Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Carbonara: A Rapid Method for SAXS-Based Refinement of Protein Structures

**Matteo Degiacomi**

matteo.degiacomi@ed.ac.uk

U. Edinburgh

**Joshua Mckeown**

Durham University

**Arron Bale**

Durham University

**Cameron Brown**

University of Southampton

**Hayden Fisher**

European Synchrotron Radiation Facility https://orcid.org/0000-0003-0093-0921

**Robert Rambo**

Diamond Light Source Ltd https://orcid.org/0000-0003-2546-1615

**Jonathan Essex**

University of Southampton https://orcid.org/0000-0003-2639-2746

**Christopher Prior**

Durham University https://orcid.org/0000-0003-4015-5106

**Article**

**Keywords:**

**Posted Date:** May 2nd, 2025

**DOI:** https://doi.org/10.21203/rs.3.rs-6447099/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** There is **NO** Competing Interest.

# Carbonara: A Rapid Method for SAXS-Based Refinement of Protein Structures

Josh McKeown,[†] Arron Bale,[†] Cameron Brown,[‡] Hayden Fisher,[‡,‖] Robert P. Rambo,[¶] Jonathan W. Essex,[‡] Matteo T. Degiacomi,[*,§,⊥] and Christopher Prior[*,†]

†*Department of Mathematical Sciences, Durham University, South Road, Durham, DH1 3LE, United Kingdom*[1]

‡*School of Chemistry and Chemical Engineering, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom*

¶*Diamond Light Source, Harwell Science and Innovation Campus, Didcot, United Kingdom*

§*EaStCHEM School of Chemistry, University of Edinburgh, Edinburgh, UK*

‖*ESRF - The European Synchrotron, 71 Avenue des Martyrs, 38043, Grenoble, France*

⊥*School of Informatics, University of Edinburgh, Edinburgh, UK*

E-mail: matteo.degiacomi@ed.ac.uk; christopher.prior@durham.ac.uk

## Abstract

Generative machine learning models for protein structure prediction are primarily trained on X-ray crystallography data, which captures proteins in crystal lattices that can deviate significantly from their native solution conformations. Biological small angle X-ray scattering (bioSAXS) offers valuable solution-state insights, but creating atomic models that rationalise this data remains challenging.

Here we present Carbonara, a rapid computational pipeline that combines coarse-grained sampling with experimental constraints to efficiently identify solution-state conformations from an initial atomic model. We demonstrate Carbonara's effectiveness by refining an AlphaFold-predicted model of the DNA repair helicase, SMARCAL1, and a crystallographically determined structure of the antigen binding domains of the anti-*h*CD40 monoclonal antibody, ChiLob7/4, a clinically relevant immunostimulatory antibody. In both cases, Carbonara identifies physiologically relevant solution-state conformations separated from crystal-like predictions by large energy barriers, achieving in minutes what traditional MD simulations might not accomplish in weeks.

# 1 Introduction

The biological function of any protein emerges from its molecular structure and associated dynamics. For the past 50 years, this paradigm has been validated through the atomistic details provided by macromolecular X-ray crystallography [1], Nuclear Magnetic Resonance (NMR)[2] and cryo-electron microscopy (cryo-EM)[3]. These techniques have served as our primary window into the molecular mechanisms underpinning life. The field of structural biology has once again been transformed, noted by the 2024 Nobel Prize for AI-based tools AlphaFold [4] and RoseTTAfold [5], that predict protein structures directly from their primary sequence. These computational tools have accelerated our progress in structural genomics by enabling the prediction of atomic structures from genes encoding millions of unobserved proteins.

However, AI-based methods inherit limitations from the data on which they are primarily trained, which to-date is gathered from the Protein Data Bank (PDB) and is mostly crystallographic in nature. Since crystal structures capture proteins within packed lattices, the PDB provides limited sampling of the conformational space proteins occupy in solution [6]. This restricted representation of protein dynamics in the training data impacts the ability of AI models to capture the full range of conformational dynamics proteins exhibit in their native, non-crystallographic, environments. Furthermore, proteins may feature disordered regions hard to resolve experimentally, yet fundamental for their function [7, 8]. Recent studies have shown that AI-based protein structure predictions often fail to reflect the dynamic nature of protein structures [9]. This limitation is particularly evident when proteins undergo large conformational changes, which are often crucial for their biological function. While ensemble adaptations such as AlphaFlow [10] and P2DFlow [11], aim to enhance structural sampling, they are most effective at capturing local protein dynamics. These methods fall short in representing the conformational changes that proteins often exhibit in solution, highlighting the need for complementary experimental approaches.

Small-angle X-ray scattering (SAXS) offers such a complementary approach to structure determination. While its low-resolution and associated challenges in data interpretation limit its use as a stand-alone technique for structure determination [12], SAXS offers the key advantage of providing information on protein shape and conformational changes in solution [13]. SAXS also allows the study of flexible structures that may be difficult to crystallize, while not requiring the monodisperse and isotopically labelled samples needed for NMR [14]. Furthermore, it is applicable to a broad range of molecular weights, whereas NMR and cryo-EM are typically limited to small ($< 50$ kDa) and large ($> 100$ kDa) protein samples, respectively [15, 16].

By capturing information on proteins structure and dynamics in their near-native environment, SAXS can provide key insights into their functionality. For instance, SAXS was used to characterise conformational changes in dengue virus proteins (NS3/NS5), aiding understanding viral replication [17]. It was also used to clarify apolipoprotein flexibility in lipid transport [18], to determine the dynamic behaviour of matrix metalloproteases (MMPs) in tissue remodeling [19] and highlight an evolutionary link between bacterial and eukaryotic lipid metabolism [20]. Crucially, discrepancies between SAXS data and both crystallographically-determined and AI-predicted structures are often observed [21, 22, 23]. This underscores the importance of developing methods that can reliably interpret SAXS data and enable the refinement of high-resolution static models to better reflect solution-state dynamics.

However, the interpretation of protein SAXS data remains challenging, owing to its lower resolution. Current approaches to incorporate SAXS data in structural refinement fall into several categories, each with its own limitations. Validation methods [24, 25, 26, 27, 28, 29, 30] are useful to compare available high-resolution structures with SAXS data, but do not provide a means to refine structures that deviate significantly from the data. Protein structure refinement can be carried out via rigid body modelling methods, [31] which approximate proteins as collections of hinging rigid bodies. However, these often struggle to generate physically realistic conformers. Molecular Dynamics (MD) simulations [32], al-

though capable of producing realistic structural variations, are computationally expensive and may not efficiently sample large conformational changes, especially when starting from initial predictions that significantly differ from the final in-solution states. Ab initio methods, allow for SAXS interpretation without initial models, such as bead models [33] and density approximation [34]. While computationally efficient, these methods produce low-resolution models that are difficult to interpret and may not be constrained to be physically realistic.

To go beyond the limitations of existing SAXS-guided computational refinement methods, we present Carbonara, a modelling pipeline which enables the production of ensembles of physically viable proteins conformers representative of their solution state. Carbonara builds upon an ab initio scattering model [35] of the protein's alpha carbon trace, and combines secondary structure restraints with a rapid conformational search to generate models matching available SAXS data, starting from a model with poor agreement to the data. We demonstrate the utility of Carbonara through two case studies involving the refinement of an AI-predicted and a crystallographically-determined protein structure that disagree with SAXS data. In the first case, we demonstrate how Carbonara refinement provides insight into the mode of action of SMARCAL1, a protein that can bind simultaneously to single and double stranded DNA to stabilize DNA at stalled replication or transcription forks. In the second case, we show how refinement of the antigen binding arms of the ChiLob7/4 IgG2 antibody enables seeding MD simulations able to sample otherwise unreachable solution-state conformations. Overall, our results demonstrate how refining experimentally or AI-generated protein models to agree with in-solution SAXS data provides new insights into protein structure, dynamics, and function.

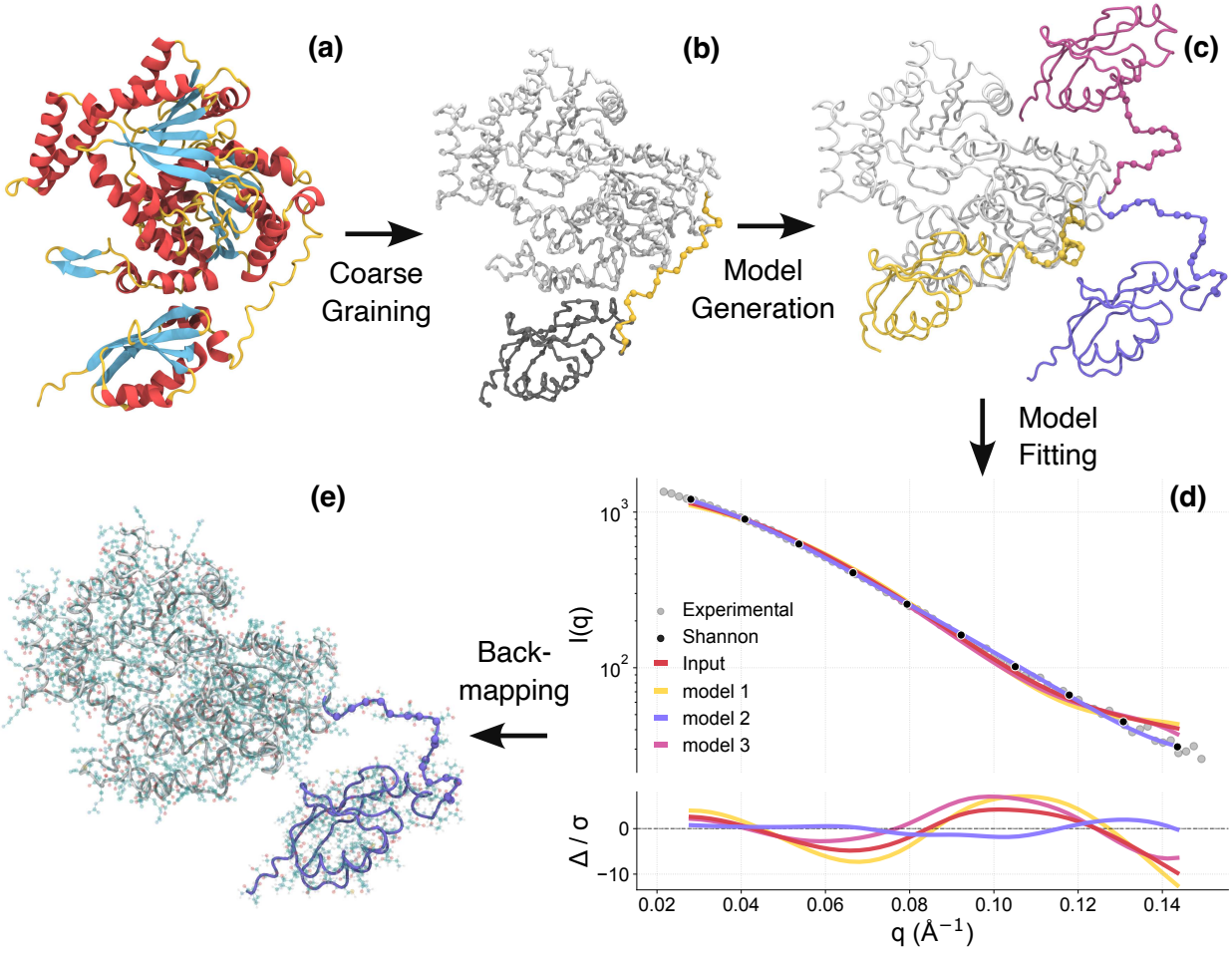



Figure 1: **Illustration of the Carbonara modelling process**. (a) the secondary structure of the input protein structure is calculated (sheets in blue, helices in red, coils in yellow). (b) the coordinates of C$\alpha$ atoms are extracted, and linker sections (an example shown in yellow) identified. (c) new protein conformers are generated by altering the linker section so as to improve the fit with SAXS data subject to statistical and user-defined constraints. (d) the models yielding the best fit, here "model 2", are selected (e) the backmapped all-atomic coordinates are superimposed onto the protein C$\alpha$ trace.

# 2 Theory

Carbonara is a computational pipeline that refines protein models against SAXS data using a coarse-grained representation, enabling efficient exploration of folded protein conformations. It integrates a validated mathematical framework for analysing bioSAXS data [35, 36] (see Supplementary Methods S1) in a modelling pipeline incorporating a range of features including progressive q-range fitting, integration of additional structural constraints, and back-mapping to all-atom models which can be assessed with MD simulations. Hereafter, we summarise its key elements, see Methods for further details.

## 2.1 Coarse-grained conformational sampling

Carbonara takes as input protein structures composed of one or more subunits, and an experimental SAXS profile. The protein structure is reduced to a coarse-grain representation using the positions of its alpha carbons (C$\alpha$) to form a chain that is segmented into secondary structure elements ($\alpha$-helices, $\beta$-sheets, and linkers, see Figure 1a). Linker sections are then altered to sample the protein's potential conformational space. The linker segments which are permitted to change during the sampling process are either automatically determined or user-defined. Carbonara offers an automated selection process that identifies flexible regions while preserving all existing inter $\beta$-sheet structures present in the initial AI-based model or crystal structure. This can then be accepted or edited to user preference.

To generate new prospective configurations with realistic protein architecture, the local geometry of the model's secondary structure sections are characterised and constrained by two quantities: the curvature $\kappa$, which characterises the tightness of coiling of subsections, and the torsion $\tau$, which additionally characterises the helical pitch of this section (Figure S1). Carbonara constructs new structures for the flexible segments using empirically obtained probability distributions for their $(\kappa, \tau)$ values from a large database of known (distinct) structures [35, 36] (Figure S2). The three distinct density peaks of these distributions were

shown to correspond to the three main Ramachandran densities in [35], so this coarse grained model captures the local geometrical constraints on the C$\alpha$ trace imposed by the Ramachandran constraints 1(a)-(c). For multimeric structures, rigid body rotations of the monomer units can also be applied. This conformational sampling algorithm is also used in the initial selection process to test which sections, if altered, will break any existing $\beta$-sheet bonds. If there exist additional known constraints, such as disulfide bonds, NMR distance constraints or FRET data, these can be imposed by restraining the distance of user-defined C$\alpha$ -C$\alpha$ pairs.

The plausibility of each generated protein conformation is then graded using a combination of an amino acid overlap metric [35], a distance constraint metric (*e.g.*, for disulfide bonds), and a fold complexity metric based on the writhe of the curve's backbone, which has been shown empirically to have a strict lower bound as a function of the size of the molecule [36].

## 2.2 Fitting models to SAXS data

The SAXS profile is calculated for C$\alpha$ conformations using the Debye scattering formula with an explicit protein hydration layer model (see Figure S2). Figure 1(d) illustrates the scattering assessment of the proposed structural changes shown in panel (c). The comparison of model and theory is made at a number of equally spaced scattering data points given by the Shannon sampling theorem [29, 35, 37] with a $\chi^2$ metric used to judge the fit including the penalties for the plausibility of the structure which are discussed above. The scattering model used in Carbonara, based solely on the $C_\alpha$ trace, has been shown to be consistent with the established FoXS protein training set ([24]) in [35] and further details are given the methods section.

In Carbonara, we introduce a novel progressive fitting procedure for SAXS data. The scattering angle, represented by the range of $q$ values, provides structural insights: low $q$ values reveal the overall shape (globular or elongated), while higher $q$ values resolve finer details

like super-secondary and secondary structures [13]. SAXS experiments capture the averaged scattering from highly flexible protein ensembles, making the higher $q$ range challenging to represent with a single structure. While MD simulations can generate such ensembles, they are computationally expensive, particularly if the starting structure poorly represents the protein's solution-state ensemble. In such cases, generating suitable ensembles within accessible simulation timescales may be infeasible. Carbonara addresses this by progressively refining protein models. It begins with fitting low to medium $q$ data, capturing the protein's larger-scale arrangement, and iteratively expands the fitting window to include higher $q$ data, ultimately providing an atomistic model refined for all relevant scattering ranges.

### 2.3 All-atom backmapping

An all-atom structure is needed to infer key biomolecular processes and functions of the model, limiting the value of C$\alpha$ backbone predictions alone. MODELLER [38] was used to reconstruct the full atomic structure from Carbonara's C$\alpha$ trace predictions [39] (see example in Figure 1e). This step includes energy minimisation, allowing the C$\alpha$ chain to relax into more energetically favourable conformation upon side chain placement. In preliminary benchmarks, we found that this only led to minimal adjustments to the C$\alpha$ trace (0.475 Å mean RMSD). This small magnitude demonstrates that the $\kappa$ - $\tau$ sampling produces C$\alpha$ geometries compatible with realistic protein structures. Once reconstructed, the all-atom model can be compared with the full range of experimental scattering data using standard techniques and serve as a starting structure for MD methods.

## 3 Results

To demonstrate Carbonara's efficacy in refining structures using experimental bioSAXS data, we selected two systems that undergo significant conformational changes to carry out their biological function: the catalytic domain (SMARCAL1$^{CD}$) of the monomeric, multi-domain

ATP-dependent chromatin remodelling protein SMARCAL1, and a disulfide engineered variant of the anti-*h*CD40 antibody ChiLob7/4 IgG2.

We use these examples to demonstrate how Carbonara can be leveraged as part of two protocols commonly used to atomistically rationalise the conformational ensemble occupied by a bioSAXS sample. The first protocol, applied to SMARCAL1$^{CD}$, involves producing an ensemble of conformers used to seed a swarm of relatively short (30x200 ns) MD simulations. The second, for the ChiLob7/4 antibody, is to produce a single starting structure for longer (3x2 $\mu$s) simulations.

## 3.1 Refinement of an AlphaFold model: Human SMARCAL1

Human SMARCAL1 is a DNA-dependent, ATP motor protein for which no experimentally determined atomic-resolution structure is currently available. SMARCAL1$^{CD}$, which consists of residues 325-870 of this protein, is a truncated form of the enzyme which is comprised of an N-terminal HARP domain (residue 328 to 396) and a C-terminal, bi-lobed ATPase motor domain related to SNF2 family of helicases. SMARCAL1$^{CD}$ has low affinity for double-stranded (ds) DNA substrates but high affinity for mixed dsDNA substrates with single-stranded (ss) regions. This affinity is enabled by the HARP domain which is critical to the ATPase activity of the associated motor domain. Previous solution-state studies of SMARCAL1$^{CD}$ demonstrated SMARCAL1$^{CD}$ undergoes significant conformational changes in the presence of dsDNA with a 10-nucleotide ss overhang after binding and hydrolysis of ATP [40].

An AI-generated model (AlphaFold3) of SMARCAL1$^{CD}$ fit the SAXS data with a $\chi^2$ of 13.52, suggesting the model poorly describes the observed solution state. To improve the fit, we used Carbonara to refine the AI-model to produce an ensemble of 30 candidate models. All of the refined models had a $\chi^2 \leq 1.32$ from experimental data (Figure S3). We selected the model with $\chi^2$ closest to 1, which had a value of $\chi^2 = 1.01$ (Figure 2c, in blue).

All improved fits required domain rearrangements of the AI-model that supported an elon-

gated conformation of SMARCAL1$^{CD}$, with the HARP2 domain extended away from the main body of the protein (yellow region in Figure 2b,c).

To explore the conformational landscape accessible from different starting structures, we performed molecular simulations on both the original AlphaFold model (three 2$\mu$s trajectories) and each of the 30 selected Carbonara models (200 ns each). Both approaches provided 6$\mu$s of cumulated MD time. To identify the key structural differences between the conformational spaces explored by the MD simulations seeded from the AlphaFold and Carbonara models, we aggregated all the produced trajectories and carried out a principal component analysis (see Figure 2c). Results revealed that the first Principal Component PC1, capturing 69.87% of the overall variance, describes variations in the position of the HARP2 domain relative to the rest of the protein. Notably, the Carbonara and AlphaFold conformational spaces are non-overlapping, revealing that the two are likely separated by a high energy barrier. The AlphaFold model remains compact throughout every simulation, and all the sampled conformations poorly match experimental SAXS data (see Figure S4). Conversely, with the same computational effort, the conformations sampled from the Carbonara model initialisations explore a broad range of more extended conformations, with much lower $\chi^2$ from experimental data. A significant proportion of these features $\chi^2 < 1.5$, a typical criterion used in the literature to imply that a model is reflective of the solution structure [24].

## 3.2 Insight into the functionality of SMARCAL1

To understand the biological implications of the Carbonara-refined SMARCAL1$^{CD}$ structure, we analysed it in the context of the related DNA-bound SNF2 family of proteins. SNF2 proteins are ATP-dependent motor proteins with a core helicase-like domain and flanking auxiliary domains tuned to their respective biological functions. The SNF2 family of proteins remodel chromatin structure and in the case of SMARCAL1, DNA binding is mediated by a critical N-terminal HARP domain. In the absence of DNA or a functional DNA-binding HARP domain, SMARCAL1 does not hydrolyze ATP and likewise, in the absence of a

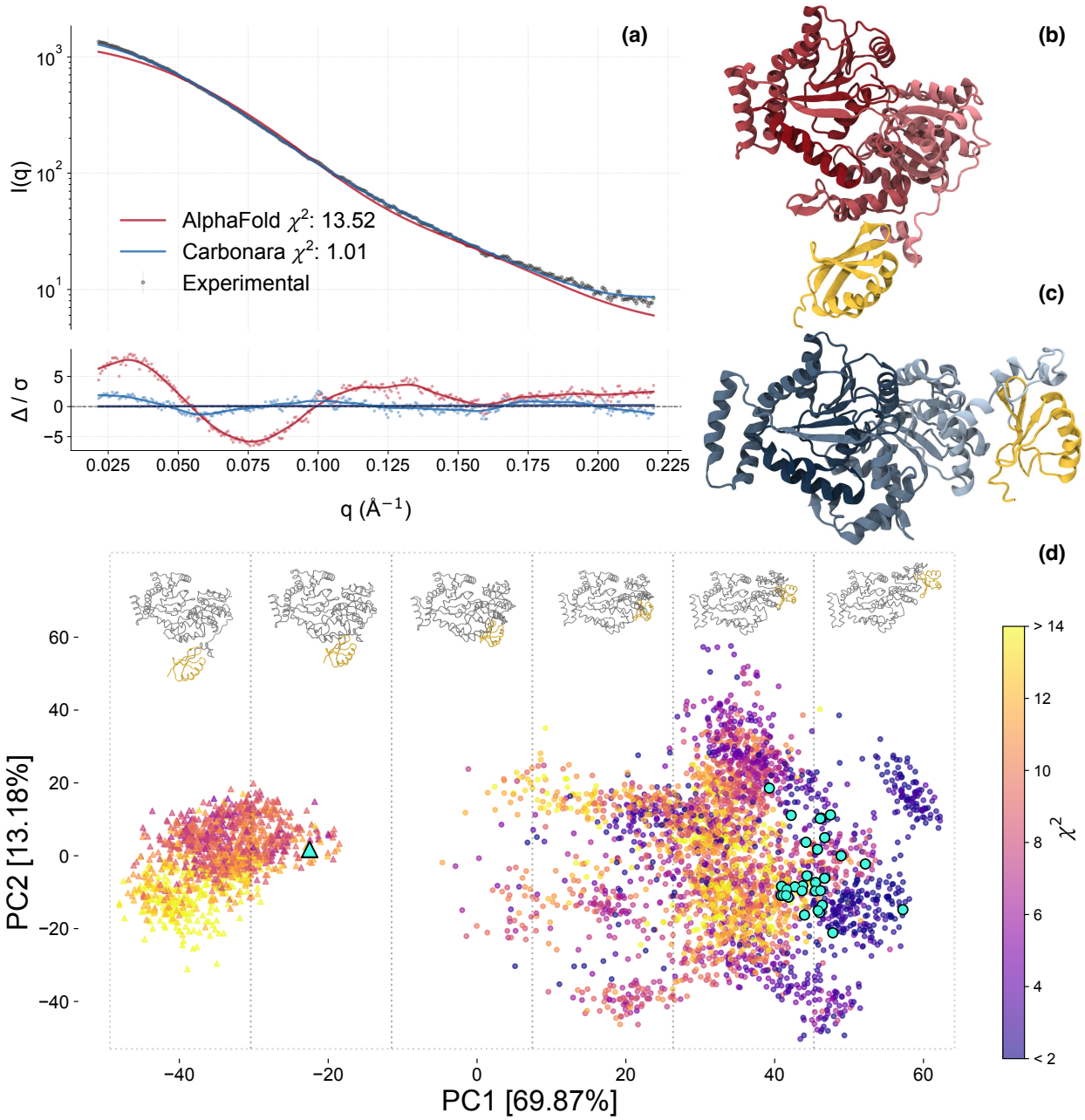



Figure 2: **Refining the AlphaFold model of SMARCAL1**. (a) Top, comparison of experimental SAXS data (gray points) and CRYSOL-computed SAXS from the AI prediction (red) and the $\chi^2$ closest to 1 Carbonara model (blue). Bottom, the error-weighted residuals between data and models. (b,c) the AI model is shown in red, the best Carbonara model in blue (light colours denoting the beginning of the chains, and dark ones the end), HARP2 domain shown in yellow. (d) Projection on the first two principal components (PC, with explained variance reported in square brackets) of all the conformers produced by MD trajectories seeded from the AlphaFold (cyan triangle) and Carbonara (cyan circles) models. Data points are coloured according to their $\chi^2$ fit to the SAXS data. Configurations representing the motion captured by the first principal component (PC1) are shown along the top of the figure, demonstrating that SMARCAL1 elongates via a rearrangement of its HARP2 domain (shown in yellow) with respect of the ATPase domain (in white).

functional HARP domain, SMARCAL1 does not bind DNA substrates (see [41] for a review). These observations suggest the structure of the ATPase core is not in a hydrolysis competent state until a suitable DNA substrate is present to promote the requisite conformational change. Previous studies showed SMARCAL1$^{CD}$ requires DNA with a minimal length of 15 base-paired dsDNA with a greater than 5 nucleotide single-stranded overhang [40]. By aligning the N-terminal lobe of the ATPase core of our structure with SNF2h and SSO1653 in their DNA-bound states we identified key features of the refined SMARCAL1 solution state structure that inform on how the ATPase core and DNA binding are coupled.

The AlphaFold structure predicts a compact state that is not conducive to DNA-binding with the ATP binding cleft of the motor domain preformed. However, SAXS-based refinement by Carbonara revealed an extended state that positions the HARP domain away from the ATPase core with the ATPase core opened such that the ATP binding site is not preformed (Figure 3). In this state, SMARCAL1$^{CD}$ is prevented from futile catalysis, as previously reported [42]. Furthermore, the HARP2 and ATPase domains are linked by a partially formed helix and previous studies have shown that binding of DNA substrates is independent of the ATPase core. This partially formed helix suggests a flexibility between the two domains that may be required for recognition of the varying branched DNA substrates. Our refined structure supports a previous SAXS study of SMARCAL1$^{CD}$ using a dsDNA substrate with a 10-nt overhang as the length of the overhang is within the 36 Å (centre-of-mass) from the ATPase core.

Comparison with DNA-bound structures of SNF2h and SSO1653 reveals difference in the position of the C-terminal lobes. In these related proteins, this domain makes direct contact with the double-stranded DNA substrate [42]. However, in our refined structure the C-terminal lobe adopts a different orientation that would require substantial repositioning in order to achieve similar DNA contacts. This observation aligns with previous SAXS studies that demonstrate that SMARCAL1 undergoes a major conformational change with ATP hydrolysis [42]. Unlike SNF2h and SSO1653, which readily interact with double-stranded

DNA, SMARCAL1 specifically requires the presence of single-stranded DNA for effective binding. We aligned the MD conformers produced from AlphaFold and Carbonara models to structures of SNF2h in complex with dsDNA (PDB: 1Z63) and a Holliday junction (PDB: 3CRX). We observed that, in the case of Carbonara models, the HARP2 and ATPase domains of SMARCAL1 are both favourably oriented with respect to their respective binding partners, ssDNA and dsDNA (see Figure 3). Thus, our structural analysis suggests a sequential mechanism for SMARCAL1's DNA interaction: initial recognition and binding of ssDNA by the HARP2 domain facilitates the positioning of the N-terminal lobe of the ATPase core on the dsDNA. Subsequent ATP hydrolysis can then drive a conformational change, translocating the C-terminal lobe of the ATPase core thereby promoting strand annealing.

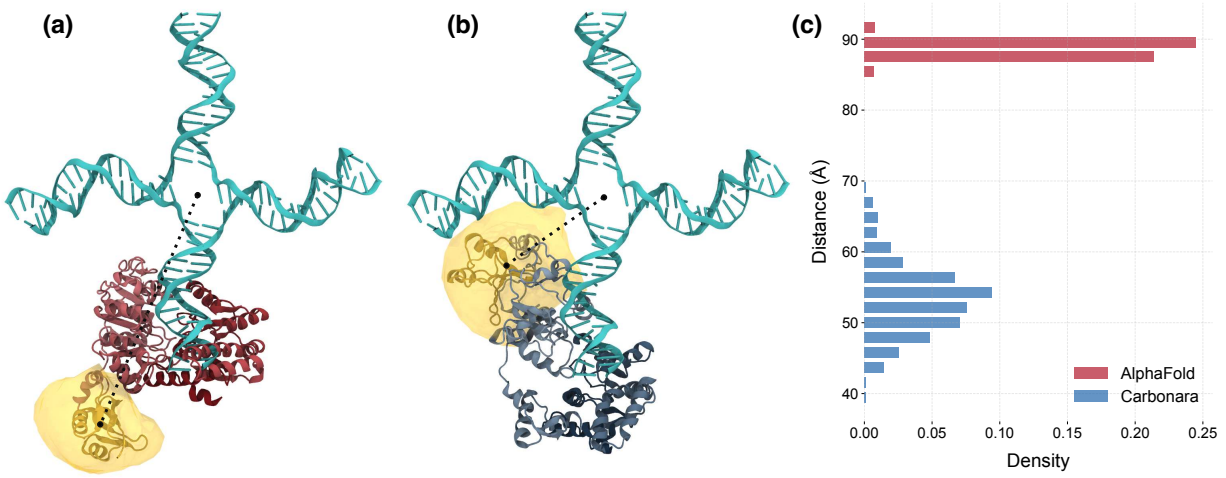



Figure 3: **Interaction of Human SMARCAL1 with DNA**. All SMARCAL1 conformations obtained from MD simulations seeded from AlphaFold (a) and Carbonara (b) models are aligned to the ATPase domain of SNF2 in complex with dsDNA, to which a Holliday junction is also aligned. In both panels, the protein conformer best matching SAXS data of apo SMARCAL1 is shown in a cartoon representation, while the region explored by the HARP2 domain in each simulation is shown as a transparent yellow surface. In (c), the distribution of distances between the centres of mass of HARP2 in MD simulations and the aligned Holliday junction is represented. Carbonara models demonstrate that in solution SMARCAL1's HARP2 domain explores states facilitating subsequent binding of the ATPase domain.

## 3.3 Refinement of a crystal structure using alternative disulfide bond constraints: ChiLob7/4 IgG2

The C239S variant of ChiLob7/4 IgG2 $F(ab)_2$, comprised of four protein chains, constitutes the two antigen-binding arms of an engineered immunostimulatory antibody with potential for cancer therapeutics. This variant is particularly notable for its capacity to activate signalling (agonism) for the key immune receptor, CD40 [43]. Its crystal structure (PDB: 6TKD, 1.90 Å resolution), shown in Figure 4b, fails to reflect the solution scattering, as reflected in a poor experimental fit $\chi^2 = 4.29$ with significant error in the low-q (large scale) portion of the scattering curve (Figure 4a, further comparisons are given in Table S1). This system features a distinct hinge region joining the two antigen-binding arms with disulfide bonds (indicated in the side panels of Figure 4b). Previous MD studies, initiated from crystal structures, failed to sample conformations with sufficiently high radii of gyration to match the corresponding experimental SAXS [43]. The restricted X-ray conformation was unable to open in a manner more reflective of the solution scattering data, a property found to be regulated in the antibody hinge region.

We applied Carbonara refinement with flexibility provided to the hinge region of this multi-domain system. The refinement was constrained to maintain the specific pattern of disulfide bonds which characterise the linking of the two F(ab) arms for the C239S variant. The intra-F(ab) structure described by the crystal structure was maintained by not allowing any flexibility in these domains (see Methods). Linker sections connecting the variable and constant regions were also trialled, but this led to undesirable unfolding of $\beta$-sheets within the variable region architecture. The Carbonara output is shown in Figure 4c (blue). The Carbonara-refined structure describes a more open hinge angle between the two F(ab) arms, with an increase from 116.4° in the starting model to 125.3°, yielding an improved agreement with the experimental data ($\chi^2 = 1.95$) and an increased radius of gyration from 36.6 Å to 38.9 Å (Table S1).

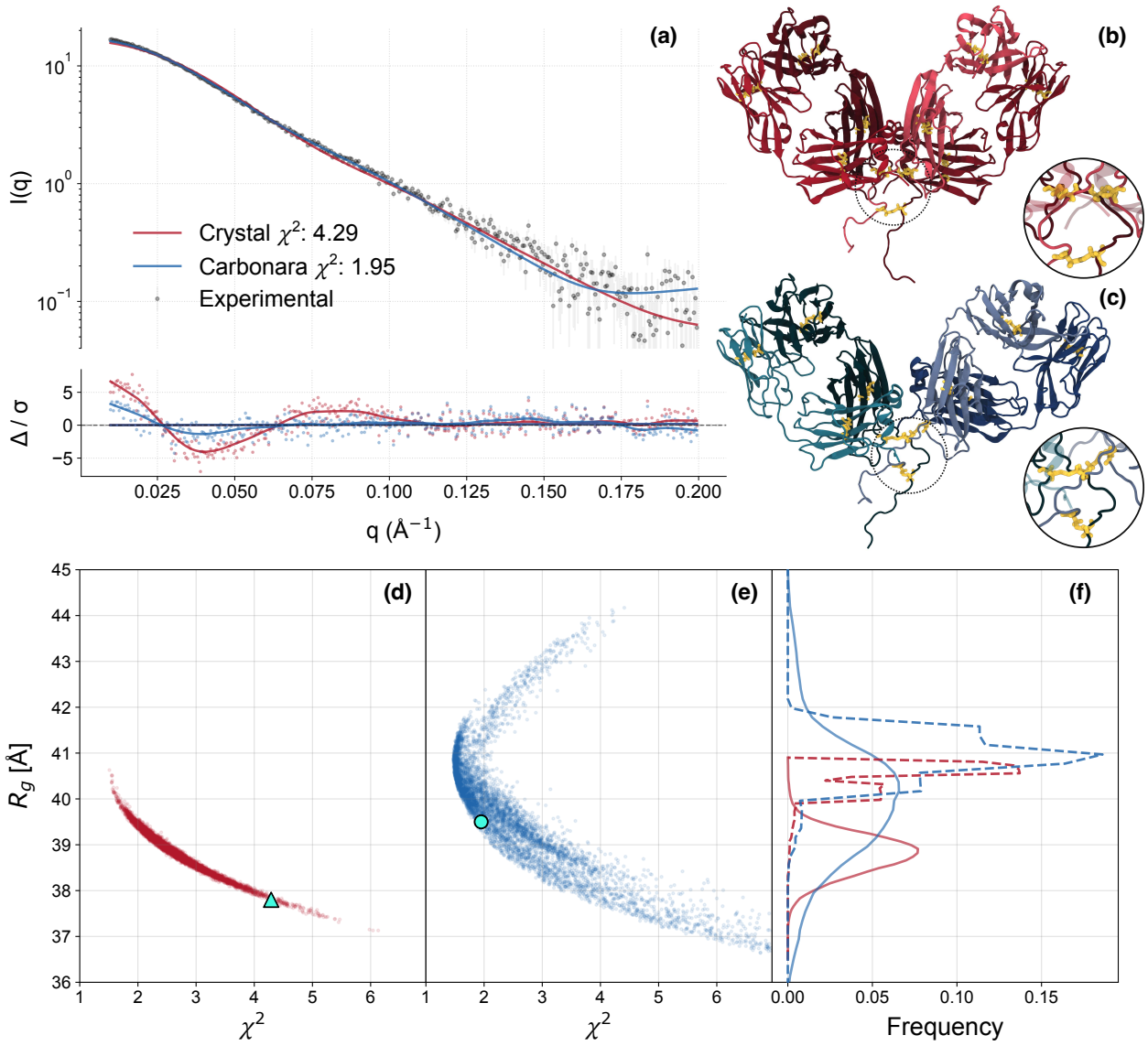



Figure 4: **Refinement of the crystal structure of the C239S IGg2 variant** (a) CRYSOL-computed SAXS of initial crystal structure (red) and the best fitting Carbonara refined structure (blue). The $\chi^2$ values are calculated for both; crystal $\chi^2$ = 4.29 and Carbonara $\chi^2$ = 1.95, up to $q_{max}$=0.2. Bottom, error weighted residuals. (b) Crystal structure of C239S IgG2 F(ab)$_2$ (red) (Protein Data Bank [PDB] ID: 6TKD). (c) Best-fitting all-atom model refined with Carbonara (blue). Hinge disulfide bonding patterns are highlighted in circles, disulfides shown in yellow. (d,e) $R_g$ and $\chi^2$ distributions for structures extracted from MD simulations seeded by the crystal structure (d, cyan triangle) and Carbonara-refined (e, cyan circle) structure. (f) $R_g$ density distributions generated by GAJOE. Red colour refers to the crystal-seeded MD ensemble, blue refers to the Carbonara seeded MD ensemble. Solid lines describe the $R_g$ distribution calculated from the MD pools without reweighting. Dotted lines represent the SAXS-optimised $R_g$ distributions selected by GAJOE.

To compare the MD conformational exploration capability of both the crystal structure and Carbonara-refined model, we performed three 2 $\mu$s-long simulations starting from each structure shown in Figure 4b and c. Following MD relaxation, the crystal structure and Carbonara output maintain comparable Ramachandran plots (Figure S5), while two of the three MD repeats rapidly settle on relatively stable low $\chi^2$ and high $R_g$ values (Figure S6). The MD sampling for the crystal structure remained confined to a relatively narrow $R_g$ range centred around 39 Å (see Figure 4f, solid red line). GAJOE re-weighting (see Methods) of this ensemble (dashed red line) shows a strong bias towards the most extended, and rarely sampled conformations with Rg values around 40.5 Å, indicating that the MD sampling lacks the extended states to accurately describe the in-solution conformations observed in SAXS. In contrast, MD sampling of the Carbonara-refined model readily explored a wider range of $R_g$ values, centred around 40 Å and extending as high as 44 Å (solid blue line). The GAJOE re-weighting (dashed blue line) selects conformations well within this sampled distribution, indicating that the simulations seeded from Carbonara models naturally access the conformational states needed to describe the solution scattering data.

The distribution of $\chi^2$ versus $R_g$ (Figure 4d,e) reveals that only 7.4% of the conformations extracted from the crystal structure simulations achieve $\chi^2 < 2$, compared to 39.7% for the Carbonara-initiated simulations. Even when initiating new simulations from the most extended conformations seen in the crystal structure trajectories (those with highest $R_g$ and lowest $\chi^2$ values), the resulting conformational exploration remained limited [43], failing to access the broader space explored by Carbonara-initiated simulations. This suggests the presence of significant energy barriers in the crystal structure that prevent adequate sampling of the solution-states. This demonstrates that Carbonara refinement of crystal structures can successfully generate starting models from which the full range of conformational states present in solution can be more efficiently sampled by MD simulations.

# 4 Discussion and Conclusions

Proteins function emerges from the Boltzmann-weighted ensemble of states they occupy in their native environment. X-ray crystallography provides a detailed potential starting point for investigating the solution state. However, while providing atomic-resolution, crystal structures can only capture a single protein conformation within a crystal lattice. AI-based methods for protein structural prediction, such as AlphaFold, are primarily trained on crystallographic data and therefore inherit this bias.

To specifically address this challenge of bridging crystal-like and solution-state conformations, we present Carbonara, a method that rapidly refines structures using SAXS data. By leveraging the atomic-level detail of existing structural models, whether from experimental techniques or computational predictions, and combining it with solution scattering data, Carbonara provides efficient access to the solution-state conformations that better reflect the biological reality.

Carbonara advances protein structure prediction via: (i) a coarse-grained sampling strategy that maintains secondary structure elements while exploring flexible regions; (ii) a topological constraint that ensures physically realistic protein folding; (iii) a progressive q-range fitting procedure that systematically refines structure from large-scale changes to local features; (iv) a backmapping protocol for generating all-atom models. The integration of these components creates a refinement pipeline that can efficiently identify conformations that bypass large energy barriers separating static predictions from solution-states, achieving in minutes what traditional MD simulations might not accomplish in weeks (see Figure S8).

This approach offers distinct advantages over existing methods for integrating SAXS data with structural prediction. Current approaches rely either on long MD simulations with posterior ensemble reweighting to match experimental data [44] or on directly incorporating SAXS as a biasing term in MD protocols [45]. To explore new protein conformational states, physics-based models such as MD simulations need to reproduce the process of the molecule

overcoming an energy barrier. When the energy barriers are high, the timescales associated with sampling a transition between states might make MD impractical. This limitation can be overcome by geometry-based conformational sampling methods such as Carbonara. While missing the atomic-level accuracy of an MD simulation, Carbonara can provide seeds where MD simulations can be initialised, bypassing the need to sample rare, high-energy transition events.

Our case studies demonstrate Carbonara's ability to capture large-scale conformational changes that are inaccessible with traditional methods. For SMARCAL1, AlphaFold yielded well-folded subdomains, but their relative arrangement was incompatible with SAXS experimental data. Long simulations initialised from this model remained confined to its immediate conformational neighbourhood. Carbonara refinement revealed an ensemble of more elongated models, featuring a substantial rearrangement of the HARP2 domain. MD simulations seeded from this ensemble sampled a conformational space distinct from that of the AlphaFold model. Besides being in good agreement with SAXS data, our models showed a possible mechanism by which SMARCAL1 can facilitate the remodelling of DNA structure and promote strand annealing. This result may provide insight into pathological conditions associated with SMARCAL1 interaction with DNA, such as cancer immune evasion [46].

The ChiLob7/4 case study demonstrated a complementary application of Carbonara, using a single refined structure rather than an ensemble approach. Starting from a crystallographic structure, Carbonara increased the $R_g$ in line with the SAXS by opening and twisting the structure, providing conformational changes beyond the reach of even microsecond-scale MD simulations. While MD simulations initiated from the crystal structure required extreme re-weighting to match SAXS data, indicating rare sampling of relevant states, simulations initiated from the Carbonara-refined structure naturally and frequently explored these solution-state conformations with minimal re-weighting needed. Owing to their flexibility and complex disulphide bonding patterns, especially in the hinge region, antibodies are chal-

lenging for both experimental structure determination and computational prediction. Our results show that Carbonara can identify an appropriate starting point for exploration of the physiologically relevant conformational space of an antibody, providing insights into the relationship between its rigidity and activity.

Understanding and predicting protein dynamics in solution remains a major challenge in structural biology and is essential for the next generation of AI-driven predictive methods [47]. Carbonara helps bridging the gap between crystal-like structural predictions and solution-state conformations. In this context, while primarily addressing the crystal-to-solution transition, the conformers it generates can serve as starting points for MD simulations to explore Boltzmann-weighted ensembles. It should be noted that, like all SAXS-based methods, Carbonara cannot guarantee the uniqueness of its solutions due to the inherent ambiguity of scattering data. Complementary experimental approaches could help reduce this ambiguity. FRET spectroscopy, for example, can provide precise distance constraints ($10\text{Å} - 100\text{Å}$) between specific residues [48], further narrowing the conformational search space. Similarly, X-ray footprinting, regularly collected at synchrotrons alongside SAXS, can offer residue-level insights into protein dynamics and solvent accessibility [49]. Including additional experimental constraints will enable Carbonara to produce conformations that even more closely reflect the states proteins take in solution.

# 5 Data and Software availability

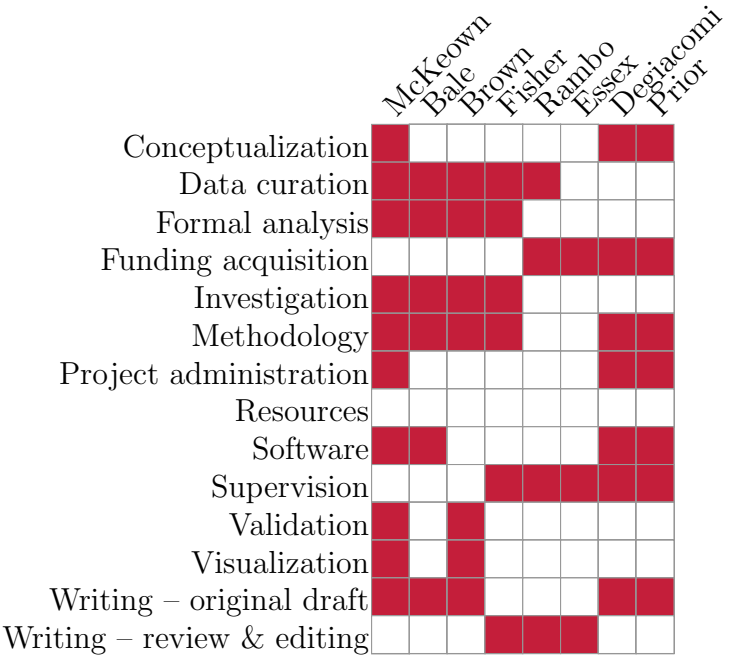
Carbonara is available open-source at: `https://github.com/Prior-Lab-Durham-University/carbonara`

# 6 Acknowledgements

JM acknowledges the EPSRC and the SOFI2 CDT (grant EP/S023631/1) for financial support. AB acknowledges an EPSRC MoSMed CDT (EP/S022791/1) and a Diamond Light

Source Ltd funded grant for financial support. CB acknowledges the EPSRC DTP and Diamond Light Source for financial support. HF acknowledges CRUK for financial support. This work used the ARCHER2 UK National Supercomputing Service (www.archer2.ac.uk) and the Durham HPC Hamilton and NCC. This project made use of time on HPC granted via the UK High-End Computing Consortium for Biomolecular Simulation, HECBioSim (http://hecbiosim.ac.uk), supported by EPSRC (grant no. EP/X035603/1).

# 7 Contributions




# 8 Methods

## 8.1 The Carbonara fitting protocol

The modelling pipeline builds upon, and extends, work presented in [35] and [36]. A brief summary of the validations performed in those studies can be found in Supplementary Methods S2. Here we outline their crucial elements and how they are used within our modelling

pipeline.

### 8.1.1 Explicit hydration layer model

To capture the hydration layer effect, crucial to match SAXS data, solvent molecules are placed on a cylindrical shell surrounding each section of the molecule. This layer is then checked for overlap of the solvent and the backbone, as well as likely shared solvent positions and any redundancy or forbidden overlap is removed until the final shell is created. An example is shown in Figure S2.

### 8.1.2 The scattering model

Within Carbonara we use the Debye scattering formula [50],

$$I(q) = \sum_{i=1}^{N} \sum_{j=1}^{N} f_i(q) f_j(q) \frac{\sin(q r_{ij})}{q r_{ij}}, \tag{1}$$

for the scattering intensity $I(q)$ as a function of momentum transfer $q = \pi \sin(\theta)/\lambda$. Here $N$ is total number of C$\alpha$ 's and solvent molecules, $f_i(q)$ the form factor for residue $i$ and $r_{ij}$ is the distance between points i and j. The amino acid form factors take the standard form

$$f_{am}(q) = f_b(q) - \rho_{ex} f_{ex}(q). \tag{2}$$

Here $f_b$ is the scattering form factor of the amino acid in a vacuum and $f_{ex}$ is a scattering form factor adjustment due to the excluded volume of solvent, with $\rho_x \in [0.95, 1.12]$ a commonly use scaling parameter [26, 24]. Each amino acid is assigned the same scattering function $f_b(q)$, a five-factor exponential representation

$$f_b = \sum_{i=1}^{5} A_i \mathrm{e}^{-B_i q^2} + C, \tag{3}$$

where $\{A_i, B_i\}_{i=1}^5$ and $C$ (a standard form used to fit molecular form factors [51]). These constants were are empirically determined in [35] using the scattering data of a number of known structures and their BioSAXS data. The excluded volume effect is captured using an exponential model in the form

$$f_{ex}^a(r_w, q) = v(r_w)\mathrm{e}^{\pi q^2 v(r_w)^{3/2}}, \quad v(r_w) = \frac{4\pi}{3} r_w^3, \tag{4}$$

where $r_w$ is the average atomic radius of the atom; this is also a standard form, see *e.g.* [26, 52, 24]. The parameter $\rho_{ex}$, the relative weighting of the excluded volume effect was also determined empirically in [35] using a similar constraint range as in [26, 52, 24].

The scattering form for an individual water molecule in the hydration layer is

$$f_h(q) = \rho_h(2f_{hy}(q) + f_{ox}(q)), \tag{5}$$

where $f_{hy}$ and $f_{ox}$ are the vacuum scattering of hydrogen and oxygen respectively [51]. The constant $\rho_h$ is varied over the same range $\rho_h \in [-2, 4]$, as in the FoXS package [31].

### 8.1.3 Fitting to experimental data

We use a mean squared distance metric $\chi_f^2$ to assess the fit of the model to the BioSAXS data:

$$\begin{aligned}
\chi_f^2 &= \frac{1}{n_s - 1} \sum_{i=1}^{n_s} \left[\log(I_m(q_i)) - \log(I_e^s(q_i)) - L_d\right]^2, \\
L_d &= \frac{1}{n_s} \sum_{i=1}^{n_s} \log(I_m(q_i)) - \log(I_e^s(q_i)). \\
n_s &= \frac{2(q_{min} - q_{max})D_{max}}{\pi}.
\end{aligned} \tag{6}$$

Where $I_m(q)$ is the model scattering calculated using the Debye formula (1), and $I_e^s(q)$ the smoothed experimental data. The values of $I_e^s(q)$ are obtained using the Shannon sampling

theorem to bin the data into $n_s$ bins and then taking the median value from each bin, a procedure described in [37, 53, 29]; we use twice the number of bins as suggested by the Shannon sampling theorem as advocated in [29], while the value $D_{max}$ is the maximum distance of any pair of coordinates in the model (C$\alpha$ coordinates or solvent coordinates). The factor $L_d$, which will superimpose identical curves which differ by a translation is used because the protein concentration can only be measured with relatively low accuracy [26, 24] (when taking a logarithm of the data a scaling factor becomes a vertical translation). Note that unlike for CRYSOL, FoXS, and a number of other packages, the metric $\chi^2_f$ is not a true $\chi^2$ measure as the noise inherent to the data is smoothed using the Shannon sampling principle before fitting, as advocated in [37]. Thus, an ideal score is 0.

#### 8.1.4 Overlap prevention measure $\chi_o$

To prevent physically unreasonable conformations, a penalty $\chi_o$ is applied if the C$\alpha$-C$\alpha$ distance of $\leq 4$Å is found for for any pair of non-adjacent C$\alpha$ positions. Previously a harmonic potential was used in [35], but in this work the penalty takes the following form:

$$\chi_o = \frac{0.1}{m} \sum_{i=1}^{m} (\mathrm{e}^{4.0-d_i} - 1) \tag{7}$$

where $d_i$ is any of $m$ non-adjacent C$\alpha$ coordinates which are less than 4.0Å, providing a similar penalty for distances close to 4Å to the harmonic potential, but a more dramatic repulsive effect if smaller distances are sampled.

#### 8.1.5 Custom distance restraints $\chi_d$

If the user has some additional knowledge of their structure, e.g., contact predictions or disulfide bonds, then Carbonara can constrain these distances. The restraints are placed on the C$\alpha$ coordinates. If there are $n_p$ pairs of C$\alpha$ coordinates, with idealised distances $d_i^0$, then

the following penalty is used to restrain these values:

$$\chi_d = 0.0001 \sum_{i=1}^{n_p} d_f^4 \quad d_f = \frac{1}{w_0} \frac{d_i - d_i^0}{d_i^0}. \tag{8}$$

Here $d_i$ is the distance between pair $i$. If $w_0 = 1$ the quantity $d_f$ represents the discrepancy from this ideal value $d_i^0$ as a fraction of $d_i^0$. The pre-multiplier is the value at which the scattering penalty $\chi_f$ is considered a good fit. Altering the value of $w_0$ dictates the value of $d_f$ at which $\chi_d = 0.0001$, a value which is equal to a high quality scattering fit (*i.e.* $\chi_f^2 = 0.0001$). The standard value used in Carbonara is $w_0 = 0.5$.

### 8.1.6 Topological fold metric $\chi_t$

The absolute crossing number *acn*, a measure of the self entanglement of a suitably smoothed protein backbone, has a strict lower bound on its value as a function of its length for protein structures in the PDB [36] (Figure S7). Enforcing this constraint during the Carbonara search prevents the method from predicting unrealistically unfolded structures. To enforce this constraint Carbonara calculates *acn* of a smoothed representation of the structure and imposed a step function-like penalty $\chi_t$ defined as:

$$\chi_t = \frac{1}{1 + \exp\left(20\left(acn - l_b\right)\right)} \tag{9}$$

where $l_b$ is the empirically obtained lower bound of the protein (a function of its length, see Supplementary Methods 3). If the current structure's $acn > l_b$ the structure is considered sufficiently tangled and this term is exponentially small. If $acn < l_b$ then the structure is considered too open, and $\chi_t \approx 1/2$, a value which is several orders of magnitude larger than $\chi_f$ for a reasonable fit to the scattering data. Thus, in practice, it acts solely to prevent the structure from becoming too unfolded.

### 8.1.7 Fitting models to SAXS data

For a given range $q \in [q_{min}, q_{max}]$ the aim is to minimise the sum $\chi = \chi_f^2 + \chi_o + \chi_d + \chi_t$ (if no distance restraints are specified $\chi_d$ is trivially 0). A Monte Carlo method is used in which the algorithm loops over each designated flexible section. For each section a Monte Carlo algorithm is used to generate a new geometry (see Supplementary Methods 1). The value of $\chi$ is calculated and if it is less than the current minimum then the new configuration is accepted, else it is rejected.

If a value of $\chi_f^2 < 0.0001$ is reached, then the algorithm extends the range $q \in [q_{min}, q_{max}]$ being fitted by adding a value of 0.1 to $q_{max}$. New configurations are only accepted if the full penalty $\chi < 0.0001$ but it is possible that a particular prediction can have a small penalty in either the overlap or distance restraints so extra data being included is based on the scattering fit alone. The form of both $\chi_o$ and $\chi_d$ are selected such that these errors will lead to either only slightly close C$\alpha$ - C$\alpha$ distances (from $\chi_o$) or minor variations from the idealised distance restraints (from $\chi_d$).

As GAJOE (see below) is only compatible with CRYSOL, for consistency we evaluated the $\chi^2$ between experimental SAXS and every all-atom model produced via the backmapping procedure using CRYSOL [26].

## 8.2 Model generation

### 8.2.1 SMARCAL1

We analysed a truncated structure of the Human SMARCAL1 protein, spanning residues 325 to 870, which contains both the HARP2 and ATPase domains [40]. We produced an atomistic model with AlphaFold3 [4], and identified its flexible regions using Carbonara's automated procedure (see details in Supplementary Methods S4).

The model was then refined with Carbonara's progressive fitting approach beginning with the low-q interval ($q \leq 0.1\text{Å}^{-1}$) of the experimental SAXS data. We ran 200 distinct fitting

runs, each starting from the initial AlphaFold model. Each structure for which $\chi_f^2 < 0.001$ was used to generate a pool of all-atom predictions. These models were screened using CRYSOL to calculate their $\chi^2$ over the interval $q = [0.02, 0.22]$ against experimental data. This optimal q-range was determined using SHANUM [54], giving the maximum useful data range. For this dataset, the analysis indicated data above $q = 0.22\text{Å}^{-1}$ does not contribute additional structure above the level of noise. 30 models with a $\chi^2$ less than 1.6 were selected for further analysis with MD simulations.

For the Carbonara-refined model with $\chi^2$ closest to 1, comprehensive validation was performed using multiple SAXS profile calculation methods. Supplemental Table 1 provides detailed comparisons between the AlphaFold and Carbonara models using CRYSOL (shown in the main text), as well as additional validations with FoXS and WAXSiS, demonstrating significant improvement across all evaluation methods.

### 8.2.2 ChiLob 7/4 IgG2

The crystal structure of ChiLob 7/4 IgG2 (PDB: 6TKD) and SAXS data (SASBDB: SASDLJ4) of its C239S variant are reported in [43]. We identified linker regions using Carbonara's automated procedure, and manually selected as flexible those at the hinge region of the structure. Since the C239S variant of ChiLob 7/4 is a multi-chained protein, Carbonara also controlled the roto-translational degrees of freedom of each chain. To ensure the two F(ab) arms were held together, the C-$\alpha$ atoms of opposing disulfide bonding cysteine residues in the hinge region are constrained at 6 Å by Carbonara. The specific amino acid pairs subjected to disulfide bond restraints were $136-680$, $149-205$, $225-591$, $231-686$, $264-329$, $375-435$, $477-551$, $604-660$, $719-784$, and $830-890$. This distance constraint is particularly significant for disulfide bonds found within the hinge region, which confer the C239S variant its distinctive physiological properties [43].

The crystal structure was then refined with Carbonara's progressive fitting approach (see Theory), beginning with the low-q interval ($q \leq 0.1\text{Å}^{-1}$) of the experimental SAXS data.

The best fitting Carbonara model was selected and used to produce an all-atom model. To ensure robust validation, the refined structure was evaluated against experimental SAXS data using multiple independent methods as detailed in Supplemental table 2, including CRYSOL (presented in the main text), FoXS, and WAXSiS, all confirming improvement over the crystal structure.

## 8.3 Molecular Dynamics simulations

### 8.3.1 SMARCAL1

MD simulations were performed using GROMACS 2023.4 [55], employing the CHARMM36m force field [56] along with the TIP3P(m) water models. These force fields were chosen for their improved accuracy in simulating non-globular protein conformations [56], making them suitable for validating the more extended conformations suggested by SAXS.

The protein was solvated with TIP3P water molecules in a dodecahedron box with a minimum distance of 1.0 nm between the protein surface and the edge of the box. A physiological 0.15 M concentration on $Na^+$ and $Cl^-$ ions was used to charge balance the simulation box. Energy minimisation was performed via steepest descent with a step size of 0.01 nm, continued until the maximum force fell below 1000 kJ/mol/nm or 50,000 steps were reached. The minimisation used a Verlet scheme with a cutoff of 1.2 nm for both van der Waals and Coulomb interactions. Energy minimisation was followed with a two-phase equilibrium process. First an NVT equilibration was performed for 150 ps with a 1 fs timestep, the temperature was maintained at 303.15 K using the V-scale thermostat with a coupling constant of 1.0 ps and position restraints were applied to the backbone atoms. Following this, an NPT equilibration was performed for 100 ps, maintaining the same temperature while coupling the pressure to 1 bar using the Parrinello-Rahman barostat, with a coupling constant of 5.0 ps.

### 8.3.2 ChiLob7/4 IgG2

MD simulations were performed using GROMACS v2021.2, employing the Amberff14SB force field [57] for protein atoms and Joung-Cheatham parameters for ions [58]. These parameters were selected to maintain parity with previous studies of the C239S IgG2 variant, specifically those used in simulations of the crystal structure [43]. The protein structure was prepared using Protein Preparation Wizard [59], capping chain termini with ACE and NMA residues and protonating at pH 7.0, followed by a restrained minimisation. The system was solvated in a TIP3P water box with a 2.5 nm buffer between the protein and box edges. System neutralisation was achieved by addition of $Na^+$ and $Cl^-$ ions to reach a physiological concentration of 0.15 mM. Energy minimisation proceeded in two phases: 1000 steps of steepest descent followed by 1500 steps of conjugate gradient protocol. The system was then equilibrated in two phases: first, a 50 ps NVT equilibration heating to 300 K, followed by a 100 ps NPT equilibration at 1 bar. Temperature was maintained using a Nose-Hoover thermostat [60] with a 2 $ps^{-1}$ collision frequency, while pressure was controlled with a Parinello-Rahman barostat [61] during equilibration using a 2.0 ps relaxation time. Production simulations used periodic boundary conditions with an 8 Å direct space cut-off for non-bonded interactions. Long-range electrostatics were handled by Particle Mesh Ewald summation [62], while long-range van der Waals interactions used an analytical dispersion correction. LINCS constraints [63] were applied to all bonds involving hydrogen, allowing a 2 fs timestep. Pressure was maintained at 1 bar with a Parinello-Rahman barostat throughout production simulations. Three independent production repeats were performed using distinct random seeds for the thermostat.

## 8.4 Simulation analysis

### 8.4.1 SMARCAL1

For all MD simulation of AlphaFold and Carbonara-refined models, we extracted snapshots at 1 ns intervals, generating a total of 12,000 structures. We concatenated and aligned all these trajectories on their C$\alpha$ coordinates (residues 392-432), and then performed a Principal Component Analysis (PCA) on the C$\alpha$ Cartesian coordinates.

A subset of 55 conserved residues in the N-terminal lobe of the ATPase core domain from both AlphaFold and Carbonara-refined SMARCAL1 model were structurally aligned to the corresponding domain of SNF2h in complex with dsDNA (PDB: 1Z63) using a custom Python. A Holliday junction structure (PDB: 3CRX) was then aligned to the dsDNA in the SNF2h complex by superimposing the C1' atoms on the DNA were aligned (Holliday: chain C, residues 20-35, chain F residues 3-18; SNF2h chain C residues 9-24 and chain D residues 3-18). The alignment was repeated for all 12,000 conformations produced in the simulations of AlphaFold and Carbonara models. The HARP2 domain conformational space was visualised using VMD's VolMap tool (residues 325-398, resolution 2 Å, atom size 1 Å, weighted by mass density) and rendered with an isosurface value of 0.025, a step size of 3 (solid surface) and step 1 (isosurface).

For each aligned conformation, we measured the distance between the centres of mass of the HARP2 domain and the Holliday junction (defined as the centre of mass for DNA strands C, D, E, and F, residues 17-20).

### 8.4.2 ChiLob7/4 IgG2

Structural snapshots were extracted every 1 ns from simulation trajectories, generating a pool of 6,000 structures. Terminal caps were removed and titratable residues renamed to match CRYSOL naming conventions. Radii of gyration, $R_g$, and $\chi^2$ values were calculated using CRYSOL (ATSAS 3.0.4) [64], with $R_g$ values taken from the slope of net intensity.

When applying ensemble reweighting, conformational ensembles were selected from the MD-derived pool using the genetic algorithm implemented in the GAJOE software (ATSAS 3.0.4) [65]. Ensembles were selected to reduce the difference between the experimental and computed SAXS intensity of the ensemble, calculated as an average of the summed individual synthetic scattering intensities, and the experimental scattering. As GAJOE is not deterministic, to ensure convergence, GAJOE was executed independently 10 times using default settings with $q_{max}$=0.2, using the crystal-seeded and Carbonara-seeded MD pools as inputs. The resulting $R_g$ frequency distributions were found to be consistent across all GAJOE runs.

## 8.5 Visualisation

All plots were generated in Python (3.11) with the matplotlib (3.8.2) library. Molecular visualisations were produced with VMD 1.9.4 [66].

# References


1. Maveyraud, L. & Mourey, L. Protein X-ray Crystallography and Drug Discovery. *Molecules* **25,** 1030 (2020).
2. Takeuchi, K., Baskaran, K. & Arthanari, H. Structure determination using solution NMR: Is it worth the effort? *Journal of Magnetic Resonance* **306,** 195–201 (2019).
3. Saibil, H. R. Cryo-EM in molecular and cellular biology. *Molecular Cell* **82,** 274–284 (2022).
4. Ruff, K. M. & Pappu, R. V. AlphaFold and implications for intrinsically disordered proteins. *Journal of Molecular Biology* **433,** 167208 (2021).
5. Baek, M., DiMaio, F., et al. & Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373,** 871–876 (2021).
6. Marino-Buslje, C., Monzon, A. M., Zea, D. J., Fornasari, M. S. & Parisi, G. On the dynamical incompleteness of the Protein Data Bank. *Briefings in Bioinformatics* **20,** 356–359 (2017).
7. DeForte, S. & Uversky, V. N. Resolving the ambiguity: Making sense of intrinsic disorder when PDB structures disagree. *Protein Science* **25,** 676–688 (2016).
8. Terwilliger, T. C. *et al.* AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods* **21,** 110–116 (2024).
9. Saldaño, T. *et al.* Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics* **38,** 2742–2748 (2022).
10. Jing, B., Berger, B. & Jaakkola, T. *AlphaFold Meets Flow Matching for Generating Protein Ensembles* 2024. arXiv: `2402.04845 [q-bio.BM]`. `https://arxiv.org/abs/2402.04845`.
11. Jin, Y. *et al.* P2DFlow: A Protein Ensemble Generative Model with SE(3) Flow Matching. *Journal of Chemical Theory and Computation* **21,** 3288–3296 (2025).
12. Hub, J. S. Interpreting solution X-ray scattering data using molecular simulations. *Current Opinion in Structural Biology* **49,** 18–26 (2018).
13. Svergun, D. I., Koch, M. H., Timmins, P. A. & May, R. P. *Small angle X-ray and neutron scattering from solutions of biological macromolecules* (Oxford University Press, 2013).

14. Delhommel, F., Martínez-Lumbreras, S. & Sattler, M. Combining NMR, SAXS and SANS to characterize the structure and dynamics of protein complexes. *Methods in Enzymology* **678,** 263–297 (2023).

15. Vijayakrishnan, S. *et al.* Solution structure and characterisation of the human pyruvate dehydrogenase complex core assembly. *Journal of molecular biology* **399,** 71–93 (2010).

16. Kim, J. S., Afsari, B. & Chirikjian, G. S. Cross-validation of data compatibility between small angle X-ray scattering and cryo-electron microscopy. *Journal of Computational Biology* **24,** 13–30 (2017).

17. Choi, K. H. & Morais, M. Use of small-angle X-ray scattering to investigate the structure and function of dengue virus NS3 and NS5. *Dengue: Methods and Protocols,* 241–252 (2014).

18. Yang, Y. *et al.* SAXS characterization of the interactions among digested food compounds and the anti-oxidant and anti-inflammatory activities of the formed nanocomplexes. *Food & function* **9,** 3408–3418 (2018).

19. Butt, L. E., Holland, R. A., Khunti, N. S., Quinn, D. L. & Pickford, A. R. Using Small Angle X-Ray Scattering (SAXS) to Characterize the Solution Conformation and Flexibility of Matrix Metalloproteinases (MMPs). *Matrix metalloproteases: methods and protocols,* 87–108 (2017).

20. Mina, J. G. *et al.* Functional and phylogenetic evidence of a bacterial origin for the first enzyme in sphingolipid biosynthesis in a phylum of eukaryotic protozoan parasites. *Journal of Biological Chemistry* **292,** 12208–12219 (2017).

21. Brookes, E., Rocco, M., Vachette, P. & Trewhella, J. AlphaFold-predicted protein structures and small-angle X-ray scattering: insights from an extended examination of selected data in the Small-Angle Scattering Biological Data Bank. *Journal of Applied Crystallography* **56,** 910–926 (2023).

22. Brookes, E. & Rocco, M. A database of calculated solution parameters for the AlphaFold predicted protein structures. *Scientific Reports* **12,** 7349 (2022).

23. Anami, Y. *et al.* Apo- and Antagonist-Binding Structures of Vitamin D Receptor Ligand-Binding Domain Revealed by Hybrid Approach Combining Small-Angle X-ray Scattering and Molecular Dynamics. *Journal of Medicinal Chemistry* **59,** 7888–7900 (2016).

24. Schneidman-Duhovny, D., Hammel, M. & Sali, A. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic acids research* **38,** W540–W544 (2010).

25. Gorba, C. & Tama, F. Normal mode flexible fitting of high-resolution structures of biological molecules toward SAXS data. *Bioinformatics and biology insights* **4,** BBI–S4551 (2010).

26. Svergun, D., Barberato, C. & Koch, M. H. CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of applied crystallography* **28,** 768–773 (1995).

27. Wright, D. W. & Perkins, S. J. SCT: a suite of programs for comparing atomistic models with small-angle scattering data. *Journal of applied crystallography* **48,** 953–961 (2015).

28. Poitevin, F., Orland, H., Doniach, S., Koehl, P. & Delarue, M. AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformally hydrated atomic models. *Nucleic acids research,* gkr430 (2011).

29. 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 research* **43,** W225–W230 (2015).

30. Grudinin, S., Garkavenko, M. & Kazennov, A. Pepsi-SAXS: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Crystallographica Section D: Structural Biology* **73,** 449–464 (2017).

31. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Research* **44,** W424–W429 (2016).

32. Huggins, D. J. *et al.* Biomolecular simulations: From dynamics and mechanisms to computational assays of biological activity. *WIREs Computational Molecular Science* **9,** e1393 (2019).

33. Franke, D. & Svergun, D. I. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *Journal of Applied Crystallography* **42,** 342–346 (2009).

34. Grant, T. D. Ab initio electron density determination directly from solution scattering data. *Nature Methods* **15,** 191–193 (2018).

35. Prior, C., Davies, O. R., Bruce, D. & Pohl, E. Obtaining tertiary protein structures by the ab initio interpretation of small angle X-ray scattering data. *Journal of chemical theory and computation* **16,** 1985–2001 (2020).

689 36. Bale, A., Rambo, R. & Prior, C. The SKMT algorithm: a method for assessing and comparing
690 underlying protein entanglement. *PLoS computational biology* **19,** 1–27 (2023).
691 37. Rambo, R. P. & Tainer, J. A. Accurate assessment of mass, models and resolution by small-
692 angle scattering. *Nature* **496,** 477–481 (2013).
693 38. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Current*
694 *Protocols in Bioinformatics* **54,** 5.6.1–5.6.37 (2016).
695 39. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints.
696 *Journal of Molecular Biology* **234,** 779–815 (1993).
697 40. Mason, A. C. *et al.* A structure-specific nucleic acid-binding domain conserved among DNA
698 repair proteins. *Proceedings of the National Academy of Sciences* **111,** 7618–7623 (2014).
699 41. Dürr, H., Flaus, A., Owen-Hughes, T. & Hopfner, K.-P. Snf2 family ATPases and DExx box
700 helicases: differences and unifying concepts from high-resolution crystal structures. *Nucleic*
701 *acids research* **34,** 4160–4167 (2006).
702 42. Bétous, R. *et al.* SMARCAL1 catalyzes fork regression and Holliday junction migration to
703 maintain genome stability during DNA replication. *Genes & development* **26,** 151–162 (2012).
704 43. Orr, C. M. *et al.* Hinge disulfides in human IgG2 CD40 antibodies modulate receptor signaling
705 by regulation of conformation and flexibility. *Science immunology* **7,** eabm3723 (2022).
706 44. Boldon, L., Laliberte, F. & Liu, L. Review of the fundamental theories behind small angle
707 X-ray scattering, molecular dynamics simulations, and relevant integrated application. *Nano*
708 *reviews* **6,** 25661 (2015).
709 45. Zhao, C. & Shukla, D. SAXS-guided Enhanced Unbiased Sampling for Structure Determina-
710 tion of Proteins and Complexes. *Scientific Reports* **8,** 17748 (2018).
711 46. Nagai, T. H. *et al.* Chromatin regulator SMARCAL1 modulates cellular lipid metabolism.
712 *Communications Biology* **6,** 1298 (2023).
713 47. Reardon, S. Five protein-design questions that still challenge AI. *Nature* **635,** 246–248 (Nov.
714 2024).
715 48. Sekar, R. & Periasamy, A. Fluorescence Resonance Energy Transfer (FRET) Microscopy Imag-
716 ing of Live Cell Protein Localizations. *The Journal of cell biology* **160,** 629–33 (Apr. 2003).

49. Gupta, S., Feng, J., Chan, L.-J., Petzold, C. & Ralston, C. Synchrotron X-ray footprinting as a method to visualize water in proteins. *Journal of Synchrotron Radiation* **23,** 1056–1069 (July 2016).

50. Debye, P. Zerstreuung von röntgenstrahlen. *Annalen der Physik* **351,** 809–823 (1915).

51. Brown, P. J., Fox, A. G., Maslen, E. N., O'Keefe, M. A. & Willis, B. T. M. in *International Tables for Crystallography* (eds Fuess, H. *et al.*) 554–595 (2006).

52. Svergun, D. I., Petoukhov, M. V. & Koch, M. H. Determination of domain structure of proteins from X-ray solution scattering. *Biophysical journal* **80,** 2946–2953 (2001).

53. Konarev, P. V. & Svergun, D. I. A posteriori determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems. *IUCrJ* **2,** 352–360 (2015).

54. Konarev, P. V. & Svergun, D. I. *A posteriori* determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems. *IUCrJ* **2,** 352–360 (2015).

55. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2,** 19–25 (2015).

56. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14,** 71–73 (2017).

57. A. Maier†, J. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **11,** 3696–3713 (8 2015).

58. Joung, I. S. & Cheatham III, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *The Journal of Physical Chemistry B* **112,** 9020–9041 (30 2008).

59. Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design* **27,** 221–234 (2013).

60. Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics* **52,** 255–268 (2 1983).

61. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **51,** 7182–7190 (12 1981).

62. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *The Journal of Chemistry Physics* **98,** 10089–10092 (12 1993).

63. Hess, B., Bekker, H., J. C. Berendsen, H. & G. E. M. Fraaije, J. LINCS: A Linear Constraint Solver for molecular simulations. *Journal of Computational Chemistry* **18,** 1463–1472 (51 1998).

64. Manalastas-Cantos, K. *et al.* ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *Journal of Applied Crystallography* **54,** 343–355 (2021).

65. Tria, G., Mertens, H. D., Kachala, M. & Svergun, D. I. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* **2,** 207–217 (2015).

66. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *Journal of molecular graphics* **14,** 33–38 (1996).

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- carbonaraappendix.pdf