among the M_4 generation of the chemically mutagenized TILLMore population (Talamè et al. 2008), which is derived from the 'Morex' cultivar background. Around 4000 M4 lines, grown under field conditions at the experimental farming facility in Cadriano, Bologna, Italy (44°33′00.0″N 11°23′39.0″E) during the growth season 2018–2019, were screened based on their photosynthetic performance [Y(II) values], using the Handy PEA fluorometer (Hansatech Instruments Ltd., UK), and on their leaf apparent chlorophyll content, using the SPAD-502 chlorophyll meter (Konica-Minolta, Tokyo, Japan).

The F_2 segregating population, generated for mapping purposes, was obtained by manually crossing the *TM2490* line with the cv. Barke. The *xan-h.chli-1* line was isolated from an F_2 population obtained by backcrossing *TM2490* with the barley *cv.* Morex (BC_2F_2). The Arabidopsis *Atchli1/Atchli1* T- DNA insertion mutant (*SAIL_230_D11*) was identified by searching the T-DNA Express database (signal.salk.edu/cgi-bin/tdnaexpress), while the homozygous line *cs/cs* was provided by Professor Tatsuru Masuda (Kobayashi et al. 2008). The transgenic Arabidopsis lines *Atchli1/Atchli1 + 35S::Xan-h* and *Atchli1/Atchli1 + 35S::xan-h.chli-1* were generated by *Agrobacterium*-mediated transformation of the heterozygous *Atchli1/AtCHLI1* mutant line with either the wild-type *Xan-h* or the mutant *xan-h.chli-1* coding sequence from barley, respectively, using the *pB2GW7* plasmid (VIB-UGhent for Plant Systems Biology). Primers used for mutant isolation and cloning procedures are listed in Table S1.

The phenotyping platform

The functional-phenotyping platform Plantarray (FPP; PlantDitech Ltd.; Yavne, Israel) was used to monitor plant growth and water balance. The Plantarray is a high-throughput functional phenotyping platform that continuously and simultaneously measures water flux in the soil–plant–atmosphere continuum. The system consists of individual, highly sensitive balances, each connected to its own control unit. As each measurement unit is connected to the water and fertilizer tank separately, individual irrigation and fertilization regimes are controlled. Every 3 min, the weight of the whole system (i.e., pot, plant, and sensors) is recorded and through internal calculations plant net weights, and a set of additional physiological plant parameters [e.g., daily transpiration (dTR), transpiration rate, and volumetric soil water content (SWC)] is obtained. Moreover, environmental factors are monitored to calculate vapor pressure deficit (VPD) throughout the experiment and to understand the influence of these environmental factors in the transpiration and other physiological parameter measured on the plants. The data are made accessible in real time via the online analysis tool (SPAC Analytics), which can also be used for data visualization and analysis. The installed feedback irrigation system allows the user to establish a standardized drought treatment allowing for comparisons between the plants. Exposing all plants to similar drought stress is possible by taking into account each plant's transpiration rate, e.g., by re-irrigating only a certain percentage of the previous day transpiration. This mimics the gradual development of soil water deficits in the field (Dalal et al. 2020). To ensure that occurring water loss was solely due to plant transpiration, we covered the soil with a styrofoam sheet to prevent soil evaporation. A more detailed description of the system and the underlying theory can be found in Dalal et al. (2020). The sensors include the HC2-S3-L meteo probe for relative humidity and temperature in the greenhouse (Rotronic, Crawley, United Kingdom), LI-COR 190 Quantum Sensor for photosynthetically active radiation measurements (Lincoln, NE, United States), and a soil moisture, electro-conductivity and temperature sensor (5 T, Decagon devices, Pullman, WA, United States) incorporated in every pot. All plants were exposed to comparable drought stress by taking each plant's transpiration rate into account, as previously described (Dalal et al. 2020). Plants were grown on the Plantarray system (pot volume 3 L; one seedling per pot; potting mix Ökohum" containing plant compost, peat, and perlite; Organic matter = 80%, 160 mg/L N, 120 mg/L P2O5, and 320 mg/L K20, pH = 5.8), under greenhouse semi-controlled conditions (Rehovot—31°53′53.7″N 34°48′24.9″E), for 43 days from January 19 to March 2, 2023. During the whole experiment, artificial light (400-W MT400DL/BH) was provided from 5:30 to 18:00 and the temperature was maintained at 23 °C/18 °C. The VPD fluctuated between 0.8 kPa and 2.4 kPa (mean, 1.7 kPa). During the pre-drought phase of 24 days, all plants were well-watered at pot capacity through nocturnal irrigation. Drought conditions were progressively imposed from February 12 to March 2 (18 days) by gradually reducing the daily irrigation to 80% of the plants' own previous day transpiration level. After 10 more days on the Plantarray (recovery phase, in which irrigation followed the well-watered regime again), plants were moved to the greenhouse and, upon harvest, total dry biomass weight was measured. To determine the dry weight, the plant material was dried at 60 °C for 72 h (Appiah et al. 2023). Whole-plant transpiration rates were derived by multiplying the first derivative of the measured load-cell time series by − 1. The transpiration was then normalized to the plant's fresh weight. Water-use efficiency (WUE) was calculated as harvest product dry weight (g)/water transpired over the entire Plantarray growth period (ml) (Jaramillo Roman et al. 2021).

RNAseq analyses for mapping and differential gene expression

For gene mapping analyses, RNA was isolated from 100 wild-type-like and 100 *TM2490*-like F_2 plants (*TM2490* × *cv* Barke) and extracted as previously described (Verwoerd et al. 1989). Independent RNA samples were bulked in an equal ratio to generate two pools. RNA pools were subjected to poly-A capture and paired-end sequencing, producing approximately 100 million 2 × 150-bp read pairs per pool (47.97 M and 48.89 M 2 × 150 nt paired-end reads, for wild-type-like and *TM2490*-like pools, respectively). Reads were mapped to the *H. vulgare* Morex v3 genome sequence (Mascher et al. 2021). Coherent mapping was obtained for 84.5% and 86.1% of pairs for the wild-type-like and *TM2490*-like pools, respectively. Samtools (Danecek et al. 2021) was used to sort and index the resulting alignment files, and FreeBayes (v1.3.2) (Garrison and Marth 2012) was used to call variants, employing default parameters except for requiring a minimum mapping quality of 20 and a minimum base-call quality of 30. Custom Python scripts were employed to identify variants segregating between pools (allele frequency > 0.1 in both pools and < 0.9 in the wild-type-like pool). Variants were binned in 2-Mb intervals along the barley genome. Variants in this region (chr7H: 592500000..605500000) were analysed using the Ensembl-vep pipeline (McLaren et al. 2016).

For quantitative transcriptomic analyses, total RNA was isolated from 14-day-old leaf samples obtained from four biological replicates each of *Xan-h* and *xan-h.chli-1* plants. Differential gene expression analysis was performed in R using the DESeq2 package (Love et al. 2014). DEGs were filtered for log of fold change (logFC) > 0.5 and an adjusted p-value (padj) < 0.05. The ncbi-blast-2.14.0 + tool (ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) was used to perform identifier mapping on the barley genes (i.e., proteins in *A. thaliana* that appear to match the input protein sequences of *H. vulgare cv*. Morex with a given percentage of identity). The subcellular localization was predicted with SUBA5 (suba.live/index.html). Biological Process Gene Ontology (GO) term enrichment was performed with agriGO v2.0 (systemsbiology.cau.edu.cn/agriGOv2/species_analysis.php?SpeciseID = 1&latin = Arabidopsis_thaliana) (Tian et al. 2017). The raw RNASeq data have been deposited in the NCBI data repository (https://submit.ncbi.nlm.nih.gov/subs/bioproject/) under the bioproject identifier PRJNA1052990.

Assay for *genomes uncoupled* phenotype

Barley and Arabidopsis seeds were surface-sterilized and grown for 6 days (100 μmol photons m^{-2} s^{-1} on a 16 h/8 h light/dark cycle) on Murashige and Skoog medium (Duchefa, Haarlem, The Netherlands), supplemented with 2% (w/v) sucrose and 1.5% (w/v) Phyto-Agar (Duchefa). To discriminate the homozygous *xan-h.56* mutants, the barley seedlings were then transferred onto MS media supplemented with 5 μM NF, while Arabidopsis seeds were grown directly on NF-supplemented media. RNA was extracted from the seedlings, and cDNA was obtained using the iScript™ gDNA Clear cDNA Synthesis Kit (Bio-Rad). The *genomes uncoupled* (*gun*) phenotype was identified by monitoring the expression of *Rbcs* and *Lhcb3* genes in *Xan-h* and *xan-h.chli-1*, together with Arabidopsis Col-0, *cs/cs, 35S::Xan-h* and *35S::xan-h.chli-1* lines, using RT-qPCR. Primers are listed in Table S1. Arabidopsis *gun5* and barley *xan-h.56* mutants were used as positive controls for the *genomes uncoupled* phenotype.

Yeast two-hybrid assay

Coding sequences for *Xan-h*, *xan-h.chli-1*, *xan-h.clo125*, *xan-h.clo157* and *xan-h.clo161*, devoid of cTPs, were cloned into *pGBKT7-GW* and *pGADT7-GW* (Takara Bio) vectors through Gateway cloning. Primer sequences are listed in Table S1. Yeast strains Y187 and AH109 were transformed according to the Clontech User's Manual (PT1172-1) with the vectors *pGBKT7* and *pGADT7*, respectively, harbouring the WT *Xan-h* and the mutant variants. Each Y187 strain was mated with the respective AH109 strain and plated on synthetic drop-out (SD) medium lacking tryptophan (-W) and leucine (-L), to select for positive diploids. To test *Xan-h* and mutant variants homodimerization, overnight liquid cultures were normalized to OD 0.5 and plated on selective media lacking histidine (-W-L–H) and histidine and adenine (-W-L-H-A). The growth of yeast culture dilutions was observed after three days.

Immunoblot analyses

Thylakoids and total protein extracts were prepared from equal amounts of barley leaves (fresh weight) collected from 2-week-old seedlings as described previously (Bassi and Simpson 1987). Protein extracts were fractionated on denaturing 12% (w/v) acrylamide Tris–glycine SDS-PAGE (Schägger and von Jagow 1987) and transferred to polyvinylidene-difluoride (PVDF) membranes (Ihnatowicz et al. 2004). Three replicate filters were probed with specific antibodies. Signals were detected by enhanced chemiluminescence (GE Healthcare). Antibodies directed against Lhca1 (AS01 005), Lhca2 (AS01 006), Lhca3 (AS01 007), Lhcb1 (AS01 004), Lhcb2 (AS01 003), Lhcb3 (AS01 002), Lhcb4 (AS04 045), Lhcb5 (AS01 009), D1 (AS05084), D2 (AS06 146), CP43 (AS111787), CP47 (AS04 038), PsaA (AS06172), PsaD (AS09461), PetA (AS08 306), PetE (AS06 141), PsbO (AS05092), PsbQ (AS06 142–16), PsbR (AS05 059), PsbS (AS09533), H3 (AS10710) were obtained from Agrisera (Vännäs, Sweden). Antibodies were raised against *Hv*CHLI, *Hv*CHLD, *Hv*CHLH as previously described (Lake et al. 2004). The *Hv*GUN4-specific antibody was kindly provided by Professor Mats Hansson (Lund University, Sweden). Three biological replicates were analysed for each SDS-PAGE and immunoblot.

Mg-chelatase activity assay

In-vitro Mg-chelatase activity assays were performed according to Hansson et al. (1999). WT and *xan-h.chli-1* seeds were sown in vermiculite and grown in the dark for 10 days. Etiolated seedlings were then homogenized in 0.4 M mannitol, 20 mM Tricine-NaOH pH 9 and 1 mM DTT. Intact chloroplasts were enriched by 15-min centrifugation at 3000 g and loaded onto a 40% (vol/vol) Percoll cushion in homogenization buffer. Gradients were centrifuged for 15 min at 13,000 g. After washing steps in homogenization buffer, chloroplasts were resuspended in 200 µL of lysis buffer (20 mM Tricine-NaOH pH 9, 1 mM DTT, 1 mM PMSF). After a centrifugation step at 11,000 g for 5 min, the recovered supernatants containing the Mg-chelatase subunits were adjusted to the same protein concentration. The enzymatic assay was carried out by adding 1 µL of the reaction cocktail (50 mM ATP, 250 mM creatine phosphate, 250 mM $MgCl_2$, and 0.06 mM deuteroporphyrin). Reactions were stopped by adding 1 mL acetone/water/25% ammonia (80/20/1, vol/vol/vol) and 200 µL heptane was added to remove chlorophyll from samples. To measure the relative amount of Mg-deuteroporphyrin, the emission spectrum of the acetone phase was recorded from 550 to 600 nm using an excitation wavelength of 408 nm. Excitation and emission slits were set to 5 nm.

Pigment extraction and quantification

Pigments from Arabidopsis and barley were extracted from fresh leaves with 90% acetone. To determine Chl*a* and Chl*b* concentrations, spectrophotometric measurements were carried out according to Porra et al. (1989) and normalized relative to fresh leaf weight. Barley leaf pigment content was also estimated by reversed-phase HPLC (Färber et al. 1997) normalised to fresh weight. Measurements were performed on five biological replicates for each genotype. Apparent chlorophyll content was also measured in vivo at different development stages using the SPAD-502 chlorophyll meter (Konica-Minolta, Tokyo, Japan).

Chlorophyll fluorescence measurements

In-vivo Chl*a* fluorescence was recorded on second barley leaves with a Dual PAM 100 (Walz, Effeltrich, Germany) according to Barbato et al. 2020. After 30 min of dark adaptation, leaves were exposed to increasing actinic light intensities (0–1287 µmol photons m^{-2} s^{-1}) and the following thylakoid electron-transport parameters were determined: the effective quantum yields of PSII [Y(II)] and PSI [Y(I)], the PSII quantum yield of non-regulated energy dissipation [Y(NO)], the PSII quantum yield of regulated energy dissipation [Y(NPQ)], the quantum yield of non-photochemical energy dissipation in PSI due to acceptor side limitation [Y(NA)], and the non-photochemical PSI quantum yield of donor-side limited heat dissipation [Y(ND)] parameters. An imaging Chl fluorometer (Imaging PAM; Walz) was used to measure Chl*a* fluorescence and for *in-vivo* imaging. Dark-adapted plants were exposed to the blue measuring beam (1 Hz, intensity 4; F_0) and a saturating light flash (intensity 10) was used to determine Fv/Fm values. A 5-min exposure to actinic light (56 µmol photons m^{-2} s^{-1}) was then used to calculate Y(II). The Handy PEA fluorometer (Hansatech Instruments Ltd., UK) was used to measure Fv/Fm values in barley plants grown under greenhouse conditions during plant growth.

Transmission electron microscopy (TEM)

TEM analyses were performed as described previously (Tadini et al. 2020). Portions (2 mm × 3 mm) of the second leaves of *Xan-h* and *xan-h.chli-1* barely plants were manually dissected and fixed under vacuum in 2.5% (w/v) glutaraldehyde and 0.1 M sodium cacodylate buffer. After washing with water several times, samples were counterstained with 0.5% uranyl acetate (w/v) overnight at 4 °C. Tissues were then dehydrated in increasing concentrations of ethanol (70%, 80%, 90%, 100% v/v) and permeated twice with 100% (v/v) propylene oxide. Samples were gradually infiltrated first with a 1:2 mixture of Epon-Araldite and propylene oxide for 2 h, then with Epon-Araldite and propylene oxide (1:1) for 1 h and left in a 2:1 mixture of Epon-Araldite and propylene oxide overnight at room temperature. Epon-Araldite resin was prepared by mixing Embed-812, Araldite 502, dodecenylsuccinic anhydride (DDSA) and Epon Accelerator DMP-30 according to the manufacturer's specifications. Ultra-thin sections of 70 nm were cut with a diamond knife (Ultra 45°, DIATOME) and collected on copper grids (G300-Cu, Electron Microscopy Sciences). Samples were observed by transmission electron microscopy (Talos L120C, Thermo Fisher Scientific) at 120 kV. Images were acquired with a digital camera (Ceta CMOS Camera, Thermo Fisher Scientific).

*Hv*CHLI hexamer structure prediction

The homo-hexameric ring model of the barley *Hv*CHLI ATPase subunit of Mg-chelatase was generated using version 3 of Multimer from DeepMind Alphafold2 (Evans et al. 2022), allowing for template search with HH-suite (Steinegger et al. 2019). The structure underwent relaxation by the gradient-descent method using the Amber (Hornak et al. 2006) force-field. Additional refinement was performed with Protein Preparation Wizard (Madhavi Sastry et al. 2013) from the Schrödinger Maestro suite, version 13.7.125, release 2023-3 (Maestro, Schrödinger, LLC, New York, NY, 2023). From the same suite, the module Residue Scanning and Mutation was used to replace R298 with lysine (K298), allowing for side-chain prediction with backbone sampling up to 2.5 Å from the mutation site, and Glide XP (Friesner et al. 2006) to perform the docking of ATP. The open-source software PyMOL (Schrödinger, LLC, The PyMOL Molecular Graphics System, Version 2.6) was used for the visualisation of the molecular structures and rendering of the Figures. The WT model of the *Hv*CHLI subunit is available at https://www.modelarchive.org/doi/https://doi.org/10.5452/ma-xoqwu, while the R298K model is available at https://www.modelarchive.org/doi/https://doi.org/10.5452/ma-tvik6.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00299-024-03328-2.

Acknowledgements We thank Tatsuru Masuda for providing the Arabidopsis *cs/cs* homozygous line and we are grateful to Paul Hardy for critical reading of the manuscript. The NoLimits platform at University of Milano is acknowledged for the TEM analysis. Technical staff at botanical garden "Città Studi" is acknowledged for barley and Arabidopsis plant cultivation.

Author contributions AP, LT, LR, LC, AlTo and PP designed the study. SS, LT, SR and PP took care of *TM2490* mutant isolation. AP, LR, MH and VT performed the molecular, biochemical and physiological characterization of the mutants described. LT, VT, CB and AnTa contributed to the yeast two-hybrid assay. AP and LR were responsible for the TEM images. LR, DH and AP took care of sequencing data analysis and identification of *xan-h.chli-1* mutation. AP, VT and PJ performed pigment analyses. AD and MM conducted the drought-stress tests. FB and CC took care of protein structure prediction and modeling. All authors helped draft the manuscript. PP coordinated the study and took care of the final version of the manuscript. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

Funding Open access funding provided by Università degli Studi di Milano within the CRUI-CARE Agreement. This study was carried out within the Agritech National Research Center, spoke 1 and received funding from the European Union (Next-Generation EU: PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)—MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4—D.D. 1032 17/06/2022, CN00000022). The experiments performed using the functional-phenotyping platform Plantarray were supported by the project Plant-RED, funded by the Italian Ministry of Foreign Affairs and International Cooperation and the Ministry of Science and Technology of the State of Israel (Italy–Israel Joint Call for Proposals on Scientific and Technological Cooperation 2021). The Mg-chelatase activity assay performed in the lab of Mats Hansson was supported by the Horizon Europe programme under Grant Number 101082091—BEST-CROP. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Data availability The raw RNASeq data were deposited to the NCBI data repository: https://submit.ncbi.nlm.nih.gov/subs/bioproject/ under the bioproject identifier PRJNA1052990 SubmissionID: SUB14068056. https://dataview.ncbi.nlm.nih.gov/object/PRJNA1052990?reviewer=mr5orheor72jk9eq701efugt4d. The WT model of the *Hv*CHLI subunit is available at https://www.modelarchive.org/doi/https://doi.org/10.5452/ma-xoqwu (access code: ZDAt3B0X0S), while the R298K model is available at https://www.modelarchive.org/doi/https://doi.org/10.5452/ma-tvik6 (access code: oNI2rKctqS).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Adams NBP, Bisson C, Brindley AA, Farmer DA, Davison PA, Reid JD, Hunter CN (2020) The active site of magnesium chelatase. Nat Plants 6(12):1491–1502. https://doi.org/10.1038/s41477-020-00806-9

Appiah M, Abdulai I, Schulman AH, Moshelion M, Dewi ES, Daszkowska-Golec A, Bracho-Mujica G, Rötter RP (2023) Drought response of water-conserving and non-conserving spring barley cultivars. Front Plant Sci 14(October):1247853. https://doi.org/10.3389/FPLS.2023.1247853/BIBTEX

Araus JL, Slafer GA, Reynolds MP, Royo C (2002) Plant breeding and drought in C3 cereals: what should we breed for? Ann Bot 89(7):925–940. https://doi.org/10.1093/AOB/MCF049

Barbato R, Tadini L, Cannata R, Peracchio C, Jeran N, Alboresi A, Morosinotto T et al (2020) Higher order photoprotection mutants reveal the importance of ΔpH-dependent photosynthesis-control in preventing light induced damage to both photosystem II and photosystem I. Sci Rep 10(1):1–14. https://doi.org/10.1038/s41598-020-62717-1

Bassi R, Simpson D (1987) Chlorophyll-protein complexes of barley photosystem I. Eur J Biochem 163(2):221–230. https://doi.org/10.1111/J.1432-1033.1987.TB10791.X

Bochtler M, Hartmann C, Song HK, Bourenkov GP, Bartunik HD, Huber R (2000) The structures of HslU and the ATP-dependent protease HslU–HslV. Nature 403(6771):800–805. https://doi.org/10.1038/35001629

Braumann I, Stein N, Hansson M (2014) Reduced chlorophyll biosynthesis in heterozygous barley magnesium chelatase mutants. Plant Physiol Biochem 78(May):10–14. https://doi.org/10.1016/J.PLAPHY.2014.02.004

Brzezowski P, Sharifi MN, Dent RM, Morhard MK, Niyogi KK, Grimm B (2016) Mg chelatase in chlorophyll synthesis and retrograde signaling in Chlamydomonas Reinhardtii: CHLI2 cannot substitute for CHLI1. J Exp Bot 67(13):3925–3938. https://doi.org/10.1093/JXB/ERW004

Caffarri S, Tibiletti T, Jennings RC, Santabarbara S (2014) A Comparison between plant photosystem I and photosystem II architecture and functioning. Curr Protein Pept Sci 15(4):296–331. https://doi.org/10.2174/1389203715666140327102218

Canham CD, Finzi AC, Pacala SW, Burbank DH (2011) Causes and consequences of resource heterogeneity in forests: interspecific variation in light transmission by canopy. Trees 24(2):337–349. https://doi.org/10.1139/X94-046

Cattivelli L, Rizza F, Badeck FW, Mazzucotelli E, Mastrangelo AM, Francia E, Marè C, Tondelli A, Michele Stanca A (2008) Drought tolerance improvement in crop plants: an integrated view from breeding to genomics. Field Crop Res 105(1–2):1–14. https://doi.org/10.1016/J.FCR.2007.07.004

Cha SS, An YJ, Lee CR, Lee HS, Kim YG, Kim SJ, Kwon KK et al (2010) Crystal structure of lon protease: molecular architecture of gated entry to a sequestered degradation chamber. EMBO J 29(20):3520–3530. https://doi.org/10.1038/EMBOJ.2010.226

Colombo M, Suorsa M, Rossi F, Ferrari R, Tadini L, Barbato R, Pesaresi P (2016) Photosynthesis control: an underrated short-term regulatory mechanism essential for plant viability. Plant Signal Behav 11(4):e1165382. https://doi.org/10.1080/15592324.2016.1165382

Cutolo EA, Guardini Z, Dall'Osto L, Bassi R (2023) A paler shade of green: engineering cellular chlorophyll content to enhance photosynthesis in crowded environments. New Phytol 239(5):1567–1583. https://doi.org/10.1111/NPH.19064

Dalal A, Shenhar I, Bourstein R, Mayo A, Grunwald Y, Averbuch N, Attia Z, Wallach R, Moshelion M (2020) A telemetric, gravimetric platform for real-time physiological phenotyping of plant-environment interactions. JoVE 2020(162):e61280. https://doi.org/10.3791/61280

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM (2021) Twelve years of SAMtools and BCFtools. GigaScience 10(2):1–4. https://doi.org/10.1093/GIGASCIENCE/GIAB008

Drewry DT, Kumar P, Long SP (2014) Simultaneous improvement in productivity, water use, and albedo through crop structural modification. Glob Change Biol 20(6):1955–1967. https://doi.org/10.1111/GCB.12567

Du S-Y, Xiao-Feng Z, Zekuan Lu, Qi X, Zhen Wu, Tao J, Yan Lu, Xiao-Fang W, Da-Peng Z (2012) Roles of the different components of magnesium chelatase in abscisic acid signal transduction. Plant Mol Biol 80:519–537. https://doi.org/10.1007/s11103-012-9965-3

Edgar RC (2022) High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. bioRxiv. https://doi.org/10.1101/2021.06.20.449169

Elmlund H, Lundqvist J, Al-Karadaghi S, Hansson M, Hebert H, Lindahl M (2008) A new Cryo-EM single-particle Ab initio reconstruction method visualizes secondary structure elements in an ATP-fueled AAA+ motor. J Mol Biol 375(4):934–947. https://doi.org/10.1016/J.JMB.2007.11.028

Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A et al (2022) Protein complex prediction with AlphaFold-Multimer. bioRxiv. https://doi.org/10.1101/2021.10.04.463034

Färber A, Young AJ, Ruban AV, Horton P, Jahns P (1997) Dynamics of xanthophyll-cycle activity in different antenna subcomplexes in the photosynthetic membranes of higher plants (the relationship between zeaxanthin conversion and nonphotochemical fluorescence quenching). Plant Physiol 115(4):1609–1618. https://doi.org/10.1104/PP.115.4.1609

Farmer DA, Brindley AA, Hitchcock A, Jackson PJ, Johnson B, Dickman MJ, Neil Hunter C, Reid JD, Adams NBP (2019) The ChlD subunit links the motor and porphyrin binding subunits of magnesium chelatase. Biochemical Journal 476(13):1875–1887. https://doi.org/10.1042/BCJ20190095

Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. J Med Chem

49(21):6177–6196. https://doi.org/10.1021/JM051256O/SUPPL_FILE/JM051256OSI20060602_023733.PDF

Gadjieva R, Axelsson E, Olsson U, Hansson M (2005) Analysis of gun phenotype in barley magnesium chelatase and Mg-Protoporphyrin IX monomethyl ester cyclase mutants. Plant Physiol Biochem 43(10–11):901–908. https://doi.org/10.1016/J.PLAPHY.2005.08.003

Galkin E, Dalal A, Evenko A, Fridman E, Kan I, Wallach R, Moshelion M (2018) Risk-management strategies and transpiration rates of wild barley in uncertain environments. Physiol Plant 164(4):412–428. https://doi.org/10.1111/PPL.12814

Gao YS, Wang YL, Wang X, Liu L (2020) Hexameric structure of the ATPase motor subunit of magnesium chelatase in chlorophyll biosynthesis. Protein Sci 29(4):1026–1032. https://doi.org/10.1002/PRO.3816

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. Preprint at https://doi.org/10.48550/arXiv.1207.3907

Genesio L, Bassi R, Miglietta F (2021) Plants with less chlorophyll: a global change perspective. Glob Change Biol 27(5):959–967. https://doi.org/10.1111/GCB.15470

Guo FQ, Jared Y, Nigel CM (2023) The nitrate transporter AtNRT11 (CHL1) functions in stomatal opening and contributes to drought susceptibility in Arabidopsis. Plant Cell 15(1):107–117. https://doi.org/10.1105/tpc.006312

Hansson A, Gamini Kannangara C, Von Wettstein D, Hansson M (1999) Molecular basis for semidominance of missense mutations in the XANTHA-H (42-KDa) subunit of magnesium chelatase. Proc Natl Acad Sci USA 96(4):1744–1749. https://doi.org/10.1073/PNAS.96.4.1744/ASSET/4B10415B-E7E4-4F60-A717-3A8A046A7C22/ASSETS/GRAPHIC/PQ0494843003.JPEG

Hansson A, Willows RD, Roberts TH, Hansson M (2002) Three semidominant barley mutants with single amino acid substitutions in the smallest magnesium chelatase subunit form defective AAA+ hexamers. Proc Natl Acad Sci USA 99(21):13944–13949. https://doi.org/10.1073/PNAS.212504499/ASSET/8275538E-B000-4459-B541-4407B9E360C2/ASSETS/GRAPHIC/PQ2125044007.JPEG

Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins 65(3):712–725. https://doi.org/10.1002/PROT.21123

Huang YS, Li HM (2009) Arabidopsis CHLI2 can substitute for CHLI1. Plant Physiol 150(2):636–645. https://doi.org/10.1104/PP.109.135368

Ihnatowicz A, Pesaresi P, Varotto C, Richly E, Schneider A, Jahns P, Salamini F, Leister D (2004) Mutants for photosystem I subunit D of Arabidopsis Thaliana: effects on photosynthesis, photosystem I stability and expression of nuclear genes for chloroplast functions. Plant J 37(6):839–852. https://doi.org/10.1111/J.1365-313X.2004.02011.X

Jansson C, Wullschleger SD, Kalluri UC, Tuskan GA (2010) Phytosequestration: carbon biosequestration by plants and the prospects of genetic engineering. Bioscience 60(9):685–696. https://doi.org/10.1525/BIO.2010.60.9.6

Jaramillo R, van de Zedde VR, Peller J, Visser RGF, van der Linden CG, van Loo EN (2021) High-resolution analysis of growth and transpiration of quinoa under saline conditions. Front Plant Sci 12:634311. https://doi.org/10.3389/FPLS.2021.634311/BIBTEX

Kirst H, Gabilly ST, Niyogi KK, Lemaux PG, Melis A (2017) Photosynthetic antenna engineering to improve crop yields. Planta 245(5):1009–1020. https://doi.org/10.1007/S00425-017-2659-Y/TABLES/3

Kirst H, Shen Y, Vamvaka E, Betterle N, Dongmei Xu, Warek U, Strickland JA, Melis A (2018) Downregulation of the CpSRP43 gene expression confers a truncated light-harvesting antenna (TLA) and enhances biomass and leaf-to-stem ratio in Nicotiana Tabacum canopies. Planta 248(1):139–154. https://doi.org/10.1007/S00425-018-2889-7/TABLES/3

Klimyuk VI, Persello-Cartieaux F, Havaux M, Contard-David P, Schuenemann D, Meiherhoff K, Gouet P, Jones JDG, Hoffman NE, Nussaume L (1999) A chromodomain protein encoded by the Arabidopsis CAO gene is a plant-specific component of the chloroplast signal recognition particle pathway that is involved in LHCP targeting. Plant Cell 11(1):87–99. https://doi.org/10.1105/TPC.11.1.87

Kobayashi K, Mochizuki N, Yoshimura N, Motohashi K, Hisabori T, Masuda T (2008) Functional analysis of arabidopsis thaliana isoforms of the Mg-Chelatase CHLI subunit. Photochem Photobiol Sci 7(10):1188–1195. https://doi.org/10.1039/B802604C

Lake V, Obson U, Willows RD, Hansson M (2004) ATPase activity of magnesium chelatase subunit I is required to maintain subunit D in vivo. Eur J Biochem 271(11):2182–2188. https://doi.org/10.1111/J.1432-1033.2004.04143.X

Larkin RM, Alonso JM, Ecker JR, Chory J (2003) GUN4, a regulator of chlorophyll synthesis and intracellular signaling. Science 299(5608):902–906. https://doi.org/10.1126/SCIENCE.1079978/SUPPL_FILE/LARKIN.SOM.PDF

Long SP, Marshall-Colon A, Zhu XG (2015) Meeting the global food demand of the future by engineering crop photosynthesis and yield potential. Cell 161(1):56–66. https://doi.org/10.1016/J.CELL.2015.03.019

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):1–21. https://doi.org/10.1186/S13059-014-0550-8/FIGURES/9

Lundqvist J, Elmlund H, Wulff RP, Berglund L, Elmlund D, Emanuelsson C, Hebert H et al (2010) ATP-induced conformational dynamics in the AAA+ motor unit of magnesium chelatase. Structure 18(3):354–365. https://doi.org/10.1016/J.STR.2010.01.001

Lundqvist J, Braumann I, Kurowska M, Müller AH, Hansson M (2013) Catalytic turnover triggers exchange of subunits of the magnesium chelatase AAA+ motor unit. J Biol Chem 288(33):24012–24019. https://doi.org/10.1074/JBC.M113.480012

Lupo Y, Moshelion M (2024) The balance of survival: comparative drought response in wild and domesticated tomatoes. Plant Sci 339(February):111928. https://doi.org/10.1016/J.PLANTSCI.2023.111928

Ma YY, Shi JC, Wang DJ, Liang X, Wei F, Gong CM, Qiu LJ et al (2023) A point mutation in the gene encoding magnesium chelatase I subunit influences strawberry leaf color and metabolism. Plant Physiol 192(4):2737–2755. https://doi.org/10.1093/PLPHYS/KIAD247

Madhavi Sastry G, Adzhigirey M, Day T, Annabhimoju R, Sherman W (2013) Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des 27(3):221–234. https://doi.org/10.1007/S10822-013-9644-8/TABLES/9

Marino G, Naranjo B, Wang J, Penzler JF, Kleine T, Leister D (2019) Relationship of GUN1 to FUG1 in chloroplast protein homeostasis. Plant J 99(3):521–535. https://doi.org/10.1111/TPJ.14342

Mascher M, Wicker T, Jenkins J, Plott C, Lux T, Koh CS, Ens J et al (2021) Long-read sequence assembly: a technical evaluation in barley. Plant Cell 33(6):1888–1906. https://doi.org/10.1093/PLCELL/KOAB077

Masuda T (2008) Recent overview of the mg branch of the Tetrapyrrole biosynthesis leading to chlorophylls. Photosynth Res 96(2):121–143. https://doi.org/10.1007/S11120-008-9291-4/FIGURES/2

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F (2016) The ensembl variant effect predictor. Genome Biol. https://doi.org/10.1186/S13059-016-0974-4

Melis A (2009) Solar energy conversion efficiencies in photosynthesis: minimizing the chlorophyll antennae to maximize efficiency. Plant Sci 177(4):272–280. https://doi.org/10.1016/J.PLANTSCI.2009.06.005

Miller JM, Arachea BT, Epling LB, Enemark EJ (2014) Analysis of the crystal structure of an active MCM hexamer. Elife 3:e03433. https://doi.org/10.7554/ELIFE.03433

Mochizuki N, Brusslan JA, Larkin R, Nagatani A, Chory J (2001) Arabidopsis genomes uncoupled 5 (GUN5) mutant reveals the involvement of Mg-chelatase H Subunit in plastid-to-nucleus signal transduction. Proc Natl Acad Sci USA 98(4):2053–2058. https://doi.org/10.1073/PNAS.98.4.2053

Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A, Ens J et al (2019) TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. Genome Biol 20(1):1–18. https://doi.org/10.1186/S13059-019-1899-5/FIGURES/6

Niu Y, Tianxiao C, Zhi Z, Chenchen Z, Chunji L, Jizeng J, Meixue Z (2022) A new major QTL for flag leaf thickness in barley (Hordeum vulgare L.). BMC Plant Biol 22:305. https://doi.org/10.1186/s12870-022-03694-7

Ogura T, Wilkinson AJ (2001) AAA+ superfamily ATPases: common structure-diverse function. Genes Cells 6(7):575–597. https://doi.org/10.1046/J.1365-2443.2001.00447.X

Porra RJ, Thompson WA, Kriedemann PE (1989) Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls a and b extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. Biochim Biophys Acta Bioenerg 975(3):384–394. https://doi.org/10.1016/S0005-2728(89)80347-0

Rotasperti L, Tadini L, Chiara M, Crosatti C, Guerra D, Tagliani A, Forlani S et al (2022) The barley mutant happy under the sun 1 (Hus1): an additional contribution to pale green crops. Environ Exp Bot 196(April):104795. https://doi.org/10.1016/J.ENVEXPBOT.2022.104795

Sadras VO, Richards RA (2014) Improvement of crop yield in dry environments: benchmarks, levels of organisation and the role of nitrogen. J Exp Bot 65(8):1981–1995. https://doi.org/10.1093/JXB/ERU061

Sakowska K, Alberti G, Genesio L, Peressotti A, Vedove GD, Gianelle D, Colombo R et al (2018) Leaf and canopy photosynthesis of a chlorophyll deficient soybean mutant. Plant Cell Environ 41(6):1427–1437. https://doi.org/10.1111/PCE.13180

Schägger H, von Jagow G (1987) Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 KDa. Anal Biochem 166(2):368–379. https://doi.org/10.1016/0003-2697(87)90587-2

Shen YY, Wang XF, Wu FQ, Du SY, Cao Z, Shang Y et al (2006) The Mg-chelatase H subunit is an abscisic acid receptor. Nature. https://doi.org/10.1038/nature05176

Slattery RA, Ort DR (2021) Perspectives on improving light distribution and light use efficiency in crop canopies. Plant Physiol 185(1):34–48. https://doi.org/10.1093/PLPHYS/KIAA006

Song L, Chen Z, Larkin RM (2018) The genomes uncoupled mutants are more sensitive to norflurazon than wild type. Plant Physiol 178(3):965–971. https://doi.org/10.1104/PP.18.00982

Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J (2019) HH-Suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20(1):1–15. https://doi.org/10.1186/S12859-019-3019-7

Tadini L, Peracchio C, Trotta A, Colombo M, Mancini I, Jeran N, Costa A et al (2020) GUN1 influences the accumulation of NEP-dependent transcripts and chloroplast protein import in Arabidopsis cotyledons upon perturbation of chloroplast protein homeostasis. Plant J 101(5):1198–1220. https://doi.org/10.1111/TPJ.14585

Talamè V, Bovina R, Sanguineti MC, Tuberosa R, Lundqvist U, Salvi S (2008) TILLMore, a resource for the discovery of chemically induced mutants in barley. Plant Biotechnol J 6(5):477–485. https://doi.org/10.1111/J.1467-7652.2008.00341.X

Tardy F, Créach A, Havaux M (1998) Photosynthetic pigment concentration, organization and interconversions in a pale green syrian landrace of barley (Hordeum Vulgare L., Tadmor) adapted to harsh climatic conditions. Plant Cell Environ 21(5):479–489. https://doi.org/10.1046/J.1365-3040.1998.00293.X

Tian T, Liu Y, Yan H, You Qi, Yi X, Zhou Du, Wenying Xu, Zhen Su (2017) AgriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res 45(W1):W122–W129. https://doi.org/10.1093/NAR/GKX382

Verwoerd TC, Dekker BMM, Hoekema A (1989) A small-scale procedure for the rapid isolation of plant RNAs. Nucleic Acids Res 17(6):2362–2362. https://doi.org/10.1093/NAR/17.6.2362

Walker BJ, Drewry DT, Slattery RA, VanLoocke A, Cho YB, Ort DR (2018) Chlorophyll can be reduced in crop canopies with little penalty to photosynthesis. Plant Physiol 176(2):1215–1232. https://doi.org/10.1104/PP.17.01401

Watanabe N, Nakada E (1999) Seasonal variation of leaf colour in syrian barley and its association with photosynthetic electron transport rate. Cereal Res Commun 27(1–2):171–178. https://doi.org/10.1007/BF03543934/METRICS

Willows RD, Beale SI (1998) Heterologous expression of the Rhodobacter Capsulatus BchI, -D, and -H genes that encode magnesium chelatase subunits and characterization of the reconstituted enzyme. J Biol Chem 273(51):34206–34213. https://doi.org/10.1074/JBC.273.51.34206

Wu F-Q, Qi X, Zheng C, Zhi-Qiang L, Shu-Yuan D, Chao M, Chen-Xi Z et al (2009) The magnesium-chelatase H subunit binds abscisic acid and functions in abscisic acid signaling: new evidence in Arabidopsis. Plant Physiol 150:1940–1954. https://doi.org/10.1104/pp.109.140731

Wu CJ, Wang J, Zhu J, Ren J, Yang YX, Luo T, Xu LX et al (2022) Molecular characterization of Mg-Chelatase CHLI subunit in pea (Pisum Sativum L.). Front Plant Sci. https://doi.org/10.3389/FPLS.2022.821683

Zhang H, Li J, Yoo JH, Yoo SC, Cho SH, Koh HJ, Seo HS, Paek NC (2006) Rice Chlorina-1 and Chlorina-9 encode ChlD and ChlI subunits of mg-chelatase, a key enzyme for chlorophyll synthesis and chloroplast development. Plant Mol Biol 62(3):325–337. https://doi.org/10.1007/S11103-006-9024-Z/METRICS

Zhang C, Liu H, Wang J, Li Y, Liu D, Ye Y, Huang R et al (2023) A key mutation in magnesium chelatase I subunit leads to a chlorophyll-deficient mutant of tea (Camellia Sinensis). Jf Exp Bot. https://doi.org/10.1093/JXB/ERAD430

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Figures

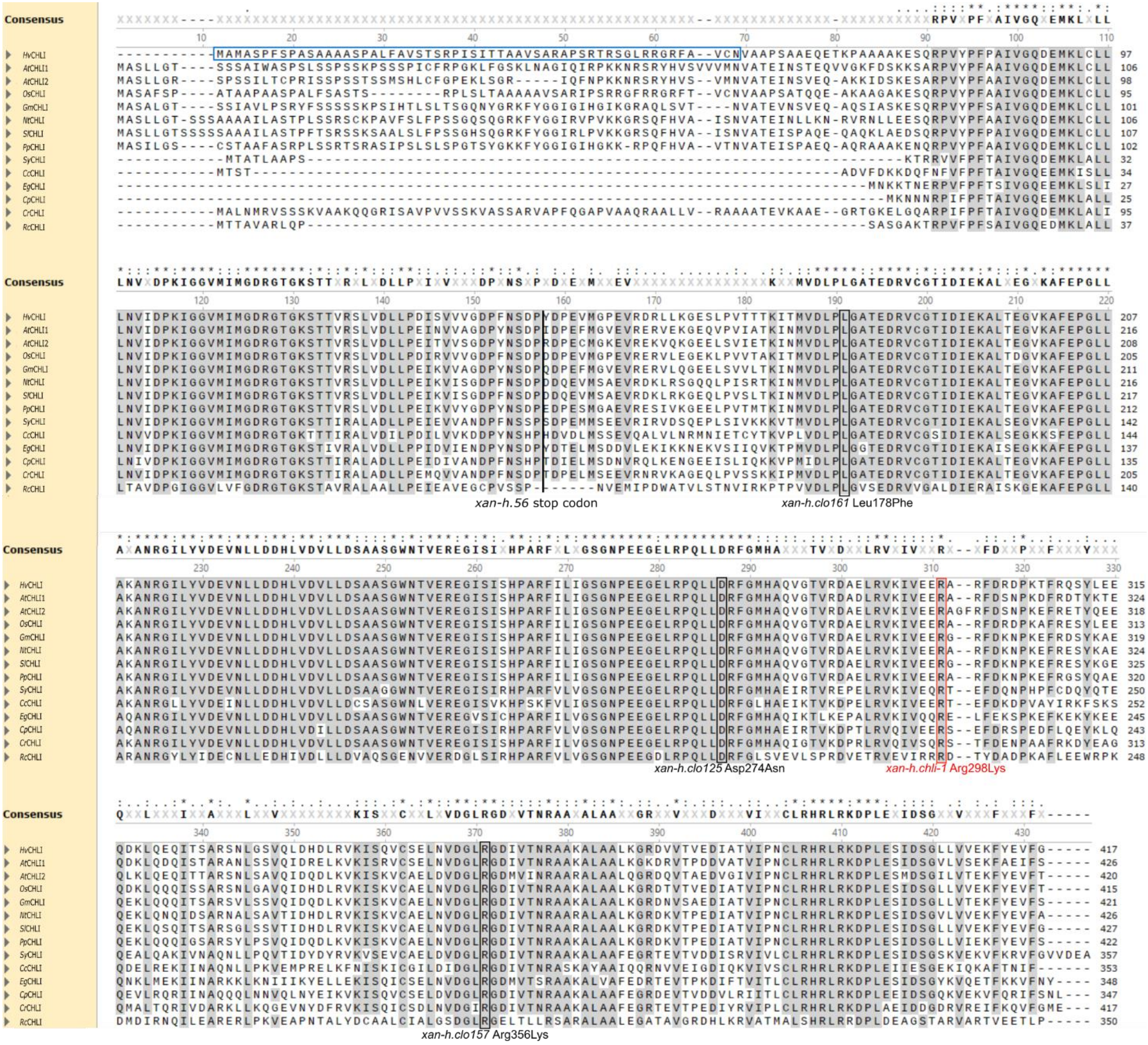


Figure S1. Alignment of CHLI sequences with Musclev5. The chloroplast transit peptide (cTP) of the *Hv*CHLI protein, as predicted by TargetP-2.0 (services.healthtech.dtu.dk/services/TargetP-2.0), is highlighted with a light-blue box. Amino-acid substitutions reported as SNPs in barley mutants are indicated. Amino-acid positions refer to the barley sequence, including the R298K missense mutation as main candidate for the *TM2490* pale green phenotype (in red). The degrees of identity between the *Hordeum vulgare* sequence and the analyzed sequences are the following: *Arabidopsis thaliana At*CHLI1 78%; *Arabidopsis thaliana At*CHLI2 81%; *Oryza sativa subsp. Japonica Os*CHLI 90%; *Glycine max Gm*CHLI 77%; *Nicotiana tabacum Nt*CHLI 76%; *Solanum lycopersicum Sl*CHLI 78%; *Prunus persica Pp*CHLI 78%; *Synechocystis sp.* (strain PCC 6803 / Kazusa) *Sy*CHLI 73%; *Cyanidium caldarium Cc*CHLI 62%; *Euglena gracilis Eg*CHLI 69%; *Cyanophora paradoxa Cp*CHLI 70%; *Chlamydomonas reinhardti Cr*CHLI 66%; *Rhodobacter capsulatus Rc*CHLI 49%.

Amino acids showing 100% conservation among the sequences considered are indicated by asterisks, while dots and colons indicate degrees of amino-acid conservation greater than 40% and 60%, respectively.

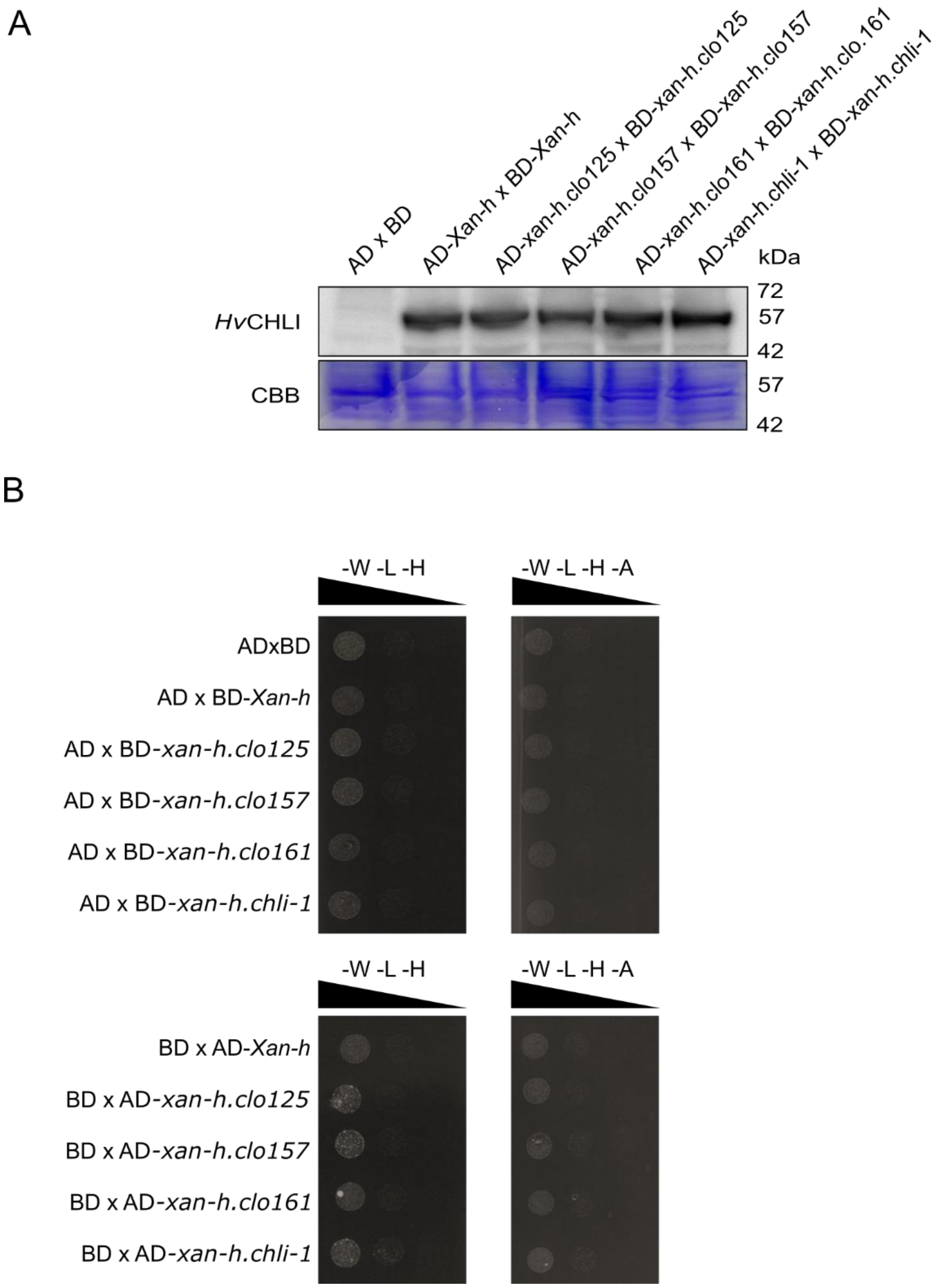


Figure S2. Western blots and negative controls to validate the yeast two-hybrid data. (A) Western blot analysis performed by using a *Hv*CHLI-specific antibody on diploid yeast cells (AH109xY187) to confirm the expression of Gal4AD and Gal4BD fusions to *Xan-h* and its allelic variants *xan-h.chli-1*, *xan-h.clo125*, *xan-h.clo157* and *xan-h.clo161* (MW between 53 and 57 kDa). Empty plasmids expressing Gal4AD and Gal4BD (AD x BD) were used as controls. CBB, Coomassie Brilliant Blue staining of a replica SDS-PAGE. (B) To exclude possible growth on selective media due to non-specific interaction between the different variants and the Gal4 domains used for the assay, each yeast strain expressing wild-type and mutant variants of *HvCHLI* was alternatively mated and tested for interaction against either the Gal4BD or the Gal4AD alone. No interaction between *Hv*CHLI and mutant variants with Gal4BD or Gal4AD could be detected, as shown by the lack of yeast growth on selective media, devoid of either Trp, Leu and His (-W -L -H) or Trp, Leu, His and Ade (-W -L -H -A).

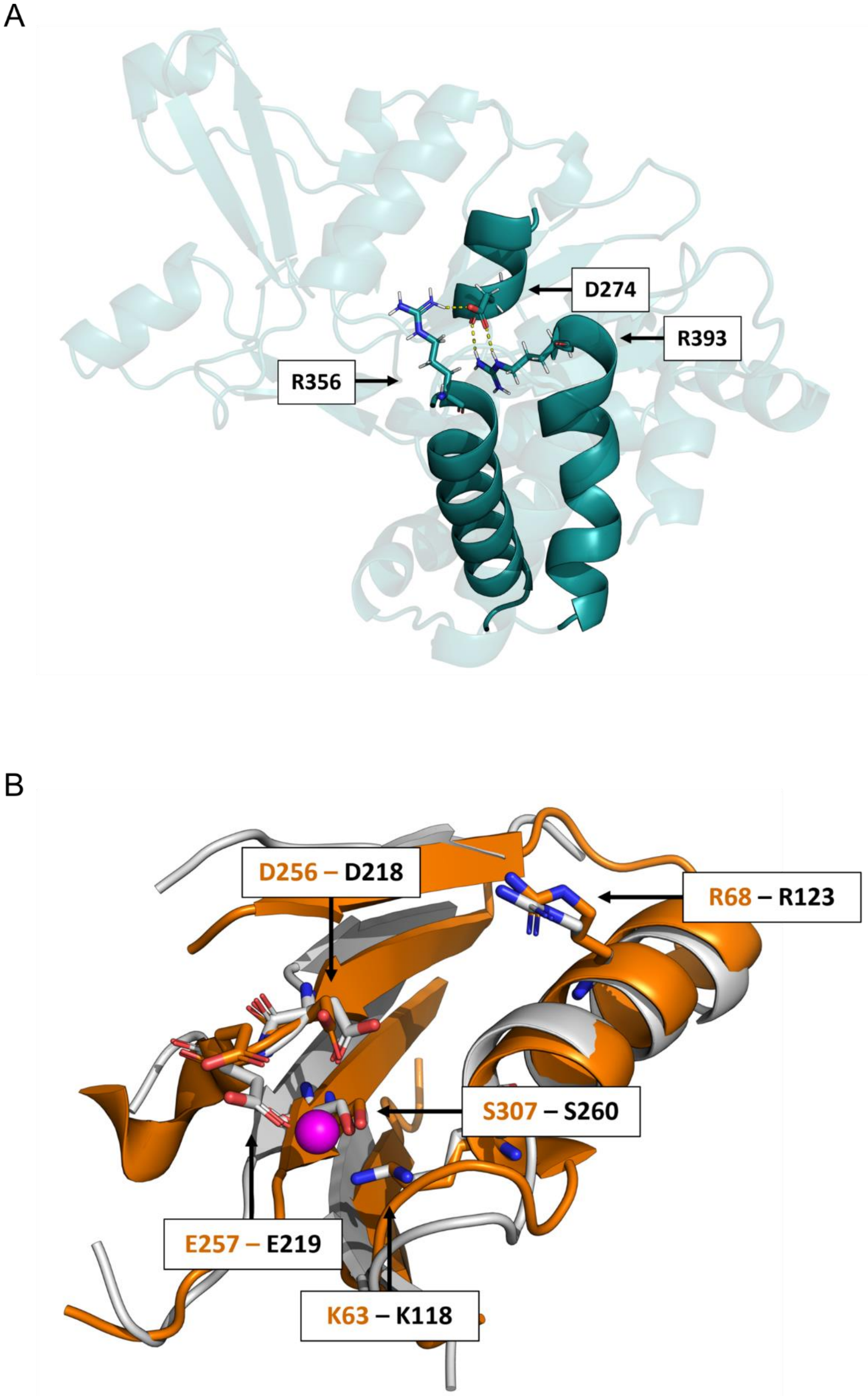


Figure S3. Details of specific properties of the *Hv*CHLI monomer. (A) Detailed view of the D274-R356-R393 interaction within the barley *Hv*CHLI monomer. The overall monomer structure is represented in transparent cyan cartoon, except for the alpha helices encompassing D274, R356 and R393, which are highlighted in solid cyan cartoon. D274, R356 and R393 are depicted as solid sticks and coloured: cyan for C atoms, blue for N, red for O, and white for H-. The hydrogen bond is represented as a dashed yellow line. (B) The ATP-binding domain from the protein Heat Shock Locus U [HSLU, PDB ID: 1DO0; (Bochtler et al. 2000)] superimposed on the barley *Hv*CHLI model. The HSLU domain is presented in solid orange cartoon, whereas the domain from the barley *Hv*CHLI model is in light grey. Selected conserved residues from HSLU are represented as solid sticks: C atoms are shown in orange, N atoms in blue, O atoms in red, and H atoms in white. Conserved residues of the barley model are depicted in the same colour scheme, except that C atoms are shown in light grey.

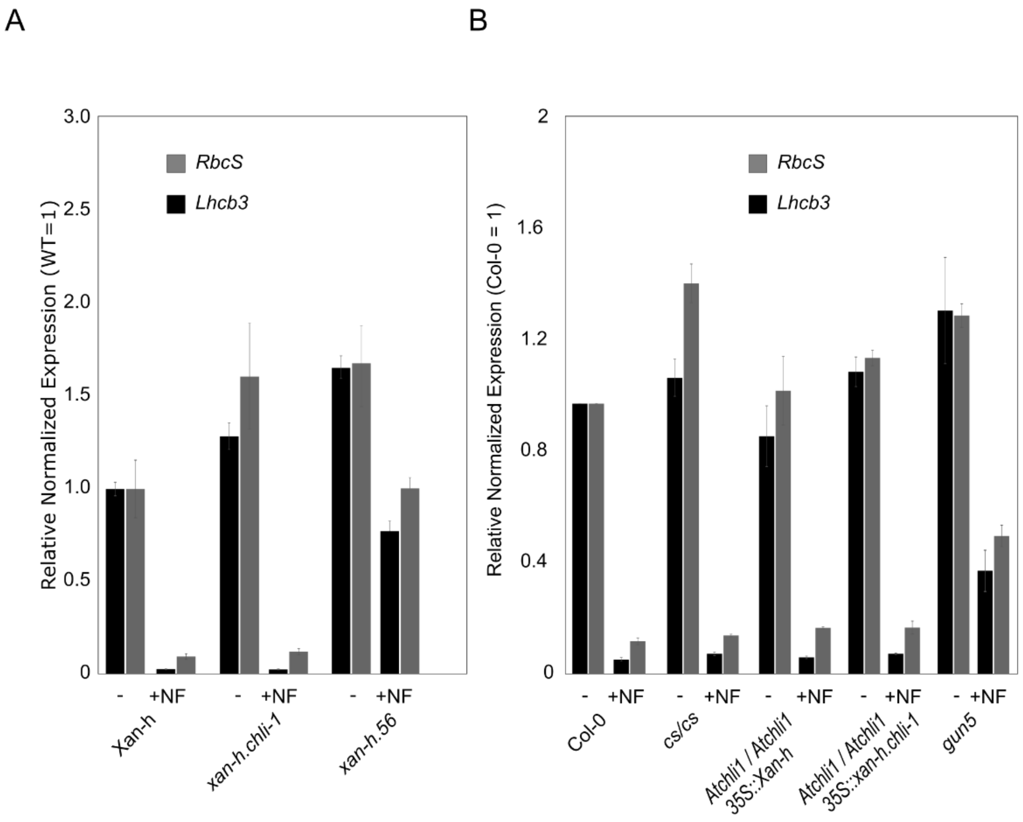


Figure S4. Assay for the *genomes uncoupled* (*gun*) phenotype in barley and Arabidopsis lines. (A) RT-qPCR expression analyses of the photosynthesis-associated nuclear genes *Rbcs* and *Lhcb3* were performed on barley *Xan-h, xan-h.chli-1* and *xan-h.56* lines grown for 6 days under sterile conditions, either in the absence of Norflurazon (NF) or on NF-supplemented medium (5 µM) for 4 days. (B) The expression of the same genes was also monitored in Arabidopsis Col-0, *cs/cs, Atchli1/Atchli1* + *35S::Xan-h* and *Atchli/Atchli* + *35S::xan-h.chli-1* mutant lines under the same conditions in presence or absence of 5 µM Norflurazon. The retrograde-signalling-defective mutant *gun5* (Arabidopsis) and *xan-h.56* (barley) were used as controls for the *gun* phenotype.

3.3.3 Future perspectives

The characterisation of the barley *TM2490* mutant continues in two main directions:

1. The *TM2490* mutants exhibit a particular sensitivity to low temperatures. While the wild-type population can grow at temperatures as low as 4°C, the *TM2490* mutants do not show this capability. To investigate this phenomenon, we are performing molecular dynamics simulations, collecting microsecond trajectories at both 4°C and 22°C in across multiple replicates. We expect that analysis of the ATP binding site under these different conditions will provide insight into this phenomenon.

2. The information gained from the previous point will drive a rational design of the ATP binding pocket with the aim of retaining the pale green phenotype characteristic of *TM2490*, but also allowing the plant to thrive at low temperatures.

4 — Additional Contributions

During my PhD, I had the opportunity to collaborate with research groups across different areas of biology, ranging from plant science to human immunology. In these projects, I applied computational structural biology techniques, including AI-based protein prediction and design, molecular docking, and molecular dynamics simulations. Regardless of the specific approach, data analysis and data visualisation have always played a primary role.

Several of these collaborative efforts have resulted in published manuscripts. My main contributions to these works were in the computational structural biology sections, where I was involved in both writing and designing figures, particularly those illustrating data analysis and visualisation of molecular structures. In all the presented works, the integration of computational approaches with experimental data has enabled the achievement of otherwise unattainable results. Indeed, while the static structure of a biomolecule can provide insights, it is often insufficient for interpreting biological phenomena. Combining computational techniques with experimental data could overcome this limitation, allowing for a deeper understanding of the function of biomolecules. *In silico* data can drive experimental research forward by generating hypotheses and guiding experimental analyses.

The following section presents the works to which I have contributed.

Molecular Therapy
Original Article

Spike mutation resilient scFv76 antibody counteracts SARS-CoV-2 lung damage upon aerosol delivery

Ferdinando M. Milazzo,[1,10] Antonio Chaves-Sanjuan,[2,3,10] Olga Minenkova,[1] Daniela Santapaola,[1] Anna M. Anastasi,[1] Gianfranco Battistuzzi,[1] Caterina Chiapparino,[1] Antonio Rosi,[1] Emilio Merlo Pich,[1] Claudio Albertoni,[4] Emanuele Marra,[5] Laura Luberto,[5] Cécile Viollet,[6] Luigi G. Spagnoli,[7] Anna Riccio,[8] Antonio Rossi,[9] M. Gabriella Santoro,[8,9] Federico Ballabio,[2] Cristina Paissoni,[2] Carlo Camilloni,[2] Martino Bolognesi,[2,3] and Rita De Santis[1]

[1]Biotechnology R&D, Alfasigma SpA, Via Pontina Km 30.400, Pomezia, 00071 Rome, Italy; [2]Department of Biosciences, University of Milan, Via Celoria 26, 20133 Milan, Italy; [3]Cryo-EM Lab, Pediatric Research Center, Fondazione Romeo e Enrica Invernizzi, University of Milan, Via Celoria 26, 20133 Milan, Italy; [4]Studio E Roma S.r.l., Via di Tor Vergata 434, 00133 Rome, Italy; [5]Takis Srl, Via di Castel Romano 100, 00128 Rome, Italy; [6]Texcell, Génavenir 5, Rue Pierre Fontaine 1, 91058 Evry Cedex, France; [7]Histo-Cyto Service Srl, Via Bernardino Ramazzini 93, 00151 Rome, Italy; [8]Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica 1, 00133 Rome, Italy; [9]Institute of Translational Pharmacology, CNR, Via Fosso del Cavaliere 100, 00133 Rome, Italy

The uneven worldwide vaccination coverage against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and emergence of variants escaping immunity call for broadly effective and easily deployable therapeutic agents. We have previously described the human single-chain scFv76 antibody, which recognizes SARS-CoV-2 Alpha, Beta, Gamma and Delta variants. We now show that scFv76 also neutralizes the infectivity and fusogenic activity of the Omicron BA.1 and BA.2 variants. Cryoelectron microscopy (cryo-EM) analysis reveals that scFv76 binds to a well-conserved SARS-CoV-2 spike epitope, providing the structural basis for its broad-spectrum activity. We demonstrate that nebulized scFv76 has therapeutic efficacy in a severe hACE2 transgenic mouse model of coronavirus disease 2019 (COVID-19) pneumonia, as shown by body weight and pulmonary viral load data. Counteraction of infection correlates with inhibition of lung inflammation, as observed by histopathology and expression of inflammatory cytokines and chemokines. Biomarkers of pulmonary endothelial damage were also significantly reduced in scFv76-treated mice. The results support use of nebulized scFv76 for COVID-19 induced by any SARS-CoV-2 variants that have emerged so far.

INTRODUCTION

Lung infection from emerging viruses can raise serious public health concern in the case of pandemics. From the coronavirus disease 2019 (COVID-19) pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), we learned how a broad and timely vaccination campaign, together with adoption of prevention measures like mask wearing and social distancing and use of antiviral medications, can reduce deaths and intensive care pressure. The relatively milder disease recently associated with emergence of the Omicron BA.1 and BA.2 variants is raising hope for a weakening of the pandemic.[1] However, because of the uneven worldwide vaccination coverage and possible emergence of new viral variants escaping immunity, the evolution of COVID-19 is unpredictable, and reoccurrence of severe pulmonary diseases cannot be ruled out.[2] The observation of several threatening post-acute sequelae of SARS-CoV-2 infection particularly affecting the nervous and cardiovascular systems,[3] urgently necessitates easily deployable therapeutic measures able to control the infection in the early stages. With prospective COVID-19 pandemic re-exacerbation, and even in the case of transition into an endemic phase, two types of interventions are being envisaged: first, to improve vaccine equity worldwide with a possible update against SARS-CoV-2 variants and second, to validate early-stage therapeutic protocols preventing worsening of the disease and ultimately hospitalizations and post-acute sequelae. As of today, Omicron variants are challenging the efficacy of most injected antibodies.[4–10] Because the Omicron variants apparently remain confined mainly to the upper respiratory tract,[11] use of systemic antibodies is becoming somehow questionable.

We recently described a cluster of human anti-SARS-CoV-2 antibodies in the format of a single-chain variable fragment (scFv) able to neutralize viral variants *in vitro* and in animal models.[12] We also showed that such an antibody format is suitable for intra-nasal or aerosol formulations that might be useful for topical treatment of upper and lower respiratory tract SARS-CoV-2 infection.[12]

Received 28 June 2022; accepted 12 September 2022;
https://doi.org/10.1016/j.ymthe.2022.09.010.

[10]These authors contributed equally

Correspondence: Rita De Santis, Biotechnology R&D, Alfasigma SpA, Via Pontina Km 30.400, Pomezia, 00071 Rome, Italy.
E-mail: rita.desantis@alfasigma.com

Table 1. ScFv76 spike/ACE2 competition by ELISA

	IC50 (nM) (±SE)		
	Delta	Omicron BA.1	Omicron BA.2
scFv76	1.64 (0.25)	1.90 (0.3)	2.4 (0.1)
scFv5	>40	>40	>40

Shown is competition of spike binding to human ACE2 by scFv antibodies, measured by ELISA. IC50 values (expressed as nanomolar concentration) are the average (±SE) from 3–4 independent experiments.

In the present work, we show that the scFv76 antibody of the cluster found previously to be able to react with SARS-CoV-2 Alpha, Beta, Gamma, and Delta is also resilient to the Omicron BA.1 and BA.2 mutations, substantially retaining neutralizing activity against these new viral variants. We provide a pre-clinical proof of concept of the efficacy of nebulized scFv76 in a mouse model of Delta infection, selected as an aggressive prototype of viral pneumonia. Finally, we prove, by single-particle cryoelectron microscopy (cryo-EM), the wide recognition properties of the scFv76 antibody at the molecular level, showing that it binds to a well-conserved epitope at the tip of the spike protein in the receptor binding domain (RBD) with an architecture that is able to accommodate the mutations found in all SARS-CoV-2 variants known to date. Our results support use of the scFv76 antibody for aerosol therapy of COVID-19 induced by all variants of concern.

RESULTS

ScFv76 efficiently neutralizes SARS-CoV-2 Delta and Omicron

The scFv76 antibody has been described previously to be able to neutralize the SARS-CoV-2 Alpha, Beta, Gamma, and Delta viral variants *in vitro* and in animal models.[12] To evaluate its reactivity with the recently emerged Omicron variants, the ability to compete the binding of the Omicron BA.1 and BA.2 spikes to human ACE2 was tested by ELISA. The results in Table 1 show that scFv76 can inhibit Omicron BA.1 and BA.2 spike binding to ACE2 at half maximal inhibitory concentration (IC50) concentrations of less than 2.5 nM, which is like the potency against Delta. The binding affinity of scFv76 to the Delta and Omicron spikes was then tested by Surface Plasmon Resonance (SPR) showing K_D values of 0.6 nM for Delta and 6.3 and 14.5 nM for BA.1 and BA.2, respectively (Table 2). Neutralizing activity against SARS-CoV-2 Omicron BA.1 and BA.2 pseudotyped viruses was also exhibited by scFv76 but not by scFv5 (an anti-RBD antibody shown previously to be devoid of neutralizing activity and used as a negative control),[12] with IC50 values of 2.84 and 2.47 nM, respectively (Figure 1A).

Neutralization of infectivity was further tested against authentic SARS-CoV-2 Delta and Omicron BA.1 viruses by microneutralization assay of cytopathic effects (CPEs) in Vero E6 cells. In this assay, scFv76 exhibited IC50 values of 1.99 and 6.38 nM against the Delta and Omicron BA.1 variants, respectively, whereas the non-neutralizing antibody scFv5 showed no anti-viral activity, as expected (Figure 1B).

Table 2. ScFv76 surface plasmon resonance (SPR) data

Spike trimer	k_a (10^5 M^{-1} s^{-1})	k_d (10^{-5} s^{-1})	K_D (nM)
Delta	1.1	6.1	0.6
Omicron BA.1	0.7	41.6	6.3
Omicron BA.2	1.2	174.9	14.5

The Omicron BA.2 spike has been shown recently to be more pathogenic and more efficient in mediating syncytium formation than the BA.1 spike.[13] The ability of scFv76 to prevent SARS-CoV-2 Omicron BA.1 or BA.2 spike-induced fusion of pulmonary cells was therefore tested *in vitro*. As shown in Figure 1C, incubation with nanomolar concentrations of the scFv76 antibody proved to be significantly effective at inhibiting fusion between BA.1 and BA.2 spike-expressing human HEK293T cells and human lung A549 cells stably expressing the hACE2 receptor (A549 hACE2).

Before an *in vivo* pharmacology study of nebulized scFv76 in a severe Delta-induced pneumonia mouse model, its antiviral neutralization potency was tested *in vitro* by qRT-PCR in Delta-infected pulmonary Calu-3 cells in comparison with the non-neutralizing control antibody scFv5. As shown in Figure 1D, scFv76 was found to inhibit infection with an IC50 of 13.5 nM, whereas no activity of the control antibody at a concentration greater than 200 nM was observed.

Therapeutic efficacy of nebulized scFv76 in a severe SARS-CoV-2 Delta interstitial pneumonia model

We previously established the biochemical suitability of scFv76 to aerosol delivery by a mesh nebulizer.[12] To test the pharmacological efficacy of the nebulized antibody, pneumonia infection was established in transgenic hACE2 mice by intranasal challenge with 1×10^5 50% tissue culture infectious dose (TCID50) SARS-CoV-2 (strain Delta B.1.617.2). The overall experimental design is shown in Figure 2A. Different from infected mice treated with vehicle, the group of mice treated with scFv76 showed significant body weight recovery 4 days after infection (Figure 2B). This result correlated with an about 100-fold reduction in lung viral RNA copy numbers, as assessed by qRT-PCR (Figure 2C), and with a reduction of infectious viral particles, as measured by TCID50 (Figure 2D). Nebulized scFv76 reduced infectious virus titers in the lungs to undetectable levels in three of five mice; significant viral RNA reduction was also observed in the nasal turbinates (Figure 2E). Histopathological analysis of lung sections showed a significant reduction of lung interstitial edema and hematic endoalveolar extravasation, a reduction of cellular inflammatory infiltrates in the alveolar/interstitial space, and a reduction of alveolar septal thickening (Figure 3A). Overall, treatment with nebulized scFv76, but not phosphate-buffered-saline (PBS), was significantly effective at counteracting the lung inflammation and damage induced by the Delta virus, as shown in Figure 3B. To further evaluate the extent of protection conferred by scFv76 nebulization in Delta-infected mice, qRT-PCR analyses were performed to measure the mRNA expression of several inflammatory effectors in lung homogenates harvested 4 days after infection. Data indicate that aerosol

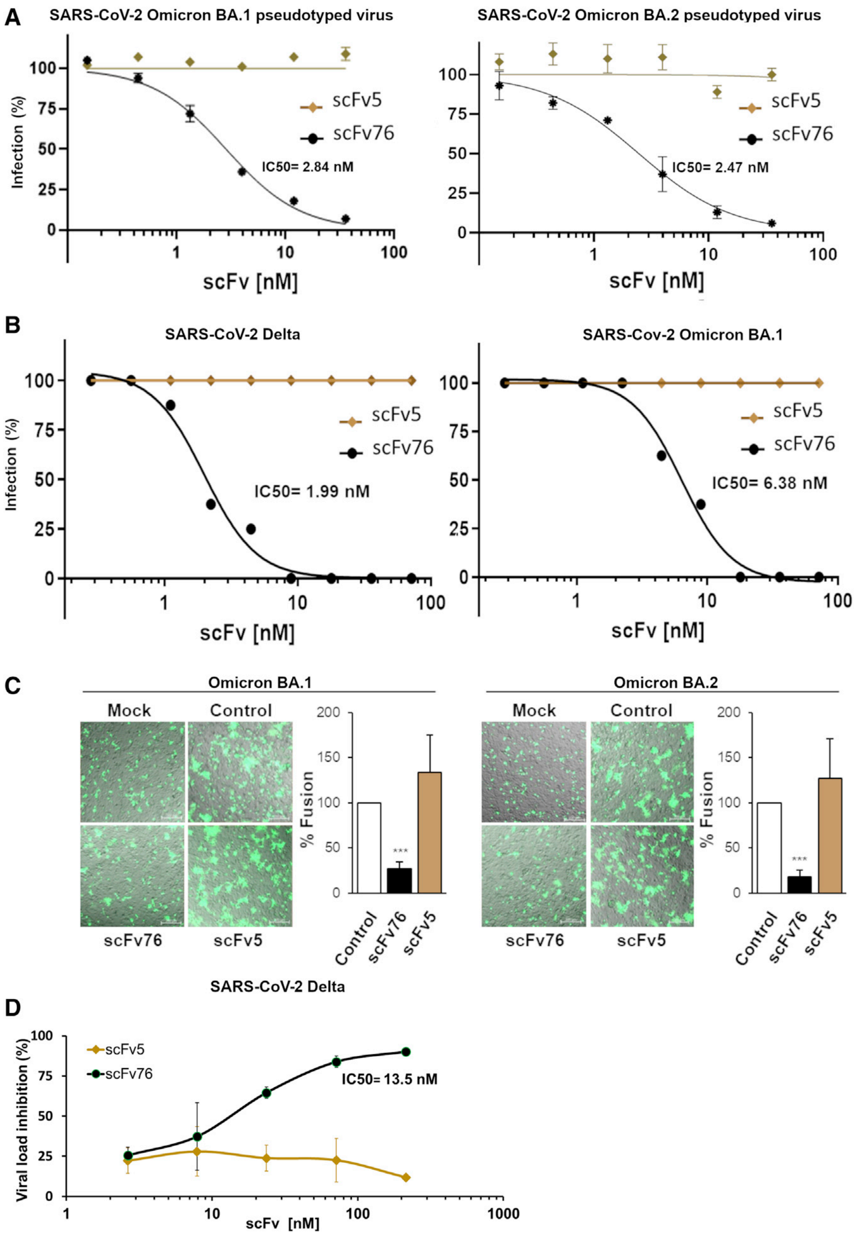


Figure 1. Resilience of scFv76 reactivity to the Omicron BA.1 and BA.2 variants

(A) Neutralization of pseudotyped virus expressing the SARS-CoV-2 Omicron (B.1.1.529) BA.1 or BA.2 spike, assessed by luciferase assay in hACE2-expressing Caco-2 cells. Data are the average (±SD) of two replicates from one representative experiment. (B) Neutralization activity of scFv antibodies assessed by viral titration (Delta and Omicron strains) on Vero E6 cells by microneutralization assay. Data are the average (±SD) of eight replicates from one representative experiment. (C) Inhibition of SARS-CoV-2 spike-mediated cell-cell fusion using HEK293T donor cells expressing green fluorescent protein (GFP) and Omicron BA.1 or BA.2 spike or GFP only (mock), incubated for 1 h with scFv76 or scFv5 (360 nM) and then overlaid on monolayers of hACE2-expressing A549 cells for 24 h. The overlay of bright-field and fluorescence images is shown. Scale bar, 200 μm. Cell-cell fusion quantification is expressed as percentage relative to control (average ± SD of 5 fields from two biological replicates). ***$p < 0.001$ (ANOVA). (D) Neutralization of the authentic SARS-CoV-2 Delta virus in Calu-3 cells. Serially diluted (3-fold) Abs were added to cells 1 h after infection. Quantification of viral load was done by qRT-PCR 72 h after infection. Data are the average (±SD) of two independent experiments. The IC50 value (expressed as nanomolar concentration) is also shown in (A), (B), and (D).

treatment with scFv76 induced significant reduction of key pro-inflammatory cytokines like the interleukins IL6, IL1B, IL21, IL10, IL4, and tumor necrosis factor alpha (TNF) and the chemokines CCL2, CCL20, CXCL1, and CXCL10 (Figures 4A and 4B). The lungs of infected and vehicle-treated mice showed upregulated transcription levels of type I interferon (IFNA1 and especially IFNB1) and type II interferon (IFNG) and of key IFN-modulated genes (IFIT1, ISG15 and MX1). All of these genes were significantly reduced in lungs of mice treated with scFv76 (Figures 4A and 4C). Finally, we evaluated some biomarkers of pulmonary vascular damage, and the data indicated that the treatment also counteracted upregulation of infection-induced tissue damage molecules, including adhesion molecules, angiopoietin 2, and inflammasome effectors such as NLRP3 (Figures 4D and 4E).

Structural bases for broad RBD recognition of SARS-CoV-2 variants by scFv76

To explore the recognition principles and rationalize the broad cross-reactivity of scFv76 toward SARS-CoV-2 variants, we determined the 3D structure of the spike:scFv76 complex using single-particle cryo-EM. We used a SARS-CoV-2 Wuhan-Hu-1 6P-stabilized glycoprotein (native antigen)[14] incubated with scFv76 to assemble the complex. Our single-particle cryo-EM analysis revealed a homogeneous population of the spike:scFv76 complex displaying two RBDs in the up conformation and one down, with one scFv76 fragment bound to the tip of each RBD (Figure 5A). The final 3D reconstruction had an overall resolution of 3.5 Å (Figures S1A and S2); nevertheless, the epitope-paratope interface regions were less clearly resolved compared with the main spike component because of flexibility of the RBDs. To gain better insight into the recognition interface structure, we applied a focused refinement procedure[15] to the RBD-down fragment region that brought the local resolution to 4.0 Å (Figures 5A, S1B, and S2) and subsequently based our analysis on this structure. The

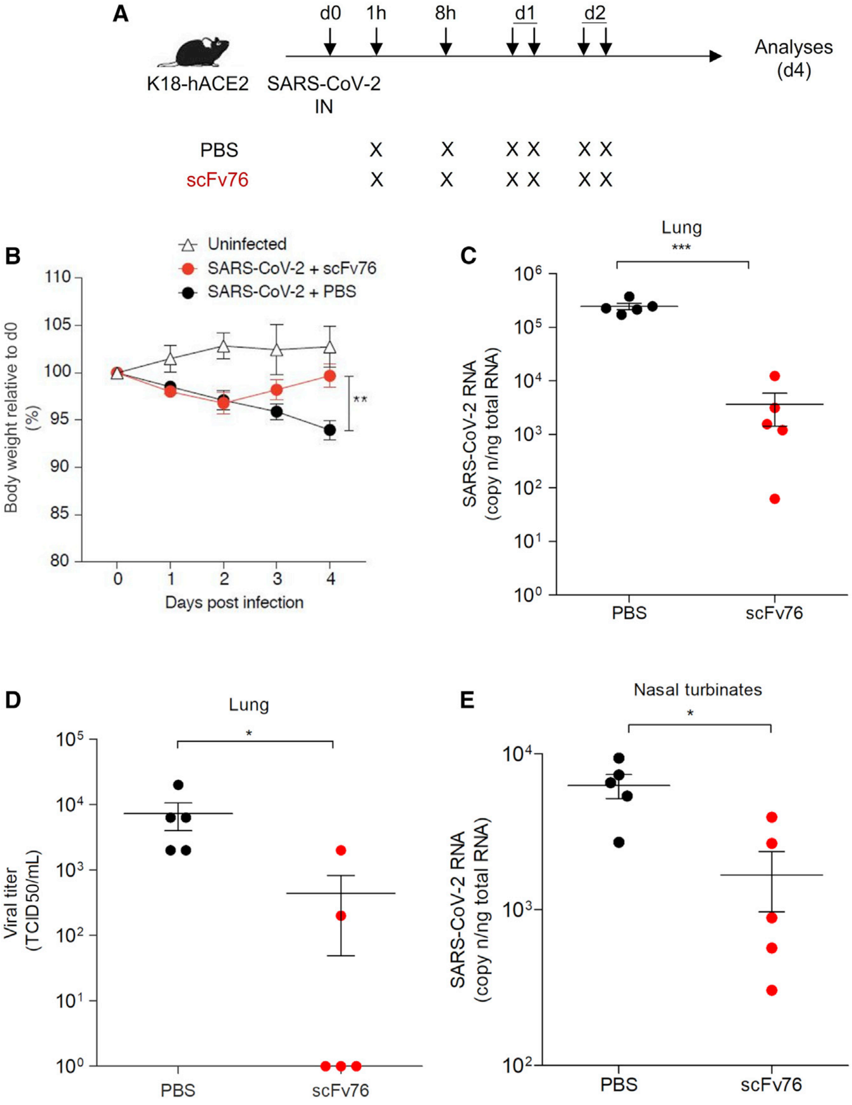


Figure 2. Therapeutic efficacy of nebulized scFv76 in a mouse SARS-CoV-2 Delta pneumonia model

(A) Study design. Human ACE2 transgenic mice were exposed by nose only to 2.5 mL of 3 mg/mL scFv76 solution or PBS (as vehicle control) 1 h and 8 h after SARS-CoV-2 Delta intranasal infection (1 × 10^5 TCID50/mouse) and twice per day for 2 additional days. (B) Body weight changes. Daily body weight from days 0–4 were recorded for each group and plotted as a percentage with respect to day 0. Data are the average (±SE). The day when there was a significant difference in average percentage of body weight between scFv76-treated or PBS-treated animals is denoted by $^{**}p < 0.01$. (C) Lung viral RNA quantification. On day 4 after infection, lungs were collected for viral RNA quantification by qRT-PCR. Each dot represents one mouse. Data are expressed as copy number per nanogram of total RNA (n = 5). (D) Lung virus titration. Viral titers in the lung 4 days after infection were determined by viral 50% tissue culture infectious dose (TCID50) assay. Each dot represents one mouse. Data are expressed as TCID50 per milliliter. (E) Viral RNA quantification in nasal turbinates (NTs). Quantification by qRT-PCR in NTs and data representation were done as in (C). Statistical differences in (B–E) were assessed by Mann-Whitney U test. Significance is indicated as follows: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

scFv76:RBD refined structure showed that the light and heavy chains of scFv76 contact the tip of the RBD in the up and down conformations. ScFv76 buries a surface area of ∼1,082 Å^2, based on an approximately equal contributions of the light and heavy chain components. The scFv76 paratope comprises residues of all three heavy-chain complementarity-determining regions (CDRs) and two from the light-chain CDRs (Table 3; Figure 5B). Conversely, the recognition site lies on the RBD receptor-binding ridge and surrounding areas (Figure 5C), in full agreement with an alanine scanning analysis reported previously,[12] where RBD L455A, F456A, Y473A, N487A and Y489A mutations strongly reduced scFv76 binding. F456 is located in the deep groove created between CDRH1 and CDRH2 on one side and CDRH3 and CDRL3 on the other. The scFv76:RBD interaction may also be stabilized by several hydrogen bonds (as evaluated on a 4.0-Å-resolution structure). Among these, RBD D420 interacts with S56 (CDRH2), and the carbonyl groups of RBD residues L455 and A475 interact with Y33 and T28 (both in CDRH1), respectively; the side chain of RBD Y421 falls close to the P53 backbone carbonyl in CDRH2. The angle of approach of scFv76 to the RBD resembles that of ACE2; the scFv76 fragment contact region overlaps with the ACE2 binding interface, matching the location of 13 of 17 ACE2-binding residues on the RBD (Figure 5D). In this respect, the close resemblance of scFv76 and ACE2 binding modes to the RBD and the ensuing competition for binding explain, on structural grounds, the potent scFv76 neutralizing activity. The scFv76:RBD pose resembles closely that observed for most antibodies from the VH3-53/VH3-66 germline. Not all such antibodies show neutralizing activity across SARS-CoV-2 variants, again stressing the key role of subtle and specific structural variations in the outcome of epitope-paratope interaction.

A core of 28 epitope residues recognized by scFv76 is conserved in SARS-CoV-2 Alpha, Beta, Gamma, and Delta variants carrying the important K417N, E484K, and N501Y mutations (Table 3). In our refined model, E484 does not directly contact scFv76, consistent with the previously shown reactivity of scFv76 with E484-mutated variants.[12] We also predict low susceptibility to mutations at K417 and N501 because they are not involved in any polar contact with scFv76. Both residues are in solvent-exposed regions, allowing

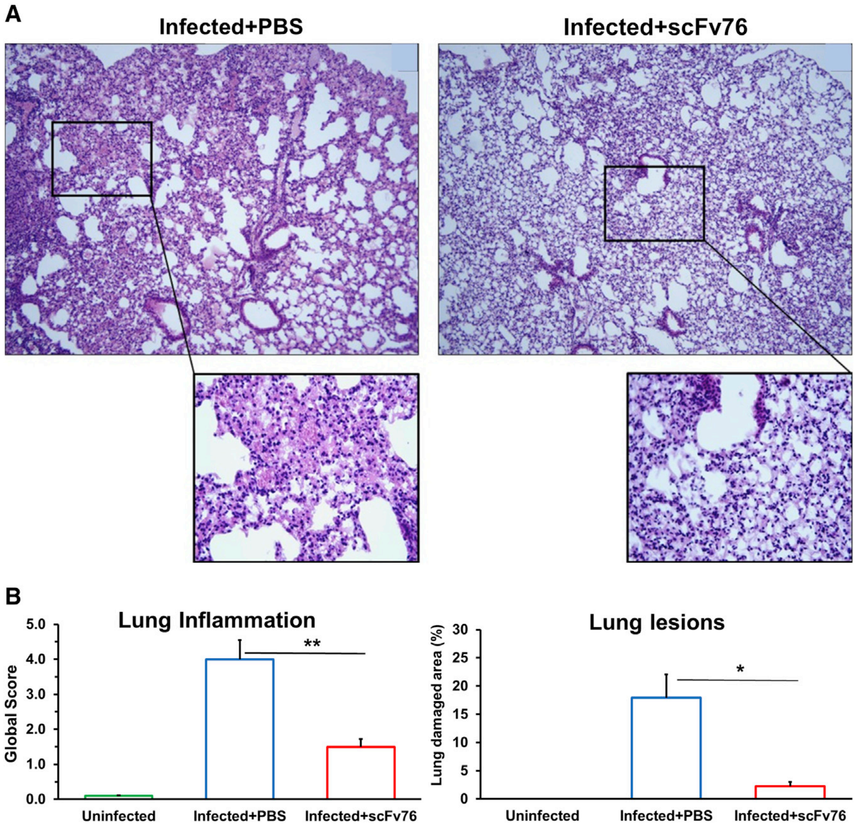


Figure 3. Therapeutic efficacy of nebulized scFv76 correlates with reduction of inflammatory scores

(A) Histopathological analysis of lung tissue sections from mice challenged with SARS-CoV-2 Delta and treated by aerosol with scFv76 or PBS, as described in Figure 2. Shown are representative pictures of lung sections, stained with hematoxylin and eosin (H&E), from PBS-treated (left panel) or scFv76-treated (right panel) mice. Scale bar, 200 μm; 10× magnification. Inset: 40× magnification. (B) Scores of overall lung inflammation (top panel) and lung lesion (bottom panel), measured on lung sections as in (A). Data are the average (±SE) (n = 5) and are expressed as global score and lung damaged area (percent), respectively (see scoring details in Materials and Methods). Statistical analysis was by Student's t test. Significance is indicated as follows: *p < 0.05, **p < 0.01.

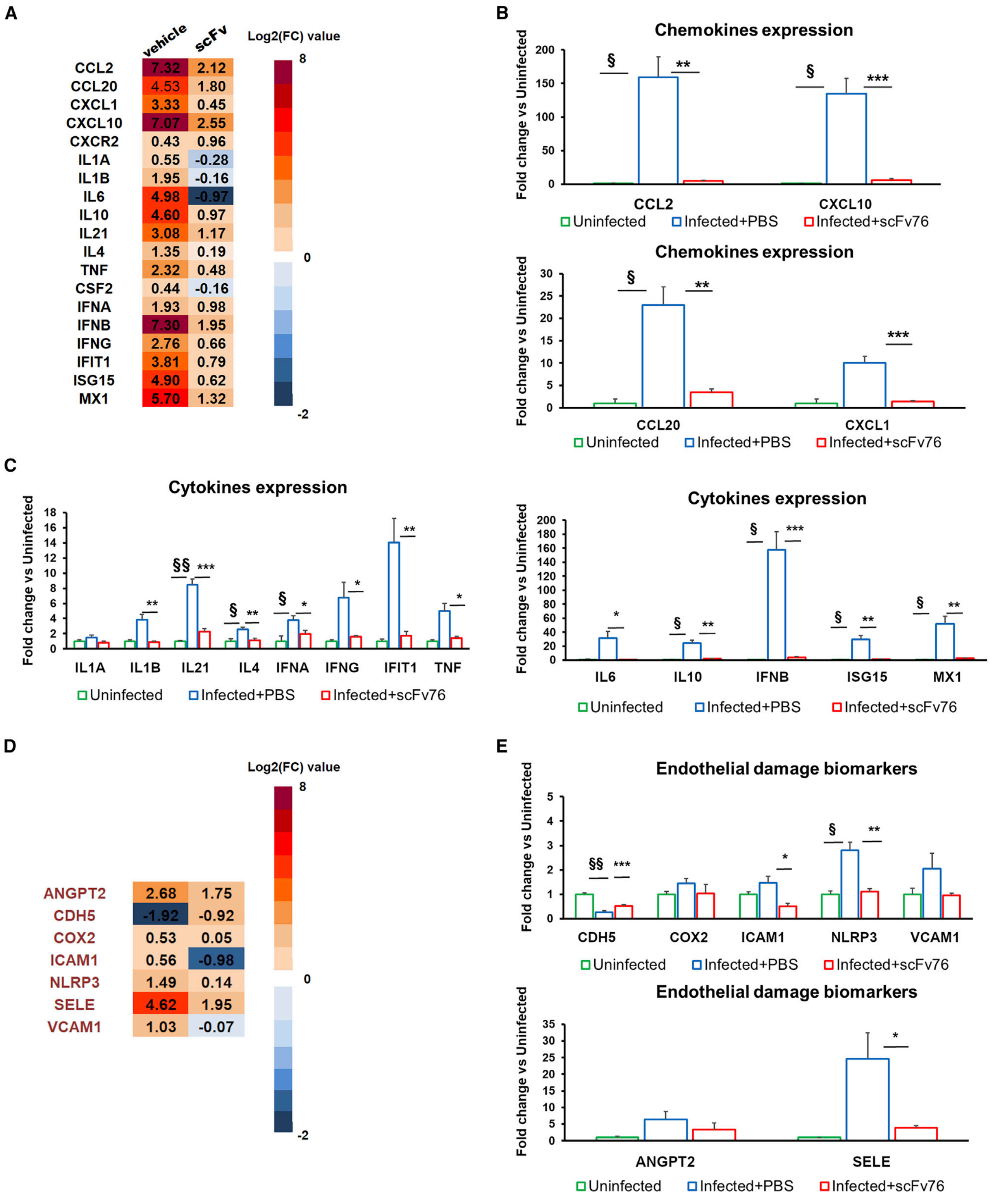
conformational flexibility; 417N would insert into a large groove created among CDRs, and 501Y would position the aromatic side chain beyond scFv76 CDRL3 (Figures 5C and S3). Both mutations are indeed known to have a limited effect on scFv76 neutralizing power,[12] in keeping with the broad SARS-CoV-2 recognition properties displayed by scFv76.

Our modeling exercise suggests that the residues building the RBD epitope, recognized by scFv76, should drop to 23 in the Omicron BA.1 and BA.2 spike variants (Table 3). This could explain the 10- and 20-fold affinity reduction for Omicron BA.1 and BA.2, respectively, compared with the Delta variant. In our model, the five Omicron-unique side-chain substitutions, occurring in the RBD epitope region in these variants, are predicted to marginally affect scFv76 binding, as confirmed by the RBD/ACE2 competition and virus neutralization data presented here. The S477N, Q493R, G496S, and Q498R substitutions in particular would place the mutated residues in solvent-exposed regions (Figures 5C and S3). Residue Y505 is located between CDRL1 and CDRL3 in the RBD:scFv76 complex and mostly participates in hydrophobic contacts (Figures 5C and S3); the Omicron Y505H mutation may follow the same scheme. Such considerations, supporting substantial conservation of the RBD:scFv76 interface in all variants, are in agreement with the functional data reported in this paper that highlight the resilience of scFv76 to the main SARS-CoV-2 variants, including Omicron BA.1 and BA.2.

DISCUSSION

In the search of easily deployable therapeutic measures against COVID-19, we recently described 76clAbs, a cluster of human single-chain antibody fragments that, in principle, could bypass all limitations of traditional monoclonal antibodies. Use of monoclonal antibodies for COVID-19 therapy is being challenged by several issues: (1) difficulties with deployment of therapy, being monoclonal antibodies parenteral drugs to be administered in a hospital environment; (2) the risk of antibody-dependent enhancement (ADE) that can be ignited by different routes involving the immunoglobulin Fc interaction with the Fc receptor[16] or with ACE2, found recently to possibly act as a secondary receptor,[17] or with Fcγ-expressing cells, including monocytes and macrophages that, by triggering inflammatory cell death, need to abort production of infectious virus and cause systemic inflammation that contributes to the severity of COVID-19 pathogenesis;[18] and (3) evasion properties of SARS-CoV-2 variants, particularly recently emerged Omicron lineages for which most approved and investigational antibodies have lost their neutralization activity.[4–10]

The single-chain antibody format, because of its high stability, can be easily used for self-administrable aerosol treatment. Single-chain antibodies are, in principle, devoid of ADE risk because of lack of an Fc sequence. 76clAbs, which were selected on the original SARS-CoV-2 Wuhan strain, were found to be resilient to Alpha, Beta, Gamma, and Delta variant mutations.[12] We show that the scFv76 antibody of the cluster can also recognize and neutralize the infectivity and fusogenic activity of Omicron BA.1 and BA.2 variants. Single-particle cryo-EM results point to the peculiar property of this antibody to bind to the up and down conformations of the spike RBD, recognizing a well-conserved epitope located at the ACE2 binding interface, thus accounting for its neutralization properties. All mutations in the RBD of the known SARS-CoV-2 variants are predicted to marginally affect scFv76 recognition, as confirmed by experimental results. We



(legend on next page)

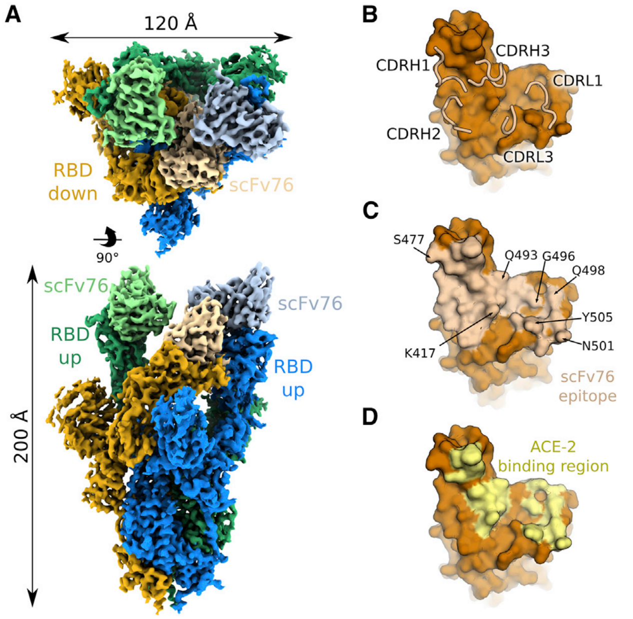


Figure 5. ScFv76 broad recognition of SARS-CoV-2 variants

(A) Composite cryo-EM map of the SARS-CoV-2 spike protein with the locally refined RBD:scFv76 in two orientations. The RBD up or down conformations with their corresponding scFv76 fragments are labeled. The spike subunits are highlighted in green, blue, and yellow, respectively, and the scFv76 fragments bound to each RBD are shown in light green, light blue, and light yellow. (B) ScFv76 CDR loops overlaid on the surface representation of the RBD. (C) RBD surface showing epitope residues as colored in (B). (D) RBD surface showing the ACE2 binding region in yellow.

hypothesize that the Omicron variants BA.4 and BA.5 might be still neutralized by scFv76. The spike mutated residues in these two variants, relative to Omicron BA.2, are 69-70del, L452R, F486V, and wild-type amino acid Q493,[19] with F486V the only mutation at the scFv76 binding interface. F486V replaces a bulky apolar side chain, closing a hydrophobic patch, with a smaller one, possibly affecting, to a limited degree, the stability of the complex.

Significant therapeutic efficacy of nebulized scFv76 is shown here in a severe model of SARS-CoV-2 Delta pneumonia. The present data indicate that aerosol treatment with scFv76 can efficiently control virus proliferation, significantly reducing lung inflammation and damage. These results encourage further clinical development of scFv76 antibody aerosol therapy as a new opportunity for treatment of COVID-19, regardless of the variant causing the disease.

Table 3.

Residues from the scFv76 heavy chain component contacting the RBD	
G26, F27, T28, A31, N32, Y33	from CDRH1
Y52, P53, G54, S56, F58	from CDRH2
R97, L99, S100, V101, A102, D106, I107	from CDRH3
Residues from scFv76 light chain component contacting the RBD	
Q160, S161, V162, S163, S164, Y166	from CDRL1
G226, S227, Y230	from CDRL3

Shown are SARS-CoV-2 RBD residues that are in contact with scFv76. Residues mutated in Alpha, Beta, Gamma, and Delta variants are shown in italics; residues mutated in Omicron BA.1 and BA.2 are displayed in italics and bold black, respectively: R403, T415, G416, *K417*, D420, Y421, T453, L455, F456, R457, K458, S459, N460, Y473, Q474, A475, G476, **S477**, F486, N487, Y489, **Q493**, S494, Y495, **G496**, **Q498**, T500, *N501*, V503, **Y505**.

MATERIALS AND METHODS

Spike/ACE2 binding competition

For competition experiments, Nunc MaxiSorp plates with 96 wells were coated with 100 μL/well of SARS-CoV-2 spike1 variant B.1.617.2 (Delta) protein (His tag) and SARS-CoV-2 spike S1+S2 trimer variant B.1.1.529 (Omicron) protein (ECD, His tag), both from Sino Biological, and SARS-CoV-2 spike trimer variant BA.2 (Omicron) protein His tag verified by Multiangle Light Scattering (MALS), from Acro Biosystems, in PBS at a final concentration of 0.5 μg/mL overnight (ON) at +4°C. Plates were blocked with 300 μL/well blocking solution for 2 h at room temperature (RT). After washing, dilutions of antibodies were added in a volume of 50 μL/well at double concentration, and after 30-min incubation at 37°C, 1.0 μg/mL human ACE2 protein mouse Fc tag (Sino Biological) was added and incubated for 1 h at 37°C. Plates were washed 4 times with PBS/Tween and then incubated for 1 h at RT with 100 μL/well of an anti-mouse Fc conjugated to alkaline phosphatase (Sigma-Aldrich), diluted 1:1,000 in blocking buffer. After washing 4 times, 100 μL/well p-nitrophenyl phosphate (pNpp) substrate was added, and plates were incubated at RT in the dark. Absorbance was recorded at 405 nm using a Sunrise Tecan spectrophotometer.

SPR

Kinetic constants were determined using SPR experiments with a Biacore T200 instrument (Cytiva). SARS-CoV-2 Spike trimer (T19R,

Figure 4. The therapeutic efficacy of nebulized scFv76 correlates with the reduction of pulmonary inflammatory and vascular damage biomarkers

(A) Heatmap of differential gene expression for inflammatory effectors, as determined by qRT-PCR, in lung homogenates of mice challenged with SARS-CoV-2 Delta and treated by aerosol with scFv76 or PBS. Data are the average of the log2 expression fold change (FC) obtained from each experimental group (n = 5) with respect to uninfected mice. Up-regulation appears as shades of red, and down-regulation appears as shades of blue. (B) The mRNA expression level of genes encoding for key chemokines, assessed by qRT-PCR in samples as in (A). Results are the average (±SE) of expression FC with respect to uninfected mice. (C) The mRNA expression levels of genes encoding for key inflammatory effectors, assessed by qRT-PCR as above. Results are expressed as in (C). (D) Heatmap of differential gene expression analysis for pulmonary vascular damage biomarkers. (E) The mRNA expression levels of key pulmonary vascular damage-related genes, assessed and represented as above. Statistical differences in (B), (C), and (E) were assessed by Student's t test. Significance is indicated as follows: $^{\S}p < 0.05$ and $^{\S\S}p < 0.01$ infected + PBS-treated versus uninfected mice; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$ infected + PBS-treated versus infected + scFv76-treated mice.

G142D, EF156-157del, R158G, L452R, T478K, D614G, P681R, and D950N) His tag (MALS verified) protein (Acro Biosystems), SARS-CoV-2 spike S1+S2 trimer variant B.1.1.529 (Omicron) protein (ECD, His tag, Sino Biological), and SARS-CoV-2 spike trimer variant BA.2 (Omicron) protein His tag (MALS verified, Acro Biosystems), all 1.25 μg/mL in buffer containing 0.01M Hepes pH 7.4, 0.15M NaCl, 0.005% w/w surfactant P20 (HBS-P+ from Cytiva), were immobilized at 1,000 Resonance Units (RU) level on the surface of a flow cell of a Series S sensor chip nitrilotriacetic acid (NTA from Cytiva) using Ni^{2+}-mediated capture followed by an amine coupling procedure, and another flow cell surface was blank immobilized by amine coupling with ethanolamine to be used as a control surface. Then scFv76 was flowed at 30 μL/min on all flow cells at 0.47, 1.40, 4.19, 12.56, 37.67, and 113 nM concentrations in buffer containing 10mM Hepes, 0.15M NaCl, 3mM EDTA disodium dihydrate, 0.005% w/w surfactant P20, pH 7.4 (HBS-EP+ buffer from Cytiva) for a contact time of 480 s. After a dissociation time of 900 s, all flow cell surfaces were regenerated by flowing a solution of 4 mM glycine-HCl and 0.1% sodium dodecyl sulphate (SDS) w/w at 30 μL/min for 30 s. Double-referenced sensorgrams were obtained by subtraction of blank-immobilized flow cell curves and of zero concentration curves from derivatized surface flow cell curves. Kinetic constants were obtained by BIAevaluation 3.2 software (Cytiva) fitting with a 1:1 binding model.

Virus neutralization in Calu-3 cells

To measure the SARS-CoV-2-neutralizing capability of scFv76, a live SARS-CoV-2 assay was performed by measuring the viral load in human lung adenocarcinoma Calu-3 cells by qRT-PCR 72 h after virus infection. The experiments were carried out at the François Hyafil Research Institute (Oncodesign; Villebon-sur-Yvette, France). Calu-3 cells were seeded in 96-well plates in complete cell culture medium consisting of Minimal Essential Medium (MEM), 1% pyruvate, 1% glutamine, and 10% fetal bovine serum and then infected, at a multiplicity of infection of 0.01, with SARS-CoV-2 Delta virus provided by the National Institute of Infectious Diseases (NIID) Japan, strain hCoV-19/Japan/TY11-330-P1/2021; originally provided by the Global Initiative on Sharing Avian Influenza Data (GISAID): EP-I_ISL_ 2158613. One hour after infection, the virus solution was discarded and replaced by a volume of growth medium containing scFv76 or non-neutralizing scFv5 antibody at a concentration ranging from 214–2.6 nM in triplicate. The plates were then transferred to a 37°C incubator for 72 h. Finally, the cell culture supernatants were collected for viral RNA extraction (Macherey Nagel Viral RNA Kit), and viral RNA copy number was quantified by qRT-PCR, targeting a region in the viral ORF1ab gene and using a QuantStudio 7 Real-Time PCR System (Applied Biosystems). Data were processed using GraphPad Prism software (v.8.0), and the IC50 values were calculated using a four-parameter logistic curve fitting approach.

SARS-CoV-2 S-pseudovirus neutralization assays

Generation of SARS-CoV-2 S-pseudovirus and SARS-CoV-2 S-pseudovirus neutralization assays were performed as described previously.[12] The vectors expressing Omicron SARS-CoV-2-spike (S1+S2)-long (B.1.1.529) and SARS-CoV-2-spike (S1+S2)-long (B.1.1.529 sublineage BA.2) were obtained from GenScript and Sino Biological, respectively. Serial (1:3) dilutions (ranging from 35.7–0.14 nM final concentration) of scFvs were tested in duplicate. Luciferase activity (relative luciferase units [RLU]) was detected 72 h after infection using the Bright-Glo Luciferase Assay System Kit (Promega) in a microplate luminometer (Wallac-PerkinElmer).

Microneutralization assay

Neutralizing antibody titers were tested using a live-virus assay as follows. ScFv samples were pre-diluted in inoculation medium consisting of Dulbecco's Modified Eagle's Medium (DMEM), 2% fetal calf serum, 1% glutamine), followed by 9 serial dilutions in inoculation medium. Each serial dilution was then mixed 1:1 with 2,000 TCID50/mL SARS-CoV-2 variant virus (Delta variant strain hCoV-19/USA/MD-HP05647/2021 and Omicron variant strain hCoV-19/USA/MD-HP20874/2021) and incubated for 1 h at +37°C ± 2°C and 5% ± 0.5% of CO_2. Thirty-five microliters of each diluted sample/virus mix were then applied in octuplicate to Vero E6 cells seeded at a density of 10^4 cells/well in a 96-well plate on day −1. After 1 h of incubation at +37°C ± 2°C and 5% ± 0.5% CO_2, 65 μL of inoculation medium (DMEM, 2% fetal calf serum, 1% glutamine) was added per well. Plates were incubated for 6 days at +37°C ± 2°C, 5% ± 0.5% CO_2. After this incubation, the cells were inspected for CPEs, and the number of positive wells (that is, exhibiting CPEs) was recorded. Data were processed using GraphPad Prism software (v.8.0), and the IC50 values were calculated using a four-parameter logistic curve fitting approach.

Cell-cell fusion assay

Human alveolar type II-like epithelial A549 cells and HEK293T cells were obtained from the ATCC (Manassas, VA). Cells were grown at 37°C and 5% CO_2 in RPMI-1640 (A549 cells) or DMEM (HEK293T cells) (Euroclone) supplemented with 10% fetal calf serum (FCS), 2 mM glutamine, and antibiotics. Generation of A549 cells stably expressing the human ACE2 receptor (A549-hACE2 cells) has been described previously.[20] The vectors expressing Omicron SARS-CoV-2-spike (S1+S2)-long (B.1.1.529) and SARS-CoV-2-spike (S1+S2)-long (B.1.1.529 sublineage BA.2) were obtained from GenScript and Sino Biological, respectively. Transfections were performed using Lipofectamine 2000 (Invitrogen, Thermo Fisher Scientific) according to the manufacturer's instructions. The donor-target cell fusion assay has been described previously.[12] Transmission and fluorescence images were taken using a Carl Zeiss Axio Observer inverted microscope, and the extent of fusion was quantified as described previously.[12] Images shown in all figures are representative of at least five random fields (scale bars are indicated). Statistical analysis was performed using one-way ANOVA (GraphPad Prism 6.0 software, GraphPad). All experiments were done in duplicate and repeated at least twice.

In vivo pharmacological evaluation of nebulized scFv76 in a model of SARS-CoV-2 Delta pulmonary infection

The animal study was carried out at the San Raffaele Scientific Institute (Milan, Italy) and performed in accordance with European

Directive 2010/63/EU for protection of animals used for scientific purposes, applied in Italy by Legislative Decree 4 March 2014, n. 26. All experimental animal procedures were approved by the Institutional Animal Committee of San Raffaele Scientific Institute. Female transgenic K18-hACE2 mice, aged 8–10 weeks, were infected via the intranasal route with 1 × 10^5 TCID50/mouse of SARS-Cov-2 variant Delta B.1.617.2 virus (hCoV-19/Italy/LOM-Milan-UNIMI9615/2021 [GISAID: EPI_ISL_3073880]), obtained from the Laboratory of Microbiology and Virology of San Raffaele Scientific Institute. One hour and 8 h after infection and twice per day for 2 additional days, infected mice (5/group) were treated by nose-only nebulization with 2.5 mL of scFv76 (3.0 mg/mL in PBS) or PBS using an Aerogen Pro (Aerogen) mesh nebulizer and a nose-only inhalation chamber suitable for delivering the nebulized antibodies contemporarily for up to 8 mice, as described previously.[12] Mice were monitored for appearance, behavior, and weight. On day 4 after infection, they were euthanized by inhalation of 5% isoflurane, followed by gentle cervical dislocation, and lungs and nasal turbinates were explanted and then fixed by 4% paraformaldehyde for histopathological analyses or snap-frozen (in liquid nitrogen) and stored at -80°C until further analyses.

Tissue homogenization and viral titer determination

For viral titer determination in the lungs, tissue homogenates were prepared by homogenizing perfused lungs using a gentleMACS Octo dissociator (Miltenyi) in M tubes containing 1 mL of DMEM. Samples were homogenized three times with program m_Lung_01_02 (34 s, 164 rpm). The homogenates were centrifuged at 3,500 rpm for 5 min at 4°C. The supernatant was collected and stored at −80°C until use for viral isolation and viral load detection. Viral titer was calculated by TCID50. Briefly, Vero E6 cells were seeded at a density of 1.5 × 10^4 cells per well in flat-bottom 96-well tissue culture plates. The following day, 2-fold dilutions of the homogenized tissue were applied to confluent cells and incubated for 1 h at 37°C. Then cells were washed with phosphate-buffered saline (PBS) and incubated for 72 h at 37°C in DMEM and 2% FBS. Cells were fixed with 4% paraformaldehyde for 20 min and stained with 0.05% (w/v) crystal violet in 20% methanol. The plate analysis was carried out by qualitative visual assessment of CPEs. TCID50 was determined using the Reed and Muench method.

qRT-PCR for viral copy quantification and gene expression analysis

For viral copy quantification and gene expression analysis, tissue homogenates were prepared by homogenizing perfused lungs or nasal turbinates (NTs) using a gentleMACS dissociator (Miltenyi) with program RNA_02 in M tubes in 1 mL or 500 μL Trizol (Invitrogen) for lungs or NTs, respectively. The homogenates were centrifuged at 2,000 × *g* for 1 min at 4°C, and then the supernatant was collected. RNA extraction was performed by combining phenol/guanidine-based lysis with silica membrane-based purification. Briefly, 100 μL of chloroform was added to 500 μL of homogenized sample; after centrifugation, the aqueous phase was added to 1 volume of 70% ethanol and loaded on a ReliaPrep RNA Tissue Miniprep column (Promega, catalog number Z6111). Total RNA was isolated according to the manufacturer's instructions. For viral copy quantification, quantitative polymerase chain reaction (qPCR) was performed using TaqMan Fast Virus 1 Step PCR Master Mix (Applied Biosystems); a standard curve was drawn with 2019_nCOV_N Positive control (Integrated DNA Technologies), and the following primers and probe were used: 2019-nCoV_N1 forward primer (5′-GAC CCC AAA ATC AGC GAA AT-3′), 2019-nCoV_N1 reverse primer (5′-TCT GGT TAC TGC CAG TTG AAT CTG-3′), and 2019-nCoV_N1 probe (5′-FAM-ACC CCG CAT TAC GTT TGG TGG ACC-BHQ1-3′) (Centers for Disease Control and Prevention [CDC] Atlanta, GA). All experiments were performed in duplicate.

For gene expression analysis of inflammation and endothelium-related genes, total RNA was retrotranscribed using SuperScript IV VILO Mastermix (Invitrogen, Thermo Fisher Scientific) according to the manufacturer's instructions. Quantitative real-time PCR was performed using TaqMan Fast Advanced Master Mix and specific TaqMan gene expression assays (listed in Table S1), both from Applied Biosystems (Thermo Fisher Scientific). The 7900HT Sequence Detection System instrument and software (Applied Biosystems) were used to quantify the mRNA levels of the target genes according to a six-point serial standard curve generated for each gene. The results were ultimately expressed, after normalization to the housekeeping gene Rlp32, as relative expression (fold change) compared with uninfected animals.

Histopathological analysis

PBS-perfused lungs were fixed in Zn-formalin for 24 h and then stored in 70% ethanol until trimming for paraffin wax embedding and the following histological examination. Consecutive sections (20 μm) were prepared and stained by the classic hematoxylin and eosin (H&E) method, and then microscopic observation was performed using a Nikon Eclipse 80i microscope equipped with a DXM1200F microscope camera. Pathological features in lung sections were scored as follows: inflammation-related parameters, including congestion of the alveolar septa, lymphomonocyte interstitial (alveolus) infiltrate, alveolar hemorrhage, interstitial edema, and platelet microthrombi, were evaluated separately by two independent pathologists, and the extent of these findings was scored arbitrarily using a two-tiered system: 0 (negative), 1 (moderate), and 2 (severe). All scores for each animal were ultimately summed up, and a global score was calculated for each group and expressed as the average ± SE. The percentage of pulmonary area affected by lesions in each section was also measured, and results for each group were reported as average percentage ± SE.

Electron microscopy sample preparation

A sample of SARS-CoV-2 Wuhan-Hu-1 6P-stabilized glycoprotein (native antigen) was incubated with scFv76 at a final concentration of 0.65 mg/mL and 0.22 mg/mL, respectively, for 1 h at RT. A 4-μL droplet of the sample was applied onto an R1.2/1.3 300-mesh copper holey carbon grid (Quantifoil) previously glow discharged for 30 s at 30 mA using a GloQube system (Quorum Technologies). The sample was incubated on the grid for 60 s at 4°C and 100% relative humidity, blotted, and plunge-frozen in liquid ethane using a Vitrobot Mk IV (Thermo Fisher Scientific).

EM data collection and image processing

Cryo-EM data were acquired on a Talos Arctica (Thermo Fisher Scientific) transmission electron microscope operated at 200 kV. The data were acquired using EPU-2.8 automated data collection software (Thermo Fisher Scientific). Movies were collected at a nominal magnification of 120,000x, corresponding to a pixel size of 0.889 Å/pixel at the specimen level, with applied defocus values between −0.8 and −2.2 μm. A total of 4,211 movies were acquired using the Falcon 3 direct electron detector (Thermo Fisher Scientific) operating in electron counting mode, with a total accumulated dose of 40 e^{-}/A2 distributed over 40 movie frames.

Movies were preprocessed with WARP 1.0.9.[21] A 5 × 5 × 40 model was used for motion correction using a 35-7 Å resolution range weighted with a −500 Å^2 B factor. Contrast transfer function (CTF) was estimated using the 40-3.5 Å resolution range and a 5 × 5 patch model. Particle picking was performed using the deep convolutional neural network BoxNet2Mask_20180918, resulting in 490,614 particles that were extracted in 400-pixel boxes and imported into CRYOSPARC-3.3.1[15] for further processing.

2D classification was used to select 366,967 particles that were 3D aligned using the SARS-CoV-2 spike glycoprotein model (EMDB: 21452) low pass filtered at 30 Å as a reference. Particles were subjected to 3D classification to select the final set of 87,623 particles that yielded an overall 3.5-Å resolution reconstruction based on the gold-standard criterion of 0.143 Fourier shell correlation (FSC) value. The initial reconstruction displayed two spike RBDs in the up conformation and one down, with all three showing one bound scFv76 fragment. Particles were subtracted with a mask comprising the entire spike molecule without the RBD in the down conformation and its corresponding scFv76 fragment and then locally refined to 4.0-Å resolution according to an FSC of 0.143.

Model building, refinement and validation, and structural analysis

The scFv76 structure was modeled with the Antibody Structure Prediction module using Schrödinger Maestro Bioluminate Suite 4.5.137, release 2021-4.[22] The modeling was performed with antigen-binding fragment (Fv) as antibody format. We used "EVQLLQ SAGGLVQPGGSLRLSCAASGFTVSANYMSWVRQAPGKGLEWV SVIYPGGSTFYADSVKGRFTISRDNSKNTLYLQMNSLRVEDTAV YYCARDLSVAGAFDIWGQGTLVTVSSGG" as the target sequence for the heavy chain (HC) and "IVLTQSPGTLSLSPGERATLSCRA SQSVSSSYLAWYQQKPGQAPRLLIYGASSRATGIPDRFSGSGSGT DFTLTISRLEPEDFAVYYCQQYGSSPYTFGQGTKLEIKRAAAGD YK" for the light chain (LC). The tool identifies the best matching framework templates for the queried sequences and builds the CDR loops based on the cluster analysis performed on the default antibody loop database, sieved using the selected framework template. To select a suitable structural reference framework, we filtered the possible candidates according to the scFv76 HC and LC germlines, IGHV3-66 and IGKV3-20, respectively.[12] We analyzed 130 CoV-AbDab structures of antibodies (Abs) bound to RBDs[23] and found that 39 of 42 entries with IGHV3-53/IGHV3-66 HCs, usually coupled with IGKV1-9 (16 Abs) or IGKV3-20 (10 Abs) LCs, share a common binding mode to the RBD. The three outliers are characterized by longer CDR-H3 (17–25 residues compared with 8–15 residues), and they are bound to IGLV2-14/IGLV2-23 LCs. Based on these findings, and considering the good resolution (2.03 Å) and the average of HC and LC similarity scores (0.99 of 1.00), we selected PDB: 7N3I as the reference framework, whose HC and LC are a combination of IGHV3-53 and IGKV3-20. To model the scFv76 CDR loops as well as their interaction with the antigen, the CDRs were grafted into a homology modeled structure built based on the reference framework template that also included the N-terminal domain of the betacoronavirus-like trimeric spike glycoprotein S1 (PDB: 7N3I). The generated scFv76:RBD model was subsequently superimposed on the three RBDs of a SARS-CoV-2 HexaPro S cryo-EM structure, with two RBDs in the up and one in the down conformation (PDB: 7N0H). Finally, the three cryo-EM RBDs in complex with the modeled scFv76 were refined with Schrödinger Protein Preparation Wizard[24] to remove clashes and optimize side chains.

The generated model was split into two parts: the first comprised the whole spike protein without the RBDs; the second consisted of the RBD in the down conformation in complex with scFv76. Both models were independently refined with COOT[25] and PHENIX[26] using the full reconstruction and the local refined map at 3.5 Å and 4.0 Å resolution, respectively. Subsequently, the local refined RBD:scFv76 complex was rigid body fitted in the full spike:scFv76 reconstruction. All data collection, image processing, and final model statistics are summarized in Table S2. The images were prepared using ChimeraX[27] and Pymol (http://www.pymol.org/pymol).

Statistical analysis

Statistical analyses were performed using Prism software (v.6.0 or v.8.0, GraphPad). Data are presented as average ±SE or SD. Statistical significance was analyzed with unpaired two-tailed Student's t test, Mann-Whitney U test, or one-way analysis of variance (ANOVA). p values below 0.05 were considered statistically significant.

DATA AND CODE AVAILABILITY

The antibodies described in the paper can be provided upon material transfer agreement (MTA) subscription. The full spike:scFv76 and the RBD:scFv76 cryo-EM volumes and the structure coordinates have been deposited in the Electron Microscopy Data Bank and the Protein

Data Bank under accession codes EMDB: 14628 and EMDB: 14629 and PDB: 7ZCE and PDB: 7ZCF, respectively. Cryo-EM videos were deposited in the Electron Microscopy Public Image Archive under accession code EMPIAR: 10990.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ymthe.2022.09.010.

ACKNOWLEDGMENTS

We thank Prof. Alessandro Rambaldi (Bergamo Hospital and Milan University) for suggestions and encouragement, the San Raffaele team for the effort and dedication devoted to the aerosol treatment of Delta-infected mice, and Bruno Bruni Ercole and Evelyn Vaccaro (Alfasigma) for excellent technical assistance. This work was funded by Alfasigma SpA.

AUTHOR CONTRIBUTIONS

Conceptualization, R.D.S., F.M.M., and O.M.; investigation, F.M.M., O.M., D.S., A.M.A., G.B., C. Chiapparino, A. Rosi, C.A., E.M., L.L., C.V., A. Riccio, A. Rossi, F.B., A.C.-S., C.P., and C. Camilloni; writing, F.M.M., R.D.S., and M.B.; supervision, R.D.S., M.B., E.M.P., L.G.S., and M.G.S.

DECLARATION OF INTERESTS

O.M., E.M.P., and R.D.S. are employees of Alfasigma SpA and are named as inventors in a patent application on the name of the same company.

REFERENCES

1. Araf, Y., Akter, F., Tang, Y.D., Fatemi, R., Parvez, M.S.A., Zheng, C., and Hossain, M.G. (2022). Omicron variant of SARS-CoV-2: genomics, transmissibility, and responses to current COVID-19 vaccines. J. Med. Virol. *94*, 1825–1832. https://doi.org/10.1002/jmv.27588.
2. Markov, P.V., Katzourakis, A., and Stilianakis, N.I. (2022). Antigenic evolution will lead to new SARS-CoV-2 variants with unpredictable severity. Nat. Rev. Microbiol. *20*, 251–252. https://doi.org/10.1038/s41579-022-00722-z.
3. Merad, M., Blish, C.A., Sallusto, F., and Iwasaki, A. (2022). The immunology and immunopathology of COVID-19. Science *375*, 1122–1127. https://doi.org/10.1126/science.abm8108.
4. Iketani, S., Liu, L., Guo, Y., Liu, L., Chan, J.F.-W., Huang, Y., Wang, M., Luo, Y., Yu, J., Chu, H., et al. (2022). Antibody evasion properties of SARS-CoV-2 Omicron sublineages. Nature *604*, 553–556. https://doi.org/10.1038/s41586-022-04594-4.
5. Liu, L., Iketani, S., Guo, Y., Chan, J.F.-W., Wang, M., Liu, L., Luo, Y., Chu, H., Huang, Y., Nair, M.S., et al. (2022). Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. Nature *602*, 676–681. https://doi.org/10.1038/s41586-021-04388-0.
6. Planas, D., Saunders, N., Maes, P., Guivel-Benhassine, F., Planchais, C., Buchrieser, J., Bolland, W.-H., Porrot, F., Staropoli, I., Lemoine, F., et al. (2022). Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. Nature *602*, 671–675. https://doi.org/10.1038/s41586-021-04389-z.
7. Hoffmann, M., Krüger, N., Schulz, S., Cossmann, A., Rocha, C., Kempf, A., Nehlmeier, I., Graichen, L., Moldenhauer, A.-S., Winkler, M.S., et al. (2022). The Omicron variant is highly resistant against antibody-mediated neutralization: implications for control of the COVID-19 pandemic. Cell *185*, 447–456.e11. https://doi.org/10.1016/j.cell.2021.12.032.
8. Cao, Y., Wang, J., Jian, F., Xiao, T., Song, W., Yisimayi, A., Huang, W., Li, Q., Wang, P., An, R., et al. (2022). Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. Nature *602*, 657–663. https://doi.org/10.1038/s41586-021-04385-3.
9. VanBlargan, L.A., Errico, J.M., Halfmann, P.J., Zost, S.J., Crowe, J.E., Jr., Purcell, L.A., Kawaoka, Y., Corti, D., Fremont, D.H., and Diamond, M.S. (2022). An infectious SARS-CoV-2 B.1.1.529 Omicron virus escapes neutralization by therapeutic monoclonal antibodies. Nat. Med. *28*, 490–495. https://doi.org/10.1038/s41591-021-01678-y.
10. Dejnirattisai, W., Huo, J., Zhou, D., Zahradník, J., Supasa, P., Liu, C., Duyvesteyn, H.M.E., Ginn, H.M., Mentzer, A.J., Tuekprakhon, A., et al. (2022). SARS-CoV-2 Omicron-B.1.1.529 leads to widespread escape from neutralizing antibody responses. Cell *185*, 467–484.e15. https://doi.org/10.1016/j.cell.2021.12.046.
11. Shuai, H., Chan, J.F.-W., Hu, B., Chai, Y., Yuen, T.T.-T., Yin, F., Huang, X., Yoon, C., Hu, J.-C., Liu, H., et al. (2022). Attenuated replication and pathogenicity of SARS-CoV-2 B.1.1.529 Omicron. Nature *603*, 693–699. https://doi.org/10.1038/s41586-022-04442-5.
12. Minenkova, O., Santapaola, D., Milazzo, F.M., Anastasi, A.M., Battistuzzi, G., Chiapparino, C., Rosi, A., Gritti, G., Borleri, G., Rambaldi, A., et al. (2022). Human inhalable antibody fragments neutralizing SARS-CoV-2 variants for COVID-19 therapy. Mol. Ther. *30*, 1979–1993. https://doi.org/10.1016/j.ymthe.2022.02.013.
13. Yamasoba, D., Kimura, I., Nasser, H., Morioka, Y., Nao, N., Ito, J., Uriu, K., Tsuda, M., Zahradnik, J., Shirakawa, K., et al. (2022). Virological characteristics of the SARS-CoV-2 Omicron BA.2 spike. Cell *185*, 2103–2115.e19. https://doi.org/10.1016/j.cell.2022.04.035.
14. Hsieh, C.-L., Goldsmith, J.A., Schaub, J.M., DiVenere, A.M., Kuo, H.C., Javanmardi, K., Le, K.C., Wrapp, D., Lee, A.G., Liu, Y., et al. (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. Science *369*, 1501–1505. https://doi.org/10.1126/science.abd0826.
15. Punjani, A., Rubinstein, J.L., Fleet, D.J., and Brubaker, M.A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat. Methods *14*, 290–296. https://doi.org/10.1038/nmeth.4169.
16. Okuya, K., Hattori, T., Saito, T., Takadate, Y., Sasaki, M., Furuyama, W., Marzi, A., Ohiro, Y., Konno, S., Hattori, T., and Takada, A. (2022). Multiple routes of antibody-dependent enhancement of SARS-CoV-2 infection. Microbiol. Spectr. *10*. e0155321-21. https://doi.org/10.1128/spectrum.01553-21.
17. Wang, Z., Deng, T., Zhang, Y., Niu, W., Nie, Q., Yang, S., Liu, P., Pei, P., Chen, L., Li, H., and Cao, B. (2022). ACE2 can act as the secondary receptor in the FcγR-dependent ADE of SARS-CoV-2 infection. iScience *25*, 103720. https://doi.org/10.1016/j.isci.2021.103720.
18. Junqueira, C., Crespo, Â., Ranjbar, S., de Lacerda, L.B., Lewandrowski, M., Ingber, J., Parry, B., Ravid, S., Clark, S., Schrimpf, M.R., et al. (2022). FcγR-mediated SARS-CoV-2 infection of monocytes activates inflammation. Nature *606*, 576–584. https://doi.org/10.1038/s41586-022-04702-4.
19. Tegally, H., Moir, M., Everatt, J., Giovanetti, M., Scheepers, C., Wilkinson, E., Subramoney, K., Moyo, S., Amoako, D.G., Baxter, C., et al. (2022). Continued emergence and evolution of Omicron in South Africa: new BA.4 and BA.5 lineages. Preprint at medRχiv. https://doi.org/10.1101/2022.05.01.22274406.
20. Riccio, A., Santopolo, S., Rossi, A., Piacentini, S., Rossignol, J.F., and Santoro, M.G. (2022). Impairment of SARS-CoV-2 spike glycoprotein maturation and fusion activity by nitazoxanide: an effect independent of spike variants emergence. Cell. Mol. Life Sci. *79*, 227. https://doi.org/10.1007/s00018-022-04246-w.
21. Tegunov, D., and Cramer, P. (2019). Real-time cryo-electron microscopy data preprocessing with Warp. Nat. Methods *16*, 1146–1152. https://doi.org/10.1038/s41592-019-0580-y.
22. Zhu, K., Day, T., Warshaviak, D., Murrett, C., Friesner, R., and Pearlman, D. (2014). Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. Proteins *82*, 1646–1655. https://doi.org/10.1002/prot.24551.
23. Raybould, M.I.J., Kovaltsuk, A., Marks, C., and Deane, C.M. (2021). CoV-AbDab: the coronavirus antibody database. Bioinformatics *37*, 734–735. https://doi.org/10.1093/bioinformatics/btaa739.
24. Sastry, G.M., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual

screening enrichments. J. Comput. Aided Mol. Des. *27*, 221–234. https://doi.org/10.1007/s10822-013-9644-8.

25. Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of coot. Acta Crystallogr. D Biol. Crystallogr. *66*, 486–501. https://doi.org/10.1107/S0907444910007493.

26. Liebschner, D., Afonine, P.V., Baker, M.L., Bunkóczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.W., Jain, S., McCoy, A.J., et al. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr. D Struct. Biol. *75*, 861–877. https://doi.org/10.1107/s2059798319011471.

27. Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and Ferrin, T.E. (2021). UCSF ChimeraX: structure visualization for researchers, educators, and developers. Protein Sci. *30*, 70–82. https://doi.org/10.1002/pro.3943.

YMTHE, Volume *31*

Supplemental Information

Spike mutation resilient scFv76 antibody counteracts SARS-CoV-2 lung damage upon aerosol delivery

Ferdinando M. Milazzo, Antonio Chaves-Sanjuan, Olga Minenkova, Daniela Santapaola, Anna M. Anastasi, Gianfranco Battistuzzi, Caterina Chiapparino, Antonio Rosi, Emilio Merlo Pich, Claudio Albertoni, Emanuele Marra, Laura Luberto, Cécile Viollet, Luigi G. Spagnoli, Anna Riccio, Antonio Rossi, M. Gabriella Santoro, Federico Ballabio, Cristina Paissoni, Carlo Camilloni, Martino Bolognesi, and Rita De Santis

Table S1. TaqMan gene assays (from Thermo Fisher Scientific) used for real time qPCR analyses.

Assay ID	Gene Symbol
Mm 00657574_s1	**ANGPT2**
Mm00441242_m1	**CCL2**
Mm01268754_m1	**CCL20**
Mm00486938_m1	**CDH5**
Mm03294838_g1	**COX2**
Mm01290062_m1	**CSF2**
Mm04207460_m1	**CXCL1**
Mm00445235_m1	**CXCL10**
Mm00438258_m1	**CXCR2**
Mm00516023_m1	**ICAM1**
Mm00515153_m1	**IFIT1**
Mm03030145_gH	**IFNA**
Mm00439552_s1	**IFNB**
Mm01168134_m1	**IFNG**
Mm00439620_m1	**IL1A**
Mm00434228_m1	**IL1B**
Mm00445259_m1	**IL4**
Mm00446190_m1	**IL6**
Mm01288386_m1	**IL10**
Mm00517640_m1	**IL21**
Mm01705338_s1	**lSG15**
Mm00487796_m1	**MX1**
Mm00840904_m1	**NLRP3**
Mm02528467_g1	**RLP32**
Mm00441278_m1	**SELE**
Mm00443258_m1	**TNF**
Mm01320970_m1	**VCAM1**

Table S2. Cryo-EM data collection, image processing and model refinement statistics.

Data collection and image processing		
Structure	Spike:scFv76 full complex	RBD:scFv76 complex (Focused refinement)
Microscope	Thermo Fisher Scientific TALOS Arctica	
Voltage (kV)	200	
Camera	Falcon 3EC	
Magnification	× 120,000	
Total electron dose (e$^-$/Å^2)	40.0	
Defocus range (µm)	-0.8 and -2.2 µm	
Pixel size (Å)	0.889	
Micrographs (no.)	4,211	
Symmetry imposed	C1	C1
Initial particle images (no.)	490,614	87,623
Final particle images (no.)	87,623	87,623
Resolution (Å)	3.5	4.0
(FSC threshold)	(0.143)	(0.143)
Sharpening B-factor (Å^2)	-136.6	-157.9
EMDB code	EMD-14628	EMD-14629
Model refinement		
Protein residues	3706	426
N-acetyl-D-glucosamine molecules	51	1
r.m.s. deviations		
Bond lengths (Å)	0.004	0.002
Bond angles (°)	0.663	0.643
Ramachandran plot		
Favored (%)	91.42	85.24
Allowed (%)	8.58	14.76
Disallowed (%)	0.00	0.00
Validation		
Molprobity score	1.96	2.18
Clashscore	8.35	10.15
Poor rotamers (%)	0.03	0.56
Map-model correlation	0.84	0.72
PDB code	7ZCE	7ZCF

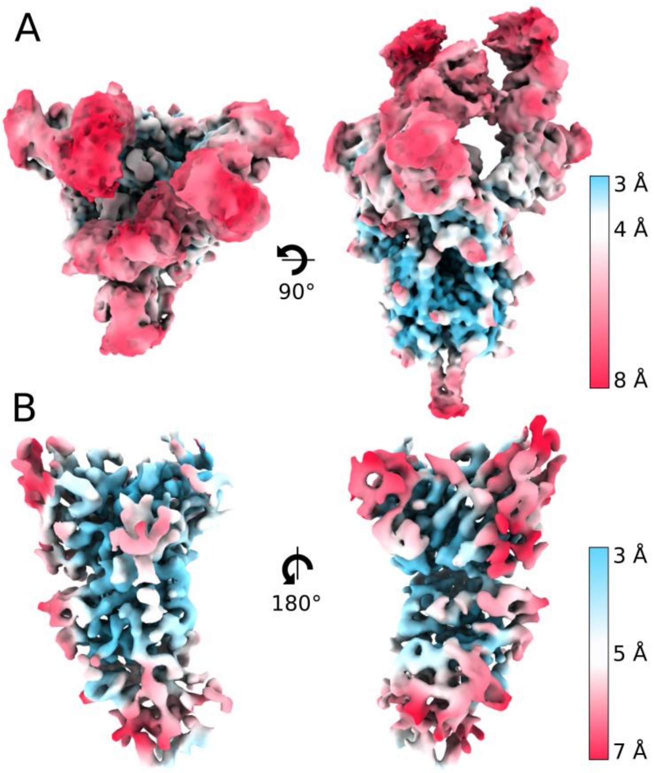


Figure S1. Local resolution cryo-EM map. A) spike:scFv76 full complex top and sides views. B) Locally refined RBD:scFv76 in the closed conformation, front and back views. Maps are colored according to the estimated local resolution.

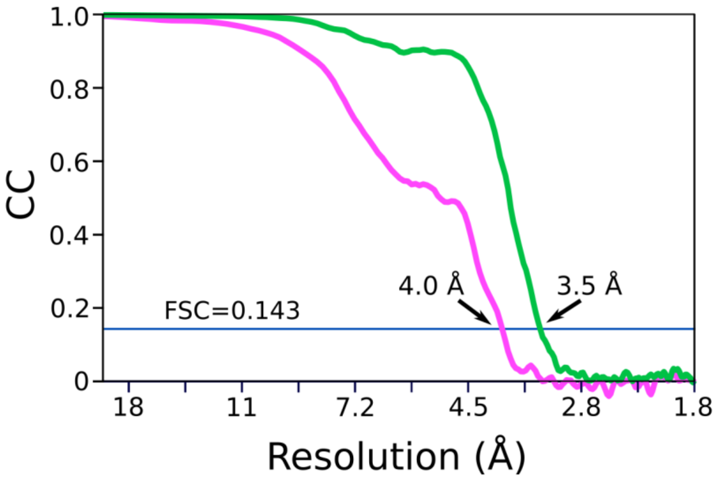


Figure S2. Cryo-EM FSC curves of spike:scFv76. The FSC curves for the spike:scFv76 full complex and the locally refined RBD:scFv76 in the closed conformation are shown in green and pink, respectively. The 0.143 threshold and the resolution cut-off for each curve are indicated.

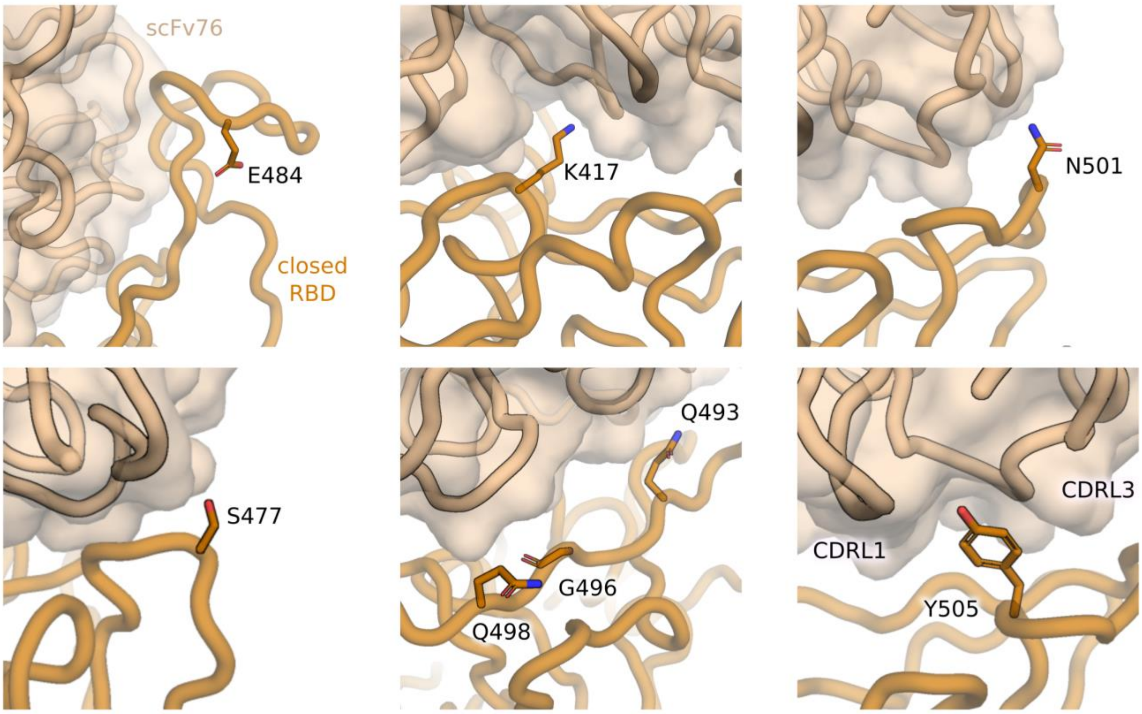


Figure S3. SARS-CoV-2 variants mutations at the RBD:scFv76 interface. Cartoon representation of scFv76 (transparent surface and worm) bound to the closed RBD (worm model); highlighted are key mutated residues (shown as stick models) in SARS-CoV-2 Omicron BA.1 and BA.2 variants.

ACS Medicinal Chemistry Letters

pubs.acs.org/acsmedchemlett

Letter

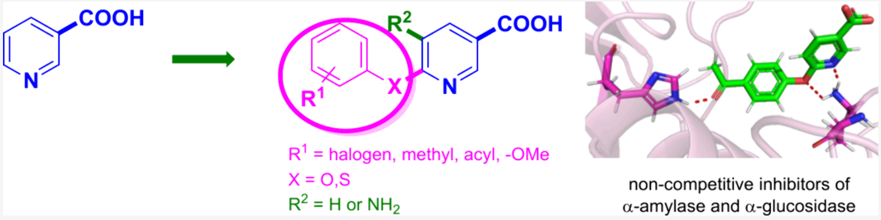
Nicotinic Acid Derivatives As Novel Noncompetitive α-Amylase and α-Glucosidase Inhibitors for Type 2 Diabetes Treatment

Andrea Citarella,* Miriam Cavinato, Elena Rosini, Haidi Shehi, Federico Ballabio, Carlo Camilloni, Valerio Fasano, Alessandra Silvani, Daniele Passarella, Loredano Pollegioni, and Marco Nardini*

Cite This: https://doi.org/10.1021/acsmedchemlett.4c00190

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information


COOH
N
R2
COOH
R1
X
N
R1 = halogen, methyl, acyl, -OMe
X = O,S
R2 = H or NH2
non-competitive inhibitors of
α-amylase and α-glucosidase

ABSTRACT: A library of novel nicotinic acid derivatives, focusing on the modification of position 6 of the pyridine ring with (thio)ether functionalities, was mostly produced through an innovative green synthetic approach (Cyrene-based) and evaluated for their α-amylase and α-glucosidase inhibitory activity. Compounds **8** and **44** demonstrated micromolar inhibition against α-amylase (IC_{50} of 20.5 and 58.1 μM, respectively), with **44** exhibiting a remarkable ~72% enzyme inactivation level, surpassing the efficacy of the control compound, acarbose. Conversely, **35** and **39** exhibited comparable inhibition values to acarbose against α-glucosidase (IC_{50} of 32.9 and 26.4 μM, respectively) and a significant enhancement in enzyme inhibition at saturation (~80–90%). Mechanistic studies revealed that the most promising compounds operated through a noncompetitive inhibition mechanism for both α-amylase and α-glucosidase, offering advantages for function regulation over competitive inhibitors. These inhibitors may open a new perspective for the development of improved hypoglycemic agents for type 2 diabetes treatment.

KEYWORDS: *synthesis, nicotinic acid, enzyme inhibitors, medicinal chemistry, organic chemistry*

Diabetes is a chronic metabolic disorder characterized by high blood sugar levels (hyperglycemia) over a prolonged period. This occurs due to either insufficient insulin production, insulin resistance, or both. Insulin is a pancreatic hormone that regulates blood sugar levels by facilitating the uptake of glucose from the bloodstream into cells to be used for energy or stored for future use. Currently, diabetes is one of the most important focal points in medical research, considering its significant social impact. Indeed, diabetes has been identified as one of the primary risk factors contributing to mortality worldwide.[1] Inadequate or ineffective treatment protocols can lead to various complications, such as stroke, cardiac arrest, limb amputation, vision loss, nervous system damage and an elevated risk of fetal death in poorly managed gestational forms of diabetes.[2]

There are several types of diabetes, but the most common ones are types 1 and 2. They share the similar symptom of elevated blood sugar levels but have different etiology. Type 1 is an autoimmune condition where the immune system mistakenly attacks and destroys the insulin-producing beta cells in the pancreas.[3] In type 2, the body's cells become resistant to the action of insulin, and the pancreas may gradually lose its ability to produce enough insulin to compensate. As a result, unlike type 1, type 2 diabetes is considerably more manageable and preventable through lifestyle interventions, which could delay absorption of glucose after meals.[4] In recent years, researchers actively investigated hypoglycemic agents with several mechanisms of action, with the aim of identifying molecules able to balance blood sugar uptake and insulin secretion during the postprandial stage. In this context, two pivotal targets for antidiabetic therapy have long been considered: α-amylase and α-glucosidase, which are key enzymes involved in saccharide hydrolysis.[5−7] Thus, the quest for an ideal hypoglycemic agent may revolve around inhibition of these enzymatic targets to improve glucose regulation in type 2 diabetes, leading to a significant amelioration in lifestyle and increased patient's life expectancy. Nicotinic acid, also known as niacin or vitamin B3, is a pyridine derivative showing a plethora of biological activities and therapeutic effects, including lipid-lowering activity, anti-inflammatory effects, vasodilatory effects and treatment of

Received: April 25, 2024
Revised: July 22, 2024
Accepted: July 31, 2024

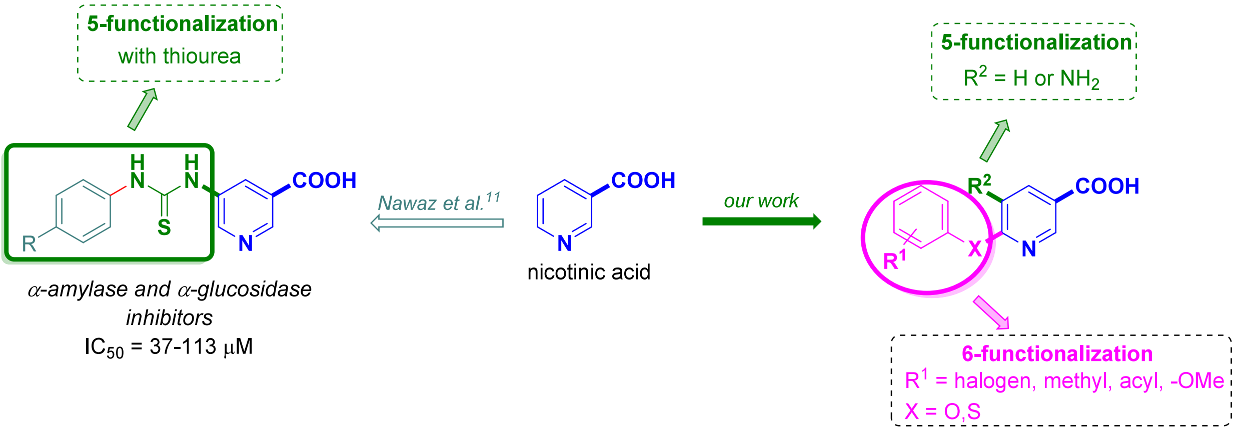
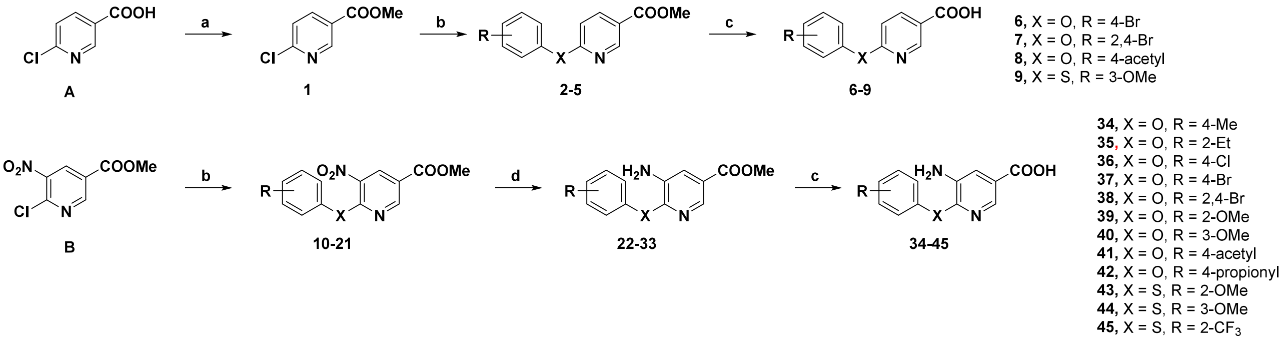


Figure 1. Design of the inhibitors. Chemical exploration of positions 5 and 6 of the nicotinic acid scaffold, proposed in this work, are shown in green and magenta, respectively (right side). Previously published thiourea moiety functionalization is shown in green (left side).

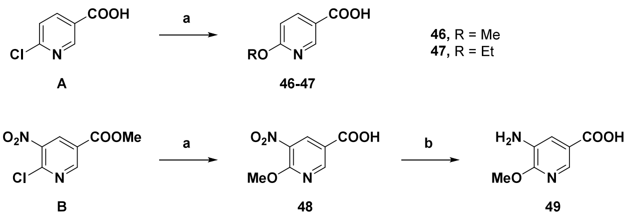
Scheme 1. Synthesis of Final Compounds 6−9 and 34−45[a]



[a]Reagents and conditions: (a) $TMSCHN_2$, toluene:MeOH 2:1 (*v*/*v*), rt, overnight, 98% yield; (b) appropriate phenol or thiophenol, NEt_3, Cyrene, 150 °C, sealed tube, 15−30 min, 60−95% yield; c) 1 M NaOH, MeOH, rt, 3−7 h, then 1 M $NaHSO_4$, 29−96% yield; d) Fe, NH_4Cl, EtOH:water 1:1 (*v*/*v*), 85 °C, 4 h, 23−88% yield.

pellagra.[8−10] Very recently nicotinic acid derivatives functionalized at position 5 with a thiourea moiety have been proposed as novel interesting α-amylase and α-glucosidase inhibitors.[11] Following our ongoing research about nicotinic acid derivatives,[12] we synthesized a library of 19 novel compounds and tested them for *in vitro* inhibition of α-amylase and α-glucosidase activity. Our chemical exploration was attempted both to position 5, keeping the hydrogen bond donor ($-NH_2$) as observed in the parental thiourea derivatives or removing it, and position 6 that was modified introducing (thio)ether moieties in order to explore available chemical space (Figure 1). In the case of aromatic ethers, phenols and thiophenols substituted with a plethora of electron withdrawing or electron donating groups were taken into consideration. Small substituents directly connected to the pyridine system, such as -OMe- or -OEt, were considered as aliphatic ether examples.

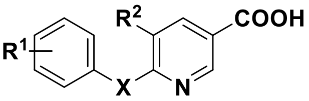
Scheme 2. Synthesis of Final Compounds 46, 47, and 49[a]



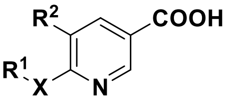
[a]Reagents and conditions: (a) MeONa or EtONa sol., refl, 24 h, *then* 1 M $NaHSO_4$, 17−49% yield; (b) Fe, NH_4Cl, EtOH:water 1:1 (*v*/*v*), 85 °C, 4 h, 35% yield.

The synthesis of the target compounds bearing 6-phenoxy or 6-phenylthio fragments was conveniently carried out following the pathways reported in Scheme 1. The presence of a free carboxylic group functionality troubled the introduction of the phenol or thiophenol nucleophile at position 6, thus the starting material 6-chloro-nicotinic acid (A) was first converted into the corresponding methyl ester **1** using MeOH in the presence of (trimethylsilyl)diazomethane ($TMSCHN_2$). For the series of 5-amino nicotinic acids, the starting material 6-chloro-5-nitro nicotinic ester (B) was used directly as the methyl ester. A S_NAr reaction allowed the synthesis of **2−5** and **10−21** pyridyl-phenyl ethers or thioethers through an innovative green synthetic approach recently published by our research group.[12] Briefly, starting compounds A or B were reacted with the appropriate phenol or thiophenol in the presence of NEt_3 as the base using the green solvent Cyrene under heating at 150 °C for 15−30 min in a sealed tube, affording the products **2−5** and **10−21** in good to optimal yields (60−95%). The advantage of this Cyrene-mediated methodology is to avoid the use of toxic and dangerous solvents (such as DMF or DMSO) and reduce reaction times. Moreover, the use of tedious chromatographic purification techniques was circumvented, as the pure products precipi-

Table 1. *In Vitro* Inhibitory Activity (IC_{50}, μM) and Inactivation % of Compounds **6**–**9**, **34**–**45**, **46**, **47**, and **49** against α-Amylase and α-Glucosidase[a]



Compound	X	R1	R2	α-amylase IC_{50} (μM)	α-amylase inactivation (%)	α-glucosidase IC_{50} (μM)	α-glucosidase inactivation (%)
Acarbose	-	-	-	11.1 ± 3.9 [0.66; 1.0; 62][6]	53.5 [99.2][6]	19.6 ± 1.4 [4300][15] [750][5] [363][16]	50.8 [71.8][15]
Miglitol	-	-	-	n.i.	-	34.5 ± 2.9 [465][16]	43.6
Voglibose	-	-	-	n.i.	-	14.7 ± 1.3 [320][16]	32.5
6	O	4-Br	H	108.5 ± 13.1	49.1	394.6 ± 23.1	38.6
7	O	2,4-diBr	H	176.9 ± 7.4	73.2	n.i.	-
8	O	4-acetyl	H	20.5 ± 2.6	38.5	n.i.	-
9	S	2-OMe	H	240.6 ± 19.1	80.7	544.6 ± 39.1	36.3
34	O	4-Me	NH_2	438.6 ± 33.6	44.3	353.2 ± 33.6	26.6
35	O	2-Et	NH_2	166.7 ± 11.6	54.9	32.9 ± 2.8	79.3
36	O	4-Cl	NH_2	363.9 ± 9.7	59.9	607.8 ± 20.7	28.3
37	O	4-Br	NH_2	331.0 ± 25.4	81.3	526.0 ± 35.3	35.6
38	O	2,4-diBr	NH_2	201.5 ± 18.9	53.1	457.8 ± 20.1	46.2
39	O	2-OMe	NH_2	162.4 ± 9.7	55.2	26.4 ± 2.0	87.3
40	O	3-OMe	NH_2	197.5 ± 13.8	75.5	565.8 ± 15.8	51.2
41	O	4-acetyl	NH_2	383.4 ± 15.4	50.6	532.9 ± 45.4	34.5
42	O	4-propionyl	NH_2	355.5 ± 10.8	52.3	n.i.	-
43	S	2-OMe	NH_2	653.7 ± 48.4	48.6	206.5 ± 18.4	56.8
44	S	3-OMe	NH_2	58.1 ± 4.1	71.5	348.5 ± 11.4	32.8
45	S	$2-CF_3$	NH_2	199.5 ± 13.4	53.5	155.1 ± 10.3	43.7



46	O	Me	H	527.5 ± 19.4	87.4	878.8 ± 37.4	14.5
47	O	Et	H	236.9 ± 18.8	69.4	442.5 ± 28.7	27.5
49	O	Me	NH_2	281.3 ± 25.0	80.3	639.0 ± 35.2	27.9

[a]Notes: n.i. = no inhibition.

tated from the reaction mixture after treatment with ice water. The reaction reached completion in 30 min whenever A was used as starting material, while 15 min of reaction time was needed for B. The different reaction time is explained by the presence of the 5-nitro group, which accelerated the S_NAr reaction. Compounds **10**–**21** were then subjected to a conventional reduction of the nitro group in the presence of iron under acid conditions, to afford methyl 5-amino nicotinate derivatives (**22**–**33**) in moderate to good yields (23–88%). In this case as well, there was no need to perform any column chromatography purification step for obtaining the clean product (except for compound **33**). After the reaction had been completed, the solvents were removed under reduced pressure; treatment of the crude mixture with a saturated solution of K_2CO_3 (until slightly alkaline pH) resulted in the precipitation of the pure products. It is noteworthy that in some instances, the addition of drops of MeOH facilitated the precipitation process and led to increased yields. Finally, alkaline hydrolysis provided the target 6-substituted nicotinic (**6**–**9**) or 5-amino-nicotinic acid compounds (**34**–**45**) in moderate to good yields (29–96%). Noteworthy, the final nicotinic acid compounds precipitated from the crude mixture after acidification with a 1 M $NaHSO_4$ solution (pH 5.5).

The synthesis of the target compounds bearing 6-methoxy or 6-ethoxy fragments was carried out following similar pathways (Scheme 2). The S_NAr reaction in this case could not be developed using the methodology seen for phenols or thiophenols because the strongly basic conditions expected

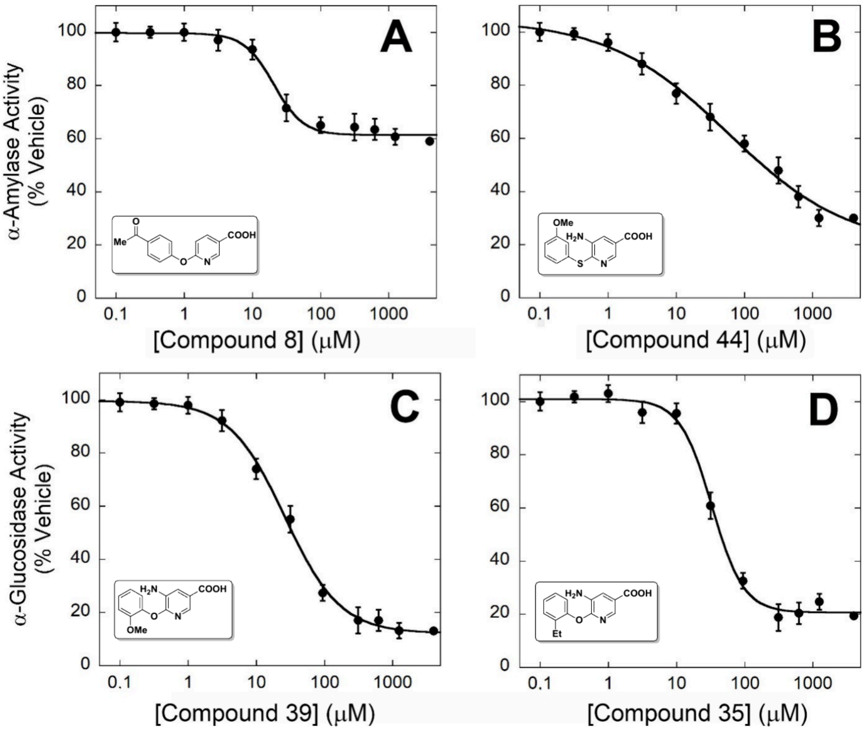


Figure 2. Enzyme inhibition assays performed on (A, B) α-amylase or (C, D) α-glucosidase by compounds **8**, **35**, **39**, and **44**. The enzymatic activity in the presence of different compound concentrations (in the 0−3500 μM range) was determined by colorimetric assays using an automated liquid-handler system. The plots display the mean values ± SD, n = 4.

from the presence of sodium alkoxides are incompatible with Cyrene, which is not stable and tends to polymerize.[13] However, it was possible to directly use the carboxylic acid as the starting material, thereby avoiding the subsequent hydrolysis step. Starting material A was reacted with alcoholic solutions of NaOMe or NaOEt at refluxing temperatures to provide **46** and **47** in low yields (17−35%). Starting material B was reacted with an alcoholic solution of NaOMe at refluxing temperatures to directly provide carboxylic acid derivative **48** in 49% yield. Finally, a reduction of the nitro group under the same reaction conditions observed before afforded **49** (35% yield). The structural confirmation and the purity of all the synthesized compounds was achieved by ^{1}H NMR, ^{13}C NMR, and HRMS. All compounds are >95% pure by HPLC analysis and representative HPLC traces are included in the Supporting Information.

The inhibitory activity of the target compounds **6**−**9**, **34**−**45**, **46**, **47** and **49** against α-amylase and α-glucosidase was measured by using an assay optimized from the one proposed by Nawaz et al.[11] The enzymes' source and the assay conditions are known to strongly affect the inhibition.[6] Acarbose was chosen as a standard control, and it showed a partial inhibition of both enzymes (~50%) and IC$_{50}$ values in the 10−20 μM range (Table 1). Regarding α-amylase inhibition, the enzymatic activity was assayed in 20 mM potassium phosphate using soluble starch, and the reaction mixture was incubated at 95 °C for 10 min before recording the absorbance intensity (see the Supporting Information for details). It was noted that the synthesized compounds demonstrated varying degrees of α-amylase inhibitory activity with IC$_{50}$ distributed over a wide range of values. Specifically, compounds **8** and **44** exhibited interesting α-amylase inhibition, with IC$_{50}$ values of 20.5 ± 2.6 and 58.1 ± 4.1 μM, respectively (Table 1 and Figure 2A, B). Interestingly, **44** resulted in a high degree of inactivation of ~71.5% (*i.e.*, ~25% of residual enzymatic activity at saturation), a value significantly improved in comparison to that of acarbose (46.5% of residual enzymatic activity at saturation). These findings suggested that in the series of nicotinic acid derivatives, the presence of an acetyl group at the *para* position of the phenyl ring (**8**) was more conducive to α-amylase inhibition compared to other substituents. On the other hand, in the series of 5-amino nicotinic acid derivatives, a *meta*-substituted phenyl ring with an −OMe substituent (**44**), whenever connected to the pyridine core with a sulfur bridge, led to the best inhibition activity. An analogue assay against α-glucosidase showed various degrees of inhibition ranging mostly in the half millimolar range (Table 1), with the exception of ethers **35** and **39**, which showed IC$_{50}$ values of 32.9 ± 2.8 and 26.4 ± 2.0 μM, respectively (Table 1 and Figure 2C, D). Enzyme inactivation at saturation for compounds **35** and **39** was higher than for acarbose (20.7% and 12.3% of residual enzymatic activity at saturation for **35** and **39**, instead of 49.2% for acarbose), showing a consistent improvement. These data demonstrated the significance of the NH$_2$ group presence at position 5 of the nicotinic acid scaffold for activity against α-glucosidase. Functionalization of the *ortho* position of the phenyl ring with electron-donating groups such as ethyl (**35**) and its isosteric replacement −OMe (**39**) led to compounds with remarkable activity. Furthermore, compounds **8**, **35** and **39** exhibited increased potency compared to the thiourea derivative disclosed by Nawaz et al. (range 37−113 μM),[11] indicating that introducing an ether or thioether functionality at position 6 was more effective in developing potent inhibitors. As further reference, miglitol and voglibose have been also evaluated: while both compounds did not

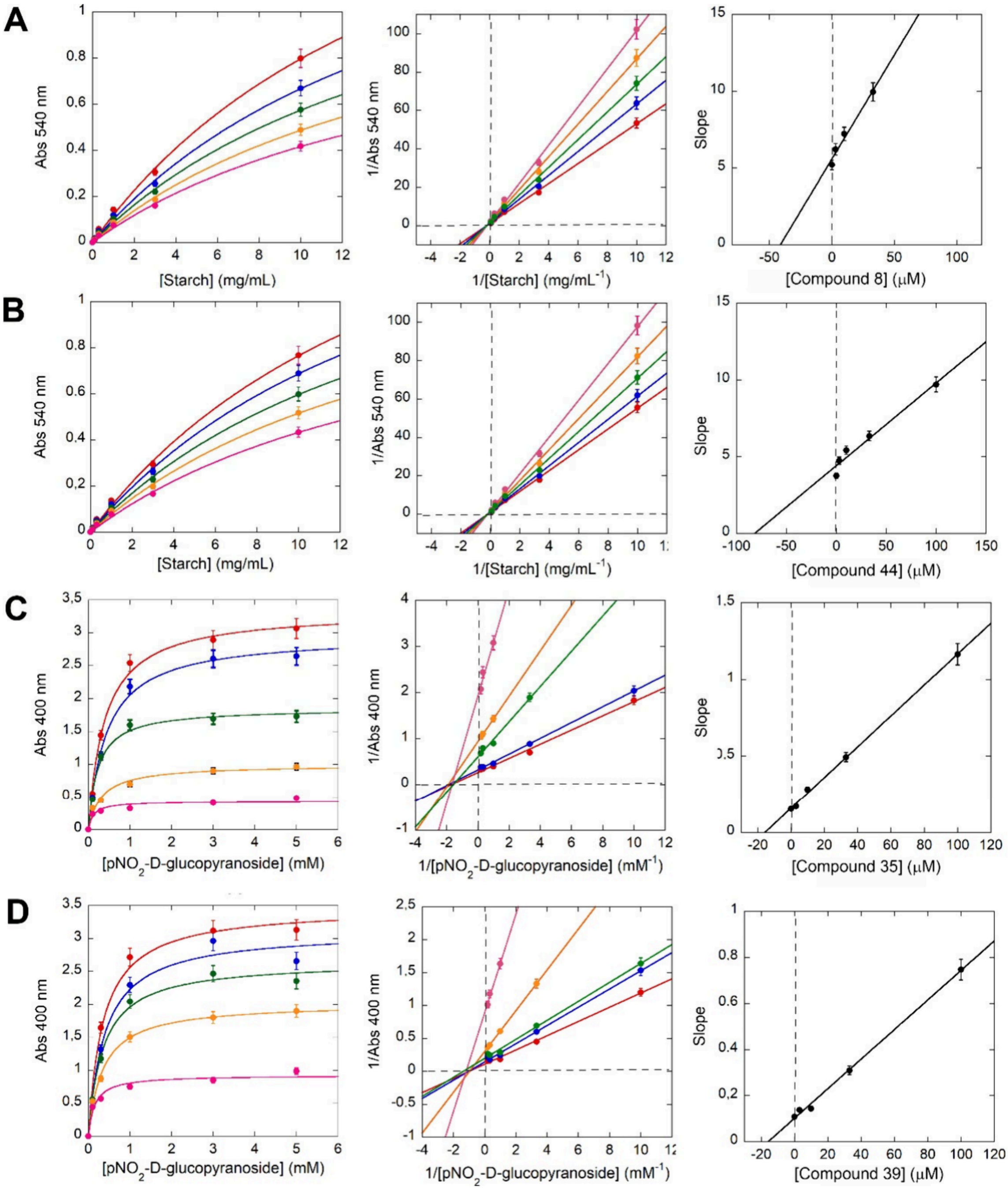


Figure 3. Inhibition of α-amylase (A,B) and α-glucosidase (C,D) by (A) **8**, (B) **44**, (C) **35**, and (D) **39**. The enzymatic activity was measured in the presence of increasing concentrations of inhibitor (0 μM, red line; 3 μM, blue line; 10 μM, green line; 33 μM, orange line; 100 μM, pink line) and different concentrations of (A, B) starch (0−10 mg/mL) or (C, D) p-nitrophenyl-α-D-glucopyranoside (0−5 mM, right panels). A Michaelis−Menten model of noncompetitive inhibition was globally fit: the Lineweaver−Burk equation was used to describe in a double reciprocal form the noncompetitive inhibition mechanism (central panels), and the K_i value was estimated by the tertiary plot of slope against compound concentration (linearly fitted, left panels). The plots display the mean values ± SD, n = 2.

modify the activity of α-amylase, in agreement with literature, a partial inhibition of α-glucosidase was determined.[14]

Then, the inhibitory mechanism of α-amylase by **8** and **44**, and of α-glucosidase by **35** and **39** was studied at different substrate and inhibitor concentrations using the same assays employed for IC$_{50}$ estimation (the starch in the amylase activity assay could not be used at a saturating concentration because of the solubility limit). In detail, all compounds tested showed a decrease in apparent V_{max} values with minimal effect on apparent K_m ones (Figure 3 and Table S1 in the Supporting Information). This result is consistent with tested compounds being noncompetitive inhibitors. Kinetic eqs (1) and (2) (see Materials and Methods Section in the Supporting Information) were used to calculate the inhibitory constant K$_i$, as reported in Table 2. A good correlation between K$_i$ and IC$_{50}$ values for each compound was observed (see Tables 1 and 2) and these values were slightly lower for α-glucosidase in comparison with α-amylase. For sake of comparison, the inhibition of α-amylase

Table 2. Inhibition constants of α-amylase by compounds 8, 44 and acarbose, and of α-glucosidase by compounds 35 and 39

Enzyme	Compound	K_i (μM)
α-Amylase	8	41.2 ± 3.1
	44	81.9 ± 4.1
	acarbose	62.4 ± 2.4
α-Glucosidase	35	17.1 ± 1.4
	39	15.9 ± 1.3
	acarbose	90.7 ± 1.8

and α-glucosidase was studied for acarbose under our experimental conditions: in both cases, a noncompetitive inhibition was apparent, with K_i values in the 60−90 μM range (Table 2 and Table S1).

Molecular docking was carried out to elucidate the binding mode of the active compounds in the putative allosteric sites of α-amylase and α-glucosidase. Probable allosteric sites were identified by using the Maestro suite (Figure S1 in the Supporting Information). Docking of **8** and **44** in a common allosteric site of α-amylase is depicted in Figure 4A and B. **8** establishes significant interactions within the pocket: a bifurcated hydrogen bond is observed between the amino group of Lys53 and the nitrogen of the pyridine ring and the ether oxygen; additionally, the carbonyl group of the acetophenone moiety of **8** interacts via a hydrogen bond with His491. **44**, on the other hand, is oriented in such a way that the carbonyl oxygen attached to the pyridine system establishes two H-bonds with the amino groups of the side chains of Lys53 and Arg392, respectively. Moreover, the NH_2 of the pyridine ring acts as a H-bond donor toward the oxygen of the amide bond of Lys457. Figure 4C and D describe the docking poses of **35** and **39** within the common allosteric site of α-glucosidase. The amino group of the pyridine ring of compound **35** forms an H-bond with the carbonyl oxygen of the peptide backbone of Phe543. In contrast, for compound **39**, a hydrogen bond is observed between the carbonyl oxygen and a hydrogen of the amino group within the side chain of Lys523. These *in silico* results provide a first picture of a likely interaction of the noncompetitive inhibitors with unprecedented allosteric sites for both target enzymes and, whenever confirmed by mutagenesis experiment, will provide the bases for a rational development of the inhibitors. In the past, acarbose was reported as uncompetitive inhibitor of barley amylase, able to bind a secondary binding site to give an abortive ESI complex.[17]

The predicted drug-like properties of compounds **8**, **35**, **39**, and **44** were examined using SwissADME (Absorption, Distribution, Metabolism, Excretion) online tool and they displayed favorable pharmacokinetic properties as shown in Table 3. Concerning their physicochemical properties, all of the synthesized compounds showed good solubility in water according to ESOL solubility this favoring drug formulation. A <5 lipophilicity (log*P*) was predicted for all molecules, indicating good permeability to the target tissue. However, the tested compounds showed no blood−brain barrier penetration and are predicted to follow the Lipinski rule of 5. Compounds **8**, **35**, **39**, and **44** are predicted to have a high GI absorbance and a comparable bioavailability score of 0.56.

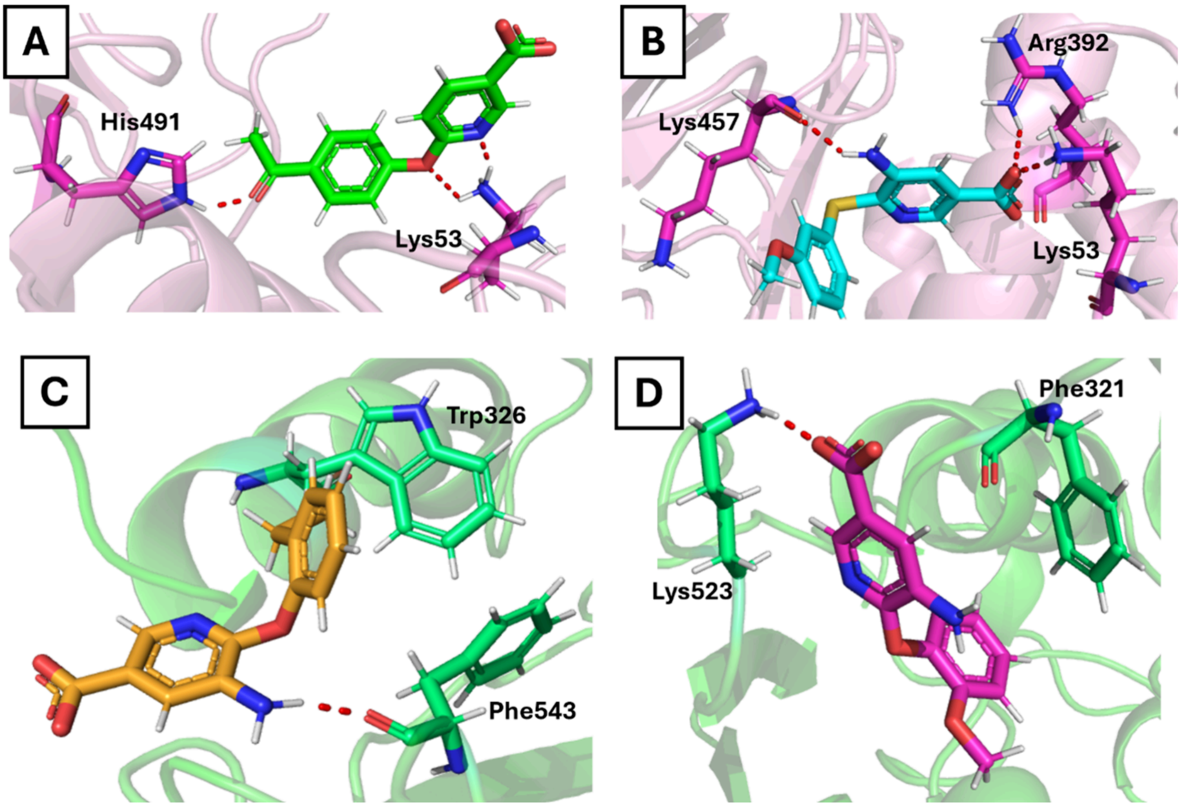


Figure 4. Binding modes of **8** (panel A) and **44** (panel B) into the putative allosteric site of α-amylase (PDB-code 1OSE) and binding modes of **35** (panel C) and **39** (panel D) into the putative allosteric site of α-glucosidase (PDB-code 3A4A). The residues involved in the interactions with the inhibitors are highlighted as sticks. H-bond interactions are displayed as red dashed lines.

Table 3. ADME and Drug-likeness of 8, 35, 39, and 44

Compound	Solubility (ESOL)	Consensus logP	H-bond acceptors	H-bond donors	Lipinski violations	GI absorption	BB Permeation	Bioavailability score
8	Soluble	1.8	5	1	No	High	No	0.56
35	Soluble	1.89	4	2	No	High	No	0.56
39	Soluble	1.25	5	2	No	High	No	0.56
44	Soluble	1.57	4	2	No	High	No	0.56

In conclusion, this work led to the discovery of novel nicotinic acid derivatives with the aim of investigating how modification of positions 5 and 6 could impact inhibitory activity against α-amylase and α-glucosidase. The target compounds were synthesized by employing a novel sustainable approach based on the use of the green solvent Cyrene. Remarkably, **8** and **44** exhibited micromolar inhibition values against α-amylase, with **44** demonstrating an ~72% enzyme inactivation level, a superior outcome compared to the control, acarbose. Concerning α-glucosidase, on the other hand, both **35** and **39** showed inhibition values comparable to acarbose but displayed a significant enhancement in their ability to strongly deactivate the enzyme at saturation, by approximately ~80−90% compared to the control. Notably, the inhibition mechanism of the most promising compounds turned out to be noncompetitive. This finding represents an important innovation since, tipically, inhibitors of α-amylase and α-glucosidase, used to alleviate postprandial glycemia, act *via* a reversible competitive mechanism.[6,7] In particular, this is also true for recently reported nicotinic-based inhibitors, showing the remarkable property to competitively inhibit both α-amylase and α-glucosidase.[11] The noncompetitive inhibition on both enzymes provided by our nicotinic acid 6-pyridine (thio)ether derivatives offers great advantages over a competitive inhibition due to the ability of the compounds to bind the enzymes at a site other than the active site, thereby not competing directly with the substrate. As a consequence, a noncompetitive inhibitor reduces the activity of the enzyme by binding equally well to the enzyme whether or not it has already bound to substrate, and its inhibition effect cannot be overcome by increasing substrate concentration. These preliminary results indicate that nicotinic acid scaffold could effectively be employed as an interesting pharmacophore in the design and optimization of new hypoglycemic drugs acting synergistically as noncompetitive inhibitors on both α-amylase and α-glucosidase, thus expanding the repertoires of potential strategies for type 2 diabetes treatment.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsmedchemlett.4c00190.

> Material and methods; procedures for the synthesis of intermediates and final compounds; α-amylase and α-glucosidase inhibition assay protocols; copies of ^{1}H and ^{13}C NMR of final compounds; HPLC traces for representative compounds (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Andrea Citarella − *Department of Chemistry, University of Milan, 20133 Milano, Italy;* orcid.org/0000-0001-5881-7142; Email: andrea.citarella@unimi.it

Marco Nardini − *Department of Biosciences, University of Milan, 20133 Milano, Italy;* orcid.org/0000-0002-3718-2165; Email: marco.nardini@unimi.it

Authors

Miriam Cavinato − *Department of Biosciences, University of Milan, 20133 Milano, Italy*

Elena Rosini − *Department of Biotechnology and Life Sciences, University of Insubria, 21100 Varese, Italy;* orcid.org/0000-0001-8384-7992

Haidi Shehi − *Department of Biosciences, University of Milan, 20133 Milano, Italy*

Federico Ballabio − *Department of Biosciences, University of Milan, 20133 Milano, Italy;* orcid.org/0000-0001-5702-3674

Carlo Camilloni − *Department of Biosciences, University of Milan, 20133 Milano, Italy;* orcid.org/0000-0002-9923-8590

Valerio Fasano − *Department of Chemistry, University of Milan, 20133 Milano, Italy;* orcid.org/0000-0003-1819-4483

Alessandra Silvani − *Department of Chemistry, University of Milan, 20133 Milano, Italy;* orcid.org/0000-0002-0397-2636

Daniele Passarella − *Department of Chemistry, University of Milan, 20133 Milano, Italy;* orcid.org/0000-0001-6180-9581

Loredano Pollegioni − *Department of Biotechnology and Life Sciences, University of Insubria, 21100 Varese, Italy;* orcid.org/0000-0003-1733-7243

Complete contact information is available at: https://pubs.acs.org/10.1021/acsmedchemlett.4c00190

Funding

This research was supported by PRIN2020 (Grant 2020K53E57) to Loredano Pollegioni and Marco Nardini. The support from Fondo di Ateneo per la Ricerca, Università degli Studi dell'Insubria, to Loredano Pollegioni and Elena Rosini is acknowledged.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Beydag-Tasöz, B. S.; Yennek, S.; Grapin-Botton, A. Towards a better understanding of diabetes mellitus using organoid models. *Nat. Rev. Endocrinol.* **2023**, *19* (4), 232−248.

(2) Papatheodorou, K.; Banach, M.; Bekiari, E.; Rizzo, M.; Edmonds, M. Complications of Diabetes 2017. *J. Diabetes Res.* **2018**, *2018*, No. 3086167.

(3) Atkinson, M. A.; Eisenbarth, G. S.; Michels, A. W. Type 1 diabetes. *Lancet (London, England)* **2014**, *383* (9911), 69−82.

(4) Zaccardi, F.; Webb, D. R.; Yates, T.; Davies, M. J. Pathophysiology of type 1 and type 2 diabetes mellitus: a 90-year perspective. *Postgrad. Med. J.* **2016**, *92* (1084), 63−9.

(5) Agrawal, N.; Sharma, M.; Singh, S.; Goyal, A. Recent Advances of α-Glucosidase Inhibitors: A Comprehensive Review. *Curr. Top. Med. Chem.* **2022**, *22* (25), 2069−2086.

(6) Kaur, N.; Kumar, V.; Nayak, S. K.; Wadhwa, P.; Kaur, P.; Sahu, S. K. Alpha-amylase as molecular target for treatment of diabetes mellitus: A comprehensive review. *Chem. Biol. Drug. Des.* **2021**, *98* (4), 539−560.

(7) Williams, L. K.; Zhang, X.; Caner, S.; Tysoe, C.; Nguyen, N. T.; Wicki, J.; Williams, D. E.; Coleman, J.; McNeill, J. H.; Yuen, V.; Andersen, R. J.; Withers, S. G.; Brayer, G. D. The amylase inhibitor montbretin A reveals a new glycosidase inhibition motif. *Nat. Chem. Biol.* **2015**, *11* (9), 691−6.

(8) Bodor, E. T.; Offermanns, S. Nicotinic acid: an old drug with a promising future. *Br. J. Pharmacol.* **2008**, *153* (Suppl 1), S68−S75.

(9) Carlson, L. A. Nicotinic acid: the broad-spectrum lipid drug. A 50th anniversary review. *J. Int. Med.* **2005**, *258* (2), 94−114.

(10) Gille, A.; Bodor, E. T.; Ahmed, K.; Offermanns, S. Nicotinic acid: pharmacological effects and mechanisms of action. *Annu. Rev. Pharmacol. Toxicol.* **2008**, *48*, 79−106.

(11) Nawaz, M.; Taha, M.; Qureshi, F.; Ullah, N.; Selvaraj, M.; Shahzad, S.; Chigurupati, S.; Waheed, A.; Almutairi, F. A. Structural elucidation, molecular docking, α-amylase and α-glucosidase inhibition studies of 5-amino-nicotinic acid derivatives. *BMC Chem.* **2020**, *14* (1), 43.

(12) Citarella, A.; Cavinato, M.; Amenta, A.; Nardini, M.; Silvani, A.; Passarella, D.; Fasano, V. A Green Approach to Nucleophilic Aromatic Substitutions of Nicotinic Esters in Cyrene. *Eur. J. Org. Chem.* **2024**, *27* (15), No. e202301305.

(13) Citarella, A.; Amenta, A.; Passarella, D.; Micale, N. Cyrene: A Green Solvent for the Synthesis of Bioactive Molecules and Functional Biomaterials. *Int. J. Mol. Sci.* **2022**, *23* (24), 15960.

(14) Poovitha, S.; Parani, M. In vitro and in vivo α-amylase and α-glucosidase inhibiting activities of the protein extracts from two varieties of bitter gourd (Momordica charantia L.). *BMC Complement. Altern. Med.* **2016**, *16* (Suppl1), 185.

(15) Feng, Y.; Nan, H.; Zhou, H.; Xi, P.; Li, B. Mechanism of inhibition of α-glucosidase activity by bavachalcone. *Food Sci. Technol.* **2022**, *42*, No. e123421.

(16) Kasturi, S.; Surarapu, S.; Uppalanchi, S.; Anireddy, J. S.; Dwivedi, S.; Anantaraju, H. S.; Perumal, Y.; Sigalapalli, D. K.; Babu, B. N.; Ethiraj, K. S. Synthesis and α-glucosidase inhibition activity of dihydroxy pyrrolidines. *Bioorg. Med. Chem. Lett.* **2017**, *27* (12), 2818−2823.

(17) Oudjeriouat, N.; Moreau, Y.; Santimone, M.; Svensson, B.; Marchis-Mouren, G.; Desseaux, V. On the mechanism of alpha-amylase. *Eur. J. Biochem.* **2003**, *270* (19), 3871−9.

A conformational fingerprint for amyloidogenic light chains.

Cristina Paissoni[1,a], Sarita Puri[1,a], Luca Broggini[2], Manoj K. Sriramoju[3], Martina Maritan[1], Rosaria Russo[1], Valentina Speranzini[1], Federico Ballabio[1], Mario Nuvolone[4], Giampaolo Merlini[4], Giovanni Palladini[4], Shang-Te Danny Hsu[3,5], Stefano Ricagno[1,2*], Carlo Camilloni[1*]

[1]Department of Bioscience, University of Milan, Milan, Italy.

[2]Institute of Molecular and Translational Cardiology, IRCCS, Policlinico San Donato, Milan, Italy.

[3]Institute of Biological Chemistry, Academia Sinica, Taipei, Taiwan.

[4]Department of Molecular Medicine, University of Pavia, Pavia, Italy; Amyloidosis Research and Treatment Center, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy.

[5]Institute of Biomedical Sciences, National Taiwan University, Taipei, Taiwan; International Institute for Sustainability with Knotted Chiral Meta Matter (SKCM2), Hiroshima University, Higashi-Hiroshima, Japan.

[a]CP and SP contributed equally to this work and share the first authorship.

*Stefano Ricagno and Carlo Camilloni

Email: stefano.ricagno@unimi.it, carlo.camilloni@unimi.it

Author Contributions: CP performed and analyzed simulations and analyzed SAXS experiments. SP prepared HDX-MS samples and analyzed the data. LB, MM, RR, and VS prepared SAXS samples and performed SAXS experiments. MKS collected HDX-MS data. FB took care of data sharing and deposition and contributed to data analysis. MV, GM, GP, SDH, SR and CC supervised the project. SP, SR and CC wrote the manuscript with contributions from all authors. SR and CC designed the project.

Competing Interest Statement: Disclose any competing interests here.

Classification: Biological sciences, Biophysics and Computational Biology.

Keywords: Amyloidogenic light chain, Conformational dynamics, Small angle X-ray scattering, Molecular dynamics, Hydrogen deuterium exchange.

This PDF file includes:

Main Text
Figures 1 to 6
Tables 1 to 2

Abstract

Immunoglobulin light chain amyloidosis (AL) shares with multiple myeloma (MM) the overproduction of one clonal light chain (LC), but whereas in MM patients LC molecules remain soluble in circulation, AL LCs misfold into toxic soluble species and amyloid fibrils that accumulate in internal organs, leading to completely different clinical manifestations. The large sequence variability of LCs has hampered our understanding of the mechanism leading to LC aggregation. Nevertheless, some biochemical properties associated with AL-LC are emerging. The stability of the dimeric LCs seems to play a role, but conformational dynamics and susceptibility to proteolysis have been identified as biophysical parameters that, under native conditions, can better distinguish AL-LCs from LCs found in MM. In this study, our goal was to delineate a conformational fingerprint that could discriminate AL from MM LCs. By subjecting four AL and two MM LCs to in vitro analysis under native conditions using small-angle X-ray scattering (SAXS), we observed that the AL LCs exhibited a slightly larger radius of gyration and greater deviation from the experimentally determined structure, indicating enhanced conformational dynamics. Integrating SAXS with molecular dynamics (MD) simulations to generate a conformational ensemble revealed that LCs can adopt multiple states, with VL and CL domains either bent or straight. AL-LCs favored a distinct state in which both domains were in a straight conformation, maximizing solvent accessibility at their relative interfaces. This unique conformation was experimentally validated by hydrogen-deuterium exchange mass spectrometry (HDX-MS). Such findings reconcile a wealth of experimental observations and provide a precise structural target for drug design investigations.

Significance Statement

The high sequence variability of antibody light chains complicates the understanding of the molecular determinants of their aggregation in AL patients. Extensive biophysical and structural analyses by our group and others have demonstrated that reduced kinetic and thermodynamic stability associated with higher conformational dynamics play a role in their amyloidogenic behavior, but specific structural elements contributing to these behaviors remain elusive. In addition, these features are not universal among all known LCs, fostering different interpretations of their aggregation mechanisms. By combining molecular dynamics simulations, small-angle X-ray scattering measurements, and hydrogen-deuterium mass exchange spectrometry, we found that enhanced conformational dynamics localized at CL-VL interface residues, coupled with structural expansion, are distinguishing features of amyloidogenic LCs.

Main Text

Introduction

Immunoglobulin light-chain (AL) amyloidosis is a systemic disease associated with the overproduction and subsequent amyloid aggregation of patient-specific light chains (LCs) (1-4). Such aggregation may take place in one or several organs, the heart and kidneys being the most affected ones (1). AL originates from an abnormal proliferation of a plasma cell clone that results in LCs overexpression and over-secretion in the bloodstream (1). LCs belonging both to lambda (λ) and kappa (κ) isotypes are associated with AL; however, λ-LCs are greatly overrepresented in the repertoire of AL patients. Specifically, AL-causing LCs (AL-LCs) most often belong to a specific subset of lambda germlines such as *IGLV6* (λ6), *IGLV1* (λ1), and *IGLV3* (λ3) (5-8).

λ-LCs are dimeric in solution with each subunit characterized by two immunoglobulin domains, a constant domain (CL) with a highly conserved sequence and a variable domain (VL) whose extreme sequence variability is the result of genomic recombination and somatic mutations (9-12). VL domains are generally indicated as the key responsible for LC amyloidogenic behavior. The observation that the fibrillar core in most of the structures of ex-vivo AL amyloid fibrils consist of VL residues further strengthens this hypothesis (13-17). However, in a recent Cryo-EM structure, a stretch of residues belonging to the CL domain is also part of the fibrillar core, and mass spectrometry (MS) analysis of several ex vivo fibrils from different patients indicates that amyloids are composed of several LC proteoforms including full-length LCs (18-21).

Interestingly the overproduction of a light chain is a necessary but not sufficient condition for the onset of AL. Indeed, the uncontrolled production of a clonal LC is often associated with Multiple Myeloma (MM), a blood cancer, but only a subset of MM patients develops AL, thus indicating that specific sequence/biophysical properties determine LC amyloidogenicity and AL onset

(12,22-24). To date, the extreme sequence variability of AL-LCs has prevented the identification of sequence patterns predictive of LC amyloidogenicity, however, it has been reproducibly reported that several biophysical properties correlate with LC aggregation propensity. AL-LCs display a lower thermodynamic and kinetic fold stability compared to non-amyloidogenic LCs found overexpressed in MM patients (named hereafter M-LCs) (12,20-24).

Interestingly, previous work on LCs has indicated how differences in conformational dynamics can play a role in the aggregation properties of AL-LCs (22-26). Oberti *et al.* have compared multiple λ-LCs obtained from either AL patients or MM patients identifying the susceptibility to proteolysis as the best biophysical parameter distinguishing the two sets (12). Weber *et al.* have shown, using a mice-derived κ-LC, how a modification in the linker region can lead to a greater conformational dynamic, an increased susceptibility to proteolysis, as well as an increased in vitro aggregation propensity (25). Additionally, AL-LC flexibility and conformational freedom have also been correlated to the proteotoxicity observed in patients affected by cardiac AL and experimentally verified in human cardiac cells and a *C. elegans* model (27,28). It is noteworthy that the amyloid LCs analyzed in this study were originally purified from patients with cardiac amyloidosis.

Here building on this previous work as well on our previous experience on β2-microglobulin, another natively folded amyloidogenic protein (29-33) we investigated the native solution state dynamics of multiple λ-LCs by combining MD simulations, SAXS, and HDX-MS. Interestingly, we found a unique conformational fingerprint of amyloidogenic LCs corresponding to a low-populated state characterized by extended linkers, with an accessible VL-CL interface and possible structural rearrangements in the CL-CL interface.

Results

SAXS suggests differences in the conformational dynamics of amyloidogenic and non-amyloidogenic LC. SAXS was acquired either in bulk or in-line with SEC for a set of LCs previously described (cf. **Table 1** and **Methods**). H3, H7, H18 (AL-LC), M7, and M10 (M-LC) were studied by Oberti, *et al.* (12) and identified in multiple AL or MM patients, while ex vivo fibrils of AL55 from heart, kidney, and fat tissue of an AL patient have been previously studied by Cryo-EM and MS (16,17,19,20). These LCs cover multiple germlines, with H18 and M7 belonging to the same germline (cf. **Table 1**). The sequence identity is the largest for H18 and M7 (91.6%) while is the lowest for AL55 and M7 (75.2%). A table showing the statistics for all pairwise sequence alignments is reported in **Table S1** in the Supporting Information. For H3, H7, and M7 a crystal structure was previously determined (12) while for H18, AL55, and M10 we obtained a model using either homology modeling (H18 and AL55) or AlphaFold2 (M10). Qualitatively, the SAXS curves in **Figure 1** did not reveal any macroscopic deviation of the solution behavior with respect to the crystal or model conformation. For each LC, we compared the experimental and theoretical curves calculated from the LC structures (cf. **Table 1**) analyzing the residuals and the associated $\mathcal{X}^2$. The analysis indicated a discrepancy between the model conformation and the data in the case of the AL-LCs, which was instead not observed in the case of the M-LCs. For AL-LC, residuals deviate from normality in the low q region, suggesting some variability in the global size of the system. Additionally, a weak trend distinguishing AL-LC from M-LCs could be identified in the radius of gyration (Rg) (cf. **Table 1**). H3, H7, H18, and AL55 display an Rg, as derived from the Guinier analysis of the SAXS curves, of 0.5 to 0.8 Å larger than M7 and M10. Overall, SAXS measurements point to less compact and more structurally heterogeneous AL-LCs compared to more compact and structurally homogeneous M-LCs.

MD simulations reveal a conformational fingerprint for amyloidogenic light chains. To investigate the conformational dynamics of the six LCs we performed Metadynamics Metainference (M&M) MD simulations employing the SAXS curves ($q<0.3$ $\mathring{A}$) as restraints (cf.

Methods) (34-37). Metainference is a Bayesian framework that allows the integration of experimental knowledge on-the-fly in MD simulations improving the latter while accounting for the uncertainty in the data and their interpretation. Metadynamics is an enhanced sampling technique able to speed up the sampling of the conformational space of complex systems. The combination of SAXS and MD simulations has been shown to be effective for multi-domain proteins as well as for intrinsically disordered proteins (38-40).

For each LC we performed two independent M&M simulations coupled by SAXS restraint, accumulating around 120-180 µs of MD per protein (cf. **Methods** and **Table S2** in the Supporting Information). The resulting conformational ensembles resulted in a generally improved agreement with the SAXS data employed as restraints (cf. **Table S2** and **Figure S1** in the Supporting Information). To investigate differences in the LC local flexibility we first analyzed the root mean square fluctuations (RMSF) for the CL and VL separately, averaging over the chains and the replicates, **Figure 2A**. The RMSF indicates comparable flexibility in most of the regions, with differences localized in the termini and in some loops. The VL of amyloidogenic LCs are generally more flexible than the M ones, but this may be associated with the lengths of their complementarity-determining regions (CDRs). Indeed, M10 has the longest and most flexible CDR1 and the shortest and least flexible CDR3. Unexpectedly, there are some differences also in the CL domains. Here, in **Figure 2A**, the AL-LCs are always more flexible in at least one region even if these differences are relatively small. Overall, the RMSF does not provide a clear indication to differentiate AL and M-LCs. To provide a global description of the dynamics of the six LC systems, we then introduced two collective variables, namely the elbow angle, describing the relative orientation of VL and CL dimers, and the distance between the VL and CL dimers center of mass, illustrated in **Figure 2B**.

In **Figure 3** we report the free energy surfaces (FES) obtained from the processing of the two replicates of each LC as a function of the elbow angle and the CL-VL distance calculated from their center of mass. The visual inspection of the FES indicates converged simulation: in all cases, the replicates explore a comparable free-energy landscape with comparable features. All six LC FES share common features: a relatively continuous low free energy region along the diagonal, spanning configurations where the CL and VL are bent and close to each other (state L_B), and configurations where the CL and VL domains are straight and at relative distance between 3.4 and 4.1 nm (state L_S). A subset of LCs, namely H18, M7, and AL55, display conformations where the domains are straight in line (elbow angle greater than 2.5 rad) and in close vicinity, with a relative distance between the center of mass of less than 3.4 nm (state G). Of note, H18 and M7 belong to the same germline, letting us speculate that this state G may be germline-specific. Most importantly, only the AL-LCs display configurations with CL and VL straight in line but well separated at relative distances greater than 4.1 nm, this state H seems to be a fingerprint specific for AL-LCs. A set of configurations exemplifying the four states is reported in **Figure 2**. The estimates of the populations for the four states L_B, L_S, G, and H are reported in **Table 2**. The quantitative analysis indicates that, within the statistical significance of the simulations, states L_B and L_S represent in all cases most of the conformational space. In the case of H18, AL55, and M7, the compact state G is also significantly populated (10-34%). The state H, associated with amyloidogenic LCs, is populated between 5 and 10% in H3, H7, H18, and AL55 and less than 1% in M7 and M10.

To identify additional differences between the conformations observed in state H and the rest of the conformational space, we focused our attention on the VL-VL and CL-CL dimerization interfaces. In **Figure 4,** we show the free energy as a function of the distance between the CL domains versus the distance between the VL domains for each of the four states for one of the two simulations performed on H3; the same analysis for all other simulations is shown in **Figures S2 to S7** in the Supporting Information. From the comparison of the FESes, it is clear that only in the conformations corresponding to the state H do the CL-CL dimers display an alternative configuration. In the case of H3, the CL-CL domains in the H state are characterized by a shift towards configurations characterized by a larger distance, the same is observed in the case of H18 and AL55, while in the case of H7 the H state is characterized by a smaller distance between the CL domains.

Our conformational ensembles allowed us to hypothesize a conformational fingerprint for AL proteins, namely the presence of a weakly but significantly populated state (H) characterized by a more extended quaternary structure, with VL and CL dimers well separated, and with perturbed CL-CL interfaces.

HDX independently validates the amyloidogenic LC conformational fingerprint. To gain further molecular insight into how the dynamics of the tertiary and quaternary structures can be differentiated in AL- and M-LCs, HDX-MS was performed on our set of proteins. HDX-MS probes the protein dynamics by monitoring the hydrogen-to-deuterium uptake over time and the obtained data well complement structural, biophysical, and computational data. Four LCs from our set (H3, H7, AL55, and M10) yielded good peptide sequence coverages of 98.6, 92.5, 98.6, and 99.1%, respectively (**Figures S8A-S11A,** and **Table S3** in the Supporting Information) while H18 and M7 were not included in this analysis due to their poor sequence coverage and were not further investigated.

HDX-MS analysis revealed subtle structural dynamics of the individual proteins. The most significant difference between the AL and M-LCs is observed for residues 34-50, which are part of both the VL-VL dimerization interface and, more importantly in the context of this work, the CL-VL interface. These residues show significantly higher deuterium uptake in all H-proteins, with H3 being the highest, implying that AL-LCs dimeric interfaces (VL-VL and CL-VL) are more dynamic and hence significantly destabilized than in M10 (**Figure 5, Figure S12** in the Supporting Information). The highly dynamic VL-VL interface of H3 also correlates well with its open VL-VL interface in a crystal structure (PDB 8P89) which houses two nanobodies interacting with each VL in a dimeric structure (28). On the other hand, residues 54-70, which are not part of either interface, show higher deuterium uptake and hence more dynamics in the M10 protein, which may be a result of redistributed dynamics due to the rigidity of its VL-VL interface, as observed previously (41,42) (**Figure 5**). In contrast, the VL-CL hinge regions (residues 100-120) show homogenous high flexibility in all the proteins due to their higher accessible surface area (**Figure 5**). As expected, the CL domains also show a similar pattern of deuterium uptake and hence flexibility in AL- and M-LCs, with a minor difference contributed by the rigid VL-CL interface containing residues 161-180 (**Figure S10** in the Supporting Information). This region shows significantly less deuterium uptake (rigid) in both AL and M proteins when compared to other peptides in the CL domain (**Figures S8B-S11B** in the Supporting Information). However, comparing the average uptake for this region (161-180) between AL and M proteins shows that H3 and AL55 have higher uptake than M10. In contrast, H7 is an exception with the lowest deuterium uptake in this region (**Figure S12** in the Supporting Information). These data are particularly interesting in the light of our simulations. The dimeric conformations identified in the H state, **Figure 3**, are characterized by higher accessibility for the CL-VL interface, which is in perfect agreement with the increased accessibility for the region 34-50 on the VL and 161-180 in the CL observed in the HDX-MS analysis. Notably, the H state of H7 is the only one in which the CL-CL interface is remarkably compact (see **Figure S3** in the Supporting Information), consistent with the lower H/D exchange for regions 161-180 observed in H7. Overall, the HDX-MS data provide an independent validation of the H-state predicted from our conformational ensembles.

Discussion

Understanding the molecular determinants of AL amyloidosis has been hampered by its high sequence variability in contrast to its highly conserved three-dimensional structure (8). In this work, building on our previous studies highlighting susceptibility to proteolysis as a property that can discriminate between AL-LC and M-LC, as well as the role of conformational dynamics in protein aggregation, we characterized LC conformational dynamics under the assumption that AL-LC proteins, despite their sequence diversity, may share a property that emerges at the level of their dynamics. We combined SAXS measurements with MD simulations under the integrative framework of Metainference to generate conformational ensembles representing the native state conformational dynamics of 4 AL-LC and 2 M-LC. While SAXS alone already indicated possible differences, its combination with MD allowed us to observe a possible low-populated state, which we refer to as state H, characterized by well-separated VL and CL dimers and a perturbed CL-CL interface, which is significantly populated in AL-LCs while only marginally populated in M-LCs. HDX measurements allowed us to independently validate this state by observing increased accessibility in CL-VL interface regions. Notably, our conformational ensembles are similar to those observed for a linker mutation in the case of a kappa LC (25). Furthermore, the presence of high-energy, so-called excited states associated with amyloidogenic proteins has been previously identified in the case of SH3 (43), and β2m (29).

Having established a conformational fingerprint for AL-LC proteins, it would be tempting to identify possible mutations that could be associated with the presence of the H state. Comparing the sequences and structures of M7 and H18, both of which belong to the IGLV3-19*01 germline, we can identify a single mutation, A40G, that could easily be associated with the appearance of the H state in H18. This mutation is located in the 37-43 loop, which H/D exchange showed to be more accessible in our three AL-LCs than in our M-LC (see **Figure 5 and Figure S12**), and it breaks a hydrophobic interaction with the methyl group of T165, as observed in the crystal structure of M7 (PDB 5MVG and **Figure S13**), potentially making T165 more accessible in H18 than in M7 (see **Figure 5, and Figure S12**). Comparing the H18 and M7 sequences with the germline reference sequence, we see that position 40 in IGLV3-19*01 is a glycine (see **Figure S13**). This would suggest the intriguing interpretation that the G40A mutation in M7 may increase the interdomain stability compared to the germline sequence, making it less susceptible to aggregation. However, it should also be noted that while this framework position is a glycine in H3, H7, and AL55, it is also a glycine in M10. Previous research has often focused on identifying, on a case-by-case basis, the key mutations that may be considered responsible for the emergence of the aggregation propensity, under the assumption that such aggregation propensity should not be present in germline sequences, but this assumption may be misleading given the observation that few germlines are strongly overrepresented in AL, suggesting that these starting germline sequences may be inherently more aggregation-prone than the germline genes that are absent or rarely found in AL patients. More generally, by comparing our AL-LC sequences with their germline references (**Figures S14 to S19** in the Supporting Information), we observe that all mutations fall exclusively in the variable domain, allowing us to exclude for these systems a direct role for residues in the linker region, as observed in ref (25,44), or in the constant domain, as observed in ref (45). Many mutations fall in the CDR regions, as expected, but others are found in the framework regions, both near the dimerization interface and in other regions of the protein. Regarding mutations in the CDRs, it has been suggested that AL-LC proteins may exhibit frustrated CDR2 and CDR3 loops, with few key residues populating the left-hand alpha helix or other high-energy conformations (46), resulting in the destabilization of the VL. In **Figures S14 to S19** in the Supporting Information, we have analyzed the Ramachandran plot obtained from our conformational ensembles, focusing only on those residues that most populate the left-hand alpha helix region, which are marked with a red circle symbol and whose Ramachandran is reported. Our data indicate the presence of residues populating left-hand alpha regions in the Ramachandran plot, but these are also found in the case of M-LCs, so our simulations do not allow to confirm or exclude this mechanism in our set of protein systems.

In conclusion, our study provides a novel, complementary, perspective on the determinants of the misfolding propensity of AL-LCs that we schematize in Figure 6. The identification of a high-energy state, with perturbed CL dimerization interfaces, extended linkers, and accessible regions in both the VL-CL and VL-VL interfaces may be the common feature interplaying with specific properties shown by previous work including the direct or indirect destabilization of both the VL-VL and CL-CL dimerization interfaces (22,23,45-48). Our conformational fingerprint is also consistent with the observation that protein stability does not fully correlate with the tendency to aggregate, whereas susceptibility to proteolysis and conformational dynamics may better to capture the differences between AL-LC and M-LC. In this context, our data allow us to rationally suggest that targeting the constant domain region at the CL-VL interface, which is more labile in the H state, maybe a novel strategy to search for molecules against LC aggregation in AL amyloidosis.

Materials and Methods

LC production and purification. Recombinant AL- (H3, H7, H18, and AL5) and M- (M7, M10) proteins were produced and purified from the host *E. coli* strain BL21(DE3). Firstly, the competent BL21(DE3) cells were transformed with plasmid pET21(b+), which contains genes encoding H3, H7, H18, AL55, M7, and M10 proteins. The transformed cells were selected for each plasmid by growing them on LB agar plates containing the antibiotic ampicillin at a final concentration of 100 µg/ml. For over-expression of protein, one colony was picked from each plate and grown overnight in 20 ml of LB broth containing ampicillin at a final concentration of 100 µg/ml. The overnight-grown cells were then used to inoculate a secondary culture in one liter of LB broth. The cells were grown until the turbidity (OD_{600nm}) reached between 0.6-0.8 and protein expression was subsequently induced by adding 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG) for 4 h. The bacterial cells containing overexpressed LCs were then harvested using a Backman Coulter centrifuge at 6000 rpm for 20 min at 4 °C. All the proteins were overexpressed as inclusion bodies. For protein purification, the inclusion bodies were isolated by cell lysis induced by sonication. The purification of inclusion bodies was performed by washing them with buffer containing 10 mM Tris (pH 8) and 1% triton X 100. The purified inclusion bodies were unfolded with buffer containing 6.0 M guanidinium hydrochloride (GdnHCl) for 4h at 4 °C. The unfolded LCs were then refolded in a buffer containing reduced and oxidized glutathione to assist in disulfide bond formation. The refolded proteins were subjected to anion exchange and size-exclusion chromatography steps for final purification. The level of protein purity was checked on 12% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) gels. The final protein concentration was measured using molecular weight and extinction coefficient of individual proteins. The purified proteins were stored at -20 °C for further use.

Additional Methods for SAXS, MD simulations, and HDX experiments are available in the Supporting Information. SAXS data are available on the SASBDB (cf. Dataset S1 in the supporting information). Simulations data are available on Zenodo (cf. Dataset S2 in the supporting information).

Acknowledgments

We acknowledge Martin A. Schroer and Dmitri Svergun at EMBL Hamburg, and Sonia Longhi at AFMB Marseille, for discussion and support on the SAXS data acquisition and analysis. We acknowledge PRACE for awarding us access to Piz Daint at CSCS, Switzerland. The authors acknowledge CINECA for an award under the ISCRA initiative for the availability of high-performance computing resources and support. This work was supported by the Italian Ministry of Research PRIN 2020 (20207XLJB2) and by CARIPLO/TELETHON Foundations (GJC23044). SP acknowledges Fondazione Veronesi for a postdoctoral fellowship. We also acknowledge Karen Hsu, and Yong-sheng Wang for protein quality check experiments, Shu-Yu Lin, and Ming-Jie Tsi of the MS core facility, IBC, Academia Sinica for their support in HDX-MS experiments.

References

1. G. Merlini, *et al.*, Systemic immunoglobulin light chain amyloidosis. *Nat Rev Dis Primers* **4**, 38 (2018).
2. L. M. Blancas-Mejia, *et al.*, Immunoglobulin light chain amyloid aggregation. *Chem. Commun.* **54**, 10664–10674 (2018).
3. P. Cascino, *et al.*, Single-molecule real-time sequencing of the M protein: Toward personalized medicine in monoclonal gammopathies. *American J Hematol* **97** (2022).
4. T. L. Poshusta, *et al.*, Mutations in specific structural regions of immunoglobulin light chains are associated with free light chain levels in patients with AL amyloidosis. *PLoS ONE* **4**, e5169 (2009).
5. V. Perfetti, *et al.*, The repertoire of λ light chains causing predominant amyloid heart involvement and identification of a preferentially involved germline gene, IGLV1-44. *Blood* **119**, 144–150 (2012).
6. T. V. Kourelis, *et al.*, Clarifying immunoglobulin gene usage in systemic and localized immunoglobulin light-chain amyloidosis by mass spectrometry. *Blood* **129**, 299–306 (2017).
7. R. L. Comenzo, Y. Zhang, C. Martinez, K. Osman, G. A. Herrera, The tropism of organ involvement in primary systemic amyloidosis: contributions of Ig VL germ line gene use and clonal plasma cell burden. *Blood* **98**, 714–720 (2001).
8. R. M. Absmeier, G. J. Rottenaicher, H. L. Svilenov, P. Kazman, J. Buchner, Antibodies gone bad – the molecular mechanism of light chain amyloidosis. *The FEBS Journal* **290**, 1398–1419 (2023).
9. P. C. Bourne, *et al.*, Three-dimensional structure of an immunoglobulin light-chain dimer with amyloidogenic properties. *Acta Crystallogr D Biol Crystallogr* **58**, 815–823 (2002).
10. M. L. Chiu, D. R. Goulet, A. Teplyakov, G. L. Gilliland, Antibody structure and function: The basis for engineering therapeutics. *Antibodies* **8**, 55 (2019).
11. J. M. Di Noia, M. S. Neuberger, Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76**, 1–22 (2007).
12. L. Oberti, *et al.*, Concurrent structural and biophysical traits link with immunoglobulin light chains amyloid propensity. *Sci Rep* **7**, 16809 (2017).
13. L. Radamaker, *et al.*, Cryo-EM reveals structural breaks in a patient-derived amyloid fibril from systemic AL amyloidosis. *Nat Commun* **12**, 875 (2021).
14. L. Radamaker, *et al.*, Role of mutations and post-translational modifications in systemic AL amyloidosis studied by cryo-EM. *Nat Commun* **12**, 6434 (2021).
15. L. Radamaker, *et al.*, Cryo-EM structure of a light chain-derived amyloid fibril from a patient with systemic AL amyloidosis. *Nat Commun* **10**, 1103 (2019).
16. P. Swuec, *et al.*, Cryo-EM structure of cardiac amyloid fibrils from an immunoglobulin light chain AL amyloidosis patient. *Nat Commun* **10**, 1269 (2019).
17. S. Puri, *et al.*, The cryo-EM structure of renal amyloid fibril suggests structurally homogeneous multiorgan aggregation in AL amyloidosis. *Journal of Molecular Biology* **435**, 168215 (2023).
18. S. Ricagno, *et al.*, Helical superstructures between amyloid and collagen VI in heart-derived fibrils from a patient with Light Chain Amyloidosis. [Preprint] (2023). Available at: https://www.researchsquare.com/article/rs-3625869/v1 [Accessed 16 April 2024].
19. F. Lavatelli, *et al.*, Mass spectrometry characterization of light chain fragmentation sites in cardiac AL amyloidosis: insights into the timing of proteolysis. *Journal of Biological Chemistry* **295**, 16572–16584 (2020).
20. G. Mazzini, *et al.*, Protease-sensitive regions in amyloid light chains: what a common pattern of fragmentation across organs suggests about aggregation. *The FEBS Journal* **289**, 494–506 (2022).
21. S. Dasari, *et al.*, Proteomic detection of immunoglobulin light chain variable region peptides from amyloidosis patient biopsies. *J. Proteome Res.* **14**, 1957–1967 (2015).
22. G. J. Rottenaicher, *et al.*, Molecular mechanism of amyloidogenic mutations in hypervariable regions of antibody light chains. *Journal of Biological Chemistry* **296**, 100334 (2021).

23. E. Rennella, G. J. Morgan, J. W. Kelly, L. E. Kay, Role of domain interactions in the aggregation of full-length immunoglobulin light chains. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 854–863 (2019).
24. E. S. Klimtchuk, *et al.*, Role of complementarity-determining regions 1 and 3 in pathologic amyloid formation by human immunoglobulin κ1 light chains. *Amyloid* **30**, 364–378 (2023).
25. B. Weber, *et al.*, The antibody light-chain linker regulates domain orientation and amyloidogenicity. *Journal of Molecular Biology* **430**, 4925–4940 (2018).
26. X. Sun, H. J. Dyson, P. E. Wright, Role of conformational dynamics in pathogenic protein aggregation. *Current Opinion in Chemical Biology* **73**, 102280 (2023).
27. M. Maritan, *et al.*, Inherent biophysical properties modulate the toxicity of soluble amyloidogenic light chains. *Journal of Molecular Biology* **432**, 845–860 (2020).
28. L. Broggini, *et al.*, Nanobodies counteract the toxicity of an amyloidogenic light chain by stabilizing a partially open dimeric conformation. *Journal of Molecular Biology* **435**, 168320 (2023).
29. L. Visconti, *et al.*, Investigating the molecular basis of the aggregation propensity of the pathological D76N mutant of beta-2 microglobulin: Role of the denatured state. *IJMS* **20**, 396 (2019).
30. B. M. Sala, *et al.*, Conformational stability and dynamics in crystals recapitulate protein behavior in solution. *Biophysical Journal* **119**, 978–988 (2020).
31. C. Camilloni, *et al.*, Rational design of mutations that change the aggregation rate of a protein while maintaining its native structure and stability. *Sci Rep* **6**, 25559 (2016).
32. T. Le Marchand, *et al.*, Conformational dynamics in crystals reveal the molecular bases for D76N beta-2 microglobulin aggregation propensity. *Nat Commun* **9**, 1658 (2018).
33. A. Achour, *et al.*, Biochemical and biophysical comparison of human and mouse beta-2 microglobulin reveals the molecular determinants of low amyloid propensity. *The FEBS Journal* **287**, 546–560 (2020).
34. M. Bonomi, C. Camilloni, M. Vendruscolo, Metadynamic metainference: Enhanced sampling of the metainference ensemble using metadynamics. *Sci Rep* **6**, 31232 (2016).
35. M. Bonomi, C. Camilloni, A. Cavalli, M. Vendruscolo, Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2**, e1501177 (2016).
36. C. Paissoni, A. Jussupow, C. Camilloni, Determination of protein structural ensembles by hybrid-resolution SAXS restrained molecular dynamics. *J. Chem. Theory Comput.* **16**, 2825–2834 (2020).
37. C. Paissoni, C. Camilloni, How to determine accurate conformational ensembles by metadynamics metainference: A chignolin study case. *Front. Mol. Biosci.* **8**, 694130 (2021).
38. D. Saad, *et al.*, High conformational flexibility of the E2F1/DP1/DNA complex. *Journal of Molecular Biology* **433**, 167119 (2021).
39. M. C. Ahmed, *et al.*, Refinement of α-synuclein ensembles against SAXS data: Comparison of force fields and methods. *Front. Mol. Biosci.* **8**, 654333 (2021).
40. F. E. Thomasen, K. Lindorff-Larsen, Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. *Biochemical Society Transactions* **50**, 541–554 (2022).
41. S. Puri, S.-T. D. Hsu, Oxidation of catalytic cysteine of human deubiquitinase BAP1 triggers misfolding and aggregation in addition to functional loss. *Biochemical and Biophysical Research Communications* **599**, 57–62 (2022).
42. K.-T. Ko, I.-C. Hu, K.-F. Huang, P.-C. Lyu, S.-T. D. Hsu, Untying a knotted SPOUT RNA methyltransferase by circular permutation results in a domain-swapped dimer. *Structure* **27**, 1224-1233.e4 (2019).
43. P. Neudecker, *et al.*, Structure of an intermediate state in protein folding and aggregation. *Science* **336**, 362–366 (2012).
44. C. N. Nokwe, *et al.*, The antibody light-chain linker is important for domain stability and amyloid formation. *Journal of Molecular Biology* **427**, 3572–3586 (2015).

45. G. J. Rottenaicher, R. M. Absmeier, L. Meier, M. Zacharias, J. Buchner, A constant domain mutation in a patient-derived antibody light chain reveals principles of AL amyloidosis. *Commun Biol* **6**, 209 (2023).
46. T. Pradhan, *et al.*, Mechanistic insights into the aggregation pathway of the patient-derived immunoglobulin light chain variable domain protein FOR005. *Nat Commun* **14**, 3755 (2023).
47. F. C. Peterson, E. M. Baden, B. A. L. Owen, B. F. Volkman, M. Ramirez-Alvarado, A single mutation promotes amyloidogenicity through a highly promiscuous dimer interface. *Structure* **18**, 563–570 (2010).
48. P. Kazman, *et al.*, Fatal amyloid formation in a patient's antibody light chain is caused by a single point mutation. *eLife* **9**, e52300 (2020).
49. B. Al-Lazikani, A. M. Lesk, C. Chothia, Standard conformations for the canonical structures of immunoglobulins. *Journal of Molecular Biology* **273**, 927–948 (1997).

Figures and Tables

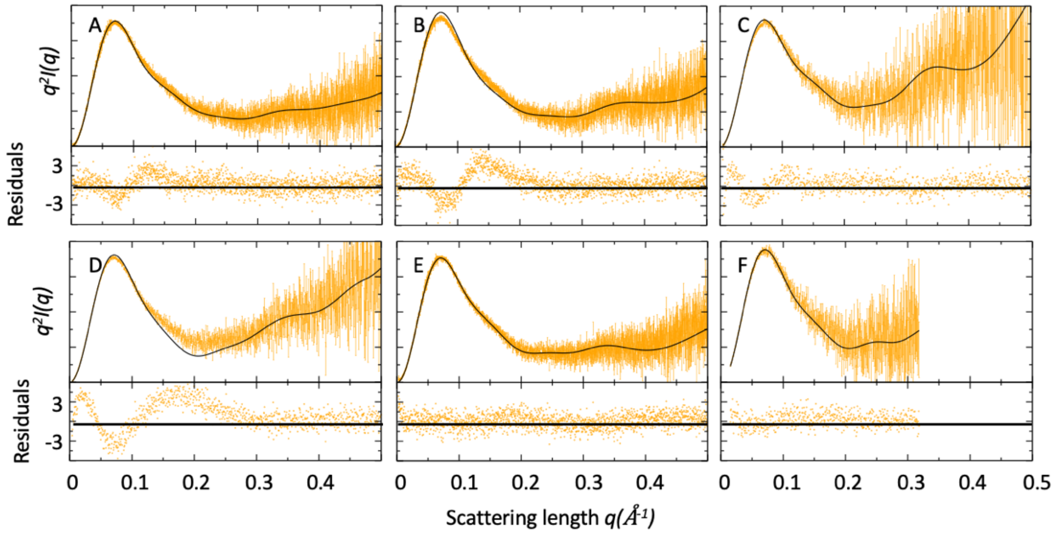


Figure 1. SAXS measurements for AL- and M light chains. Kratky plots comparing experimental (orange), and theoretical (black) curves and associated residuals (bottom panels) indicate that H LC solution behavior deviates from reference structures more than M LC. (A) H3 measured in bulk (Hamburg), 3.4 mg/ml. (B) H7 measured in bulk (Hamburg), 3.4 mg/ml. (C) H18 measured by online SEC-SAXS (ESRF), starting at 2.8 mg/ml. (D) AL55 measured in bulk (ESRF), 2.6 mg/ml. (E) M7 measured in bulk (Hamburg), 3.6 mg/ml. (F) M10 measured by online SEC-SAXS starting at 6.7 mg/ml (ESRF).

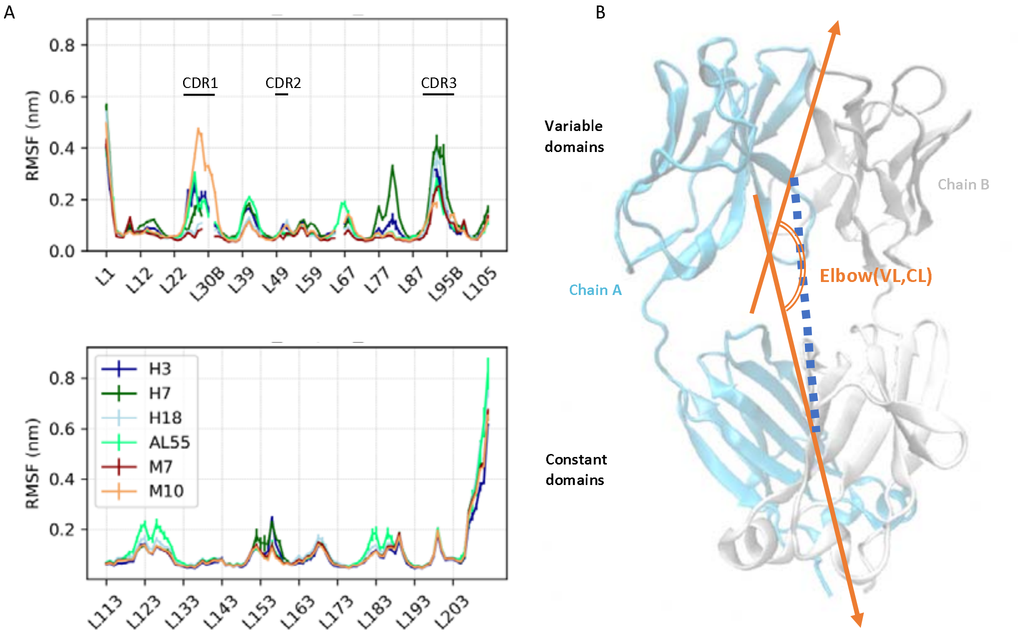


Figure 2. (A) Residue-wise root mean square fluctuations (RMSF) obtained by averaging the two Metainference replicates and the two equivalent domains for the six systems studied. The top panel shows data for the variable domains, while the bottom panel shows data for the constant domain. Residues are reported using Chothia numbering (49). (B) Schematic representation of two global collective variables used to compare the conformational dynamics of the different systems, namely the distance between the center of mass of the VL and CL dimers and the angle describing the bending of the two domain dimers.

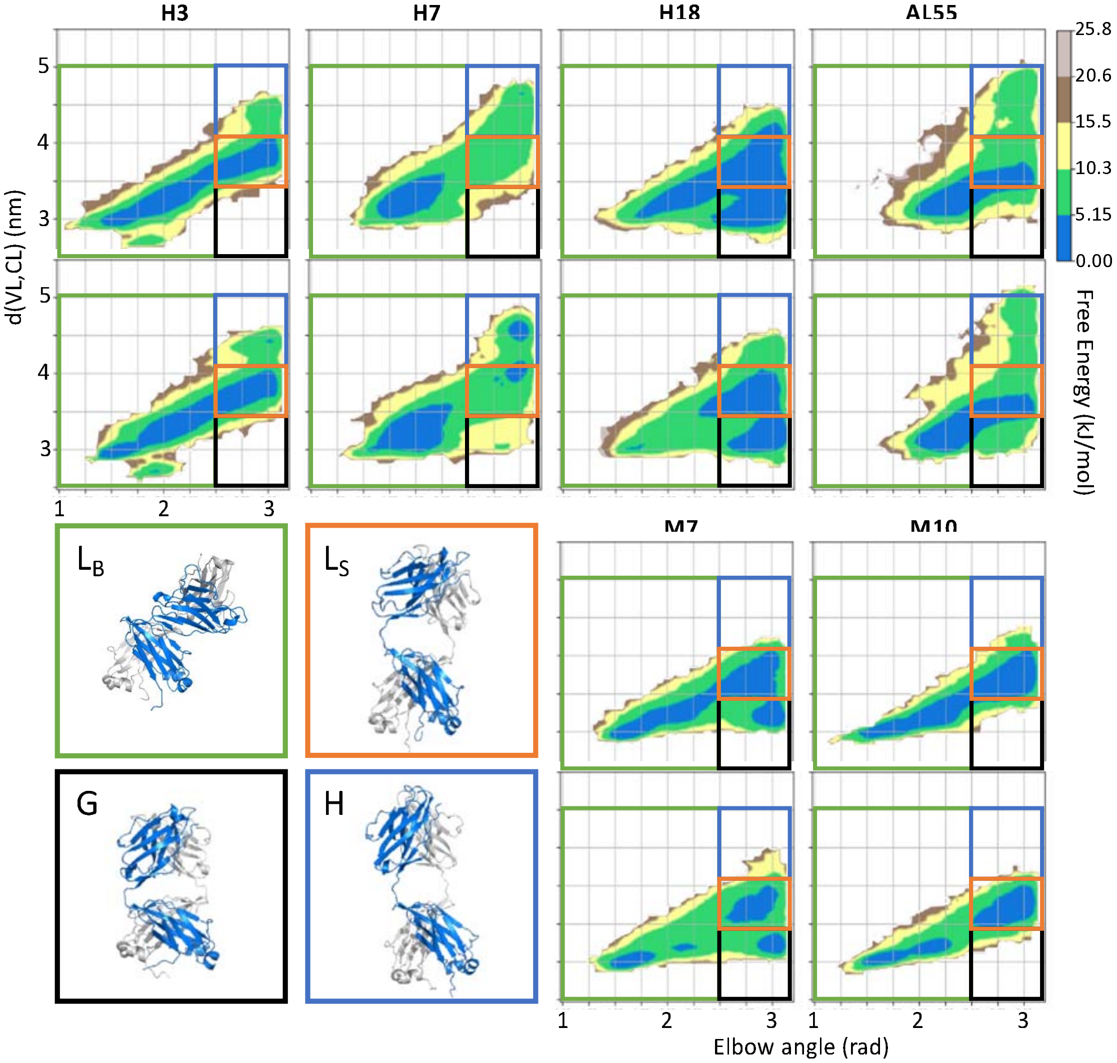


Figure 3. Free Energy Surfaces (FESes) for the six light chain systems under study by Metadynamics Metainference MD simulations. For each system, the simulations are performed in duplicate. The x-axis represents the elbow angle indicating the relative bending of the constant and variable domains (in radians), while the y-axis represents the distance in nm between the center of mass of the CL and VL dimers. The free energy is shown with color and isolines every $2k_BT$ corresponding to 5.16 kJ/mol. On each FES are represented four regions (green, red, blue, and black rectangles) highlighting their main features. For each region, a representative structure is reported.

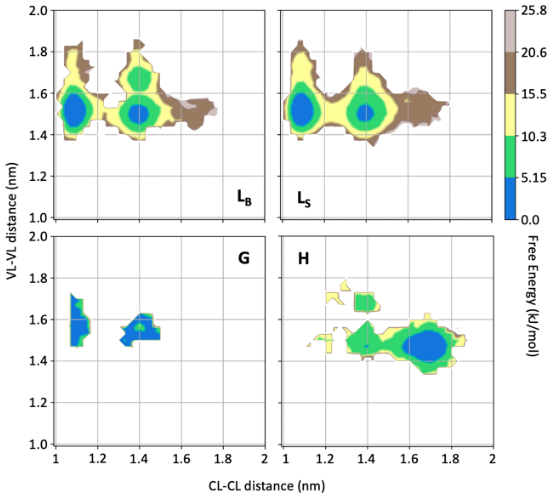


Figure 4. Free energy surfaces for the four substates identified in Figure 3 in the case of the first H3 Metainference simulation. The x-axis shows the distance between the centers of mass of the constant domains, while the y-axis shows the distance between the centers of mass of the variable domains. The free energy is shown with color and isolines every $2k_BT$ corresponding to 5.16 kJ/mol.

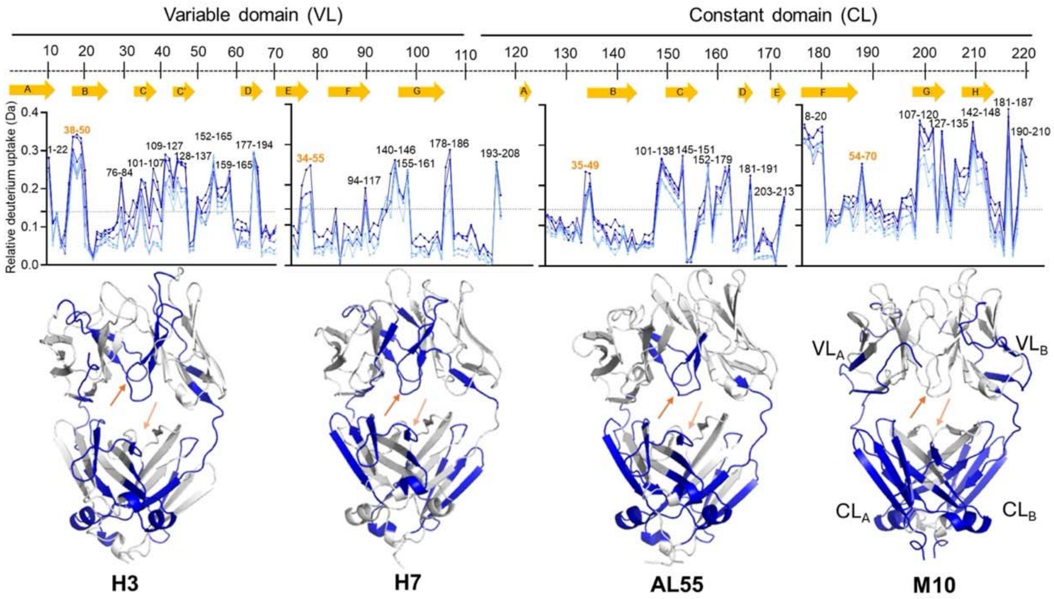


Figure 5. HDX-MS analysis. The top panel represents the simplified presentation of the primary structure of an LC including variable domain (VL) and constant domain (CL). The location of β-strands according to *Chothia and Lesk* (49). The middle panel represents the relative HDX butterfly plots of H3, H7, AL55, and M10 proteins. The peptides showing significantly higher deuterium uptake are labeled on their respective peaks. The peptide from residues 34-50 in AL-LCs and 54-70 in M-LC are labeled in orange. The lower panel represents the structural mapping of the selected peptides showing the highest deuterium uptakes using PyMOL. The VL-VL and VL-CL interfaces covering residues 34-50 and residues 161-180 are pointed with dark orange and light orange arrows, respectively.

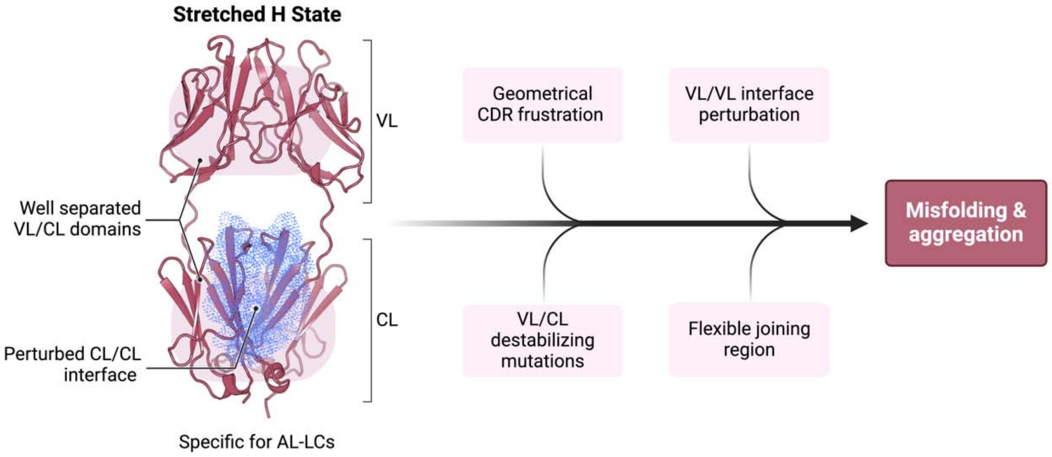

Stretched H State
VL
CL
Well separated VL/CL domains
Perturbed CL/CL interface
Specific for AL-LCs
Geometrical CDR frustration
VL/VL interface perturbation
VL/CL destabilizing mutations
Flexible joining region
Misfolding & aggregation

Figure 6. Schematic representation summarizing our findings in the context of previous work on the biophysical properties of amyloidogenic light chains. We propose that the H state is the conformational fingerprint distinguishing AL LCs from other LCs, which together with other features contributes to the amyloidogenicity of AL LCs.

Table 1. LC systems studied in this work. For each LC in the table are reported the germline, the phenotype, the structure or the method used to obtain one, the agreement between the structure and the SAXS curves, and the radius of gyration derived from the SAXS data.

LC	**Germline**	**Phenotype**	**Structure**	**SAXS χ^2 q<0.5 (q<0.3)**	**Rg (SAXS) [nm]**
H3	IGLV**1**-44*01	AL	5MTL	1.6 (1.9)	2.57 ± 0.02
H7	IGLV**1**-51*01	AL	5MUH	2.8 (4.0)	2.56 ± 0.02
H18	IGLV**3**-19*01	AL	*Homology*	1.6 (1.9)	2.56 ± 0.01
AL55	IGLV**6**-57*02	AL	*Homology*	5.1 (7.8)	2.58 ± 0.04
M7	IGLV**3**-19*01	MM	5MVG	1.2 (1.2)	2.50 ± 0.02
M10	IGLV**2**-14*03	MM	*AF2*	1.2 (1.2)	2.51 ± 0.01

Table 2. Populations of the four states shown in Figure 3 resulting from the two independent Metadynamics Metainference simulations performed for each of the 6 LCs. The population of the H state, which we supposed to be a fingerprint specific for AL-LCs, is in bold.

%	H3	H7	H18	AL55	M7	M10
L_B	62.0±0.4	72.3±2.5	22.9±2.8	48.3±0.1	46.8±0.1	48.4±0.3
L_S	33.0±0.2	15.2±2.7	38.4±2.1	32.5±2.0	35.0±0.1	49.1±1.5
G	0.2±0.1	0.8±0.4	33.8±3.8	10.5±1.4	17.6±0.1	1.8±0.6
H	**4.8±0.5**	**11.7±0.5**	**5.0±1.0**	**8.7±0.5**	0.6±0.2	0.8±0.6

Supporting Information for
A conformational fingerprint for amyloidogenic light chains.

Cristina Paissoni[1,a], Sarita Puri[1,a], Luca Broggini[2], Manoj K. Sriramoju[3], Martina Maritan[1], Rosaria Russo[1], Valentina Speranzini[1], Federico Ballabio[1], Mario Nuvolone[4], Giampaolo Merlini[4], Giovanni Palladini[4], Shang-Te Danny Hsu[3,5], Stefano Ricagno[1,2*], Carlo Camilloni[1*]

[1]Department of Bioscience, University of Milan, Milan, Italy.

[2]Institute of Molecular and Translational Cardiology, IRCCS, Policlinico San Donato, Milan, Italy.

[3]Institute of Biological Chemistry, Academia Sinica, Taipei, Taiwan.

[4]Department of Molecular Medicine, University of Pavia, Pavia, Italy; Amyloidosis Research and Treatment Center, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy.

[5]Institute of Biomedical Sciences, National Taiwan University, Taipei, Taiwan; International Institute for Sustainability with Knotted Chiral Meta Matter (SKCM2), Hiroshima University, Higashi-Hiroshima, Japan.

[a]CP and SP contributed equally to this work and share the first authorship.

*Stefano Ricagno and Carlo Camilloni

Email: stefano.ricagno@unimi.it, carlo.camilloni@unimi.it

This PDF file includes:

Supporting text
Figures S1 to S19
Tables S1 to S3
Legends for Movies S1 to S4
Legends for Datasets S1 to S2
SI References

Other supporting materials for this manuscript include the following:

Movies S1 to S4
Datasets S1 to S2

Supporting Information Text

Methods

Small-Angle X-ray scattering (SAXS)

For SAXS analysis, H3 was diluted to 3.4 mg/mL, H7 was diluted to 3.4 mg/mL, H18 was diluted to 2.8 mg/mL, AL55 was diluted to 2.6 mg/mL, M7 was diluted to 3.6 mg/mL, in 20 mM TrisHCl, 150 mM NaCl, pH 8. H3, H7 and M7 batch data were collected at the P12 BioSAXS beamline of the EMBL Hamburg Synchrotron (1), while AL55 batch data and H18 and M10 online SEC data were collected at the BM29 BioSAXS beamline of the ESRF, Grenoble (2). For SEC-SAXS, H18 and M10 were injected into a superdex 200 increase 10/300 GL column previously equilibrated in 20 mM TrisHCl, 150 mM NaCl, pH 8, at a concentration of 2.8 mg/mL and 6.7 mg/mL, respectively. SAXS data were processed using programs PRIMUS and GNOM within the ATSAS package (3). Data are deposited in the SASBDB (4) and available with accession codes.

Molecular dynamics simulations

The available crystallographic structures of H3, H7 and M7 (PDB: 5mtl, 5muh and 5mvg, respectively (5)) were used as starting conformations, using Modeller to add missing residues (6). H18 and AL55 were modelled by homology modelling using SwissModel (7), while M10 was modelled using AF2 (8). Simulations were performed using GROMACS 2019 (9) and the PLUMED2 software (10), using AMBER-DES force field and TIP4P-D water (11,12). During in-vacuum minimization RMSD-restraints were imposed to enhance the symmetry between the two constant and the two variable domains. The systems were solvated in a periodic dodecahedron box, initially 1.2 nm larger than the protein in each direction, neutralized with Na and Cl ions to reach a salt concentration of 10 mM, then minimized and equilibrated at the temperature of 310 K and pressure of 1 atm using the Berendsen thermostat and barostat. Two independent 900 ns long plain MD simulations were run to generate reliable and independent starting conformations for the Metadynamics Metainference simulations (13). 30 conformations were extracted from each simulation and duplicated by inverting the two chains, to obtain 60 starting conformations symmetrically distributed with respect to chain inversion.

Metadyamics Metainference production simulations were run in duplicate using 60 replicas, each replica evolved for ~1 μs (cf. Table S2). Simulations were performed in the NPT ensemble maintaining the temperature at 310 K with the Bussi thermostat (14) and the pressure of 1 atm with the Parrinello-Rahman barostat (15); the electrostatic was treated by using the particle mesh Ewald scheme with a short-range cut-off of 0.9 nm and van der Waals interaction cut-off was set to 0.9 nm. To reduce the computational cost, the hydrogen mass repartitioning scheme was used (16): the mass of heavy atoms was repartitioned into the bonded hydrogen atoms using the heavyh flag in the *pdb2gmx* tool, the LINCS algorithm was used to constraint all bonds, allowing to use a time step of 5 fs. In these simulations Parallel Bias Metadynamics (17) was used to enhance the sampling, combined with well-tempered metadynamics and the multiple-walker scheme, where Gaussians with an initial height of 1.0 kJ/mol were deposited every 0.5 ps, using a bias factor of 10. Five CVs were biased, including combinations of phi/psi dihedral angles of the linker regions (i.e. residues connecting variable and constant domains) in the two chains, combinations of chi dihedral angles of the linker regions in the two chains, combination of inter-domain contacts between the variable and the constant domains. The width of the Gaussians was 0.07, 0.12 and 120 for the combination of phi/psi, of chi dihedral angles and combination of contacts, respectively. Metainference was used to include SAXS restraints, using the hySAXS hybrid approach described in (18-20). A set of 13 representative SAXS intensities at different scattering angles, ranging between 0.015 Å^{-1} and 0.25 Å^{-1} and equally spaced, was used as restraints. These intensities were extracted from experimental data, after performing regularization with the Distance Distribution tool of Primus, based on Gnom (3). Metainference was applied every 5 steps, using a single Gaussian noise per data point and sampling a scaling factor between experimental and calculated SAXS intensities with a flat prior between 0.5 and 1.5. The aggregate sampling from the 60 replicas was reweighted using the final metadynamics bias to obtain a conformational ensemble where each conformation has an associated statistical weight (21). Convergence and error estimates were assessed by the inspection of the two replicated metadynamics metainference run. All relevant data are available on Zenodo.

Hydrogen-deuterium mass exchange spectrometry (HDX-MS)

SYNAPT G2-HDMS system (Waters Corporation, USA) equipped with a LEAP robotic liquid handler was used to perform HDX-MS measurements in a fully automated mode as described previously (22-25). The data collection was carried out by a 20-fold dilution of H3, H7, AL55, and M10 proteins (100 µM) with the labeling buffer containing 1X phosphate buffer saline (PBS) (pD 7.4) to trigger HDX for 0, 0.5, 1,10, 30, 120, and 240 min at 25 °C in triplicates. Each reaction was quenched by mixing the labeled protein with quench buffer (50 mM sodium phosphate, 250 mM TCEP, 3.0 M GdnHCl (pH 2)) in a 1:1 ratio at 0 °C. Online digestion was then performed using an immobilized pepsin digestion column (Waters Enzymate BEH Pepsin, 2.1 x 30 mm). The digested peptides were trapped using a C18 trapping column (Acquity BEH VanGuard 1.7 µm, 2.1 x 5.0 mm) and separated by a linear acetonitrile gradient of 5 to 40%. Protein Lynx Global Server (PLGS), and DynamX (Waters Corporation, USA) were used to identify the individual peptides, and subsequently, data processing using parameters: maximum peptide length of 25, the minimum intensity of 1000; minimum ion per amino acid of 0.1; maximum MH+ error of 5 ppm and a file threshold of three. A reference molecule [(Glu1)-fibrinopeptide B human (CAS No 103213-49-6, Merck, USA)] was used to lock mass with an expected molecular weight of 785.8426 Da. The relative deuterium uptakes of individual peptides were extracted from DynamX to generate heatmaps as a residue number function and peptides showing higher deuterium uptake are used for structural mapping on the modeled structures (26).

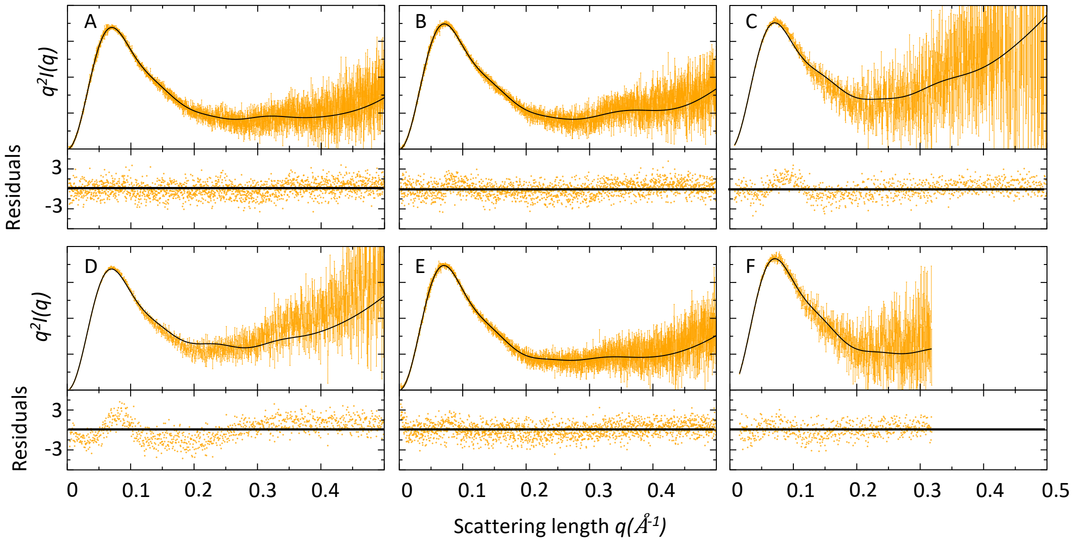


Figure S1: Kratky plots and associated residuals (bottom panels) comparing experimental (orange), and theoretical (black) curves obtained by averaging over the metainference ensemble. (A) to (F) report the data for H3, H7, H18, AL55, M7 and M10, respectively.

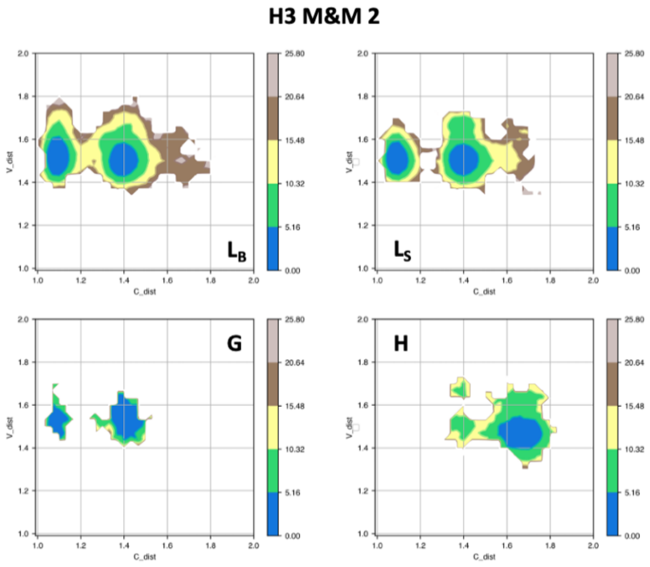


Figure S2: Free energy surfaces for the four substates identified in Figure 3 in the case of the second H3 metainference simulation. The x-axis shows the distance between the centers of mass of the constant domains, while the y-axis shows the distance between the centers of mass of the variable domains. The free energy is shown with color and isolines every $2k_BT$ corresponding to 5.16 kJ/mol.

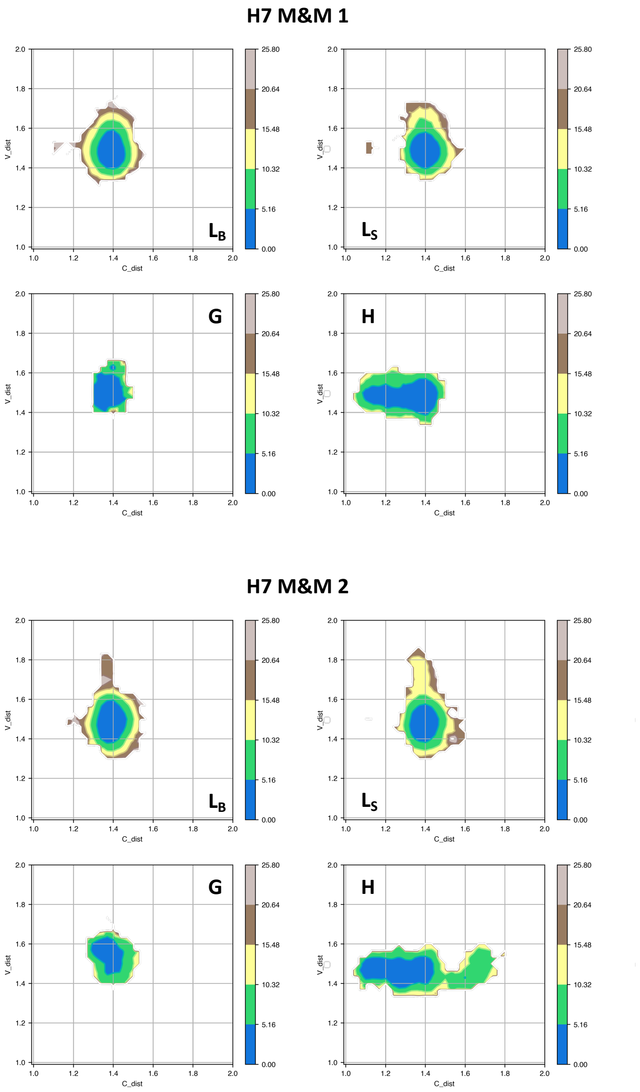


Figure S3: Free energy surfaces for the four substates identified in Figure 3 in the case of the first (top) and second (bottom) H7 metainference simulation. The x-axis shows the distance between the centers of mass of the constant domains, while the y-axis shows the distance between the centers of mass of the variable domains. The free energy is shown with color and isolines every 2k$_B$T corresponding to 5.16 kJ/mol.

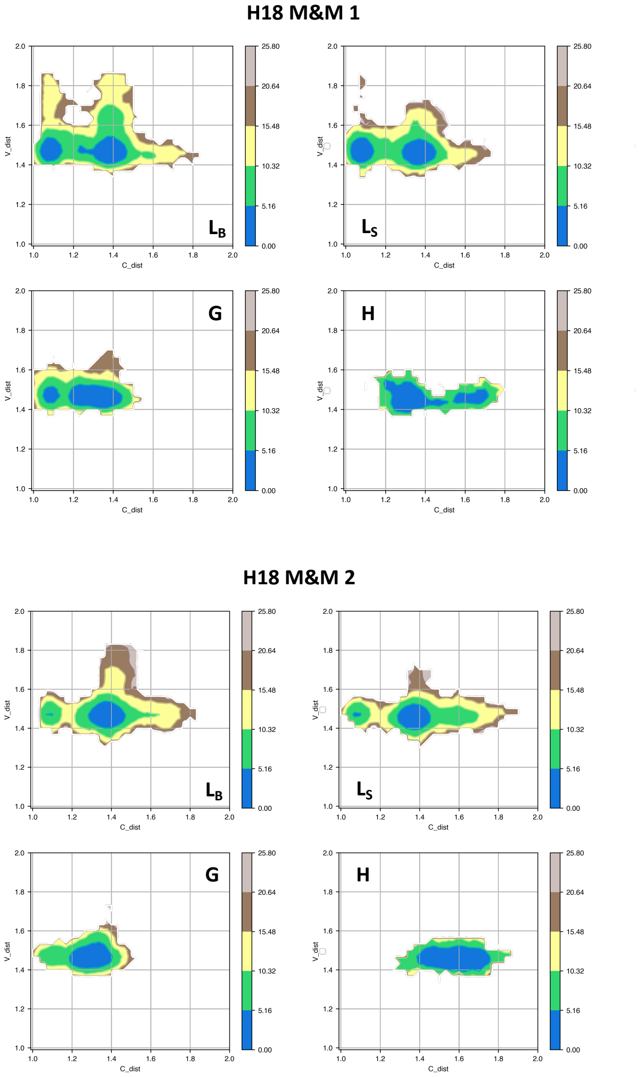


Figure S4: Free energy surfaces for the four substates identified in Figure 3 in the case of the first (top) and second (bottom) H18 metainference simulation. The x-axis shows the distance between the centers of mass of the constant domains, while the y-axis shows the distance between the centers of mass of the variable domains. The free energy is shown with color and isolines every $2k_BT$ corresponding to 5.16 kJ/mol.

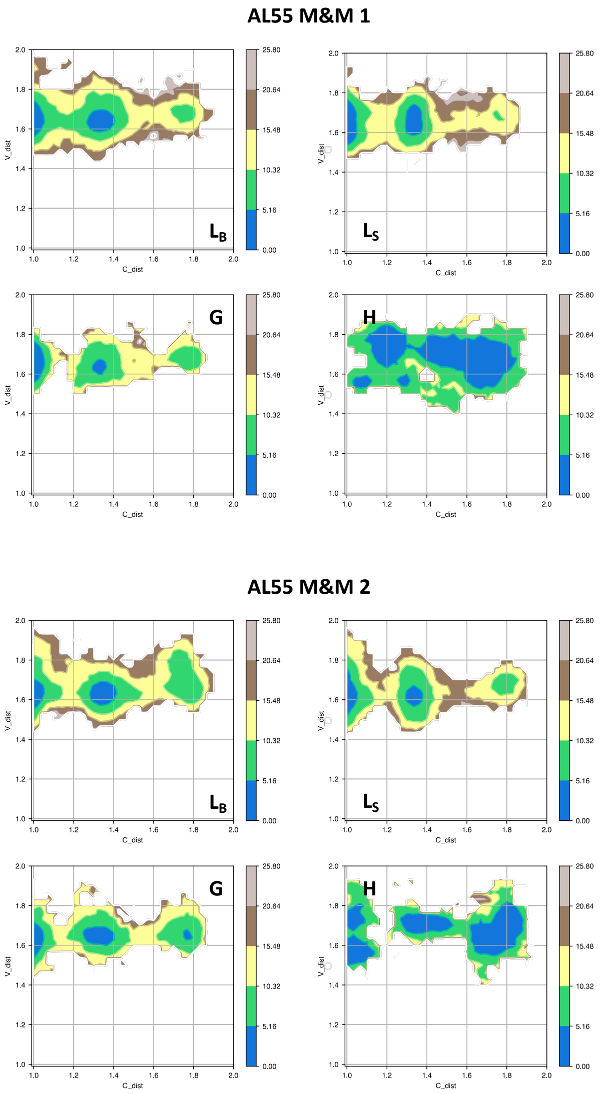


Figure S5: Free energy surfaces for the four substates identified in Figure 3 in the case of the first (top) and second (bottom) AL55 metainference simulation. The x-axis shows the distance between the centers of mass of the constant domains, while the y-axis shows the distance between the centers of mass of the variable domains. The free energy is shown with color and isolines every $2k_BT$ corresponding to 5.16 kJ/mol.

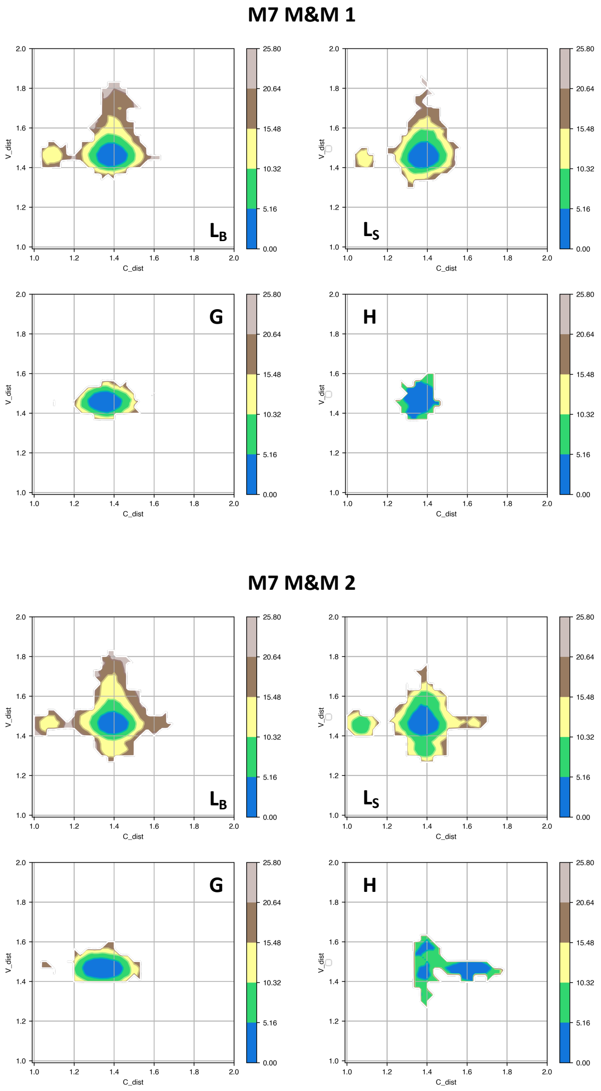


Figure S6: Free energy surfaces for the four substates identified in Figure 3 in the case of the first (top) and second (bottom) M7 metainference simulation. The x-axis shows the distance between the centers of mass of the constant domains, while the y-axis shows the distance between the centers of mass of the variable domains. The free energy is shown with color and isolines every 2k$_B$T corresponding to 5.16 kJ/mol.

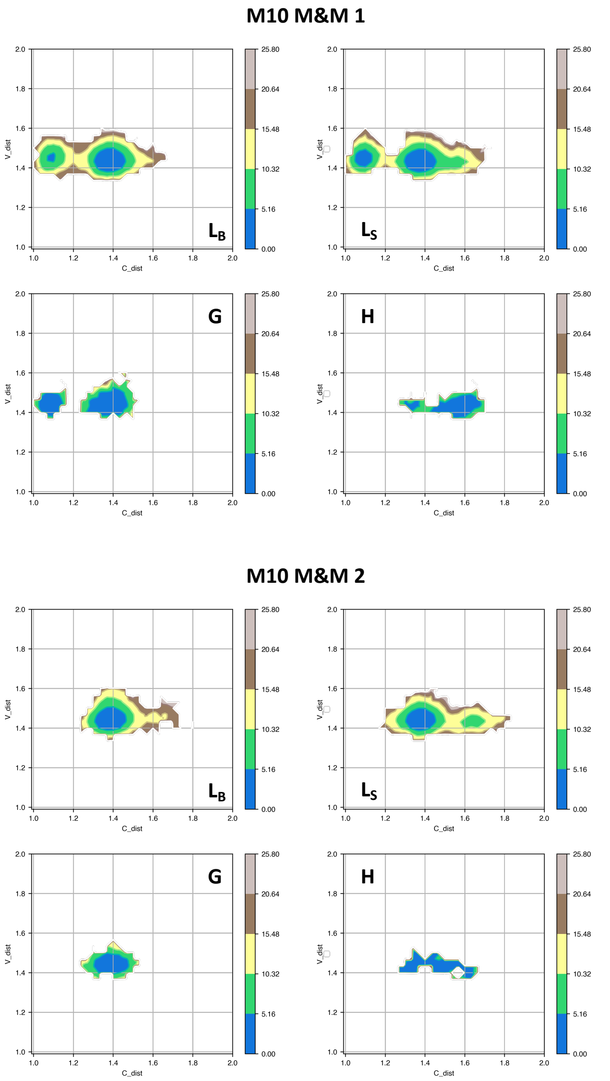


Figure S7: Free energy surfaces for the four substates identified in Figure 3 in the case of the first (top) and second (bottom) M10 metainference simulation. The x-axis shows the distance between the centers of mass of the constant domains, while the y-axis shows the distance between the centers of mass of the variable domains. The free energy is shown with color and isolines every $2k_BT$ corresponding to 5.16 kJ/mol.

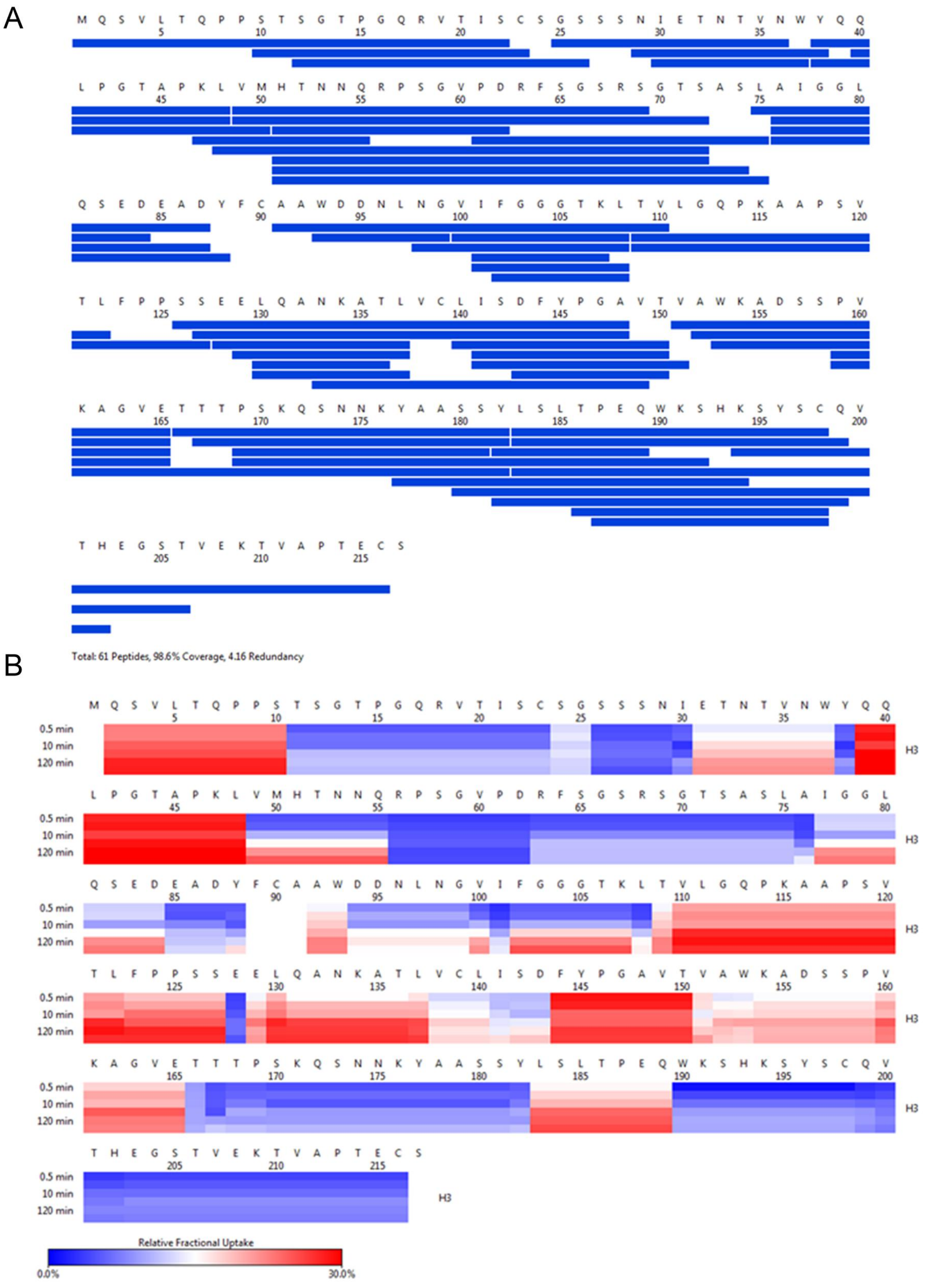


Figure S8: (A) Peptide coverage map of protein H3. The total number of peptides is 61 with a coverage of 98.6% and a redundancy of 4.16. (B) As shown on the left, heat map as a function of HDX-time at different time points. The relative deuterium uptake is color-coded from blue-to-white-to-red for 0 to 30% as indicated by the scale bar below.

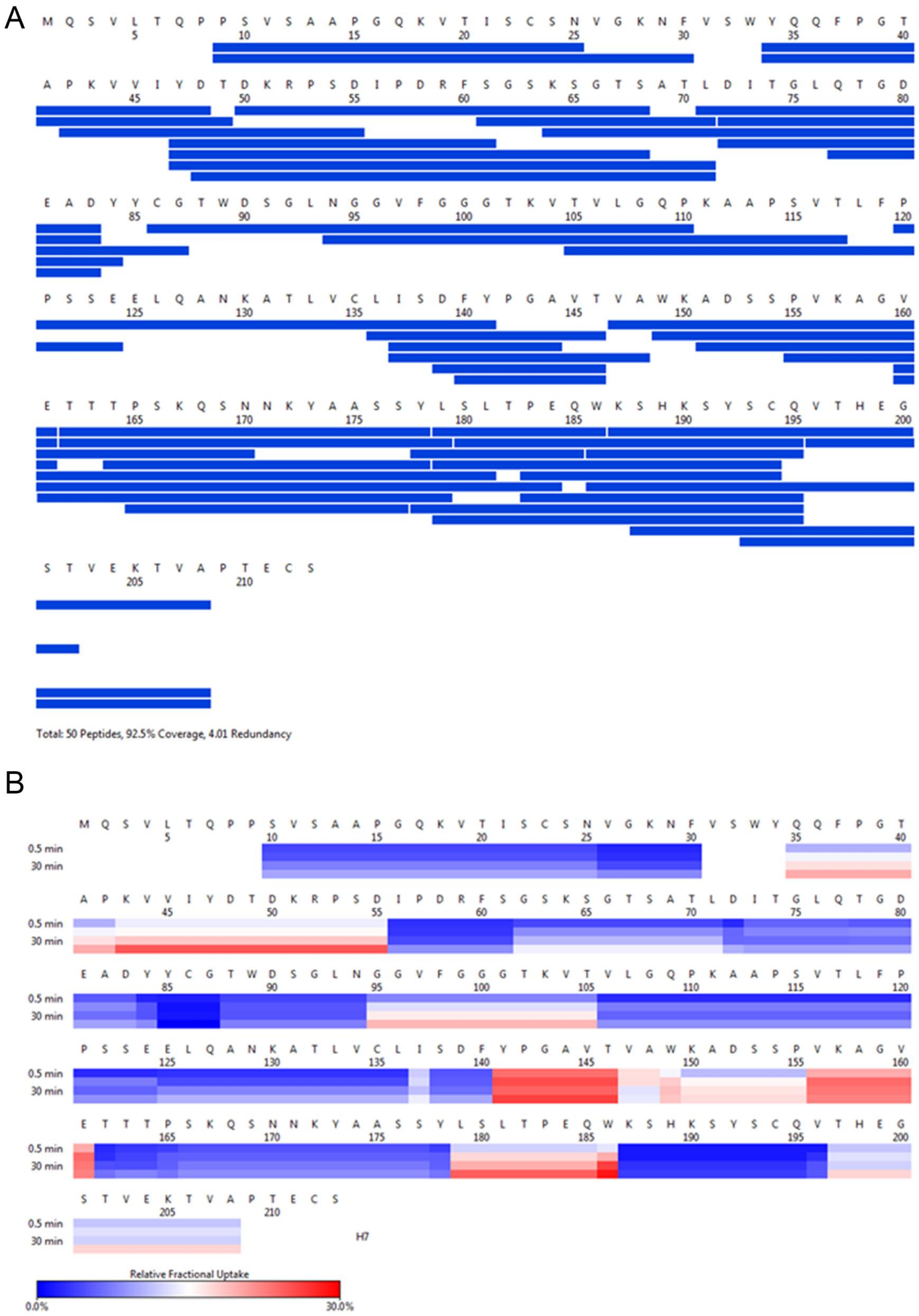


Figure S9: (A) Peptide coverage map of protein H7. The total number of peptides is 50 with a coverage of 92.5% and a redundancy of 4.01. (B) Heat map as a function of HDX-time at different time points as shown on the left. The relative deuterium uptake is color-coded from blue-to-white-to-red for 0 to 30% as shown by the scale bar below.

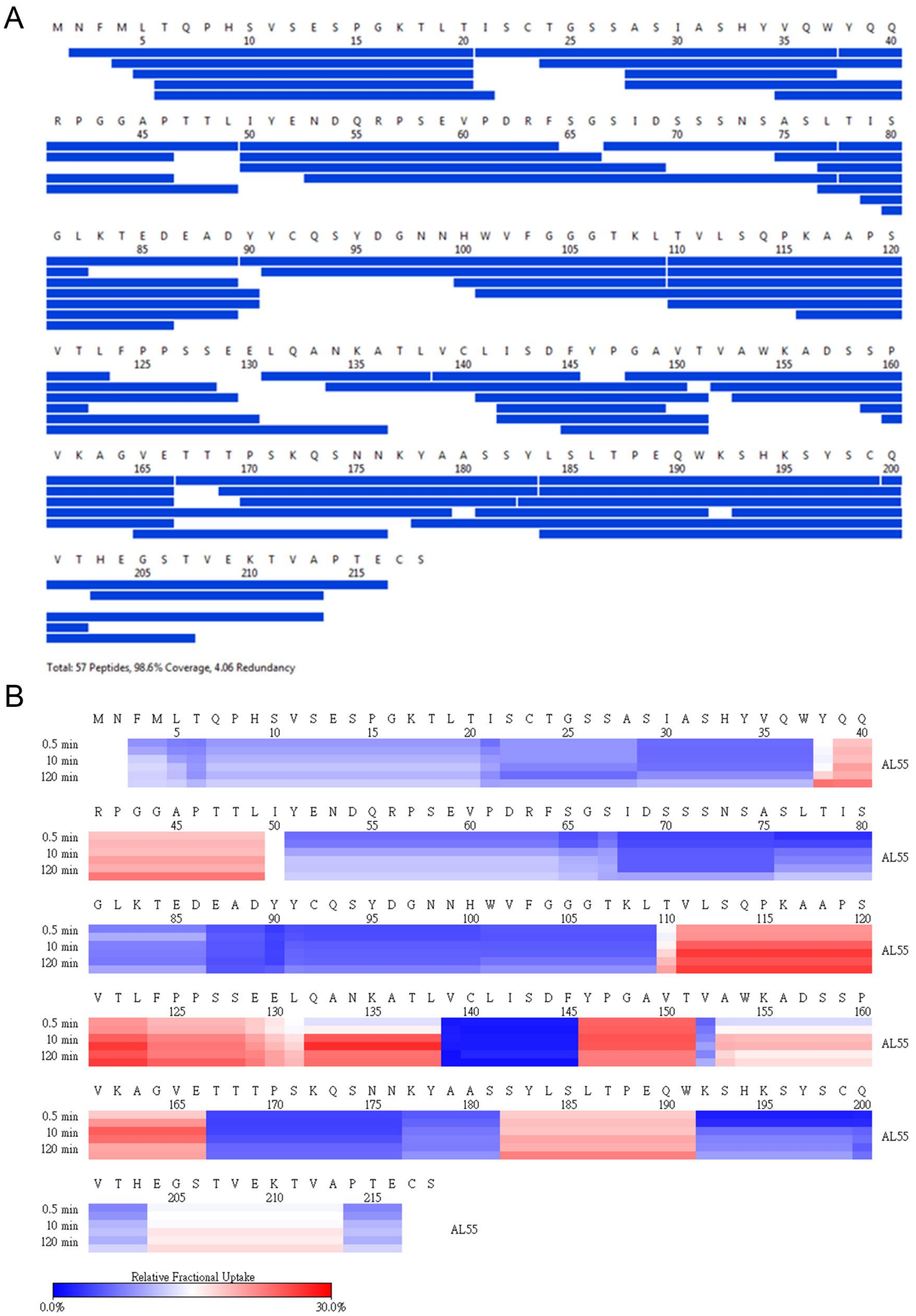


Figure S10: (A) Peptide coverage map of protein AL55. The total number of peptides is 57 with a coverage of 98.6% and a redundancy of 4.06. (B) Heat map as a function of HDX-time at different time points as shown on the left. The relative deuterium uptake is color-coded from blue-to-white-to-red for 0 to 30% as shown by the scale bar below.

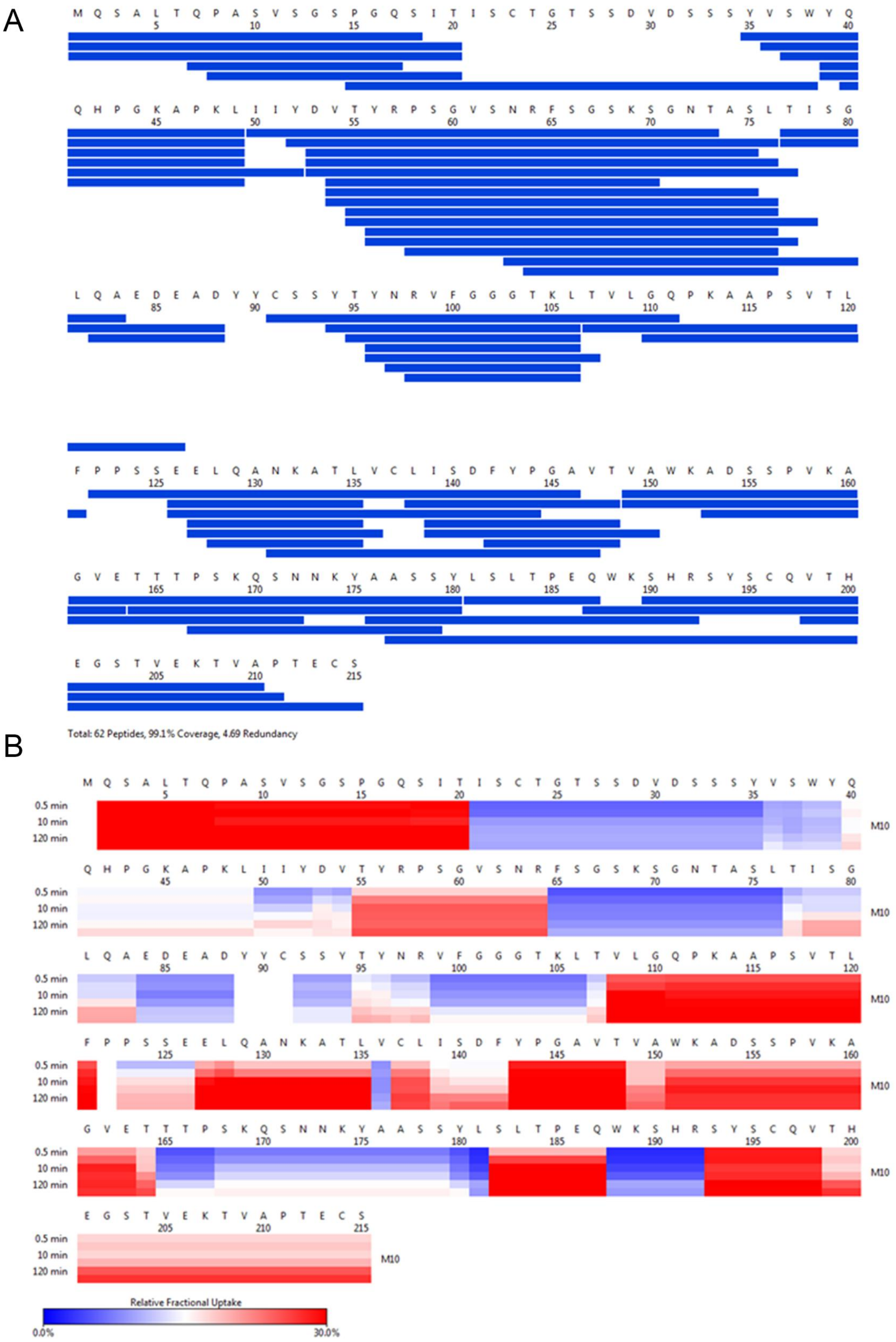


Figure S11: (A) Peptide coverage map of protein M10. The total number of peptides is 62 with a coverage of 99.1% and a redundancy of 4.69. (B) As shown on the left, heat map as a function of HDX-time at different time points. The relative deuterium uptake is color-coded from blue-to-white-to-red for 0 to 30% as indicated by the scale bar below.

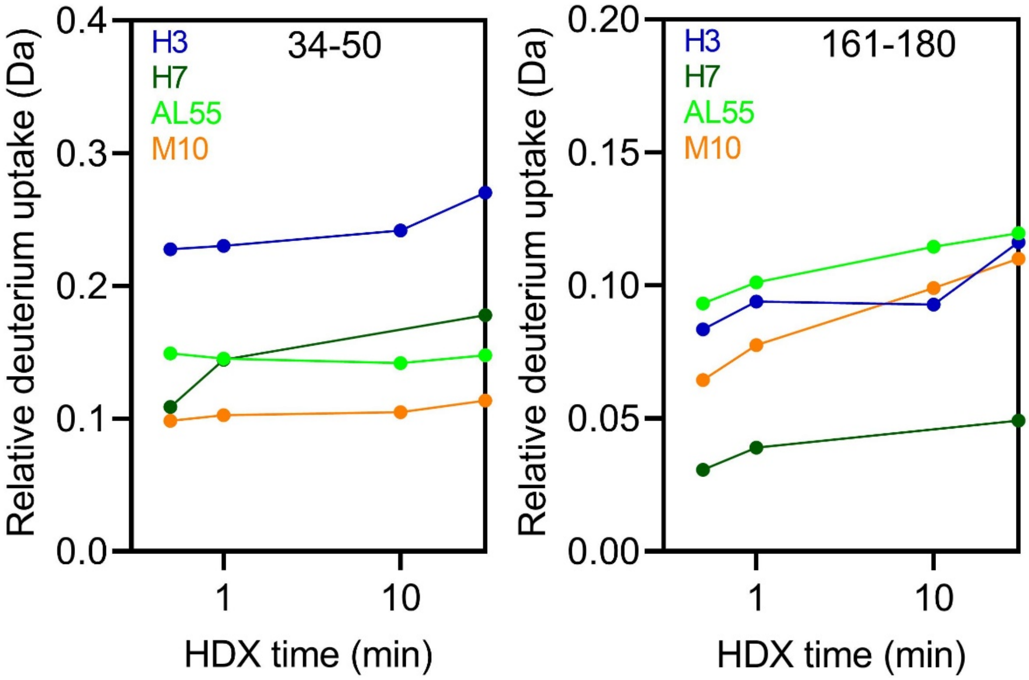
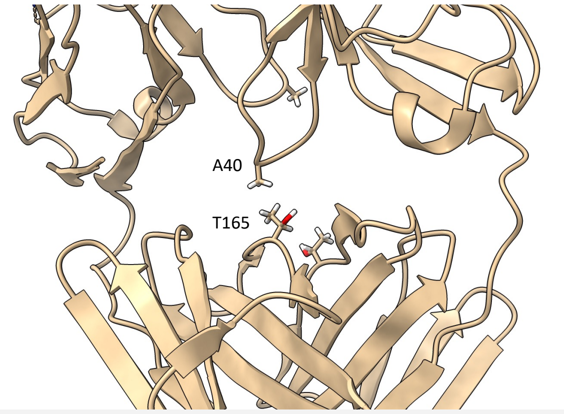
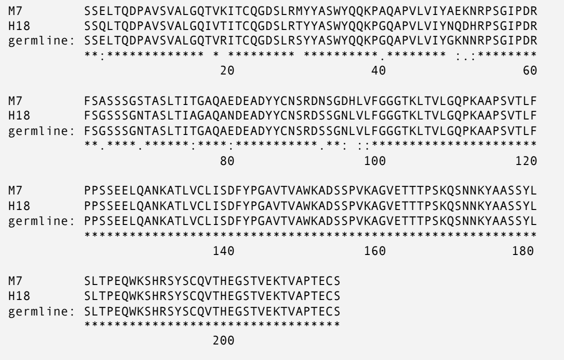


Figure S12: HDX kinetics of relative deuterium uptake of the peptide from residues 34-50 and peptides in a range of residues 161-180. The color corresponding to each protein is shown in the figure panels.





```
M7        SSELTQDPAVSVALGQTVKITCQGDSLRMYYASWYQQKPAQAPVLVIYAEKNRPSGIPDR
H18       SSQLTQDPAVSVALGQIVTITCQGDSLRTYYASWYQQKPGQAPVLVIYNQDHRPSGIPDR
germline: SSELTQDPAVSVALGQTVRITCQGDSLRSYYASWYQQKPGQAPVLVIYGKNNRPSGIPDR
          **:************* * ********* **********.******** :.:********
                            20                  40                  60

M7        FSASSSGSTASLTITGAQAEDEADYYCNSRDNSGDHLVFGGGTKLTVLGQPKAAPSVTLF
H18       FSGSSSGNTASLTIAGAQANDEADYYCNSRDSSGNLVLFGGGTKLTVLGQPKAAPSVTLF
germline: FSGSSSGNTASLTITGAQAEDEADYYCNSRDSSGNLVLFGGGTKLTVLGQPKAAPSVTLF
          **.****.******:****:***********.**: ::**********************
                            80                 100                 120

M7        PPSSEELQANKATLVCLISDFYPGAVTVAWKADSSPVKAGVETTTPSKQSNNKYAASSYL
H18       PPSSEELQANKATLVCLISDFYPGAVTVAWKADSSPVKAGVETTTPSKQSNNKYAASSYL
germline: PPSSEELQANKATLVCLISDFYPGAVTVAWKADSSPVKAGVETTTPSKQSNNKYAASSYL
          ************************************************************
                           140                 160                180

M7        SLTPEQWKSHRSYSCQVTHEGSTVEKTVAPTECS
H18       SLTPEQWKSHRSYSCQVTHEGSTVEKTVAPTECS
germline: SLTPEQWKSHRSYSCQVTHEGSTVEKTVAPTECS
          **********************************
                           200
```

Figure S13. Zoom in on the crystal structure of M7 to show the hydrophobic contact between A40 in the VL and T165 in the CL. Below is reported the multiple sequence alignment for M7, H18 and their germline sequence (IGLV3-19*01 for the VL and IGLC2*02 for the CL).

```
H3                 1 QSVLTQPPSTSGTPGQRVTISCSGSSSNIETNTVNWYQQLPGTAPKLVMH    50
                     |||||||||.|||||||||||||||||||.:||||||||||||||||:::
IGLV1-44           1 QSVLTQPPSASGTPGQRVTISCSGSSSNIGSNTVNWYQQLPGTAPKLLIY    50

                     ●●                                          ●
H3                51 TNNQRPSGVPDRFSGSRSGTSASLAIGGLQSEDEADYFCAAWDDNLNGVI   100
                     :|||||||||||||||:|||||||||.||||||||||:||||||:|||||
IGLV1-44          51 SNNQRPSGVPDRFSGSKSGTSASLAISGLQSEDEADYYCAAWDDSLNGVI   100

H3               101 FGGGTKLTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTV   150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLV1-44/C3*03   101 FGGGTKLTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTV   150

H3               151 AWKADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQVT   200
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLC3*03         151 AWKADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQVT   200

H3               201 HEGSTVEKTVAPTECS    216
                     ||||||||||||||||
IGLC3*03         201 HEGSTVEKTVAPTECS    216
```

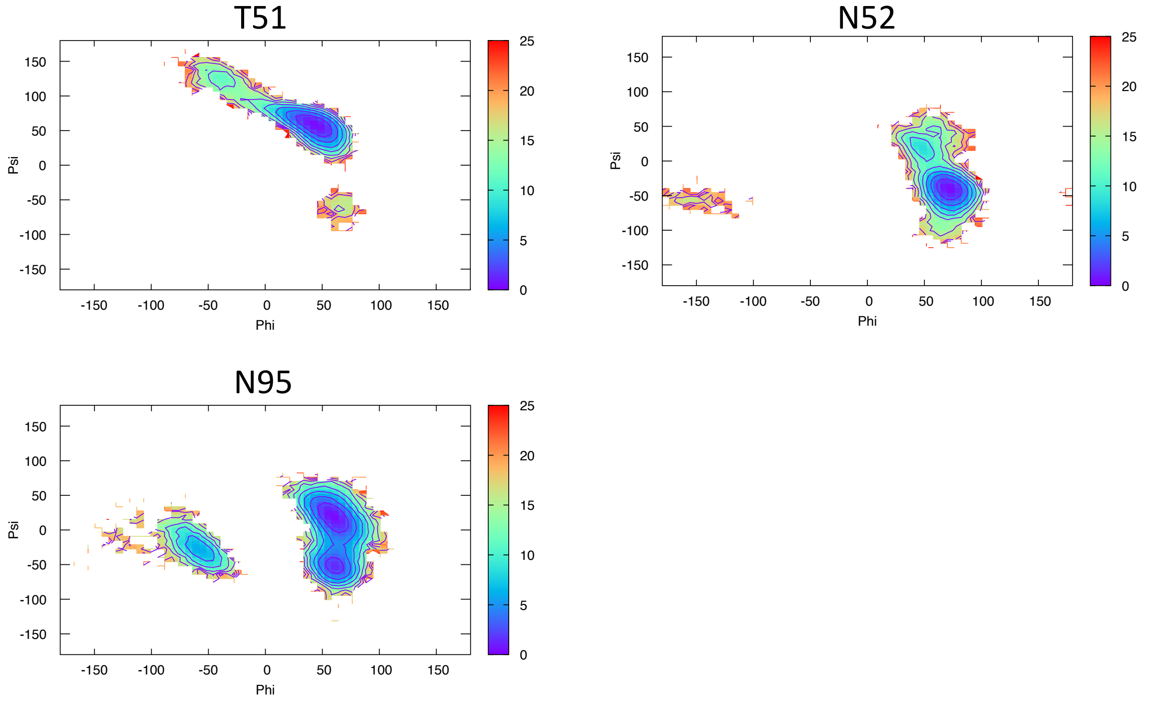


Figure S14: Pairwise sequence alignment between H3 and its corresponding germline as identified by igBLAST using the IGMT databases. The three CDRs and the linker region are highlighted in light blue and orange, respectively. The red circles indicate residues for which the left alpha is the most populated region in the Ramachandran plot. FES (in kJ/mol) representing the Ramachandran plot for the indicated residues are reported in the bottom panels.

```
H7                 1 QSVLTQPPSVSAAPGQKVTISC----SNVGKNFVSWYQQFPGTAPKVVIY    46
                       ||||||||||||||||||||||    ||:|.|:||||||.||||||::||
IGLV1-51*01        1 QSVLTQPPSVSAAPGQKVTISCSGSSSNIGNNYVSWYQQLPGTAPKLLIY    50

                       ●●
H7                47 DTDKRPSDIPDRFSGSKSGTSATLDITGLQTGDEADYYCGTWDSGLNGGV    96
                       |.:||||.||||||||||||||||.|||||||||||||||||||.|:.||
IGLV1-51*01       51 DNNKRPSGIPDRFSGSKSGTSATLGITGLQTGDEADYYCGTWDSSLSAGV   100

H7                97 FGGGTKVTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTV   146
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLV1-51/C3*03   101 FGGGTKVTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTV   150

H7               147 AWKADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQVT   196
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLC3*03 .       151 AWKADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQVT   200

H7               197 HEGSTVEKTVAPTECS    212
                     ||||||||||||||||
IGLC3*03         201 HEGSTVEKTVAPTECS    216
```

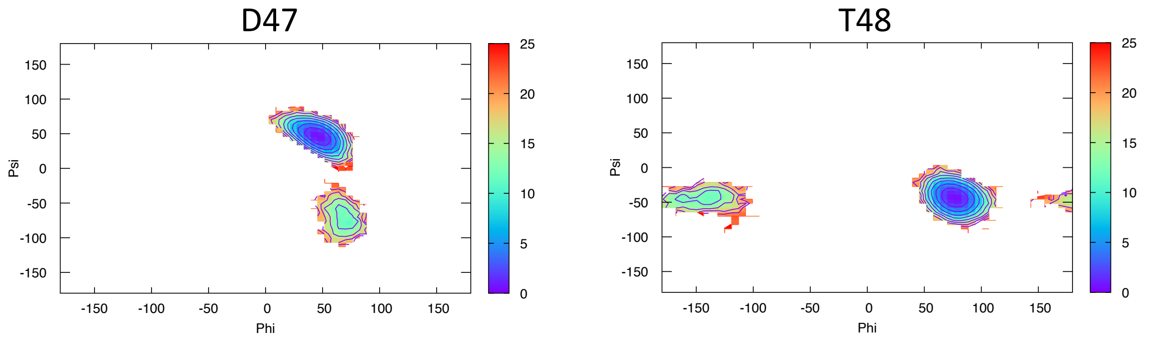


Figure S15: Pairwise sequence alignment between H7 and its corresponding germline as identified by igBLAST using the IGMT databases. The three CDRs and the linker region are highlighted in light blue and orange, respectively. The red circles indicate residues for which the left alpha is the most populated region in the Ramachandran plot. FES (in kJ/mol) representing the Ramachandran plot for the indicated residues are reported in the bottom panels.

```
                                                                    ●●
H18               1 SSQLTQDPAVSVALGQIVTITCQGDSLRTYYASWYQQKPGQAPVLVIYNQ     50
                    ||:|||||||||||||.|.|||||||||:|||||||||||||||||||.:
IGLV3-19*01       1 SSELTQDPAVSVALGQTVRITCQGDSLRSYYASWYQQKPGQAPVLVIYGK     50

H18              51 DHRPSGIPDRFSGSSSGNTASLTIAGAQANDEADYYCNSRDSSGNLVLFG    100
                    ::||||||||||||||||||||||.||||.||||||||||||||||||||
IGLV3-19*01      51 NNRPSGIPDRFSGSSSGNTASLTITGAQAEDEADYYCNSRDSSGNLVLFG    100

H18             101 GGTKLTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTVAW    150
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLV3-19/C2*02  101 GGTKLTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTVAW    150

H18             151 KADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHRSYSCQVTHE    200
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLC2*02        151 KADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHRSYSCQVTHE    200

H18             201 GSTVEKTVAPTECS    214
                    ||||||||||||||
IGLC2*02        201 GSTVEKTVAPTECS    214
```

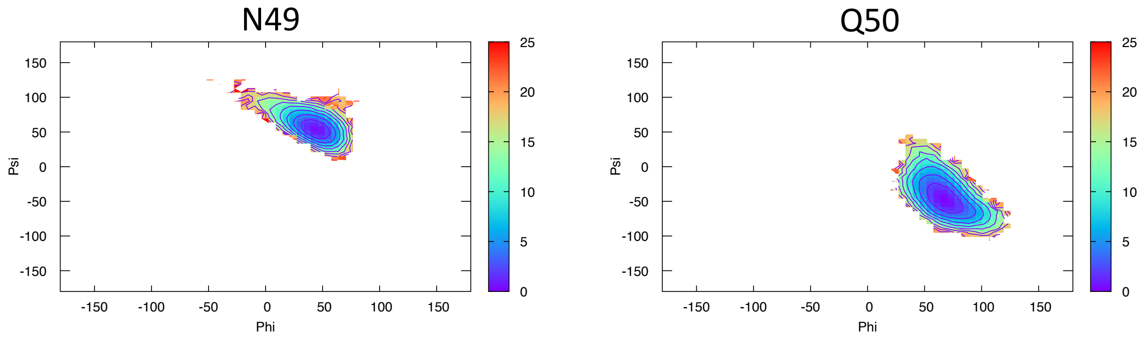


Figure S16: Pairwise sequence alignment between H18 and its corresponding germline as identified by igBLAST using the IGMT databases. The three CDRs and the linker region are highlighted in light blue and orange, respectively. The red circles indicate residues for which the left alpha is the most populated region in the Ramachandran plot. FES (in kJ/mol) representing the Ramachandran plot for the indicated residues are reported in the bottom panels.

```
AL55               1 NFMLTQPHSVSESPGKTLTISCTGSSASIASHYVQWYQQRPGGAPTTLIY    50
                     ||||||||||||||||:|||||||||.||||:||||||||||.||||:||
IGLV6-57*02        1 NFMLTQPHSVSESPGKTVTISCTGSSGSIASNYVQWYQQRPGSAPTTVIY    50
                     ●●                                             ●
AL55              51 ENDQRPSEVPDRFSGSIDSSSNSASLTISGLKTEDEADYYCQSYDGNNHW   100
                     |::||||.|||||||||||||||||||||||||||||||||||||.:|||
IGLV6-57*02       51 EDNQRPSGVPDRFSGSIDSSSNSASLTISGLKTEDEADYYCQSYDSSNHW   100

AL55             101 VFGGGTKLTVLSQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVT   150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLV6-57/C3*03   101 VFGGGTKLTVLSQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVT   150

AL55             151 VAWKADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQV   200
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLC3*03         151 VAWKADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQV   200

AL55             201 THEGSTVEKTVAPTECS    217
                     |||||||||||||||||
IGLC3*03         201 THEGSTVEKTVAPTECS    217
```

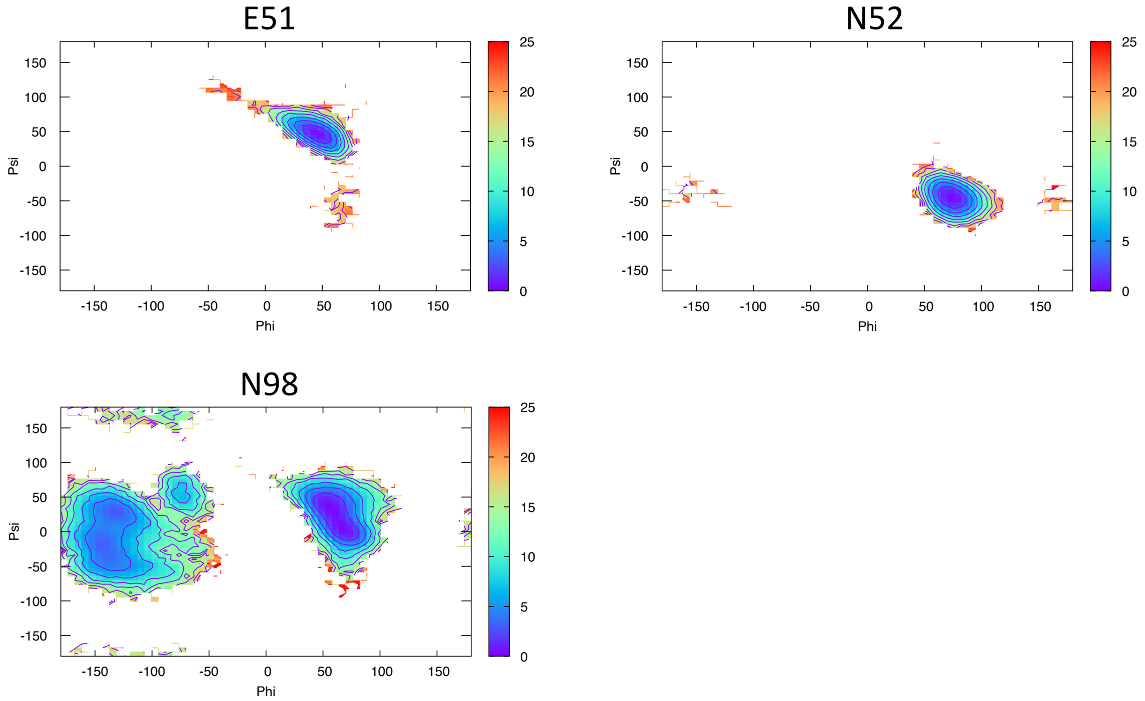


Figure S17: Pairwise sequence alignment between AL55 and its corresponding germline as identified by igBLAST using the IGMT databases. The three CDRs and the linker region are highlighted in light blue and orange, respectively. The red circles indicate residues for which the left alpha is the most populated region in the Ramachandran plot. FES (in kJ/mol) representing the Ramachandran plot for the indicated residues are reported in the bottom panels.

```
                                                                       ●●
M7               1 SSELTQDPAVSVALGQTVKITCQGDSLRMYYASWYQQKPAQAPVLVIYAE   50
                   ||||||||||||||||||:|||||||||.||||||||||.||||||||.:
IGLV3-19*01      1 SSELTQDPAVSVALGQTVRITCQGDSLRSYYASWYQQKPGQAPVLVIYGK   50

M7              51 KNRPSGIPDRFSASSSGSTASLTITGAQAEDEADYYCNSRDNSGDHLVFG  100
                   .|||||||||||.||||:|||||||||||||||||||||||:||:|||||
IGLV3-19*01     51 NNRPSGIPDRFSGSSSGNTASLTITGAQAEDEADYYCNSRDSSGNHLVFG  100

M7             101 GGTKLTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTVAW  150
                   ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLV3-19/C2*02 101 GGTKLTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTVAW  150

M7             151 KADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHRSYSCQVTHE  200
                   ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLC2*02       151 KADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHRSYSCQVTHE  200

M7             201 GSTVEKTVAPTECS    214
                   ||||||||||||||
IGLC2*02       201 GSTVEKTVAPTECS    214
```

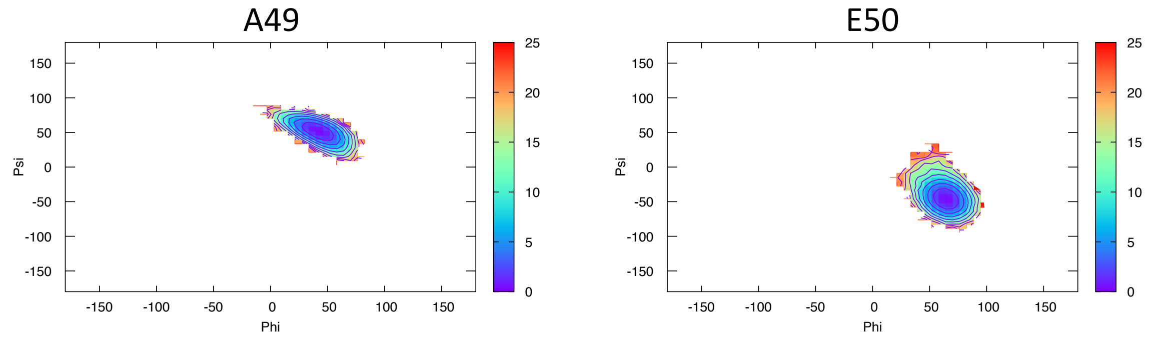


Figure S18: Pairwise sequence alignment between M7 and its corresponding germline as identified by igBLAST using the IGMT databases. The three CDRs and the linker region are highlighted in light blue and orange, respectively. The red circles indicate residues for which the left alpha is the most populated region in the Ramachandran plot. FES (in kJ/mol) representing the Ramachandran plot for the indicated residues are reported in the bottom panels.

```
M10                 1 QSALTQPASVSGSPGQSITISCTGTSSDVDSSSYVSWYQQHPGKAPKLII    50
                      |||||||||||||||||||||||||||||...:|||||||||||||||:|
IGLV2-14*01         1 QSALTQPASVSGSPGQSITISCTGTSSDVGGYNYVSWYQQHPGKAPKLMI    50
                       ●●                                          ●
M10                51 YDVTYRPSGVSNRFSGSKSGNTASLTISGLQAEDEADYYCSSYTYNRVFG   100
                      |:|:.|||||||||||||||||||||||||||||||||||||||.:||||
IGLV2-14*01        51 YEVSNRPSGVSNRFSGSKSGNTASLTISGLQAEDEADYYCSSYTSSRVFG   100

M10               101 GGTKLTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTVAW   150
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLV2-14/C3*04    101 GGTKLTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTVAW   150

M10               151 KADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHRSYSCQVTHE   200
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
IGLC3*04          151 KADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHRSYSCQVTHE   200

M10               201 GSTVEKTVAPTECSc    215
                      ||||||||||||||
IGLC3*04          201 GSTVEKTVAPTECS-    214
```

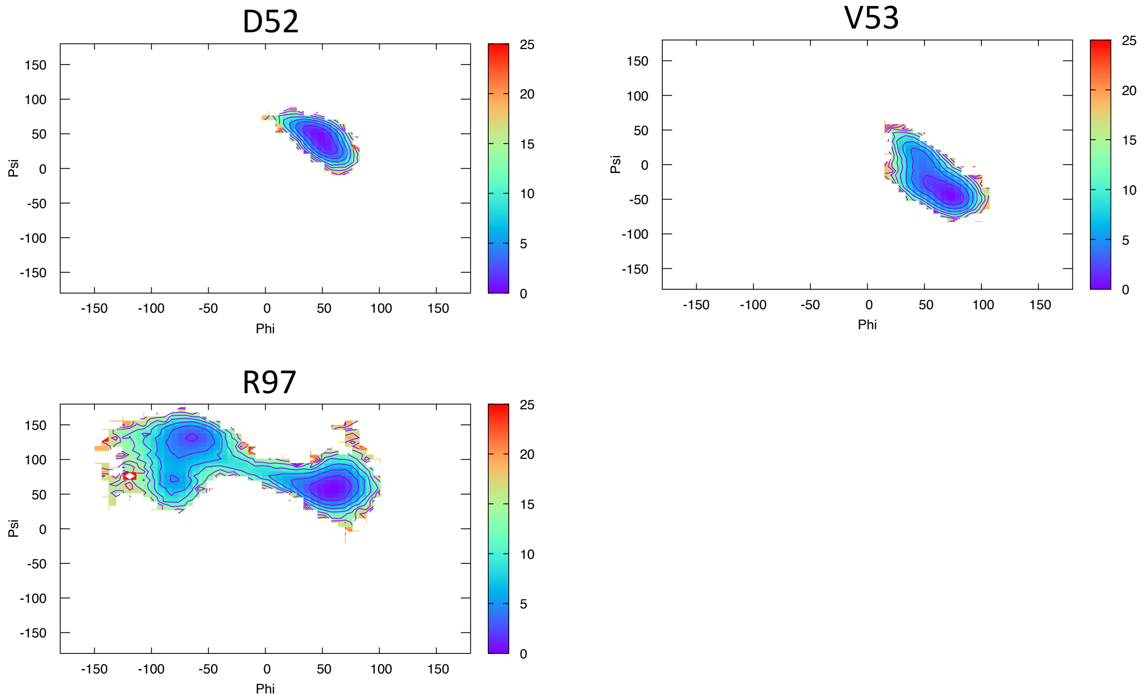


Figure S19: Pairwise sequence alignment between M10 and its corresponding germline as identified by igBLAST using the IGMT databases. The three CDRs and the linker region are highlighted in light blue and orange, respectively. The red circles indicate residues for which the left alpha is the most populated region in the Ramachandran plot. FES (in kJ/mol) representing the Ramachandran plot for the indicated residues are reported in the bottom panels.

Table S1. Pairwise sequence identity (above diagonal) and similarity (below diagonal) for the 6 systems under study. On the diagonal is reported the germline identified by igBLAST using the IGMT database.

	H3	**H7**	**H18**	**AL55**	**M7**	**M10**
H3	**IGLV1-44*01**	179/216 (82.9%)	168/216 (77.8%)	171/218 (78.4%)	163/216 (75.5%)	172/217 (79.3%)
H7	194/216 (89.8%)	**IGLV1-51*01**	169/214 (79.0%)	168/218 (77.1%)	165/214 (77.1%)	169/217 (77.9%)
H18	183/216 (84.7%)	182/214 (85.0%)	**IGLV3-19*01**	166/218 (76.1%)	**196/214 (91.6%)**	171/217 (78.8%)
AL55	190/218 (87.2%)	188/218 (86.2%)	184/218 (84.4%)	**IGLV6-57*02**	**164/218 (75.2%)**	175/218 (80.3%)
M7	181/216 (83.8%)	179/214 (83.6%)	**204/214 (95.3%)**	**182/218 (83.5%)**	**IGLV3-19*01**	169/217 (77.9%)
M10	196/217 (90.3%)	185/217 (85.3%)	186/217 (85.7%)	188/218 (86.2%)	183/217 (84.3%)	**IGLV2-14*03**

Table S2. Metainference simulations performed in this work for the 6 systems. For each simulation is reported the simulation time per replica and the final agreement of the resulting conformational ensemble with the experimental SAXS curve. The range q<0.3 is the one used as restraint in the simulation.

LC code	Simulation	Length per replica [ns]	SAXS $\mathcal{X}^2$ q<0.5 (q<0.3)
H3	M&M 1	1,530	1.2 (1.1)
H3	M&M 2	1,520	1.2 (1.1)
H7	M&M 1	1,627	1.1 (1.2)
H7	M&M 2	1,545	1.1 (1.2)
H18	M&M 1	1,643	1.4 (1.6)
H18	M&M 2	1,529	1.4 (1.7)
AL55	M&M 1	1,545	2.7 (3.0)
AL55	M&M 2	1,591	2.5 (2.7)
M7	M&M 1	1,623	1.2 (1.2)
M7	M&M 2	1,530	1.1 (1.2)
M10	M&M 1	987	1.1 (1.1)
M10	M&M 2	995	1.1 (1.1)

Table S3. HDX-MS data summary.

Datasets	**H3**	**H7**	**AL55**	**M10**
HDX reaction details	1X Phosphate buffer saline in D_2O (pD 7.0), 25°C			
HDX time course (min)	0, 0.5, 1, 10, 30, 120, and 240			
Back exchange (mean/IQR)	ND			
No. of peptides	61	50	57	62
Sequence coverage (%)	98.6	92.5	98.6	99.1
Average peptide length/ Redundancy	13.6/4.16	15.8/4.01	14.3/4.06	15.1/4.69
Replicates (technical)	3	3	3	3
Repeatability (average SD)	0.04 Da	0.04 Da	0.04 Da	0.05 Da

Legends for Movies

Movie S1 (separate file). This movie depicts the 3D structure of H3, color-coded by HDX exchnage and rotating 360 degrees.

Movie S2 (separate file). This movie depicts the 3D structure of H7, color-coded by HDX exchnage and rotating 360 degrees.

Movie S3 (separate file). This movie depicts the 3D structure of AL55, color-coded by HDX exchnage and rotating 360 degrees.

Movie S4 (separate file). This movie depicts the 3D structure of M10, color-coded by HDX exchnage and rotating 360 degrees.

Legends for Datasets

Dataset S1 (separate file).

SAXS data are available on the SASBDB with accession codes

Dataset S2 (separate file).

DOI: 10.5281/zenodo.12731283, https://dx.doi.org/10.5281/zenodo.12731283

Molecular dynamics simulation trajectories and associated statistical weights.

SI References

1. C. E. Blanchet, *et al.,* Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). *J Appl Crystallogr*. **48**(Pt 2):431-443 (2015).
2. P. Pernot, *et al.,* New beamline dedicated to solution scattering from biological macromolecules at the ESRF. *J. Phys.: Conf. Ser*. **247**, 012009 (2010).
3. K. Manalastas-Cantos, *et al.*, ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *J. Appl. Cryst*. **54**, 343-355 (2021).
4. E. Valentini, *et al.,* SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Research* **43**(D1), D357–D363 (2015).
5. L. Oberti, et al., Concurrent structural and biophysical traits link with immunoglobulin light chains amyloid propensity. *Sci Rep* **7,** 16809 (2017).
6. B. Webb, A. Sali. Comparative protein structure modeling using modeller. Current protocols in bioinformatics 54, John Wiley & Sons, Inc., 5.6.1-5.6.37, 2016.
7. A. Waterhouse, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 46, W296-W303. (2018)
8. J. Jumper, *et al.,* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).
9. M. J. Abraham, et al., GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *Softwarex* **1–2,** 19–25 (2015).
10. G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, G. Bussi, PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **185,** 604–613 (2014).
11. S. Piana, P. Robustelli, D. Tan, S. Chen, D. E. Shaw, Development of a force field for the simulation of single-chain proteins and protein-protein complexes. *J Chem Theory Comput* (2020).
12. S. Piana, A. G. Donchev, P. Robustelli, D. E. Shaw, Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J Phys Chem B* **119,** 5113–5123 (2015).
13. M. Bonomi, C. Camilloni, M. Vendruscolo, Metadynamic metainference: Enhanced sampling of the metainference ensemble using metadynamics. *Sci. Rep.* **6,** 31232 (2016).
14. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J Chem Phys* **126,** 014101 (2007).
15. M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* **52,** 7182–7190 (1981).
16. C. W. Hopkins, S. L. Grand, R. C. Walker, A. E. Roitberg, Long-time-step molecular dynamics through hydrogen mass repartitioning. *J Chem Theory Comput* **11,** 1864–1874 (2015).
17. J. Pfaendtner, M. Bonomi, Efficient sampling of high-dimensional free-energy landscapes with parallel bias metadynamics. *J Chem Theory Comput* **11,** 5062–7 (2015).
18. C. Paissoni, A. Jussupow, C. Camilloni, Determination of protein structural ensembles by hybrid-resolution SAXS restrained molecular dynamics. *J. Chem. Theory Comput.* **16,** 2825–2834 (2020).
19. C. Paissoni, C. Camilloni, How to determine accurate conformational ensembles by metadynamics metainference: A chignolin study case. *Front. Mol. Biosci.* **8,** 694130 (2021).
20. F. Ballabio, C. Paissoni, M. Bollati, M. de Rosa, R. Capelli, C. Camilloni, Accurate and efficient SAXS/SANS implementation including solvation layer effects suitable fo molecular simulations. *J. Chem. Theory Comput.* **19,** 8401–8413 (2023).
21. D. Branduardi, G. Bussi, M. Parrinello, M. Metadynamics with adaptive gaussians. *J Chem Theory Comput* **8,** 2247–54 (2012).
22. G. R. Masson, *et al.*, Recommendations for performing, interpreting and reporting hydrogen-deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat Methods* **16**, 595–602 (2019).

23. S. Puri, *et al.*, Impacts of cancer-associated mutations on the structure–activity relationship of BAP1. *Journal of Molecular Biology* **434**, 167553 (2022).
24. S. Puri, S.-T. D. Hsu, Oxidation of catalytic cysteine of human deubiquitinase BAP1 triggers misfolding and aggregation in addition to functional loss. *Biochemical and Biophysical Research Communications* **599**, 57–62 (2022).
25. K.-T. Ko, I.-C. Hu, K.-F. Huang, P.-C. Lyu, S.-T. D. Hsu, Untying a knotted SPOUT RNA methyltransferase by circular permutation results in a domain-swapped dimer. *Structure* **27**, 1224-1233.e4 (2019).
26. The PyMOL molecular graphics system, Version 1.2, *Schrödinger, LLC.*

5 — Conclusions

The combination of computational techniques and experimental data is essential to advance the understanding and characterisation of structure, dynamics, function, as well as their relationship in the context of biomolecular systems. While computational methods, such as molecular dynamics simulations, have improved greatly, they are still hampered by approximations and limited in their sampling capabilities. Experimental techniques serve as essential validation tools to ensure that computational predictions are consistent with biological observations. In addition, experimental data can be used as restraints to reduce sampling challenges and guide the modelling process towards specific or functionally relevant states.

The integration of experimental data from different methods addresses the limitations of each approach. High-resolution techniques such as X-ray crystallography provide detailed static structures but lack dynamic insight, while NMR and SAS provide information on molecular dynamics and flexibility. By combining these methods with computational techniques, integrative modelling provides a more complete view, capturing both structural details and dynamic behaviour that are often elusive in isolated experimental approaches.

Moreover, in an iterative feedback loop, computational models are improved or validated by experimental data, and in turn, the molecular details and mechanisms observed *in silico* provide novel directions and inspiration for new experiments that would otherwise be inaccessible. The power of integrative modelling lies in its ability to overcome the limitations of both fields. Ultimately, this comprehensive approach provides deeper insights into biological mechanisms and opens up new opportunities also for practical applications in areas such as drug discovery and protein engineering.

REFERENCES

[1] Watson, J. D.; Crick, F. H. C. *Nature* **1953**, *171*, 737–738.

[2] Sinden, R. R.; Pearson, C. E.; Potaman, V. N.; Ussery, D. W. In *Genes and Genomes*; Verma, R. S., Ed.; Advances in Genome Biology; JAI, 1998; Vol. 5; pp 1–141.

[3] Oliver, S. G. *Nature* **1996**, *379*, 597–600.

[4] Kornberg, A. *Trends in Biochemical Sciences* **1984**, *9*, 122–124.

[5] Aguilera, A.; Gómez-González, B. *Nature Reviews Genetics* **2008**, *9*, 204–217.

[6] Alberts, B.; Johnson, A.; Lewis, J.; Walter, P.; Raff, M.; Roberts, K. *Molecular Biology of the Cell 4th Edition: International Student Edition*; Routledge, 2002.

[7] Steitz, T. A. *Nature reviews Molecular cell biology* **2008**, *9*, 242–253.

[8] Kornberg, R. D. *Proceedings of the National Academy of Sciences* **2007**, *104*, 12955–12961.

[9] Sonenberg, N.; Hinnebusch, A. G. *Cell* **2009**, *136*, 731–745.

[10] Kozak, M. *Gene* **1999**, *234*, 187–208.

[11] Schimmel, P. R.; Söll, D. *Annual review of biochemistry* **1979**, *48*, 601–648.

[12] Crick, F. *Journal of Molecular Biology* **1966**, *19*, 548–555.

[13] Lagerkvist, U. *Proceedings of the National Academy of Sciences* **1978**, *75*, 1759–1762.

[14] Nissen, P.; Hansen, J.; Ban, N.; Moore, P. B.; Steitz, T. A. *Science* **2000**, *289*, 920–930.

[15] Bartel, D. P. *cell* **2004**, *116*, 281–297.

[16] Hannon, G. J. *nature* **2002**, *418*, 244–251.

[17] Sioud, M. In *Design and Delivery of SiRNA Therapeutics*; Ditzel, H. J., Tuttolomondo, M., Kauppinen, S., Eds.; Springer US: New York, NY, 2021; pp 1–15.

[18] Breaker, R. R. *Cold Spring Harbor perspectives in biology* **2012**, *4*, a003566.

[19] Storz, G.; Altuvia, S.; Wassarman, K. M. *Annu. Rev. Biochem.* **2005**, *74*, 199–217.

[20] Doudna, J. A.; Cech, T. R. *Nature* **2002**, *418*, 222–228.

[21] Narlikar, G. J.; Herschlag, D. *Annual review of biochemistry* **1997**, *66*, 19–59.

[22] Wolf, Y. I.; Kazlauskas, D.; Iranzo, J.; Lucía-Sanz, A.; Kuhn, J. H.; Krupovic, M.; Dolja, V. V.; Koonin, E. V. *MBio* **2018**, *9*, 10–1128.

[23] Anfinsen, C. B. *Science* **1973**, *181*, 223–230.

[24] Nelson, D. L.; Lehninger, A. L.; Cox, M. M. *Lehninger principles of biochemistry*; Macmillan, 2008.

[25] Kadler, K. E.; Baldock, C.; Bella, J.; Boot-Handford, R. P. *Journal of cell science* **2007**, *120*, 1955–1958.

[26] Pollard, T. D.; Cooper, J. A. *science* **2009**, *326*, 1208–1212.

[27] Desai, A.; Mitchison, T. J. *Annual review of cell and developmental biology* **1997**, *13*, 83–117.

[28] Blanco-Rodriguez, G.; Di Nunzio, F. *Viruses* **2021**, *13*, 1178.

[29] Storz, J. F. *Hemoglobin: insights into protein structure, function, and evolution*; Oxford University Press, 2018.

[30] Harrison, P. M.; Arosio, P. *Biochimica et biophysica acta (BBA)-bioenergetics* **1996**, *1275*, 161–203.

[31] Delves, P. J.; Roitt, I. M. *New England journal of medicine* **2000**, *343*, 37–49.

[32] Carroll, M. C. *Nature immunology* **2004**, *5*, 981–986.

[33] Opal, S. M.; DePalo, V. A. *Chest* **2000**, *117*, 1162–1172.

[34] Latchman, D. S. *The international journal of biochemistry & cell biology* **1997**, *29*, 1305–1312.

[35] Gagnidze, K.; Pfaff, D. W. *Neuroscience in the 21st Century: From Basic to Clinical*; Springer, 2022; pp 2677–2716.

[36] Clapier, C. R.; Cairns, B. R. *Annual review of biochemistry* **2009**, *78*, 273–304.

[37] Hentze, M. W.; Castello, A.; Schwarzl, T.; Preiss, T. *Nature reviews Molecular cell biology* **2018**, *19*, 327–341.

[38] Crick, F. *Nature* **1970**, *227*, 561–563.

[39] Dahm, R. *Developmental biology* **2005**, *278*, 274–288.

[40] Dahm, R. *Human genetics* **2008**, *122*, 565–581.

[41] Levene, P.; Jacobs, W. *Berichte der deutschen chemischen Gesellschaft* **1909**, *42*, 2474–2478.

[42] Griffith, F. *Epidemiology & Infection* **1928**, *27*, 113–159.

[43] Avery, O. T.; MacLeod, C. M.; McCarty, M. *Die Entdeckung der Doppelhelix* **1944**, *97*.

[44] Hershey, A. D.; Chase, M. *Journal of General Physiology* **1952**, *36*, 39–56.

[45] Franklin, R. E.; Gosling, R. G. *Nature* **1953**, *171*, 740–741.

[46] WILKINS, M. H. F.; STOKES, A. R.; WILSON, H. R. *Nature* **1953**, *171*, 738–740.

[47] Bateson, W.; Mendel, G. *Mendel's principles of heredity*; Courier Corporation, 2013.

[48] Jacob, F.; Monod, J. *Journal of molecular biology* **1961**, *3*, 318–356.

[49] Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R.; Wyckoff, H.; Phillips, D. C. *Nature* **1958**, *181*, 662–666.

[50] Perutz, M. F.; Rossmann, M. G.; Cullis, A. F.; Muirhead, H.; Will, G.; North, A. C. *Nature* **1960**, *185*, 416–422.

[51] Karplus, M.; McCammon, J. A. *Nat Struct Biol* **2002**, *9*, 646–652.

[52] Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. *J Mol Biol* **2001**, *305*, 567–580.

[53] Luscombe, N. M.; Greenbaum, D.; Gerstein, M. *Methods of information in medicine* **2001**, *40*, 346–358.

[54] Jumper, J. et al. *Nature* **2021**, *596*, 583–589.

[55] Schwede, T. et al. *Structure* **2009**, *17*, 151–159.

[56] Dobson, C. M. *Nature* **2003**, *426*, 884–890.

[57] Pauling, L.; Corey, R. B.; Branson, H. R. *Proc Natl Acad Sci U S A* **1951**, *37*, 205–211.

[58] Richardson, J. S.; Richardson, D. C. *Science* **1988**, *240*, 1648–1652.

[59] Engelman, D. M.; Steitz, T. A.; Goldman, A. *Annu Rev Biophys Biophys Chem* **1986**, *15*, 321–353.

[60] Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.

[61] Chothia, C. *Annu Rev Biochem* **1984**, *53*, 537–572.

[62] Moore, P. B. *Annu Rev Biochem* **1999**, *68*, 287–300.

[63] Brion, P.; Westhof, E. *Annu Rev Biophys Biomol Struct* **1997**, *26*, 113–137.

[64] Tinoco, I., Jr; Bustamante, C. *J Mol Biol* **1999**, *293*, 271–281.

[65] Leckband, D.; Israelachvili, J. *Quarterly reviews of biophysics* **2001**, *34*, 105–267.

[66] Dill, K. A. *Biochemistry* **1990**, *29*, 7133–7155.

[67] Richardson, J. S. *Adv Protein Chem* **1981**, *34*, 167–339.

[68] Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. *Chem Rev* **2006**, *106*, 3210–3235.

[69] Kim, S. H.; Ganji, M.; Kim, E.; van der Torre, J.; Abbondanzieri, E.; Dekker, C. *Elife* **2018**, *7*.

[70] Goodsell, D. S.; Olson, A. J. *Annu Rev Biophys Biomol Struct* **2000**, *29*, 105–153.

[71] Luger, K.; Mäder, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. *Nature* **1997**, *389*, 251–260.

[72] Kornberg, R. D. *Science* **1974**, *184*, 868–871.

[73] Felsenfeld, G.; Groudine, M. *Nature* **2003**, *421*, 448–453.

[74] Dauter, Z. *Acta Crystallogr D Biol Crystallogr* **2005**, *62*, 1–11.

[75] Smyth, M. S.; Martin, J. H. *Mol Pathol* **2000**, *53*, 8–14.

[76] Fernández, F. J.; Querol-García, J.; Navas-Yuste, S.; Martino, F.; Vega, M. C. *Adv Exp Med Biol* **2024**, *3234*, 125–140.

[77] McPherson, A.; Gavira, J. A. *Acta Crystallogr F Struct Biol Commun* **2013**, *70*, 2–20.

[78] Gonen, T. *Methods Mol Biol* **2013**, *955*, 153–169.

[79] Kay, L. E. *J Mol Biol* **2015**, *428*, 323–331.

[80] Clore, G. M.; Gronenborn, A. M. *Crit Rev Biochem Mol Biol* **1989**, *24*, 479–564.

[81] Bax, A.; Grzesiek, S. *Accounts of Chemical Research* **1993**, *26*, 131–138.

[82] Clore, G. M.; Gronenborn, A. M. *Proceedings of the National Academy of Sciences* **1998**, *95*, 5891–5898.

[83] Wüthrich, K. *Nat Struct Biol* **2001**, *8*, 923–925.

[84] Kainosho, M.; Torizawa, T.; Iwashita, Y.; Terauchi, T.; Mei Ono, A.; Güntert, P. *Nature* **2006**, *440*, 52–57.

[85] Wüthrich, K. *J Biol Chem* **1990**, *265*, 22059–22062.

[86] Weissenberger, G.; Henderikx, R. J.; Peters, P. J. *Nature Methods* **2021**, *18*, 463–471.

[87] Nogales, E.; Scheres, S. H. W. *Mol Cell* **2015**, *58*, 677–689.

[88] Bai, X.-C.; McMullan, G.; Scheres, S. H. W. *Trends Biochem Sci* **2014**, *40*, 49–57.

[89] Callaway, E. *Nature* **2015**, *525*, 172–174.

[90] Kühlbrandt, W. *Science* **2014**, *343*, 1443–1444.

[91] Egelman, E. H. *Biophys J* **2016**, *110*, 1008–1012.

[92] Nogales, E. *Nat Methods* **2016**, *13*, 24–27.

[93] Jeschke, G.; Polyhach, Y. *Phys Chem Chem Phys* **2007**, *9*, 1895–1910.

[94] Hubbell, W. L.; Gross, A.; Langen, R.; Lietzow, M. A. *Curr Opin Struct Biol* **1998**, *8*, 649–656.

[95] Roser, P.; Schmidt, M. J.; Drescher, M.; Summerer, D. *Org. Biomol. Chem.* **2016**, *14*, 5468–5476.

[96] Torricella, F.; Pierro, A.; Mileo, E.; Belle, V.; Bonucci, A. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2021**, *1869*, 140653.

[97] Jeschke, G. *Chemphyschem* **2002**, *3*, 927–932.

[98] Klare, J. P.; Steinhoff, H.-J. *Photosynth Res* **2009**, *102*, 377–390.

[99] Wolfenden, R.; Snider, M. J. *Acc Chem Res* **2001**, *34*, 938–945.

[100] Imoto, T.; Johnson, L.; North, A.; Phillips, D.; Rupley, J. In *21 Vertebrate Lysozymes*; Boyer, P. D., Ed.; The Enzymes; Academic Press, 1972; Vol. 7; pp 665–868.

[101] Blake, C. C.; Koenig, D. F.; Mair, G. A.; North, A. C.; Phillips, D. C.; Sarma, V. R. *Nature* **1965**, *206*, 757–761.

[102] Luscombe, N. M.; Laskowski, R. A.; Thornton, J. M. *Nucleic Acids Res* **2001**, *29*, 2860–2874.

[103] Perutz, M. F. *Nature* **1970**, *228*, 726–739.

[104] Ingram, V. M. *Nature* **1956**, *178*, 792–794.

[105] Sela-Culang, I.; Kunik, V.; Ofran, Y. *Frontiers in immunology* **2013**, *4*, 302.

[106] Wesolowski, J. et al. *Med Microbiol Immunol* **2009**, *198*, 157–174.

[107] Rosenbaum, D. M.; Rasmussen, S. G. F.; Kobilka, B. K. *Nature* **2009**, *459*, 356–363.

[108] Hilger, D.; Masureel, M.; Kobilka, B. K. *Nature structural & molecular biology* **2018**, *25*, 4–12.

[109] Hauser, A. S.; Attwood, M. M.; Rask-Andersen, M.; Schiöth, H. B.; Gloriam, D. E. *Nature reviews Drug discovery* **2017**, *16*, 829–842.

[110] Lagerström, M. C.; Schiöth, H. B. *Nat Rev Drug Discov* **2008**, *7*, 339–357.

[111] Thornton, J. M.; Todd, A. E.; Milburn, D.; Borkakoti, N.; Orengo, C. A. *nature structural biology* **2000**, *7*, 991–994.

[112] Henzler-Wildman, K.; Kern, D. *Nature* **2007**, *450*, 964–972.

[113] Koshland Jr., D. E. *Angewandte Chemie International Edition in English* **1995**, *33*, 2375–2378.

[114] Vogt, A. D.; Pozzi, N.; Chen, Z.; Di Cera, E. *Biophysical chemistry* **2014**, *186*, 13–21.

[115] Weikl, T. R.; Paul, F. *Protein Science* **2014**, *23*, 1508–1518.

[116] Laskowski, R. A.; Gerick, F.; Thornton, J. M. *FEBS letters* **2009**, *583*, 1692–1698.

[117] Monod, J.; Wyman, J.; Changeux, J. P. *J Mol Biol* **1965**, *12*, 88–118.

[118] Gianni, S.; Ivarsson, Y.; Jemth, P.; Brunori, M.; Travaglini-Allocatelli, C. *Biophys Chem* **2007**, *128*, 105–113.

[119] Privalov, P. L. *Journal of molecular biology* **1996**, *258*, 707–725.

[120] Hartl, F. U.; Bracher, A.; Hayer-Hartl, M. *Nature* **2011**, *475*, 324–332.

[121] Selkoe, D. J. *Nature cell biology* **2004**, *6*, 1054–1061.

[122] Sali, A.; Glaeser, R.; Earnest, T.; Baumeister, W. *Nature* **2003**, *422*, 216–225.

[123] Behrmann, E.; Loerke, J.; Budkevich, T. V.; Yamamoto, K.; Schmidt, A.; Penczek, P. A.; Vos, M. R.; Bürger, J.; Mielke, T.; Scheerer, P.; Spahn, C. M. T. *Cell* **2015**, *161*, 845–857.

[124] Yon, J. M.; Perahia, D.; Ghélis, C. *Biochimie* **1998**, *80*, 33–42.

[125] Ellis, R. J. *Curr Opin Struct Biol* **2001**, *11*, 114–119.

[126] Wright, P. E.; Dyson, H. J. *Nat Rev Mol Cell Biol* **2015**, *16*, 18–29.

[127] Dyson, H. J.; Wright, P. E. *Nat Rev Mol Cell Biol* **2005**, *6*, 197–208.

[128] Uversky, V. N. *Frontiers in Physics* **2019**, *7*.

[129] Nussinov, R.; Ma, B. *BMC Biology* **2012**, *10*, 2.

[130] Dyson, H. J.; Wright, P. E. *J Biol Chem* **2016**, *291*, 6714–6722.

[131] Kim, Y. E.; Hipp, M. S.; Bracher, A.; Hayer-Hartl, M.; Hartl, F. U. *Annu Rev Biochem* **2013**, *82*, 323–355.

[132] Peng, Z.; Yan, J.; Fan, X.; Mizianty, M. J.; Xue, B.; Wang, K.; Hu, G.; Uversky, V. N.; Kurgan, L. *Cellular and Molecular Life Sciences* **2015**, *72*, 137–151.

[133] Van Der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; others *Chemical reviews* **2014**, *114*, 6589–6631.

[134] Tompa, P. *Trends Biochem Sci* **2012**, *37*, 509–516.

[135] Tompa, P.; Fuxreiter, M. *Trends Biochem Sci* **2007**, *33*, 2–8.

[136] Doherty, A. J.; Jackson, S. P. *Curr Biol* **2001**, *11*, R920–4.

[137] Bao, L.; Zhang, X.; Jin, L.; Tan, Z.-J. *Chinese Physics B* **2015**, *25*, 018703.

[138] Kay, L. E. *J Magn Reson* **2005**, *173*, 193–207.

[139] Palmer, A. G. I. *Chemical Reviews* **2004**, *104*, 3623–3640, PMID: 15303831.

[140] Kleckner, I. R.; Foster, M. P. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2011**, *1814*, 942–968, Protein Dynamics: Experimental and Computational Approaches.

[141] Palmer, A. G., 3rd; Kroenke, C. D.; Loria, J. P. *Methods Enzymol* **2001**, *339*, 204–238.

[142] Selvin, P. R. *Nature Structural Biology* **2000**, *7*, 730–734.

[143] Stryer, L. *Annu Rev Biochem* **1978**, *47*, 819–846.

[144] Jares-Erijman, E. A.; Jovin, T. M. *Current opinion in chemical biology* **2006**, *10*, 409–416.

[145] Bastiaens, P. I. H.; Squire, A. *Trends in Cell Biology* **1999**, *9*, 48–52.

[146] Millar, D. P. *Curr Opin Struct Biol* **1996**, *6*, 637–642.

[147] Eaton, W. A.; Muñoz, V.; Hagen, S. J.; Jas, G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. *Annu Rev Biophys Biomol Struct* **2000**, *29*, 327–359.

[148] Hess, C. *Chemical Society Reviews* **2021**, *50*, 3519–3564.

[149] Putnam, C. D.; Hammel, M.; Hura, G. L.; Tainer, J. A. *Q Rev Biophys* **2007**, *40*, 191–285.

[150] Svergun, D. I.; Koch, M. H. J.; Timmins, P. A.; May, R. P. *Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules*; Oxford University Press, 2013.

[151] Heller, W. T. *Acta Crystallogr D Biol Crystallogr* **2010**, *66*, 1213–1217.

[152] Krueger, S. *Curr Opin Struct Biol* **2022**, *74*, 102375.

[153] Gabel, F. *Methods Enzymol* **2015**, *558*, 391–415.

[154] Jeffries, C. M.; Pietras, Z.; Svergun, D. I. The basics of small-angle neutron scattering (SANS for new users of structural biology). EPJ Web of Conferences. 2020; p 03001.

[155] Kirby, N. M.; Cowieson, N. P. *Current opinion in structural biology* **2014**, *28*, 41–46.

[156] Karplus, M.; Kuriyan, J. *Proc Natl Acad Sci U S A* **2005**, *102*, 6679–6685.

[157] Marais, A.; Adams, B.; Ringsmuth, A. K.; Ferretti, M.; Gruber, J. M.; Hendrikx, R.; Schuld, M.; Smith, S. L.; Sinayskiy, I.; Krüger, T. P. J.; Petruccione, F.; van Grondelle, R. *J R Soc Interface* **2018**, *15*.

[158] Soleymani, F.; Paquet, E.; Viktor, H.; Michalowski, W.; Spinello, D. *Comput Struct Biotechnol J* **2022**, *20*, 5316–5341.

[159] Dill, K. A.; MacCallum, J. L. *Science* **2012**, *338*, 1042–1046.

[160] Martí-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; Sali, A. *Annu Rev Biophys Biomol Struct* **2000**, *29*, 291–325.

[161] Krieger, E.; Joo, K.; Lee, J.; Lee, J.; Raman, S.; Thompson, J.; Tyka, M.; Baker, D.; Karplus, K. *Proteins* **2009**, *77 Suppl 9*, 114–122.

[162] Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. *Nucleic acids research* **2003**, *31*, 3381–3385.

[163] Bonneau, R.; Baker, D. *Annual review of biophysics and biomolecular structure* **2001**, *30*, 173–189.

[164] Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D. *Proteins* **1999**, *Suppl 3*, 171–176.

[165] Baek, M. et al. *Science* **2021**, *373*, 871–876.

[166] Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. *Proteins* **2021**, *89*, 1607–1617.

[167] Terwilliger, T. C.; Liebschner, D.; Croll, T. I.; Williams, C. J.; McCoy, A. J.; Poon, B. K.; Afonine, P. V.; Oeffner, R. D.; Richardson, J. S.; Read, R. J.; others *Nature Methods* **2024**, *21*, 110–116.

[168] Dror, R. O.; Jensen, M. Ø.; Borhani, D. W.; Shaw, D. E. *Journal of General Physiology* **2010**, *135*, 555–562.

[169] Hollingsworth, S. A.; Dror, R. O. *Neuron* **2018**, *99*, 1129–1143.

[170] Friesner, R. A.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389–427.

[171] Senn, H. M.; Thiel, W. *Angew Chem Int Ed Engl* **2009**, *48*, 1198–1229.

[172] Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. *Chem. Soc. Rev.* **2012**, *41*, 3025–3038.

[173] Pagadala, N. S.; Syed, K.; Tuszynski, J. *Biophys Rev* **2017**, *9*, 91–102.

[174] Evans, R. et al. *bioRxiv* **2022**,

[175] Sanavia, T.; Birolo, G.; Montanucci, L.; Turina, P.; Capriotti, E.; Fariselli, P. *Comput Struct Biotechnol J* **2020**, *18*, 1968–1979.

[176] Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J Med Chem* **2004**, *47*, 1739–1749.

[177] Cheng, T. M. K.; Lu, Y.-E.; Vendruscolo, M.; Lio', P.; Blundell, T. L. *PLoS Comput Biol* **2008**, *4*, e1000135.

[178] Thusberg, J.; Olatubosun, A.; Vihinen, M. *Human mutation* **2011**, *32*, 358–368.

[179] Niroula, A.; Vihinen, M. *PLoS computational biology* **2019**, *15*, e1006481.

[180] Glaser, J.; Nguyen, T. D.; Anderson, J. A.; Lui, P.; Spiga, F.; Millan, J. A.; Morse, D. C.; Glotzer, S. C. *Computer Physics Communications* **2015**, *192*, 97–107.

[181] Shaw, D. E.; Adams, P. J.; Azaria, A.; Bank, J. A.; Batson, B.; Bell, A.; Bergdorf, M.; Bhatt, J.; Butts, J. A.; Correia, T.; others Anton 3: twenty microseconds of molecular dynamics simulation before lunch. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2021; pp 1–11.

[182] Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. *Annu Rev Biophys* **2012**, *41*, 429–452.

[183] Sugita, Y.; Okamoto, Y. *Chemical physics letters* **1999**, *314*, 141–151.

[184] Barducci, A.; Bonomi, M.; Parrinello, M. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 826–843.

[185] Bernardi, R. C.; Melo, M. C.; Schulten, K. *Biochimica et Biophysica Acta (BBA) - General Subjects* **2015**, *1850*, 872–877, Recent developments of molecular dynamics.

[186] MacKerell Jr, A. D. *Journal of computational chemistry* **2004**, *25*, 1584–1604.

[187] Nerenberg, P. S.; Head-Gordon, T. *Current opinion in structural biology* **2018**, *49*, 129–138.

[188] Ponder, J. W.; Case, D. A. *Protein Simulations*; Advances in Protein Chemistry; Academic Press, 2003; Vol. 66; pp 27–85.

[189] Li, P.; Song, L. F.; Merz, K. M., Jr *J Phys Chem B* **2014**, *119*, 883–895.

[190] Senn, H. M.; Thiel, W. *Atomistic approaches in modern biology: from quantum chemistry to molecular simulations* **2007**, 173–290.

[191] Lin, H.; Truhlar, D. G. *Theoretical Chemistry Accounts* **2007**, *117*, 185–199.

[192] Desta, I. T.; Porter, K. A.; Xia, B.; Kozakov, D.; Vajda, S. *Structure* **2020**, *28*, 1071–1081.

[193] Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. *J Med Chem* **2006**, *49*, 5851–5855.

[194] Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins: Structure, Function, and Bioinformatics* **2002**, *47*, 409–443.

[195] Coupez, B.; Lewis, R. *Current medicinal chemistry* **2006**, *13*, 2995–3003.

[196] Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. *J Med Chem* **2006**, *49*, 5912–5931.

[197] Senior, A. W. et al. *Nature* **2020**, *577*, 706–710.

[198] Li, B.; Basu Roy, R.; Wang, D.; Samsi, S.; Gadepally, V.; Tiwari, D. Toward sustainable hpc: Carbon footprint estimation and environmental implications of hpc systems. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2023; pp 1–15.

[199] Lannelongue, L.; Aronson, H.-E. G.; Bateman, A.; Birney, E.; Caplan, T.; Juckes, M.; McEntyre, J.; Morris, A. D.; Reilly, G.; Inouye, M. *Nature Computational Science* **2023**, *3*, 514–521.

[200] Wagner, M. *Journal of Chemical Education* **1976**, *53*, A472.

[201] Hersh, R. T. *Systematic Biology* **1967**, *16*, 262–263.

[202] Needleman, S. B.; Wunsch, C. D. *Journal of Molecular Biology* **1970**, *48*, 443–453.

[203] Smith, T.; Waterman, M. *Journal of Molecular Biology* **1981**, *147*, 195–197.

[204] McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585–590.

[205] Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. *J Mol Biol* **1963**, *7*, 95–99.

[206] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Research* **2000**, *28*, 235–242.

[207] Humphrey, W.; Dalke, A.; Schulten, K. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

[208] Schrödinger, L.; DeLano, W. PyMOL. `http://www.pymol.org/pymol`.

[209] Olson, M. V. *Proceedings of the National Academy of Sciences* **1993**, *90*, 4338–4344.

[210] Collins, F. S.; Patrinos, A.; Jordan, E.; Chakravarti, A.; Gesteland, R.; Walters, L.; members of the DOE; planning groups, N. *science* **1998**, *282*, 682–689.

[211] Lander, E. S. et al. *Nature* **2001**, *409*, 860–921.

[212] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *Journal of Molecular Biology* **1990**, *215*, 403–410.

[213] Felsenstein, J. *Journal of Molecular Evolution* **1981**, *17*, 368–376.

[214] Cummings, M. P. *Dictionary of Bioinformatics and Computational Biology*; John Wiley, Ltd, 2004.

[215] Huelsenbeck, J. P.; Ronquist, F. *Bioinformatics* **2001**, *17*, 754–755.

[216] Posada, D.; Crandall, K. A. *Trends in Ecology and Evolution* **2001**, *16*, 37–45.

[217] Leaver-Fay, A. et al. *Methods Enzymol* **2011**, *487*, 545–574.

[218] Moult, J.; Pedersen, J. T.; Judson, R.; Fidelis, K. A large-scale experiment to assess protein structure prediction methods. 1995.

[219] Sayers, E. W. et al. *Nucleic Acids Res* **2022**, *50*, D20–D26.

[220] Kitano, H. *science* **2002**, *295*, 1662–1664.

[221] Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. *Genome Res* **2003**, *13*, 2498–2504.

[222] Baek, M. et al. *Science* **2021**, *373*, 871–876.

[223] Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; others *Proteins: Structure, Function, and Bioinformatics* **2021**, *89*, 1711–1721.

[224] Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. *arXiv preprint arXiv:2210.01776* **2022**,

[225] Dauparas, J. et al. *Science* **2022**, *378*, 49–56.

[226] Watson, J. L. et al. *Nature* **2023**, *620*, 1089–1100.

[227] Li, Q.; Vlachos, E. N.; Bryant, P. *bioRxiv* **2024**, 2024–06.

[228] Warth, J.; Desforges, J. F. *Br J Haematol* **1975**, *29*, 369–372.

[229] Bu, W. et al. *Sci Adv* **2023**, *9*, eade0059.

[230] Zhu, L. J.; Holmes, B. R.; Aronin, N.; Brodsky, M. H. *PLoS One* **2014**, *9*, e108424.

[231] Naito, Y.; Hino, K.; Bono, H.; Ui-Tei, K. *Bioinformatics* **2014**, *31*, 1120–1123.

[232] Braun, E.; Gilmer, J.; Nayes, H. B.; Mboley, D. L.; Monroe, J. I.; Prasa, S.; Zuckerman, D. M. *Living Journal of Computational Molecular Science* **2019**, *1*, 5957.

[233] Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Van Gunsteren, W. F.; Mark, A. E. *Angewandte Chemie International Edition* **1999**, *38*, 236–240.

[234] Adcock, S. A.; McCammon, J. A. *Chemical Reviews* **2006**, *106*, 1589–1615.

[235] Bottaro, S.; Lindorff-Larsen, K. *Science* **2018**, *361*, 355–360.

[236] Parasuraman, S. *Journal of Pharmacology and Pharmacotherapeutics* **2012**,

[237] Gonzalez, M. A. *Collection SFN* **2011**, *12*, 1269–200.

[238] Leimkuhler, B. J.; Reich, S.; Skeel, R. D. In *Mathematical Approaches to Biomolecular Structure and Dynamics*; Mesirov, J. P., Schulten, K., Sumners, D. W., Eds.; Springer New York: New York, NY, 1996; pp 161–185.

[239] Muhammed, M. T.; Aki-Yalcin, E. *Letters in Drug Design & Discovery* **2024**, *21*, 480–495.

[240] Paggi, J. M.; Pandit, A.; Dror, R. O. *Annual Review of Biochemistry* **2024**, *93*.

[241] Sahu, M. K.; Nayak, A. K.; Hailemeskel, B.; Eyupoglu, O. E. *International Journal of Pharmaceutical Research and Allied Sciences* **2024**, *13*, 24–40.

[242] Kumar, P.; Kumar, A. *Computational Drug Discovery: Molecular Simulation for Medicinal Chemistry* **2024**, 65.

[243] Morris, G. M.; Lim-Wilby, M. *Molecular Modeling of Proteins*; Humana Press, 2008; Vol. 443.

[244] Lill, M. A. *Biochemistry* **2011**, *50*, 6157–6169.

[245] Lexa, K. W.; Carlson, H. A. *Quarterly reviews of biophysics* **2012**, *45*, 301–343.

[246] Krieger, E.; Nabuurs, S. B.; Vriend, G. *Structural bioinformatics* **2003**, *44*, 509–523.

[247] Huang, P.-S.; Boyken, S. E.; Baker, D. *Nature* **2016**, *537*, 320–327.

[248] Lin, E.; Lin, C.-H.; Lane, H.-Y. *Molecules* **2020**, *25*, 3250.

[249] Lin, E.; Lin, C.-H.; Lane, H.-Y. *Journal of Chemical Information and Modeling* **2022**, *62*, 761–774.

[250] Ingraham, J.; Garg, V.; Barzilay, R.; Jaakkola, T. Generative Models for Graph-Based Protein Design. Advances in Neural Information Processing Systems. 2019.

[251] Chu, A. E.; Lu, T.; Huang, P.-S. *Nature biotechnology* **2024**, *42*, 203–215.

[252] Bottaro, S.; Bussi, G.; Kennedy, S. D.; Turner, D. H.; Lindorff-Larsen, K. *Science Advances* **2018**, *4*, eaar8521.

[253] Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. *Sci Adv* **2016**, *2*, e1501177.

[254] Löhr, T.; Jussupow, A.; Camilloni, C. *The Journal of chemical physics* **2017**, *146*.

[255] Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J. *Journal of the American Chemical Society* **2003**, *125*, 1731–1737, PMID: 12580598.

[256] van Zundert, G.; Rodrigues, J.; Trellet, M.; Schmitz, C.; Kastritis, P.; Karaca, E.; Melquiond, A.; van Dijk, M.; de Vries, S.; Bonvin, A. *Journal of Molecular Biology* **2016**, *428*, 720–725, Computation Resources for Molecular Biology.

[257] Tuukkanen, A. T.; Spilotros, A.; Svergun, D. I. *IUCrJ* **2017**, *4*, 518–528.

[258] Niebling, S.; Björling, A.; Westenhoff, S. *Journal of Applied Crystallography* **2018**, *51*, 968.

[259] Paissoni, C.; Jussupow, A.; Camilloni, C. *Journal of Applied Crystallography* **2019**, *52*, 394–402.

[260] Svergun, D. I.; Richard, S.; Koch, M. H. J.; Sayers, Z.; Kuprin, S.; Zaccai, G. *Proceedings of the National Academy of Sciences* **1998**, *95*, 2267–2272.

[261] Tribello, G.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Computer Physics Communications* **2014**, *185*, 604–613.

[262] PLUMED Tutorials. `https://www.plumed-tutorials.org/`.

[263] Scalone, E.; Broggini, L.; Visentin, C.; Erba, D.; Toplek, F. B.; Peqini, K.; Pellegrino, S.; Ricagno, S.; Paissoni, C.; Camilloni, C. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2203181119.

[264] Bačić Toplek, F.; Scalone, E.; Stegani, B.; Paissoni, C.; Capelli, R.; Camilloni, C. *Journal of Chemical Theory and Computation* **2024**, *20*, 459–468, PMID: 38153340.

[265] Stevens, R. C.; Cherezov, V.; Katritch, V.; Abagyan, R.; Kuhn, P.; Rosen, H.; Wüthrich, K. *Nature reviews Drug discovery* **2013**, *12*, 25–34.

[266] Schrödinger, L. Maestro 2023-3, 2023; New York, NY.

[267] Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *Journal of Computational Chemistry* **2008**, *29*, 1859–1865.

[268] Lee, J. et al. *Journal of Chemical Theory and Computation* **2016**, *12*, 405–413.

[269] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. *SoftwareX* **2015**, *1*, 19–25.

[270] Van Rossum, G.; Drake, F. L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.

[271] Billesbølle, C. B.; de March, C. A.; van der Velden, W. J. C.; Ma, N.; Tewari, J.; del Torrent, C. L.; Li, L.; Faust, B.; Vaidehi, N.; Matsunami, H.; Manglik, A. *Nature* **2023**, *615*, 742–749.

[272] Wolf, S.; Jovancevic, N.; Gelis, L.; Pietsch, S.; Hatt, H.; Gerwert, K. *Scientific Reports* **2017**, *7*, 16007.

[273] Abaffy, T.; Bain, J. R.; Muehlbauer, M. J.; Spasojevic, I.; Lodha, S.; Bruguera, E.; O'Neal, S. K.; Young Kim, S.; Matsunami, H. *Frontiers in Oncology* **2018**, *8*.

[274] Cutolo, E. A.; Guardini, Z.; Dall'Osto, L.; Bassi, R. *New Phytologist* **2023**, *239*, 1567–1583.

[275] Genesio, L.; Bassi, R.; Miglietta, F. *Global Change Biology* **2021**, *27*, 959–967.

[276] Talamè, V.; Bovina, R.; Sanguineti, M. C.; Tuberosa, R.; Lundqvist, U.; Salvi, S. *Plant Biotechnology Journal* **2008**, *6*, 477–485.

[277] Gao, Y.-S.; Wang, Y.-L.; Wang, X.; Liu, L. *Protein Science* **2020**, *29*, 1026–1032.

[278] Walker, J. C.; Willows, D. R. *Biochemical Journal* **1997**, *327*, 321–333.