THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

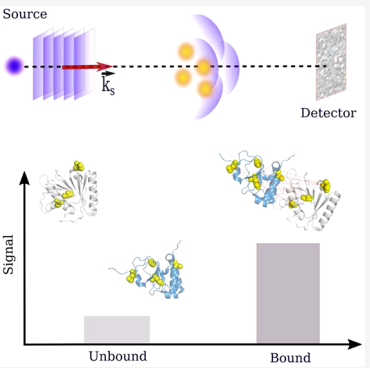
pubs.acs.org/JPCL
Letter

High-Throughput Interactome Determination via Sulfur Anomalous Scattering

Mattia Miotto,* Edoardo Milanetti, Riccardo Mincigrucci, Claudio Masciovecchio, and Giancarlo Ruocco

Cite This: J. Phys. Chem. Lett. 2024, 15, 3478−3485

ABSTRACT: We propose a novel approach for detecting the binding between proteins making use of the anomalous diffraction of natively present heavy elements, e.g., sulfurs, inside molecular three-dimensional structures. In particular, we analytically and numerically show that the diffraction patterns produced by the anomalous scattering of the sulfur atoms in a given direction depend additively on the relative distances between all couples of sulfur atoms. Thus, the differences in the patterns produced by bound proteins with respect to their nonbonded states can be exploited to rapidly assess protein complex formation. On the basis of our results, we suggest a possible experimental procedure for detecting protein−protein binding. Overall, the completely label-free and rapid method we propose may be readily extended to probe interactions on a large scale, thus paving the way for the development of a novel field of research based on a synchrotron light source.



Protein−protein interactions play a crucial role in various biological processes, including signal transduction, enzymatic regulation, and molecular recognition.[1] Understanding the mechanisms and dynamics of these interactions is essential for elucidating cellular processes and developing therapeutic interventions. Given the importance of the knowledge of protein−protein interactions (PPIs), several experimental techniques have been developed in the past several decades on the basis of biochemical and/or biophysical methods. The former include co-immunoprecipitation, bimolecular fluorescence complementation (BiFC),[2] phage display,[3] tandem affinity purification (TAP),[4−6] and a proximity ligation assay (PLA).[7,8] Techniques based on biophysical processes comprise surface plasmon resonance (SPR), dual polarization interferometry (DPI), flow-induced dispersion analysis (FIDA), fluorescence resonance energy transfer (FRET), and biolayer interferometry (BLI).[9,10]

Each of them offers different insights into the nature and characteristics of the interactions.[11] However, to date, a label-free technique that can rapidly assess whether and where two proteins bind has not been available. To obtain detailed information about the protein−protein complex structure, we rely on X-ray, nuclear magnetic resonance, and cryo-electron microscopy experiments to provide the spatial configuration of the complex up to a certain resolution.[12] Unfortunately, these experimental methodologies are time-consuming and strongly depend on the kinds of protein complexes and the experimental conditions.[13,14] For this reason, the development of new, fast, structure-driven experimental techniques for assessing protein−protein binding is of paramount importance, especially in the era of artificial intelligence aimed at predicting the three-dimensional conformations of protein structures (see, for instance, the recently developed methods of AlphaFold2[15] and RoseTTA fold[16]). Experiments will be required to test huge amounts of computationally predicted interactions as well as to increase the level of training and, therefore, the performance of all those emerging predictive methods based on machine learning data.[15]

In this Letter, we suggest using anomalous scattering from native sulfur atoms to rapidly assess whether a couple of proteins form stable complexes. Moreover, depending on the number of native sulfur atoms, we expect that combining the information coming from interference patterns with the knowledge of the two unbounded protein structures could provide insights into the binding region.

In this respect, sulfur anomalous scattering has already been applied in the context of protein structure determination. It became a valuable tool in X-ray crystallography for studying protein structure and addressing phase ambiguity or dephasing problems during structure determination.[17] In X-ray crystallography, knowing the phases of diffracted X-rays is crucial for determining the electron density of a protein crystal, which subsequently reveals the protein's atomic structure. However,

Received: December 29, 2023
Revised: February 16, 2024
Accepted: February 22, 2024
Published: March 21, 2024

https://doi.org/10.1021/acs.jpclett.3c03632
J. Phys. Chem. Lett. 2024, 15, 3478−3485

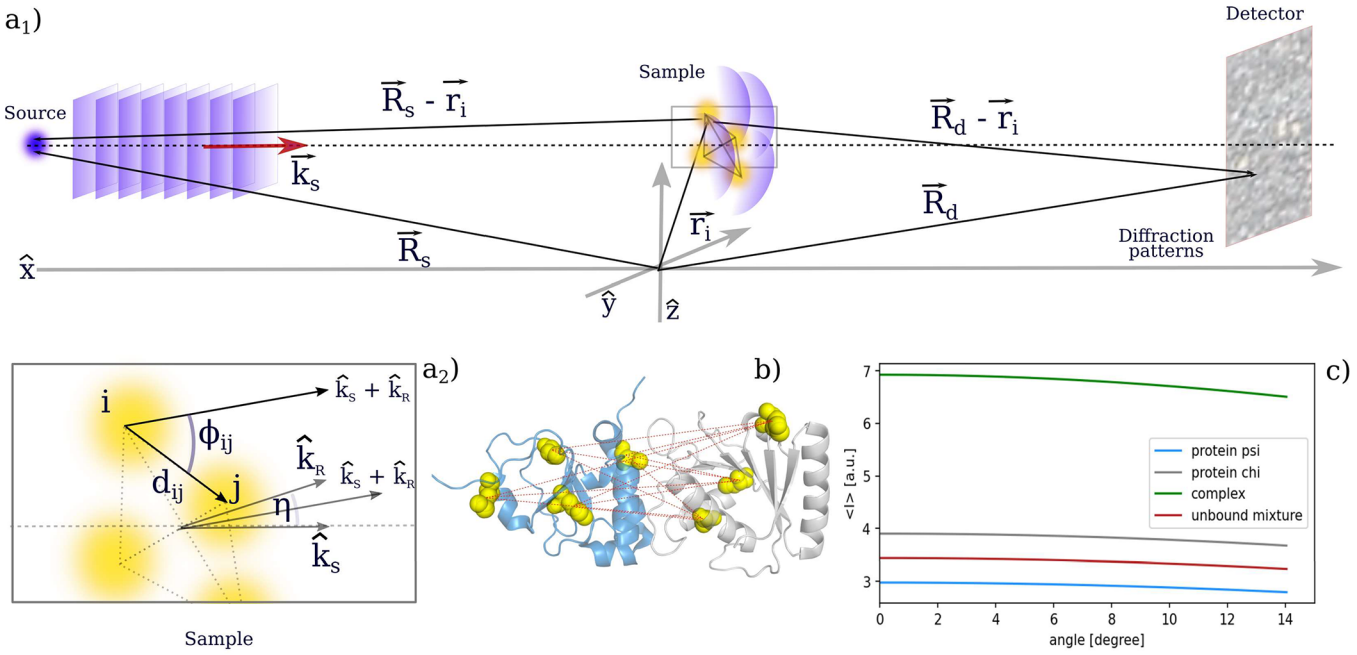


Figure 1. Scheme of the scattering process and case-of-study results. (a_1) Sketch of the scattering process. (a_2) Close-up of the sample with a schematic representation of the key geometrical quantities. (b) Cartoon representation of the simulated complex (Protein Data Bank entry 1EM8) with the position of the sulfur atoms highlighted in yellow. In particular, proteins A and B comprise three and four sulfur atoms, respectively. (c) Radial intensity of the signal registered on the detector as a function of the observation angle.

dephasing problems can arise due to the phase ambiguity inherent in the X-ray diffraction pattern, especially when the resolution of the data is limited or when the crystal lacks heavy atoms. Sulfur atoms in proteins exhibit anomalous scattering, which arises from X-ray interactions with sulfur's electron configuration. This phenomenon leads to differences in X-ray scattering between wavelengths, providing additional experimental data for solving phase ambiguity and improving electron density maps. In particular, multiwavelength anomalous dispersion (MAD) is a technique used in X-ray crystallography to overcome the problem of phase ambiguity.[18] By collecting X-ray diffraction data at multiple wavelengths around the absorption edge of sulfur, the anomalous differences between the wavelengths provide additional phase information, allowing for more accurate structure determination. Similarly, the single-wavelength anomalous dispersion (SAD) approach involves the collection of X-ray diffraction data at a single wavelength near the absorption edge of either native atoms, used in the crystallographic process, or atoms inserted *ad hoc* into the structure.[19,20] Although SAD requires fewer experimental data compared to MAD, it still offers enough anomalous signal from sulfur to deduce phase information and resolve the protein structure.[19]

Here, we propose using anomalous scattering to detect protein−protein interactions without the need for a crystallographic structure. Our intuition is to compare the sum of the signals of the photons scattered by the native sulfur atoms for ultradiluted solutions of each protein independently with that of the solution containing both proteins. Proteins that do not interact will produce an intensity made by the superposition of the patterns produced by the two proteins found alone in solution, which will differ from the pattern produced by the proteins forming a complex.

To gain analytical insights into this crucial aspect, we consider the model setup described in Figure 1a, depicting the diffusion of a coherent excitation source on a sample. The coherent excitation source can be considered as a plane wave $\phi_S = A_s e^{-i\vec{k}_s\cdot\vec{r}}$, with $|k_s| = 2\pi/\lambda$ and λ is in the range around the sulfur K edge (∼0.5 nm). Interaction between the source and the proteins sulfur atoms around their K-edge threshold produces a scattering process, which can be approximated by spherical waves originating from each sulfur atom. The electromagnetic field at position $\vec{R}_d = (x_D, y, z)$ on the detector is given by the scattered light of the N protein's sulfur atoms:

$$\psi = \sum_i^N \psi_i \tag{1}$$

where the field produced by the ith atom has the form

$$\psi_i = \frac{A_i e^{-i|\vec{k}_f^{\,i}||\vec{R}_d - \vec{r}_i|}}{|\vec{R}_d - \vec{r}_i|} e^{-i\vec{k}_s\cdot(\vec{R}_s - \vec{r}_i)} \tag{2}$$

i.e. ψ_i is the product of a spherical wave of amplitude A_i and wave vector $|\vec{k}_f^{\,i}| = \frac{2\pi}{\lambda}$ with a phase term depending on the distance between the source and the ith sulfur atom. Assuming one can measure one protein/complex at a time, the intensity at the detector will depend on the number and disposition of the sulfur atoms in the system:

$$I(\vec{R}_d) = |\psi|^2 = \frac{A^2}{|R_d|^2} \sum_{i,j}^N e^{-i(|\vec{k}_f^{\,i}||\vec{R}_d - \vec{r}_i| - |\vec{k}_f^{\,j}||\vec{R}_d - \vec{r}_j|)} e^{-i\vec{k}_s\cdot(\vec{r}_j - \vec{r}_i)} \tag{3}$$

where we assumed that spherical waves possess the same amplitude and the distance between the sample and the detector is such that $|\vec{R}_d| \gg |\vec{r}_i|$ for $i = 1, ..., N$. In the latter approximation regime, we can consider the directions of propagation of the spherical waves as parallel, i.e., $\hat{k}_f^i \sim \hat{k}_f^j$, so

https://doi.org/10.1021/acs.jpclett.3c03632
J. Phys. Chem. Lett. 2024, 15, 3478−3485

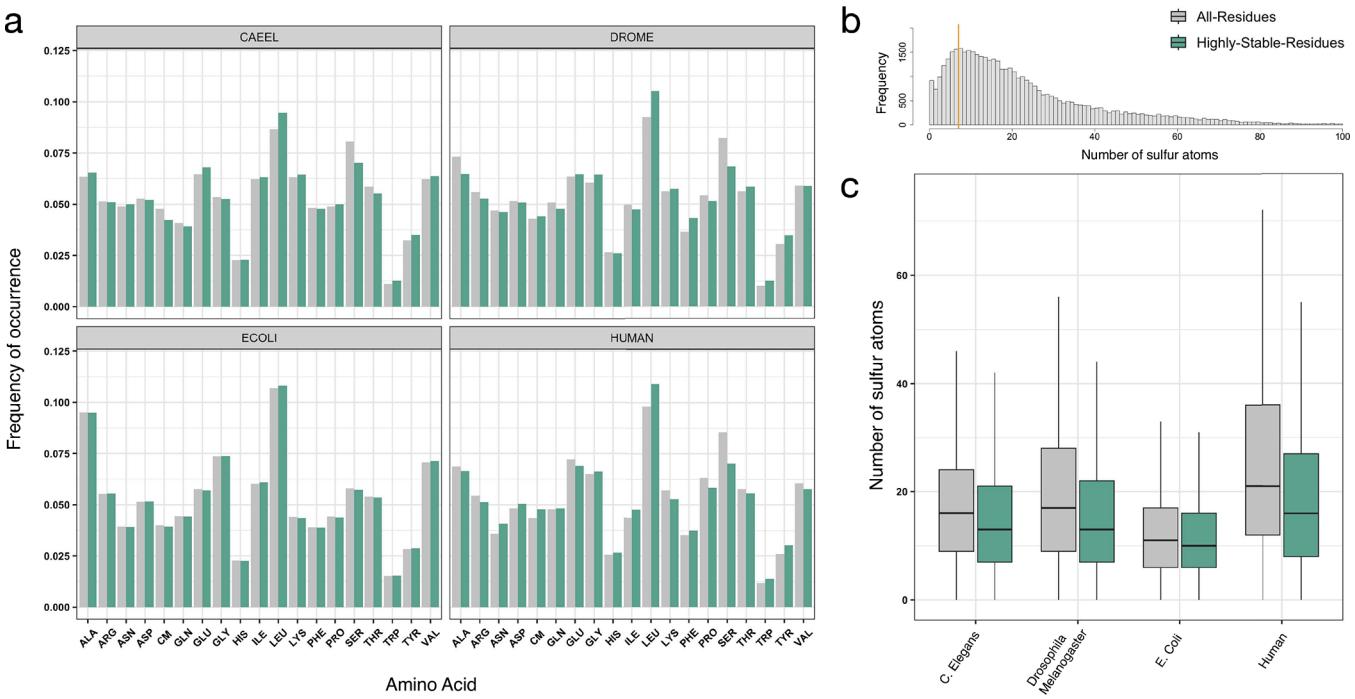


Figure 2. Abundance of sulfur atoms in protein structures. (a) Frequency of occurrence of the 20 natural amino acids in the proteome of four representative organisms, i.e., *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli*, and *Homo sapiens*. Gray bars refer to all protein residues, while green ones correspond to residues predicted to be in structured, highly stable regions, according to the AlphaFold2 score. (b) Distribution of the number of sulfur atoms per protein found in the human proteome. An orange vertical line marks the mode of distribution. (c) Box plot representation of the distributions of the number of sulfur atoms per protein found in the proteome of the four representative organisms. Gray bars refer to all protein residues, while green ones correspond to residues predicted to be in structured, highly stable regions according to the AlphaFold2 score.

that $(|\vec{k}_f^{\,i}||\vec{R}_d - \vec{r}_i| - |\vec{k}_f^{\,j}||\vec{R}_d - \vec{r}_j|) \sim \frac{2\pi}{\lambda}\hat{k}_R\cdot(\vec{r}_j - \vec{r}_i)$, where the versor $\hat{k}_R$ has the same direction of $\vec{R}_d$. In this regime, eq 3 reduces to

$$I(\vec{R}_d) = \frac{A^2}{|R_d|^2}\sum_{i,j}^{N} e^{-i\vec{q}\cdot\vec{d}_{ij}} = \frac{A^2}{|R_d|^2} \times \left\{N + 2\sum_{i<j}^{N} \cos\left[\frac{4\pi}{\lambda}\cos(\eta/2)|d_{ij}|\cos(\phi_{ij})\right]\right\} \quad (4)$$

where $d_{ij} = |\vec{r}_i - \vec{r}_j|$, $\vec{q} = \vec{k}_R + \vec{k}_S$, $|\vec{q}| = \frac{2\pi}{\lambda} \times 2\cos(\eta/2)$, η is the convex angle between versors $\hat{k}_R$ and $\hat{k}_S = \hat{x}$, and ϕ_{ij} is the convex angle between versor $\hat{q}$ and vector $\vec{d}_{ij} = (\vec{r}_i - \vec{r}_j)$.

Notably, the intensity depends on the distances between all couples of sulfur atoms in the system and on the orientation of the protein/complex. As measurements will be performed on proteins in suspension, the orientation of the protein and/or complex will be random. To remove the effect of the orientation, one can compute an average over different acquisitions, each being associated with an orientation uniformly distributed in the unit sphere. The outcome of the measure will be given by

$$\langle I\rangle = \int_0^{2\pi} d\theta \int_0^{\pi} d\phi I(\theta, \phi)\frac{\sin(\phi)}{4\pi} \quad (5)$$

where $[\sin(\phi)]/(4\pi)$ is the uniform distribution of sampling an orientation described by angles (θ, ϕ) around the observation axis, $\hat{q}$.

After some straightforward calculations, one obtains

$$\langle I\rangle = \frac{A^2}{|R_d|^2}\left\{N + 2\sum_{i<j}^{N}\frac{\sin\left[4\pi\frac{|d_{ij}|}{\lambda}\cos(\eta/2)\right]}{4\pi\frac{|d_{ij}|}{\lambda}\cos(\eta/2)}\right\} \quad (6)$$

The final expression depends on the relative distances between sulfur atoms and the observational angle (note that relative distances are assumed to be fixed; however, it could be possible to partially account for protein motion, e.g., assuming normal motion of the sulfur atoms around the structure positions). In particular, looking at eq 6, one can see that for an interacting complex made by proteins A and B, the equation can be decomposed into three contributions: $\langle I\rangle = \langle I\rangle_A + \langle I\rangle_B + S_{AB}$, where $S_{AB} = 2\sum_{i<j}\sin(x)/x$, and $x_{ij} = 4\pi|d_{ij}|/\lambda\cos(\eta/2)$. The signal of a complex is given by the sum of the signals of the two proteins when found alone in solution, plus a term given by the sum over all intermolecular sulfur−sulfur distances. The amplitude of the summation term rapidly decays toward zero for distances higher than the wavelength (see Figure S3). At distances comparable with the wavelength, the summation provides a sizable contribution to the mean intensity. As one can see from Figure S3, the probability distribution of sulfur−sulfur distances in the selected protein data set has a mean at ∼25 Å and a minimum distance of ∼1.8 Å, indicating that a small fraction of couples form a sulfur−sulfur bridge. Such couples of sulfur at small distances perturb the mean intensity in a significant manner. This is interesting as for proteins with exposed sulfurs, depending on the quality of the observed signal, it could be possible to determine whether exposed sulfurs interact on the basis of the registered signal. Large deviations of the measured intensity from the sum

https://doi.org/10.1021/acs.jpclett.3c03632
J. Phys. Chem. Lett. 2024, 15, 3478−3485

Table 1. Features of the Distributions of the Number of Sulfur Atoms per Protein for the Proteomes of the Considered Organisms[a]

organism	⟨AR⟩	σ_{AR}	⟨HSR⟩	σ_{HSR}	mode AR	mode HSR	median AR	median HST
C. elegans	16.9	13.0	15.4	12.3	14	3	14	13
D. melanogaster	18.7	15.1	16.6	13.5	10	9	15	13
E. coli	12.0	8.3	11.7	8.2	7	6	10	10
H. sapiens	23.0	18.4	20.2	17.2	8	8	18	15

[a]For each organism, the mean, standard deviation, mode, and median of the distributions are reported considering all protein residues (AR) or just highly stable ones (HSR) according to the AlphaFold confidence score.

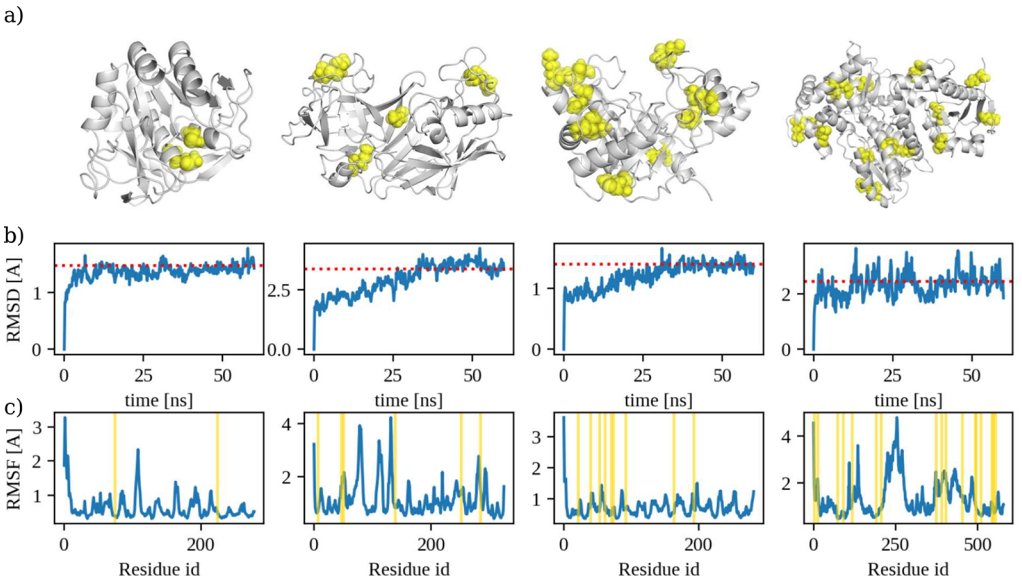


Figure 3. Analysis of molecular dynamics simulations. (a) Cartoon representation of four proteins from the selected data set containing different numbers of sulfur atoms (yellow). (b) Root-mean-square deviation (RMSD) as a function of simulation time. The red horizontal line marks the equilibrium value. (c) Root-mean-square fluctuation (RMSF) of each residue of the four chosen proteins. Yellow vertical lines mark the RMSF values of the residues containing the sulfur atoms.

of the intensity of the two proteins alone would indicate an interaction pair of sulfur atoms at a short distance, i.e., in the binding site of the complex (see Figure S4). This kind of information may be exploited to obtain information about the residues involved in the interaction.

To check the model, we ran a numerical simulation of an ideal, simplistic experimental outcome to test whether the disposition of the sulfur atoms in two interacting proteins produces a diffraction pattern distinct from the pattern produced by the two proteins alone in solution. In panels b and c of Figure 1, we consider a system of sulfur atoms mimicking that of the χ and ψ subunit heterodimers from DNA polymerase III (Protein Data Bank entry 1EM8). The two proteins contain three and four sulfur atoms, respectively (Figure 1b). Evaluating the outcome of a scattering process from a system composed of (i) just protein χ, (ii) just protein ψ, (iii) the two proteins not bound, and (iv) the complex, we obtained the signal in Figure 1c. One can see that the outcomes in the presence or absence of binding are distinguishable.

To verify the range of applicability of the proposed technique, we performed a set of analyses, evaluating at first the abundance and distribution of amino acids containing sulfur atoms (i.e., methionine and cysteine) in the proteomes of different organisms that are usually used in protein−protein interaction investigations to determine the average number of sulfur atoms; next, we analyzed the motion of sulfur atoms with respect to the motion of the whole protein structures, focusing on the relative distances between couples of sulfur atoms found in the structures of a representative set of 20 proteins. Finally, we propose an experimental apparatus and protocol for actually measuring protein−protein interactions.

To begin, we performed a statistical analysis of the number of sulfur atoms in protein structures across different organisms. In particular, we select all sequences of the proteome of *C. elegans*, *D. melanogaster*, *E. coli*, and *H. sapiens*.

Figure 2a reports the frequency of occurrence of the 20 natural amino acids. Gray bars refer to all protein residues, while green ones correspond to residues predicted to be in structured, highly stable regions, according to the AlphaFold2 score. AlphaFold2 produces a per-residue estimate of its confidence, which is called pLDDT, on a scale from 0 to 100. This confidence measure assigns to each residue a reliability value of the corresponding position prediction. The higher the pLDDT score, the greater the reliability of the prediction. The authors of the AlphaFold2 algorithm demonstrated that a pLDDT of <50 is a reasonably strong predictor of disorder, thus suggesting that this region is unstructured under physiological conditions.[15] Figure 2b displays the distribution of the number of sulfur atoms found in the human proteome. As one can see, most of the proteins have fewer than 20 sulfur atoms in their structures, while out of the entire human proteome, the most likely value is eight sulfur atoms for each protein. In Figure 2c, we reported, as box plots, the

https://doi.org/10.1021/acs.jpclett.3c03632
J. Phys. Chem. Lett. 2024, 15, 3478−3485

distributions of the number of sulfur atoms per protein found in the proteome of the four representative organisms we considered. Gray bars refer to all protein residues, while green ones correspond to residues predicted to be in structured, highly stable regions, according to the AlphaFold2 score (see Table 1 for additional details). Sulfur atoms belonging to the disordered regions will be characterized by larger movements. For instance, upon their removal from the statistical analysis conducted on the whole *C. elegans* proteome, the distribution is characterized by a mode of 3 (compared to 14 in the previous case in which all residues have been considered).

This preliminary analysis demonstrates that from a biological point of view, the technique may be applied to perform large screenings, as most of the proteins have sulfur atoms in their low-motility regions. Concurrently, most of the analyzed proteins have fewer than 10−20 sulfur atoms, which would allow one to obtain a reliable signal, as from a resolution point of view, the smaller the number of sulfur atoms for each protein considered, the higher the resolution of the signal obtained from the scattering.

Our protocol does not rely on crystallized protein structures and requires the performance of two measurements of the sample with beams having energies lower and higher than that of the sulfur K edge to exploit the anomalous dispersion. As the latter depends on the relative distances between the sulfur atoms inside the two proteins when measured alone and in complex, we must consider the relative motion between sulfur atoms, which, ideally, should be fixed both between the two measurements and in different samples. To determine the regimes in which this assumption holds, we performed molecular dynamics simulations of a set of 20 proteins extracted from a larger data set spanning different protein structures, families, and types. As one can see from Figure 3, the number of sulfur atoms ranges from one to ∼20 across the considered proteins, and their spatial disposition ranges along the whole protein structure. Note that proteins tend to form sulfur bridges; therefore, couples of sulfur atoms are frequently found at distances <3 Å. Such bridges have a local stabilizing effect, so that cysteines and methionines belonging to ordered regions tend to have smaller fluctuations than other parts of the structure. In particular, evaluating the root-mean-square fluctuation (RMSF) of all protein residues (see Figure 3) confirms this trend (see the Supporting Information for more details). Finally, to obtain an estimate of the time scales of sulfur atom relative motion, we computed the average difference in the distances of the sulfur atoms of the considered data set in time. At the picosecond time scale, the average variation of relative distances is 0.5 Å, 10-fold smaller than the needed wavelength.

Notably, other sources of significant motions are the conformational changes between the bound and unbound states of molecular partners that may occur during protein−protein interactions. Traditionally, these interactions have been categorized into three main classes. The "lock-and-key" model, characterized by minimal changes upon binding, emphasizes shape complementarity, often exhibiting RMSD values of less than 1−2 Å. However, proteins are inherently dynamic, undergoing conformational changes due to thermal noise and interactions with partners.[21] The "induced fit" model suggests a significant conformational change upon binding, while the "conformational selection" model proposes that binding stabilizes a conformation explored in the unbound state.[21]

Recent studies extensively analyzed these models across various protein−protein interactions, revealing that a substantial portion of proteins conform well to the lock-and-key model. Approximately one-third explores the bound conformation in their unbound state, aligning with the conformational selection model, while only a small percentage follows the induced fit model.[22] In scenarios involving substantial conformational changes upon binding, our proposed methodology might experience a resolution loss due to modified intraprotein sulfur distances. However, we anticipate that the method can still detect binding events in such cases by observing changes in the mean intensity resulting from the additional signals emanating from interprotein sulfur couples formed during binding.

Based on the found results, we finally propose a possible experimental setup for performing the measurements. Recent advances in X-ray sources have made available the possibility of realizing emissions of multiple colors, in the approximately millijoule range. Depending on the spectral range, few techniques have been implemented that permit two colors well separated in energy, time, or both.[23,24] Of particular interest for this work, the latter option is also available around the sulfur K edge where two pulses separated by ≤2 ps can be emitted with a temporal duration of a few femtoseconds. Using the experimental setup sketch represented in Figure 4, the two colors can be separated and recombined at the sample position. Hypothesizing to use a Si111 set of crystals, C1 can be oriented to reflect only the 2.5 keV beam at an angle θ_{b1} of

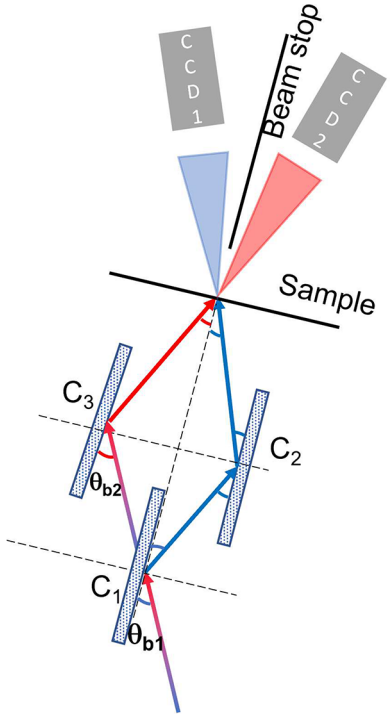


Figure 4. Sketch of the experimental setup. A two-color FEL emission (red−blue arrow) is sent to crystal C1, which is set at the Bragg angle (θ_{b1}) only for the 2.5 keV photons. The reflected beam (blue arrow) is then intercepted by the C2 crystal, which operates at the same Bragg angle, steering it toward the sample. The C3 crystal operated at the Bragg angle for the 2.4 keV θ_{b2} emission reflects toward the sample only that color. Two X-ray cameras are placed after the sample to measure the scattering from the proteins. Color cross talk is avoided by using a metal foil.

https://doi.org/10.1021/acs.jpclett.3c03632
J. Phys. Chem. Lett. 2024, 15, 3478−3485

4.53°. A second crystal (C2) can subsequently be used to steer such a beam toward the sample position. A third crystal (C3) placed to intercept the beam transmitted by C1 can be set to reflect 2.4 keV at a θ_{b2} equal to 4.72°. The setup can be designed so that the optical paths of the beams after the point of impact on C1 can compensate for the delay difference between the two X-ray pulses. However, this is not a critical parameter because delays of a few picoseconds can be tolerated. Indeed, the reorientational dynamics of proteins are in the range of nanoseconds.[25] Two CCD cameras, centered on the transmitted beams, will record the scattered photons from the two used pulses. By subtracting the two recorded signals, we obtain the scattering contribution of sulfur atoms. Because scattering is expected to extend for >15°, a beam stop between the two CCDs will be installed to ensure that each CCD will detect mainly one color. The presence of a residual overlap of the other signal in each camera, i.e., errors in the beam stop component, is expected to slightly reduce the contrast between the two signals. However, we expect this effect not to impact the overall capability of discriminating between bound and unbound proteins.

Finally, we suggest using a microfluidic system both to allow for a fast acquisition of several independent signals and to minimize radiation-induced damage of the sample. In fact, it is known that proteins within a solution context are more susceptible to radiation damage than are their crystalline counterparts.[26] However, the relevance of the damage depends on the absorbed dose, which, in turn, is dictated by the chosen experimental conditions. One of the major effects of radiation damage in our setup would be modifications that induce aggregation of the proteins in solution, which would alter the outcome of the method.[27] To address these challenges effectively, a rigorous optimization of experimental parameters, including radiation exposure time and intensity, will be essential for striking a balance between obtaining quality data and minimizing radiation-induced damage. In this respect, the use of a flow apparatus that guarantees a single exposure of the sample to radiation will help prevent aggregation and contribute to the automatization of the procedure, allowing for large screenings.

Overall, the determination of the three-dimensional protein structure and protein−protein interactions is ranked among the 125 open problems of the century.[28] After >50 years, the recent development of artificial intelligence (AI)-based platforms is allowing for the rapid computational prediction of protein structures starting from their amino acid sequence.[15] AI-based software, like the novel AlphaFold2 neural network,[15] can now predict protein structures with atomic accuracy starting from their amino acid sequences even in cases in which no similar structures are known, thus overcoming the need to use homology modeling templates, and presents an accuracy competitive with experimental data in most of the cases as assessed in the last CASP competition.[29] Indeed, recently, the entire human proteome was predicted and made available to the scientific community. This achievement paves the way to the persistent problem, i.e., determining protein−protein interactomes. Computational advances in this contest are also being made.[30−38] The abundance of predicted interactions vouches for rapid and efficient experimental techniques that can both validate and guide these predictions.

Here, we suggested using sulfur atom anomalous scattering as a way to rapidly detect binding between couples of proteins. In particular, analyzing the composition of the proteomes of four organisms widely used in protein−protein interaction studies, we found that proteins contain on average seven sulfur atoms preferentially located on low-motility regions of the protein structure, whose relative distances remain stable on the picosecond time scales. A minimal model for the photon scattering of a set of sulfur atoms predicts a difference between the signal of unbound and bound proteins that depends on the relative distances between all of the couples of sulfur atoms. Thus, the signal coming from intermolecular sulfur couples permits the rapid detection of the binding.

Building upon the performed calculations, we finally proposed an experimental setup for actually measuring the interactions. We expect that the time scale for a single measure will be in the range of second(s), which would allow for high-throughput scanning of molecule interactions, where robotics manipulate single protein solutions and mixtures and software analyzes fast automatics data.

If it is confirmed, we envisage that our proposed technique will be a determinant in addressing the future challenge of protein interactome determination as it will permit us to tackle the scanning of the tens of millions of possible couples of dimeric complexes that form the interactomes of complex organisms like humans.

METHODS

Protein Data Set. We consider the data set proposed by Hensen et al.,[39] in which a collection of 112 representative proteins for each family was reported. From this initial set, we selected the 20 proteins having (i) longer sequences and (ii) no missing or incomplete residues (see ref 40 for further details). For each protein, a molecular dynamics simulation with an explicit solvent was performed.

Molecular Dynamics Simulation. The following protocol was used for each of the 20 simulations. We used Gromacs 2020[41] and built the system topology using the CHARMM-27 force field.[42] The protein was placed in a dodecahedron simulative box, with periodic boundary conditions, filled with TIP3P water molecules.[43] We checked that each atom of the protein was at least 1.1 nm from the box borders. The system was then minimized with the steepest descent algorithm. Next, a relaxation of water molecules and thermalization of the system were run in *NVT* and *NPT* environments each for 0.1 ns with a time step of 2 fs. The temperature was kept constant at 300 K with the v-rescale algorithm;[44] the final pressure was fixed at 1 bar with the Parrinello−Rahman algorithm.[45] The LINCS algorithm[46] was used to constrain H-bonds. A cutoff of 12 Å was imposed for the evaluation of short-range nonbonded interactions and the particle mesh Ewald method[47] for the long-range electrostatic interactions. Finally, we performed 60 ns of molecular dynamics with a time step of 2 fs, saving configurations every 2 ps. We considered the last 20 ns (10 000 frames) for the analysis.

ASSOCIATED CONTENT

Data Availability Statement

All relevant data are provided in the text. Codes can be made available upon reasonable request of the authors.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.3c03632.

Figures S1−S4 (PDF)

Transparent Peer Review report available (PDF)

https://doi.org/10.1021/acs.jpclett.3c03632
J. Phys. Chem. Lett. 2024, 15, 3478−3485

AUTHOR INFORMATION

Corresponding Author

Mattia Miotto − *Center for Life Nano & Neuro Science, Istituto Italiano di Tecnologia, 00161 Rome, Italy;* orcid.org/0000-0002-0043-8921; Email: mattia.miotto@iit.it

Authors

Edoardo Milanetti − *Department of Physics, Sapienza University, 00185 Rome, Italy; Center for Life Nano & Neuro Science, Istituto Italiano di Tecnologia, 00161 Rome, Italy;* orcid.org/0000-0002-3046-5170

Riccardo Mincigrucci − *Elettra-Sincrotrone Trieste S.C.p.A. di interesse nazionale, 34149 Trieste, Italy*

Claudio Masciovecchio − *Elettra-Sincrotrone Trieste S.C.p.A. di interesse nazionale, 34149 Trieste, Italy*

Giancarlo Ruocco − *Center for Life Nano & Neuro Science, Istituto Italiano di Tecnologia, 00161 Rome, Italy; Department of Physics, Sapienza University, 00185 Rome, Italy*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.3c03632

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was partially funded by ERC-2019-Synergy Grant (ASTRA, 855923), EIC-2022-PathfinderOpen (ivBM-4PAP, 101098989), and the 'National Center for Gene Therapy and Drugs based on RNA Technology' project (CN00000041) financed by NextGeneration EU PNRR MUR-M4C2-Action 1.4-Call 'Potenziamento strutture di ricerca e creazione di campioni nazionali di R&S' (CUP J33C22001130001).

REFERENCES

(1) Bonetta, L. Interactome under construction. *Nature* **2010**, *468*, 851−852.

(2) Hu, C.-D.; Chinenov, Y.; Kerppola, T. K. Visualization of Interactions among bZIP and Rel Family Proteins in Living Cells Using Bimolecular Fluorescence Complementation. *Mol. Cell* **2002**, *9*, 789−798.

(3) Sidhu, S. S.; Fairbrother, W. J.; Deshayes, K. Exploring protein−protein interactions with phage display. *Chembiochem* **2003**, *4*, 14−25.

(4) Rohila, J. S.; Chen, M.; Chen, S.; Chen, J.; Cerny, R.; Dardick, C.; Canlas, P.; Xu, X.; Gribskov, M.; Kanrar, S.; et al. Protein−protein interactions of tandem affinity purification-tagged protein kinases in rice. *Plant Journal* **2006**, *46*, 1−13.

(5) Bailey, D.; Urena, L.; Thorne, L.; Goodfellow, I. Identification of protein interacting partners using tandem affinity purification. *J. Visualized Exp.* **2012**, e3643.

(6) Goodfellow, I.; Bailey, D. Detection of Protein−Protein Interactions Using Tandem Affinity Purification. *Protein Affinity Tags: Methods and Protocols* **2014**, *1177*, 121−133.

(7) Söderberg, O.; Gullberg, M.; Jarvius, M.; Ridderstråle, K.; Leuchowius, K.-J.; Jarvius, J.; Wester, K.; Hydbring, P.; Bahram, F.; Larsson, L.-G.; et al. Direct observation of individual endogenous protein complexes in situ by proximity ligation. *Nat. Methods* **2006**, *3*, 995−1000.

(8) Hegazy, M.; Cohen-Barak, E.; Koetsier, J. L.; Najor, N. A.; Arvanitis, C.; Sprecher, E.; Green, K. J.; Godsel, L. M. Proximity ligation assay for detecting protein-protein interactions and protein modifications in cells and tissues in situ. *Curr. Protoc. Cell Biol.* **2020**, *89*, e115.

(9) Rich, R. L.; Myszka, D. G. Higher-throughput, label-free, real-time molecular interaction analysis. *Anal. Biochem.* **2007**, *361*, 1−6.

(10) Piacentini, R.; Centi, L.; Miotto, M.; Milanetti, E.; Di Rienzo, L.; Pitea, M.; Piazza, P.; Ruocco, G.; Boffi, A.; Parisi, G. Lactoferrin Inhibition of the Complex Formation between ACE2 Receptor and SARS CoV-2 Recognition Binding Domain. *International Journal of Molecular Sciences* **2022**, *23*, 5436.

(11) Miura, K. An Overview of Current Methods to Confirm Protein- Protein Interactions. *Protein & Peptide Letters* **2018**, *25*, 728−733.

(12) Wang, H.-W.; Wang, J.-W. How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Sci.* **2017**, *26*, 32−39.

(13) Kang, C. Applications of in-cell NMR in structural biology and drug discovery. *International journal of molecular sciences* **2019**, *20*, 139.

(14) Oviedo, F.; Ren, Z.; Sun, S.; Settens, C.; Liu, Z.; Hartono, N. T. P.; Ramasamy, S.; DeCost, B. L.; Tian, S. I.; Romano, G.; et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mater.* **2019**, *5*, 60.

(15) 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*, 583−589.

(16) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.

(17) Dauter, Z.; Dauter, M.; de La Fortelle, E.; Bricogne, G.; Sheldrick, G. M. Can anomalous signal of sulfur become a tool for solving protein crystal structures?11Edited by I. A. Wilson. *J. Mol. Biol.* **1999**, *289*, 83−92.

(18) Guss, J. M.; Merritt, E. A.; Phizackerley, R. P.; Hedman, B.; Murata, M.; Hodgson, K. O.; Freeman, H. C. Phase Determination by Multiple-Wavelength X-Ray Diffraction: Crystal Structure of a Basic "Blue" Copper Protein from Cucumbers. *Science* **1988**, *241*, 806−811.

(19) Rose, J. P.; Wang, B.-C. SAD phasing: History, current impact and future opportunities. *Arch. Biochem. Biophys.* **2016**, *602*, 80−94.

(20) Panneerselvam, S.; Kumpula, E.-P.; Kursula, I.; Burkhardt, A.; Meents, A. Rapid cadmium SAD phasing at the standard wavelength 1 A. *Acta Crystallographica Section D Structural Biology* **2017**, *73*, 581−590.

(21) Csermely, P.; Palotai, R.; Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Nature Precedings* **2010**, DOI: 10.1038/npre.2010.4422.1.

(22) Stein, A.; Rueda, M.; Panjkovich, A.; Orozco, M.; Aloy, P. A Systematic Study of the Energetics Involved in Structural Changes upon Association and Connectivity in Protein Interaction Networks. *Structure* **2011**, *19*, 881−889.

(23) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; et al. Two-colour pump−probe experiments with a twin-pulse-seed extreme ultraviolet free-electron laser. *Science* **2021**, *373*, 871−876.

(24) Serkez, S.; et al. Two Colors at the SASE3 Line of the European XFEL: Project Scope and First Measurements. *Proceedings of the 39th Free Electron Laser Conference* **2019**, *FEL2019*, 4.

(25) Bashardanesh, Z.; Elf, J.; Zhang, H.; van der Spoel, D. Rotational and Translational Diffusion of Proteins as a Function of Concentration. *ACS Omega* **2019**, *4*, 20654−20664.

(26) Sliz, P.; Harrison, S. C.; Rosenbaum, G. How does Radiation Damage in Protein Crystals Depend on X-Ray Dose? *Structure* **2003**, *11*, 13−19.

(27) Kuwamoto, S.; Akiyama, S.; Fujisawa, T. Radiation damage to a protein solution, detected by synchrotron X-ray small-angle

scattering: dose-related considerations and suppression by cryoprotectants. *Journal of Synchrotron Radiation* **2004**, *11*, 462−468.
(28) Kennedy, D.; Norman, C. What Don't We Know? *Science* **2005**, *309*, 75−75.
(29) Elofsson, A. Progress at protein structure prediction, as seen in CASP15. *Curr. Opin. Struct. Biol.* **2023**, *80*, 102594.
(30) Li, Y.; Ilie, L. SPRINT: ultrafast protein-protein interaction prediction of the entire human interactome. *BMC Bioinf.* **2017**, *18*, 485.
(31) Pitre, S.; Dehne, F.; Chan, A.; Cheetham, J.; Duong, A.; Emili, A.; Gebbia, M.; Greenblatt, J.; Jessulat, M.; Krogan, N.; et al. PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinf.* **2006**, *7*, 365.
(32) Martin, S.; Roe, D.; Faulon, J.-L. Predicting protein−protein interactions using signature products. *Bioinformatics* **2005**, *21*, 218−226.
(33) Shen, J.; Zhang, J.; Luo, X.; Zhu, W.; Yu, K.; Chen, K.; Li, Y.; Jiang, H. Predicting protein−protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 4337−4341.
(34) Zhang, Y.-N.; Pan, X.-Y.; Huang, Y.; Shen, H.-B. Adaptive compressive learning for prediction of protein−protein interactions from primary sequence. *Journal of theoretical biology* **2011**, *283*, 44−52.
(35) Milanetti, E.; Miotto, M.; Di Rienzo, L.; Monti, M.; Gosti, G.; Ruocco, G. 2D Zernike polynomial expansion: Finding the protein-protein binding regions. *Computational and Structural Biotechnology Journal* **2021**, *19*, 29−36.
(36) Miotto, M.; Di Rienzo, L.; Bò, L.; Ruocco, G.; Milanetti, E. Zepyros: A webserver to evaluate the shape complementarity of protein-protein interfaces. *arXiv* **2024**, DOI: 10.48550/arXiv.2402.06960.
(37) Desantis, F.; Miotto, M.; Di Rienzo, L.; Milanetti, E.; Ruocco, G. Spatial organization of hydrophobic and charged residues affects protein thermal stability and binding affinity. *Sci. Rep.* **2022**, *12*, 12087.
(38) Grassmann, G.; Di Rienzo, L.; Gosti, G.; Leonetti, M.; Ruocco, G.; Miotto, M.; Milanetti, E. Electrostatic complementarity at the interface drives transient protein-protein interactions. *Sci. Rep.* **2023**, *13*, 10207.
(39) Hensen, U.; Meyer, T.; Haas, J.; Rex, R.; Vriend, G.; Grubmüller, H. Exploring Protein Dynamics Space: The Dynasome as the Missing Link between Protein Structure and Function. *PLoS One* **2012**, *7*, e33931.
(40) Di Rienzo, L.; Miotto, M.; Bò, L.; Ruocco, G.; Raimondo, D.; Milanetti, E. Characterizing hydropathy of amino acid side chain in a protein environment by investigating the structural changes of water molecules network. *Front. Mol. Biosci.* **2021**, *8*, 626837.
(41) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701−1718.
(42) Brooks, B. R.; et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(43) 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.
(44) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(45) Parrinello, M.; Rahman, A. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Phys. Rev. Lett.* **1980**, *45*, 1196−1199.
(46) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(47) Cheatham, T. E. I.; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA, and Proteins. *J. Am. Chem. Soc.* **1995**, *117*, 4193−4194.

https://doi.org/10.1021/acs.jpclett.3c03632
J. Phys. Chem. Lett. 2024, 15, 3478−3485