Accommodating protein dynamics in the modelling of chemical cross-links

Matteo T. Degiacomi*[#‡], Carla Schmidt[†], Andrew J. Baldwin, Justin L.P. Benesch*

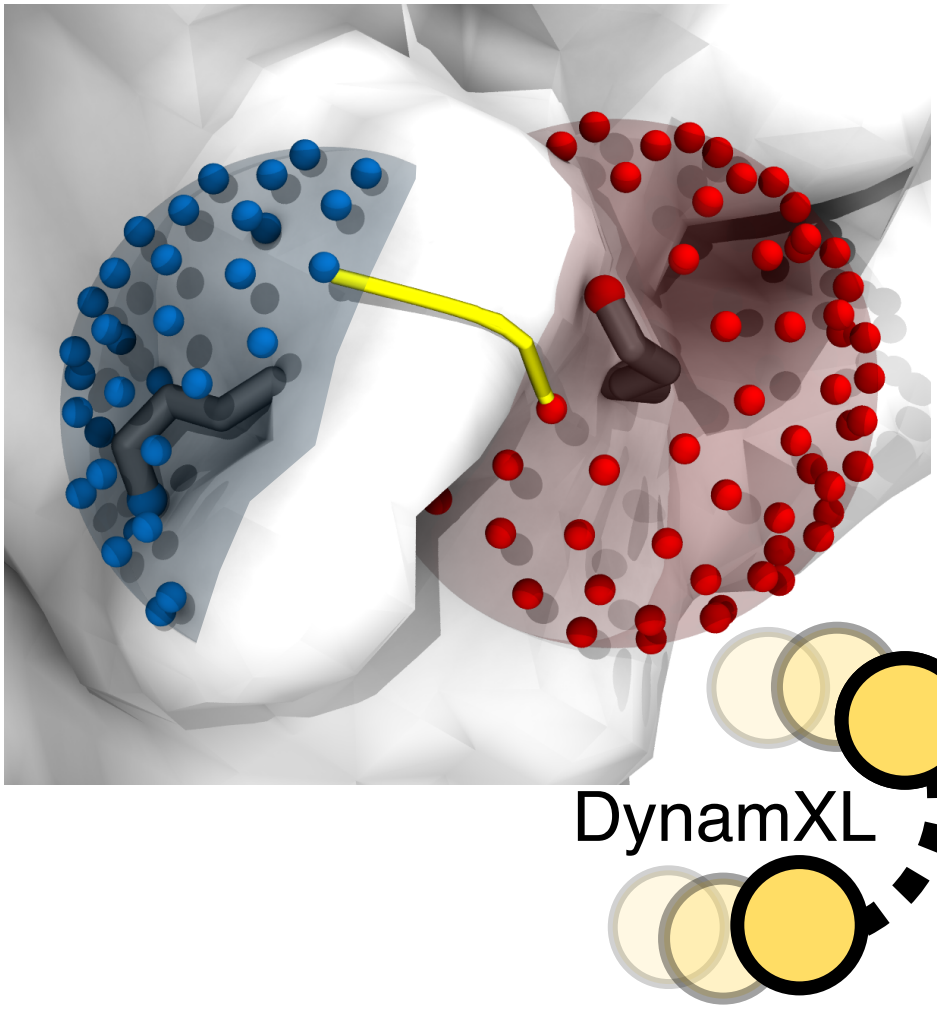
Department of Chemistry, Physical and Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford, OX1 3QZ, U.K.

* Correspondence to: matteo.t.degiacomi@durham.ac.uk and justin.benesch@chem.ox.ac.uk

[‡] Lead contact

[#] Present address: Chemistry Department, Durham University, South Road, Durham, DH1 3LE, UK

[†] Present address: Interdisciplinary research center HALOmem, Martin Luther University Halle-Wittenberg, Kurt-Mothes-Str. 3, 06120 Halle, Germany


DynamXL

Highlights

- Accounting for protein dynamics is key to measurement of cross-linking distances
- DynamXL is software that explicitly accommodates dynamics in cross-linking
- DynamXL models side-chain rotamers and allows analysis of structural ensembles
- This method improves data interpretation and protein docking performance

eToC

Degiacomi et al present a new method and associated software, DynamXL, that accommodates intrinsic protein dynamics in the modelling of cross-linking data. This approach substantially reduces error rates, leading to higher confidence in structure assessment and improved protein-protein docking.

Summary

Chemical cross-linking can identify the neighborhood relationships between specific amino acid residues in proteins. The interpretation of cross-linking data is typically performed using single, static atomic structures. However, proteins are dynamic, undergoing motions spanning from local fluctuations of individual residues to global motions of protein assemblies. Here we demonstrate that failure to explicitly accommodate dynamics when interpreting cross-links structurally can lead to considerable errors. We present a method and associated software, DynamXL, that is able to account directly for flexibility in the context of cross-linking modelling. Our benchmarking on a large data-set of model structures demonstrates significantly improved rationalization of experimental cross-linking data, and enhanced performance in a protein-protein docking protocol. These advances will provide a considerable increase in the structural insights attainable using chemical cross-linking coupled to mass spectrometry.

Keywords: cross-linking, molecular modelling, protein docking, mass spectrometry, computational structural biology

Introduction

Identifying and characterizing structurally the interactions between proteins is key to our understanding of their biological function (Landry et al., 2013; Vidal et al., 2011). Chemical cross-linking coupled to mass spectrometry (XL-MS) is revolutionizing our ability to obtain such insights by identifying covalent links made by bifunctional reagents between specific amino-acid side chains (Leitner et al., 2010; Rappsilber, 2011; Sinz, 2014). This information can be employed to identify protein assemblies and the connectivity of subunits within (Bruce, 2012; Liu et al., 2015). Furthermore, by establishing proximity relationships between amino-acid residues, XL-MS has been used to validate structures (Agafonov et al., 2016; Greber et al., 2015), examine protein conformations (Fischer et al., 2013; Schmidt and Robinson, 2014), and guide coarse-grained and atomistic structural modeling (Hall et al., 2016; Kalisman et al., 2012; Leitner et al., 2016; Politis et al., 2015; Shi et al., 2014).

In principle, because a cross-linking reagent and the side-chains it bridges have defined lengths, XL-MS data contains information that could directly enable high-resolution structure determination. A cross-link may be established during an experiment if the distance between two linking atoms is momentarily shorter than the spacer-length of the cross-linking reagent. From a modeling perspective, therefore, a significant challenge in using such data stems from complications arising from the intrinsic dynamics of proteins. These range from the reorientation of individual side-chains, to the concerted movement of large segments of the backbone (Fig. 1A)(Henzler-Wildman and Kern, 2007). As such, proteins populate a conformational ensemble in solution, rather than a static structure. Within this ensemble, the distance between two specific amino acids may vary dramatically (Fig. 1B). Because cross-linking is an essentially irreversible chemical modification, XL-MS is sensitive to the conformational heterogeneity of proteins. As such, XL-MS modeling based on a single atomic structure is fundamentally limited, and methods that accommodate protein dynamics are required.

In XL-MS modeling, dynamics are frequently accounted for implicitly, in coarse fashion. Typically, to determine whether two sites are cross-linkable, the distance between α- or β-carbon pairs rather than between the side-chain atoms that mediate cross-linking is measured. By comparison of this measure with a cut-off

defined by the length of the cross-linker plus that of the side chains, reorientation of the amino acids is, in part, accommodated (Rappsilber, 2011). In order to account for movements of the protein backbone, this cut-off is sometimes increased *ad hoc*. While simple, such blanket approximations do not account for the respective orientation of specific linkage sites, and the fact that different regions of the protein will feature different mobilities. As a result, interpretation of XL-MS data in a structural context, at atomistic resolution with a single structure, can lead to misinterpretation of the data.

Here we present a modeling strategy for cross-links that explicitly accommodates protein dynamics. We employ an efficient path-finding algorithm to define the shortest physical distance between two linkage sites, defined as the space accessible to the individual amino-acid side-chains. We also account for larger protein rearrangements by enabling the analysis of multiple protein conformations. We demonstrate that our method, implemented in the freely available software DynamXL, considerably outperforms existing approaches in the interpretation of XL-MS information, and allows us to improve the accuracy of protein-protein docking applications.

Results

We specified four principal requirements for an improved means for interpreting XL-MS data structurally: (1) to explicitly, on a case-by-case basis, accommodate the possibility of side-chain reorientation; (2) to be able to measure physical distances between linkage sites; (3) to allow the processing of multiple protein conformations concurrently; and (4) employ well-defined error estimates and confidence limits. Our solution is DynamXL, software written in Python, which can function both as a standalone package with graphical user interface, or as a library to facilitate integration with other computational structural biology tools.

Shortest paths between ensembles of amino acid side-chain conformations

Amino-acid side chains undergo rapid motions on the picosecond timescale, sampling many alternate conformations (Fig. 1A,B) (Henzler-Wildman and Kern, 2007). To accommodate this flexibility, we developed a strategy based on the linking atom being able to explore a series of concentric hemispheres defined by the plane of the protein backbone, and centered on the α-carbon (Cα) of the residue under consideration (Fig. 1C and S1). The radii of the hemispheres correspond to the probable distances, arising from side chain rotamers, between the Cα and the linking atom (Fig. S1). These hemispheres are constructed by DynamXL as a mesh of evenly distributed points, omitting those that overlap with the electron density of the remainder of the protein. This set of points thereby represents the linkage site as an ensemble explicitly accounting for the possible reorientations of the side-chains.

To measure the shortest path between two linkage ensembles, we evaluate the distances between all possible pairs of points. To ensure that the paths we measure are physically meaningful, in that they do not penetrate the protein, we calculate the shortest solvent accessible surface distance (SASD) (Bullock et al., 2016; Kahraman et al., 2011). We achieve this using a novel approach (full details in the STAR Methods), exploiting the Lazy Theta* path-planning algorithm (Nash et al., 2010). In brief, our strategy involves surrounding the protein with a fine-grained mesh, and determining the shortest route between defined start- and end-points, through a series of mesh points. The shortest path is subsequently smoothed, resulting in a shorter trajectory closer to the protein surface (Fig. S2A). We benchmarked the performance of our path-finding approach, and found it to return values that are within 5.7% of the true shortest path (Fig. S2B,C),

while the error associated with the discretization of the space accessible to side chains was estimated as 6.1% (Fig. S2D-G).

Side chain flexibility is key to effective distance estimation

In order to evaluate the benefit of accommodating side-chain dynamics and solvent-accessible paths in modeling cross-links, we used DynamXL to compare five different cross-linking distance measurement methods. These report either straight-line distances (SLDs) between static Cα-Cα or Cβ-Cβ pairs (hereon notated $SLD_{C\alpha}$ and $SLD_{C\beta}$), SASDs between static Cβ or Nζ pairs ($SASD_{C\beta}$ and $SASD_{N\zeta}$), and both SLDs and SASDs between Nζ ensembles ($\{SLD\}_{N\zeta}$ and $\{SASD\}_{N\zeta}$, where curly brackets indicate the shortest distance between two ensembles of alternative atom positions). We then performed extensive benchmarking using all 1755 PDB files in the PiQSi database of protein assemblies (Levy, 2007). Though DynamXL is capable of accommodating links between any atoms, we focused here on lysine side-chains, as their amino-nitrogen (Nζ) is targeted by many commonly used cross-linking reagents, and measured distances between each of the 70001 lysine occurrences using these different methods (Fig. S2H).

When interpreting cross-linking data using atomic coordinates, an approach widely adopted in the literature is to employ a cut-off longer than just the length of the cross-linker (plus the side-chain lengths, if not measuring between Nζ atoms) to accommodate implicitly the possibility of protein dynamics, including side-chain reorientation. The inevitable result is that, if the cut-off is not sufficiently accommodating, possible cross-links are missed (false-negatives). Conversely, if the threshold is overly generous, cross-links will be presumed when they are physically impossible (false-positives) (Fig. 2A). In our approach, where side-chain ensembles are modeled explicitly, the cut-off can instead be defined as the exact length, plus computational error, of the cross-linker used in the experiment. To describe this improvement quantitatively, we used the $\{SASD\}_{N\zeta}$ metric as reference to identify, for every alternative method, the cut-off that minimizes the classification error, that is the sum of false-positives and -negatives. This revealed that, for instance, using $SLD_{C\beta}$ in trying to explain data produced with lysine cross-linkers BS2, BS3 or EGS or the heterobifunctional SDA leads, at best, to classification errors of 15.5%, 20.9%, 17.1% and 9.2%, respectively (Fig. 2B, second column, and Table S1). When instead of measuring a straight line path we examined the

$SASD_{C\beta}$ measure, we noted better performances, especially for the longer cross-linkers. Nevertheless, we still found errors of approximately 10%. Importantly, while measuring the $SASD_{N\zeta}$ proved inferior to $SASD_{C\beta}$, measuring $\{SLD\}_{N\zeta}$ produced relatively low errors. Indeed, for short-range cross-links, this approach featured fewer errors than $SASD_{C\beta}$ measurements. For instance, for the SDA cross-linker 4.9% classification error has to be expected when using $\{SLD\}_{N\zeta}$, whereas $SASD_{C\beta}$ would lead to 8.8% (Fig. 2B). The performance obtained by $SASD_{C\beta}$ and $\{SLD\}_{N\zeta}$ shows how side-chain flexibility is essentially as important for accuracy in modeling cross-links as SASDs. Importantly, $\{SLD\}_{N\zeta}$ calculations are over 10-fold faster than $SASD_{C\beta}$, making it suitable for applications where computational speed is critical (Table S2, Fig. S2D). This combines to demonstrate that modelling alternate side-chain conformations outweighs the loss in performance caused by not measuring SASDs.

Examining ensembles of proteins allows the distinction of different conformational and ligand-bound states

To test our method on experimental data, we mined the XLdb (Kahraman et al., 2013), a database comprising experimentally determined cross-links of proteins for which a crystal structure has been deposited in the protein data bank (PDB). We calculated the $\{SASD\}_{N\zeta}$ for BS3 cross-links detected on 55 different proteins. Of 359 measures, 286 (79.7%) were found to be below the length defined by the cross-linker (Fig. S3A). In comparison, the commonly used $SASD_{C\beta}$ could explain only 69.1% of the distances (Fig. S3B), even when applying the ideal 25 Å cut-off which we determined above (Table S1). This reveals that $\{SASD\}_{N\zeta}$ is the most effective metric for explaining the available experimental data (Fig. S3C). Cross-links which could not be explained by our measurements are likely caused by a difference between the oligomerization or conformation of the protein in solution and the crystal lattice. One approach to tackle the latter error is to increase the cross-linking cut-off by a constant amount. We found that, whatever this additional increment, $\{SASD\}_{N\zeta}$ explained more data than $SASD_{C\beta}$ (Fig. S3D). This shows that protein dynamics beyond the side-chain level, have a negligible effect on the relative performance of the different distance metrics.

However proteins differ in their flexibility, both internally and between each other. To tackle this problem, DynamXL can measure the $\{SASD\}_{N\zeta}$ on multiple alternative protein conformations concurrently, thereby delivering a "distance of closest approach" (Jacobsen et al., 2006) within the multi-level ensemble for each pair of atoms under consideration. To examine the benefits of this capability, we first examined the four nuclear magnetic resonance spectroscopy (NMR) derived ensembles for which a total of 10 cross-links have been deposited in the XLdb (PDB ID: 1JM7, 2EJM, 2CQY, 1ZWV). We found that one of these cross-links could be explained by just a subset of available models, and one could not be explained by any of them (Fig. S3E). These two cross-links span notably flexible protein regions, associated with large fluctuations in $\{SASD\}_{N\zeta}$ (7.7 Å and 9.8 Å) within the respective ensemble. Since these four proteins are relatively small (6-26 kDa) and sparsely cross-linked, we next selected rhodopsin, a 39 kDa membrane protein (Jacobsen et al., 2006) for which 11 cross-links are reported in XLdb. We aggregated an ensemble of 16 rhodopsin structures from the PDB, and calculated the $\{SASD\}_{N\zeta}$ for each of them. We found that, although no single structure was able to explain more than 7 of the 11 (64%), the resulting ensemble of structures was sufficient to rationalise all the experimental cross-links (Fig. 3).

To test the selectivity of the cross-linking data, we tested them against an ensemble of opsin (rhodopsin lacking its cofactor retinal). In this case, one cross-link could not be explained from the available structures, revealing the data to, in principle, be capable of discerning conformational variations. When we examined another protein-ligand system, myoglobin and its haem group (Seebacher et al., 2006), we found that the data could not be explained entirely by an ensemble of 38 non-redundant *holo*-myoglobin structures, but necessitated the inclusion of *apo*- structures (Fig. S3F). This indicates that at least some of the myoglobin molecules in solution may have not had a haem group bound. These results reveal that the binding of a ligand, and the conformational change it induces, can be distinguishable by combining XL-MS data with our analysis approach.

Side chain flexibility for docking applications

Docking algorithms predict the arrangement of multiple proteins of known structure in a complex. Given the vast number of possible combinations, this task is extremely challenging (Bonvin, 2006). Experimental data

can however simplify the docking process, by guiding the prediction process towards regions of interest in the search space. In this context, XL-MS data have already been exploited multiple times to guide protein-protein docking protocols (Leitner et al., 2016). Given the high error-rate associated with simpler distance measures (Fig. 2), we quantified the number of false-positives and false-negatives a docking algorithm should be robust against. We extracted from our PiQSi database analysis all the inter-molecular distances obtained between lysine pairs, and simulated the identification of SDA, BS2, BS3 and EGS using as cut-off the optimal distances identified above (Fig. 2B and Table S1). We then quantified the expected number of false-positives and -negatives using the $\{SASD\}_{N\zeta}$ measures as a reference (Fig. 4A). Our results indicate that no significant difference is to be expected between the usage of $SLD_{C\alpha}$ and $SLD_{C\beta}$ metrics. $SASD_{C\beta}$ provides a substantial reduction in false-positives for long cross-linkers (5% less for BS3, 10% less for EGS), and outperforms $SASD_{N\zeta}$, which has a high false-negative rate. Notably, measuring $\{SLD\}_{N\zeta}$ yields the best performance in terms of false-negatives (always <4%), while false-positive rates are comparable to those of $SASD_{C\beta}$ for all but the longest cross-linkers. This propensity of only rarely overestimating the distance between two atoms, is important in a molecular docking context: while false-positives may lead to a larger pool of docking candidates being returned (decreasing precision), false-negatives may lead to the rejection of a correct docking pose (decreasing accuracy).

In principle, performing docking using $\{SASD\}_{N\zeta}$ should yield the best performance. However, the associated computational cost renders it currently impractical for applications where tens of thousands of measures must be performed rapidly (Table S2). Our benchmarking suggests however that measuring $\{SLD\}_{N\zeta}$ constitutes a good alternative. To test this hypothesis we simulated 1180 protein-protein docking cases restrained by sparse BS3 cross-linking distances, and 351 using SDA ones (Table S2, Fig. S4). We docked all these cases using the POW^{ER} docking algorithm (Degiacomi and Dal Peraro, 2013) that, in a total of 4593 independent runs, exploited $SLD_{C\alpha}$, $SLD_{C\beta}$, or $\{SLD\}_{N\zeta}$ to assess the distance restraints. The best models obtained with our ensemble approach have on average a lower RMSD from the crystal structure, compared with those produced by using $SLD_{C\alpha}$ or $SLD_{C\beta}$ (Fig. 4C). Furthermore, they satisfied a larger number of cross-linking restraints, and ranked higher in the list of best models produced by the docking algorithm (see Table S2 and Fig. S4). We observed that using the less accurate distance metrics resulted in

an increased likelihood of obtaining models that, despite satisfying all distance restraints, were incorrect (e.g. in Fig. 4D). Interestingly, we found that using a larger number of restraints did not necessarily lead to improved results, since the likelihood of misinterpreting at least one distance (and thus wrongly biasing the docking algorithm) increased (Fig. S4F). Notably, constraining the docking with a short- rather than long-range cross-linkers (SDA vs BS3) only improved performance when using the $\{SLD\}_{N\zeta}$ metric (Table S2). In sum, therefore, $\{SLD\}_{N\zeta}$, constitutes a fast and more accurate means for exploiting XL-MS data for protein docking.

Discussion

Recent advances in XL-MS experiments have enabled the identification of cross-links in, and between, a wide variety of proteins, even within complex mixtures. Here we have presented DynamXL, software that greatly improves the interpretation of identified cross-links in a structural context. In contrast to alternative strategies, our approach explicitly accommodates protein dynamics, allowing the interrogation of conformational ensembles rather than individual, static structures.

Other approaches accommodate side-chain and backbone motions implicitly by adding a constant to the length of the cross-linker and amino-acid residues to obtain a cut-off distance below which a cross-link is considered feasible. A variety of different values for this constant have been used in the literature. For example, in the case of BS3/DSS cross-linkers, cut-offs of 27 Å (Fritzsche et al., 2012), 28 Å (Kalisman et al., 2012) and 35 Å (Hall et al., 2016) have been used for $SLD_{C\alpha}$; 30 Å (Kahraman et al., 2013) and 34 Å (Herzog et al., 2012) for $SASD_{C\beta}$. Others have attempted to estimate the average effect of dynamics by mining a dataset of MD simulations, suggesting a criterion for $SLD_{C\alpha}$ <26-30 Å (Merkley et al., 2014), or to further adapt a static cut-off distance on the basis of specific linking pairs crystallographic beta factors (Rappsilber, 2011). Here we have shown how failing to account for side chain flexibility on a case-by-case basis will introduce large errors: for instance, the error of $SLD_{C\alpha}$ for BS3 is >20% (Fig. 2B). Our new approach, involving the measurement of a SASD on side-chain rotamers that are modeled explicitly, without the need of *ad hoc* corrections, is therefore more accurate than any previous method.

As well as accommodating the high-frequency motions of side-chains, we have specifically designed DynamXL to allow the concurrent interrogation of multiple structures differing in back-bone conformation. Such ensembles could be experimental (e.g. from NMR or collections of crystal structures in different conformations), or computational (e.g. from MD simulations or normal mode analysis) in origin. Our analyses demonstrate that incorporation of these different structures, representing low-frequency tertiary structure rearrangements, can be important when explaining experimental data. We anticipate that this capability will prove particularly important for large proteins and their assemblies, as these will typically

have an increased capacity for large amplitude motions, and necessitate multiple models to explain the observed data.

While the $\{SASD\}_{N\zeta}$ method is the most accurate in its description of protein flexibility, it demands the highest computational cost. Our results indicate that, within all alternative distance metrics, $\{SLD\}_{N\zeta}$ performs the best for cross-linkers shorter than ~12 Å. This metric features the lowest false-negative rate, and a false-positive rate comparable to that of $SASD_{C\beta}$, while requiring a fraction of its computational time. This observation shows how modeling the side-chain ensemble results in a greater improvement in accuracy over measuring SASDs instead of SLDs. In particular, we have demonstrated that using $\{SLD\}_{N\zeta}$ instead of the commonly employed $SLD_{C\alpha}$ and $SLD_{C\beta}$ leads to significant improvements in the performance of protein-protein docking restrained by XL-MS data.

In summary, we have demonstrated that our software, DynamXL, is well suited to accommodate the complications incurred by considering the dynamic structures of proteins and how they impact on XL-MS experiments. It is capable of explicitly accommodating motions at the level of both the reactive side-chains and larger-scale rearrangements of the protein backbone. By performing exhaustive benchmarks, we have revealed how these inclusions increase the accuracy of cross-link assessment, and also allow significant improvements in protein-protein docking. These more sophisticated computational approaches ultimately represent improvements in the effective resolution of structures built based on MS data. These improvements will go hand-in-hand with improved scoring metrics (Bullock et al., 2016), advances in methods for profiling interactions between proteins in the cell (Liu et al., 2015), and developments in integrative modeling (Sali et al., 2015) to cement the role of XL-MS in structural proteomics.

Author Contributions

Conceptualization, MTD, CS, AJB and JLPB; Software, MTD; Investigation, MTD; Writing – original draft, MTD and JLPB; Writing – Review & Editing, MTD, CS, AJB and JLPB; Funding acquisition, MTD and JLPB.

Acknowledgements

We thank Carol Robinson (Oxford) for helpful discussions and support, Tim Allison (Oxford) for critical review of the manuscript and testing the software, and Yusuf Ismail for preliminary work. We are grateful for the following funding sources: the Swiss National Science Foundation (P2ELP3_155339 to MTD), Impact Acceleration Awards (to JLPB) from the Biotechnology and Biological Sciences Research Council (BBSRC) and the Engineering and Physical Sciences Research Council, the BBSRC Tools and Resources Fund (BB/K004247/1 to JLPB), a BBSRC David Phillips Fellowship (BB/J014346/1 to AJB), and a Royal Society University Research Fellowship (UF120251 to JLPB).

References

Agafonov, D.E., Kastner, B., Dybkov, O., Hofele, R.V., Liu, W.T., Urlaub, H., Luhrmann, R., and Stark, H. (2016). Molecular architecture of the human U4/U6.U5 tri-snRNP. Science *351*, 1416-1420.

Bonvin, A.M. (2006). Flexible protein-protein docking. Curr Opin Struct Biol *16*, 194-200.

Bruce, J.E. (2012). In vivo protein complex topologies: sights through a cross-linking lens. Proteomics *12*, 1565-1575.

Bullock, J.M.A., Schwab, J., Thalassinos, K., and Topf, M. (2016). The Importance of Non-accessible Crosslinks and Solvent Accessible Surface Distance in Modeling Proteins with Restraints From Crosslinking Mass Spectrometry. Mol Cell Proteomics *15*, 2491-2500.

Degiacomi, M.T., and Dal Peraro, M. (2013). Macromolecular symmetric assembly prediction using swarm intelligence dynamic modeling. Structure *21*, 1097-1106.

Dijkstra, E.W. (1959). A note on two problems in connexion with graphs. Numer Math *1*, 269-271.

Fischer, L., Chen, Z.A., and Rappsilber, J. (2013). Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. J Proteomics *88*, 120-128.

Fritzsche, R., Ihling, C.H., Gotze, M., and Sinz, A. (2012). Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis. Rapid Commun Mass Spectrom *26*, 653-658.

Greber, B.J., Bieri, P., Leibundgut, M., Leitner, A., Aebersold, R., Boehringer, D., and Ban, N. (2015). Ribosome. The complete structure of the 55S mammalian mitochondrial ribosome. Science *348*, 303-308.

Hall, Z., Schmidt, C., and Politis, A. (2016). Uncovering the Early Assembly Mechanism for Amyloidogenic beta2-Microglobulin Using Cross-linking and Native Mass Spectrometry. J Biol Chem *291*, 4626-4637.

Henzler-Wildman, K., and Kern, D. (2007). Dynamic personalities of proteins. Nature *450*, 964-972.

Herzog, F., Kahraman, A., Boehringer, D., Mak, R., Bracher, A., Walzthoeni, T., Leitner, A., Beck, M., Hartl, F.U., Ban, N.*, et al.* (2012). Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. Science *337*, 1348-1352.

Jacobsen, R.B., Sale, K.L., Ayson, M.J., Novak, P., Hong, J., Lane, P., Wood, N.L., Kruppa, G.H., Young, M.M., and Schoeniger, J.S. (2006). Structure and dynamics of dark-state bovine rhodopsin revealed by chemical cross-linking and high-resolution mass spectrometry. Protein Sci *15*, 1303-1317.

Kahraman, A., Herzog, F., Leitner, A., Rosenberger, G., Aebersold, R., and Malmstrom, L. (2013). Cross-link guided molecular modeling with ROSETTA. PLoS One *8*, e73411.

Kahraman, A., Malmstrom, L., and Aebersold, R. (2011). Xwalk: computing and visualizing distances in cross-linking experiments. Bioinformatics *27*, 2163-2164.

Kalisman, N., Adams, C.M., and Levitt, M. (2012). Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling. Proc Natl Acad Sci U S A *109*, 2884-2889.

Landry, C.R., Levy, E.D., Abd Rabbo, D., Tarassov, K., and Michnick, S.W. (2013). Extracting Insight from Noisy Cellular Networks. Cell *155*, 983-989.

Leitner, A., Faini, M., Stengel, F., and Aebersold, R. (2016). Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. Trends Biochem Sci *41*, 20-32.

Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F., Rinner, O., Beck, M., and Aebersold, R. (2010). Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. Mol Cell Proteomics *9*, 1634-1649.

Levy, E.D. (2007). PiQSi: Protein quaternary structure investigation. Structure *15*, 1364-1367.

Liu, F., Rijkers, D.T., Post, H., and Heck, A.J. (2015). Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat Methods *12*, 1179-1184.

Merkley, E.D., Rysavy, S., Kahraman, A., Hafen, R.P., Daggett, V., and Adkins, J.N. (2014). Distance restraints from crosslinking mass spectrometry: mining a molecular dynamics simulation database to evaluate lysine-lysine distances. Protein Sci *23*, 747-759.

Nash, A., and Koenig, S. (2013). Any-angle path planning. AI Magazine *34*, 85-107.

Nash, A., Koenig, S., and Tovey, C. (2010). Lazy Theta*: Any-Angle Path Planning and Path Length Analysis in 3D.

Politis, A., Schmidt, C., Tjioe, E., Sandercock, A.M., Lasker, K., Gordiyenko, Y., Russel, D., Sali, A., and Robinson, C.V. (2015). Topological models of heteromeric protein assemblies from mass spectrometry: application to the yeast eIF3:eIF5 complex. Chem Biol *22*, 117-128.

Rappsilber, J. (2011). The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. J Struct Biol *173*, 530-540.

Sali, A., Berman, H.M., Schwede, T., Trewhella, J., Kleywegt, G., Burley, S.K., Markley, J., Nakamura, H., Adams, P., Bonvin, A.M.*, et al.* (2015). Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. Structure *23*, 1156-1167.

Schmidt, C., and Robinson, C.V. (2014). A comparative cross-linking strategy to probe conformational changes in protein complexes. Nat Protoc *9*, 2224-2236.

Seebacher, J., Mallick, P., Zhang, N., Eddes, J.S., Aebersold, R., and Gelb, M.H. (2006). Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing. Journal of proteome research *5*, 2270-2282.

Shi, Y., Fernandez-Martinez, J., Tjioe, E., Pellarin, R., Kim, S.J., Williams, R., Schneidman-Duhovny, D., Sali, A., Rout, M.P., and Chait, B.T. (2014). Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. Mol Cell Proteomics *13*, 2927-2943.

Sinz, A. (2014). The advancement of chemical cross-linking and mass spectrometry for structural proteomics: from single proteins to protein interaction networks. Expert Rev Proteomics *11*, 733-743.

Uras, T., and Koenig, S. (2015). An empirical comparison of any-angle path-planning algorithms. Paper presented at: Eighth Annual Symposium on Combinatorial Search.

Vidal, M., Cusick, M.E., and Barabasi, A.L. (2011). Interactome networks and human disease. Cell *144*, 986-998.

Vogel, H. (1979). A better way to construct the sunflower head. Mathematical biosciences *44*, 179-189.

Figure Legends

Figure 1: Distances between side-chains depends on both global and local protein dynamics. A A 550 ns MD simulation of monomeric HIV-1 capsomer shows that it is composed of two rigid domains connected by a flexible linker (Degiacomi and Dal Peraro, 2013). Multiple resulting structures are superimposed with the N-terminal domain aligned (and the motional extremes emphasised), revealing large relative motions of the two domains. **Inset** Examination of individual side-chains reveals them to undergo considerable local fluctuations. **B** The distribution of distances between lysine Nζ (blue spheres in **A**) pairs, separating them into inter- (red) and intra- domain pairs (blue). Inter-domain distances are affected by relative motions of the domains such that broad (up to 45 Å wide), and sometimes bimodal, distance distributions are observed. Intra-domain distances are instead mostly affected by the relative orientation of side chains, leading to variations up to 15 Å. **C** To accommodate side-chain orientations, we can consider lysine Nζ atoms as able to explore a spherical space centered on the backbone Cα, and described by spherical coordinates with respect to their backbone plane. **D** The sampling of predicted Nζ positions for all lysines in the HIV-1 capsomer using our methodology (coloured surfaces) encompass all orientations observed in the MD simulation (black points), and more. Notably, some side chains are heavily constrained by their local environment (e.g. K20), while others can fully explore the space around them (e.g. K193). See also Fig. S1.

Figure 2: Benchmarking cross-linking measures using the PiQSi database. A Simulation of putative cross-links made from a selected lysine using $\{SASD\}_{N\zeta}$ (right), $SASD_{N\zeta}$ (middle), and $SASD_{C\beta}$ (right). For each distance metric, a cross-link to a nearby lysine is considered possible (yellow lines, or overlapping spheres) if the measured distance is below the optimal, fixed cut-off (identified in our benchmarks, see **B**). Using the $\{SASD\}_{N\zeta}$ metric, the most accurate because it measures physically meaningful distances between alternative locations representing side chain flexibility, three links are identified. $SASD_{N\zeta}$ identifies five links, thus producing two false-positives (i.e. links that are not possible given the structure). $SASD_{C\beta}$ predicts only two links, as such yielding one false-negative (i.e. an unidentified link). **B** We measured distances between all ~1.5 million lysine pairs in the PiQSi database using five alternative distance metrics ($SLD_{C\alpha}$, $SLD_{C\beta}$, $SASD_{C\beta}$, $SASD_{N\zeta}$, $\{SLD\}_{N\zeta}$) and compared them to $\{SASD\}_{N\zeta}$. In this latter method, a cut-off equal

to the cross-linker length, plus computational error, is used (e.g. 13.0 Å for BS3). For the other methods, that either do not model side-chain reorientation or use straight-line distances, the cut-off is altered so that cross-linker and side-chain flexibility are implicitly accommodated for must be used. Note, this cut-off encompasses contributions from the cross-linker, the side-chains it links, computational error, and the side-chain dynamics of the residue under question). **Upper** Classification error (the sum of false-positives and -negatives) depending on the cross-linker length and choice of cut-off distance, versus the $\{SASD\}_{N\zeta}$ reference. Blue regions feature the smallest classification error, red the highest. White circles indicate the cut-off that minimizes classification error for four commonly used cross-linkers. Smaller cut-offs lead to more false-negatives, larger ones increase false-positives. By selecting cut-offs that minimize the classification error, $SASD_{C\beta}$ yields a significant improvement over a simple SLD. Nevertheless, classification errors above 10% are still expected. **Lower** Expected error for the BS3 cross-linker, depending on the cut-off distance employed, for each measurement approach. The minimum corresponds to the cut-off that minimizes the classification error in our benchmark. See also Tables S1-2, Fig. S2.

Figure 3: DynamXL allows comparison of experimental cross-links within a structural ensemble. **Upper left** The horizontal lines indicate experimentally determined shortest distances between cross-linked lysine pairs in rhodopsin (Jacobsen et al., 2006), and points indicate $\{SASD\}_{N\zeta}$ measures on an ensemble of X-ray structures in the PDB. Blue points indicate measures below the experimental cut-off, and red points those above it. **Lower left** Areas colored in blue indicate an experimental cross-link explained by a specific crystal structure, while red ones are unexplained, and white ones indicate that a measure could not be performed because one or more of the linking atoms was absent in the structure. **Lower right** Hierarchical clustering of crystal structures according to Cα RMSD. Edges are colored according to their explanatory power: the darker, the more cross-links can be explained by the underlying cluster. **Top right** The various structures are superimposed, with the position of the lysine Nζ in the different structures shown (each a different color). Notably, the experimental data for rhodopsin can only be explained by rhodopsin structures, one link cannot be explained by opsin structures (identified by the PDB codes in parentheses). See also Fig. S3.

Figure 4: Accommodating side-chain dynamics improves protein-protein docking performance. A The number of false-positives and -negatives returned for the different distance metrics, when measuring distances between inter-monomer lysine pairs on all 245 dimers in the PiQSi database. $\{SLD\}_{N\zeta}$ yields the best performance, particularly in terms of false-negatives, revealing how modeling side-chain orientations is more important in this application that SASDs. **B** RMSD of the best model obtained with each of the three SLD metrics relative to the known dimeric arrangement. Cases in the grey band are those where the two metrics perform similarly, returning a difference in RMSD <1 Å. A larger proportion of cases lie below this gray band, rather than above it, indicating that $\{SLD\}_{N\zeta}$ in general outperforms $SLD_{C\alpha}$ and $SLD_{C\beta}$. **C** Distributions of the best models' RMSD to the known structure shows that the $\{SLD\}_{N\zeta}$ yields better models on average. **D** Example docking results performed with different distance metrics are superimposed on the known dimeric structure, shown in light grey (PDB codes, upper: 1E87, middle: 1T4H, lower: 1T3C). The best model obtained using $SLD_{C\alpha}$ is shown in blue, $SLD_{C\beta}$ in red and $\{SLD\}_{N\zeta}$ in dark grey. The fraction of satisfied distance restraints, and RMSD of the best model are noted. See also Table S3, Fig. S4.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for data will be fulfilled by the lead contacts Matteo T. Degiacomi (matteo.t.degiacomi@durham.ac.uk) and Justin L.P. Benesch (justin.benesch@chem.ox.ac.uk).

METHOD DETAILS

Software implementation

DynamXL is written in Python and C (via Cython). It can be executed as a Python package to facilitate its integration in external software, or via a graphical user interface allowing the facile comparison of experimental XL data and theoretical distances. The current version is capable of handling multi-PDB as well as gro files from molecular dynamics simulations performed with Gromacs. Though in this work we have focused on the lysine Nζ, DynamXL can measure distances between any pair of atoms.

Protein density representation and generation of the accessible mesh grid

We obtain an electron density map for the protein by centering a 3D Gaussian distribution with the atom's van der Waals radius at 1 standard deviation on each atomic coordinate. We then generate a mesh grid encompassing the whole protein with a spacing of 1 Å, and within every voxel calculate the sum of all local Gaussian distributions. Mesh points associated with voxels having an electron density greater than 1 (e.g. falling within the van der Waals radius of an atom, and hence considered as occupied) are removed. When studying larger proteins, the mesh might become too large to be held in memory. In this case, a smaller cubic mesh of predefined dimensions (based on the cross-linker spacer length) is used. When studying several cross-links on the same protein, the cubic mesh is displaced between two linkable atoms, if these are at a distance shorter than the box size. When using a single mesh grid encompassing the whole protein, our shortest path algorithm will always return a distance measure between two linkable atoms, as long as they are both solvent accessible. Using a local grid may fail in the case of start- and end-points being located in disconnected mesh regions. Since the size of the local grid is determined by the longest possible distance the

cross-linker can span, failure indicates that the two linking atoms are excessively far apart, and DynamXL reports an unsatisfied cross-link.

Representing the protein as an electron density dramatically reduces the number of small internal cavities, and the likelihood that a shortest path may be identified as passing through the protein instead of on its surface. An open area for investigation is represented by large loops and tunnels. If a tunnel through a protein is sufficiently large, our path detection algorithm can in principle identify shortest paths through it. Although such a scenario is unlikely *in vitro*, it might be possible for cross-linkers to connect residues through larger tunnels, such as the central cavity of a pore.

Construction of hemispheres

The underlying hypothesis of our side-chain modelling method is that linking atoms will explore the whole accessible region, represented as multiple concentric spheres centered on the Cα (Fig. 1C, Fig. S1B). Our method discretizes the spherical accessible region, replacing it by a mesh of equally spaced points representing alternative linking atom conformations. A uniform spherical point distribution ("golden sphere") is obtained using the 3D equivalent of Vogel's spiral method used to describe the distribution of sunflower seeds (Vogel, 1979), as follows. Let the coordinates of a point distribution centered at the origin be described in cylindrical coordinates (r, α, z), where α is the azimuth, and z the elevation (Fig. 1C); θ the golden angle $\pi(3-\sqrt{5})$ equal to ~137.508°; and i an integer number from 0 to the desired number of points N. Then:

$$\begin{cases} r_i = \theta i \\ \alpha_i = \sqrt{1 - z_i^2} \\ z_i = (1 - \frac{1}{N})(1 - \frac{2i}{N-1}) \end{cases} \qquad \text{(Equation 1)}$$

We define N as the closest integer to the value necessary for a desired point density. This is found by controlling the surface area σ represented by a single point. The higher the point density (and the smaller σ), the better the accessible sphere is represented. From this resulting mesh, points are removed according to three rules. First, the backbone plane is identified (using the amide O, C, and N atoms), and only the points located on the same side of the plane as the side chain atom are retained to ensure that no alternate atom positions represent unrealistic backbone twists (see Fig. S1A). Second, points within 2 Å of other atoms are

removed as these represent clashes with the protein. Last, if multiple concentric golden spheres feature an accepted point at the same position (same elevation and azimuth), only the outermost is retained.

Shortest path algorithm

First, accessible regions surrounding the protein of interest are identified. We consider a region accessible if it does not clash with the protein's buried amino acids, as well as backbone and Cβ atoms of solvent-exposed residues. This is based on the hypothesis that solvent-exposed side chains can move away from the trajectory of the cross-linker (see below). The problem of measuring the shortest distance between two atoms is now one of finding the shortest path within the remaining mesh points. The path start- and end-points are identified as the mesh points closest to the atoms of interest. The shortest path between the two is identified using the Theta* path detection algorithm, in its "lazy" form (Nash et al., 2010). Importantly, the paths identified by Theta* are not bound to pass through all intermediate mesh points if line-of-sight between two distant points exists (any-angle path detection (Nash and Koenig, 2013)). Without smoothing, the paths identified by this algorithm are shorter than those of algorithms bound to travel solely through mesh points, and are therefore very close to the true shortest path (Uras and Koenig, 2015). In order to obtain smoother direction changes, and to better connect the actual start- and end-coordinates, the path is then smoothed. This is done by applying a two-pass smoothing algorithm on Theta* paths. First, our algorithm adds pseudo-atoms separated by 1 Å on every straight section of path. Then, in a sliding-window fashion, triplets (a_1-a_2-a_3) are read from this pseudo-atom chain. When a triplet is not collinear (i.e. forms an angle other than 180°), a_2 is first displaced in the geometrical center of atoms a_1 and a_3, and then displaced towards its initial position in 0.1 Å steps, until no clash is detected. The operation is performed first from the start to the end of the chain, and then in the opposite direction. This important step makes changes in direction less angular, and paths around the protein shorter (on average 6.5%, Fig. S2A).

The Theta* path detection algorithm is executed on a mesh grid associated with all unoccupied voxels given the following conditions: 1) the SLD between start and end points is shorter than the length of the cross-linker, 2) the distance between the linking atom and the closest mesh point is <1 Å (a longer distance indicates that the linkage site is likely buried) and 3) the two points fail a line-of-sight test (i.e. an obstacle is

located between them). If start and end points have line-of-sight, their SLD is returned, otherwise Theta* is executed. For our tests, SASDs were measured between the subset of lysine pairs being at an SLD <24 Å (a cut-off including any pair at a distance shorter than the length of a BS3 cross-linker plus twice the length of a lysine side chain).

Protein-protein docking restraints and scoring function

We extracted all 245 dimers from the PiQSi database, and simulated experimental SDA and BS3 cross-linking datasets using $\{SASD\}_{N\zeta}$. For BS3, any inter-monomer lysine distance below 13 Å (i.e. 11.4 Å plus computational error) was considered as cross-linked. For analogous reasons, for SDA the cross-linking cut-off was set to 4.4 Å. All datasets had to contain at least two cross-linking distances. Whenever possible (i.e. >2 linking pairs available), up to four random subsets were produced from each of these cross-linking datasets in order to simulate data sparsity. As such, a total of 1180 simulated cross-linking datasets were obtained for BS3, and 351 for SDA (Table S3).

All the obtained datasets were exploited as distance restraints for docking, and $SLD_{C\alpha}$, $SLD_{C\beta}$, and $\{SLD\}_{N\zeta}$ were exploited as measurement metrics. For every docking pose, the distance of each pair in the restraint set was measured, and compared with a cross-linking acceptance criterion, δ. In the case of the BS3 datasets, distances δ were equal to 18.7 Å for $SLD_{C\alpha}$ and $SLD_{C\beta}$, and 10.6 Å for $\{SLD\}_{N\zeta}$. For SDA, δ was equal to 11.0 Å for $SLD_{C\alpha}$, 12.1 Å for $SLD_{C\beta}$, and 3.8 Å for $\{SLD\}_{N\zeta}$. All these distances had been previously identified as the optimal cut-off distance (Fig. 2B, Table S1). The following penalty function p was exploited, to score how much a docking pose respects n simulated target cross-linking distances, where d are the cross-linking distances measured on the structure:

$$p = \sum_{i=0}^{n} \begin{cases} 0 \; if \; d_i < \delta_i \\ \delta_i - d_i \; otherwise \end{cases} \qquad \text{(Equation 2)}$$

Furthermore, models featuring clashes between atoms should also be penalised. Hence we added a further energy term to the penalty function. This was represented by a 9-6 Lennard-Jones potential, e, between the Cα and Cβ of both docking partners, where r_{ij} is the distance between two atoms i and j, $A = 1$ is the depth of the potential minimum, and $B = 2.7$ the distance at which the potential is zero:

$$e = A \sum_i \sum_j \left(\frac{B}{r_{ij}}\right)^9 - \left(\frac{B}{r_{ij}}\right)^6 \quad \text{(Equation 3)}$$

Energies less than 0 indicate proteins that are in contact, but do not clash. Since the objective of this test was to assess the capacity of a cross-linking metric alone to the docking process, we decided to treat the van der Waals energy solely as a penalty. As such, any clash-free model was considered as equiprobable, by setting a minimum value of 0 for *e*. We combined equations 2 and 3 to produce the following fitness function, *f*:

$$f = p + \min(0, e) \quad \text{(Equation 4)}$$

Protein-protein docking protocol

All optimizations were performed exploiting the POW^{ER} optimization environment (Degiacomi and Dal Peraro, 2013). One of the two docking partners, the "receptor", was aligned at the origin and kept fixed. The other protein, the "ligand", was roto-translated around the receptor. The search space was therefore 6-dimensional (3 translations, 3 rotations). Boundary conditions were treated as reflexive, and were equal to [-100,100] for translations and [0,360] for rotations. The aim here was to identify positions in the search space leading to dimers being clash-free while respecting all distance restraints, i.e. having fitness f=0. The Particle Swarm Optimization "kick and reseed" (PSO-KaR) optimization algorithm was used with default parameters for inertia, and weights for personal- and global-best. Each optimization featured 80 particles, exploring the dimeric conformational space for 500 steps. In the "kick and reseed" paradigm, particles that are too slow (stagnating, i.e. displacing in any dimension by less than 0.01% of the total dimension size) are restarted in a new region of the search space, and their memory erased. To avoid other particles exploring that same region of the search space, a repulsive potential is added at that location. Every optimization case was executed five times, and the list of repelling potentials passed from one instance to the next.

All measures performed by every particle were logged. In a post-processing phase, all models associated with $f = 0$ were generated. If fewer than 100 models were found to fulfill this criterion, the 100 top-scoring models were extracted. For all of these, the Cα RMSD against the reference structure was calculated. The lowest score, as well as its rank in terms of fitness and the total number of fully satisfied distance restraints, was returned. All models with f=0 were considered as equiprobable, and assigned the same rank (i.e. 1).

QUANTIFICATION AND STATISTICAL ANALYSIS

Estimation of SASD measurement accuracy

We identified two possible sources of error in our SASD measurement method: the Theta* heuristic (e_1), and the specific location of mesh points with respect of protein atoms (e_2). To assess algorithm error, we randomly selected 100000 lysine pairs from the PiQSi database, 63.3% of which had no line-of-sight. For each of these, we measured the SASDs using both the Theta* and Dijkstra algorithms. The latter, though much slower, identifies the absolute shortest path between two nodes by exploring every accessible node in the mesh (Dijkstra, 1959). Paths obtained with the Dijkstra algorithm were refined by identifying key nodes via line-of-sight tests, and removing all others. Both Theta* and Dijkstra paths were then smoothed as described in the *STAR Methods*. Results indicated that, on average, only marginal differences between the two methods exist (for links shorter than 40 Å, average measures deviate by 0.08%, Fig. S2B). Standard error e_1 appears to grow near-linearly with SASD length, and is equal to 2.1% in our region of interest (SASD <24 Å).

The mesh grid location can affect the distance of linking atoms to their closest mesh point, as well as its capacity to sample narrow gaps. To assess the magnitude of these effects, we calculated the SASD on a subset of 5000 lysine pairs, 64 times. For each of these cases, the mesh grid position was altered with 0.25 Å steps in the x, y and z direction, for maximal displacements of 0.75 Å. Results indicated that the error e_2 caused by grid location is equal to 3.6%, independent of link length (Fig. S2C). In 1.3% of pairs, at least one grid position led to a SASD distance not being measured while other positions yielded a measure. The likelihood of missing a possible SASD measurement was 0.5%, which we can consider as negligible. Overall, we assess the error in our SASD measurement method as $e_1 + e_2 = 5.7\%$.

Assessment of side chain reorientation error

To study the effect of the surface area per point, σ, on path length, we randomly selected 300 different proteins from the PiQSi database. Within this subset, 4539 lysine pairs were at a $SLD_{N\zeta}$ <24 Å. We measured the $\{SASD\}_{N\zeta}$ between every pair multiple times, with σ ranging from 1 to 100 $Å^2$, and compared them with the measures obtained from the crystal structure conformation alone (i.e. $SASD_{N\zeta}$). Our results indicated that,

the smaller σ, the shorter the measured distance (Fig. S2E). We fitted the relationship between σ and path shortening, i.e. $\{SASD\}_{N\zeta}/SASD_{N\zeta}$, with a shifted logarithmic curve. The resulting function allows us to estimate the SASD one would obtain by treating the sphere as a continuum (i.e. as σ tends to 0). In such a case, paths would be on average 41% shorter than measured from the crystal structure alone. The function can also be exploited to predict how much longer the paths obtained with a sphere discretized at an arbitrary σ will be, versus a continuous sphere. This information can be used to identify the error e_3 dependent σ:

$$e_3(\sigma) = 0.18 \log_{10}(\sigma + 9.7) - 0.41 \qquad \text{(Equation 5)}$$

By combining this equation with the errors e_1 and e_2 identified in the previous section (expressed as fractions), every measured SASD is then rescaled as follows to obtain more accurate distance estimations:

$$SASD(\sigma) = SASD[1 - e_1 - e_2 - e_3(\sigma)] = SASD[1.35 - 0.18 \log_{10}(\sigma + 9.7)] \qquad \text{(Equation 6)}$$

A cross-link is considered possible if the distance between the linking atoms is below a given cut-off distance, typically dependent on the spacer length of the cross-linker. This classification can be prone to both false-positives and -negatives. Although SASDs corrected with equation 3 better reflect the shortest possible distance, higher σ values will be inevitably connected with worse distance estimates. Indeed, the fewer sampling points on the sphere, the fewer sphere-sectors corresponding to favourable side chain conformations are likely to be sampled. To better visualize the effect of σ on cross-link classification, we calculated the classification error of measures collected with different σ values against those collected at the smallest σ (1 Å^2) for SDA, BS2, BS3 and EGS cross-linkers. Without any correction, large classification errors are observed. In particular, larger errors arise with long spacer arms and low σ (Fig. S2F). Importantly, the application of the correction factor (Equation 6) substantially decreases all classification errors (Fig. S2G). While, in principle, one should choose the smallest possible value, reducing σ leads to increasing computational times (Fig. S2D). In the following tests, we selected σ = 4 Å^2, connected to a ~2% classification error, and an error in SASD of 6.1%. Coupling this error connected to sphere point density with the SASD measurement error identified above (e_1 + e_2 = 5.7%), a total error of 11.8% has to be accounted for. We stress that this error quantifies the imprecision in distance measurement, and not the algorithm's classification error. In this work, we accounted for our estimated distance measurement error by proportionally increasing the cross-linking distance acceptance cut-off.

Testing collisions of SASD with solvent-exposed side chains

The flexibility of side-chains (in addition to those involved in the cross-link directly) may be important, as they could impact on the shortest path by reorienting themselves to avoid a clash. Our approach, and that of others (Kahraman et al., 2011), is to account for this implicitly, by measuring SASDs using paths that avoid collisions only with the protein backbone and Cβ atoms. This assumes that the side-chains are essentially invisible to the cross-linker, returning a distance at least as short as if we modeled these side-chains explicitly. We therefore verified whether it will always be possible to rearrange side chains so that the cross-linker's path is unhindered. To do so, we measured $\{SASD\}_{N\gamma}$ between lysine pairs on 80 proteins randomly extracted from the PiQSi database, representing the resulting shortest paths as trails of pseudo-atoms. We then counted how many "spectator" side-chains each trail would clash with. In order to determine how many of these clashes could be avoided by rearranging the side chain, we mined the PiQSi database to characterize the accessible rotameric space of each amino-acid type (i.e., as per Fig S1, but for all amino acids). For each side-chain involved in a clash, we were thereby able to identify an ensemble of alternative positions. We then tested whether at least one of these alternative positions would abolish the clash with the path. Our benchmark consisted of 3711 paths, of which 1541 clashed with at least one side chain. This indicates that only in 32.7% of the cases paths travel through areas occupied by protein mass. Within the subset of hindered paths, in 1346 cases side chains could be rearranged to avoid the clash, whereas in 297 the clash was inevitable (i.e. 5.2% of all paths are hindered by a side chain). This benchmark required us to test a total of 3171 "spectator" side chains (within the 1541 clashing paths), of which 2874 (89.7%) could be reoriented to accommodate the cross-linker. This is a stringent test, because it only considers the rearrangement of side-chains, and does not account for minor and local rearrangements of the backbone that also occur, and would act to increase the number of alternative locations available to the clashing side-chain.

Comparison of* $\{SASD\}_{N\zeta}$ *distances against those obtained with alternative distance metrics

A scatter plot of $\{SASD\}_{N\zeta}$, the most sophisticated and accurate measure, against $SLD_{C\alpha}$ reveals two striking features (Fig. S2H, first column): a vertical limit at $SLD_{C\alpha}$ ~4 Å, due to the shortest possible distance between Cαs in a protein structure, and a diagonal limit. The data points on the first limit correspond to a situation where two side chains point away from each other. In these cases, because an extended lysine side chain is

~6.5 Å long, the distance between linkage ensembles is ~13 Å shorter than the corresponding $SLD_{C\alpha}$ distance. Importantly, a large number of $SLD_{C\alpha}$ measures yield shorter paths (above the 1:1 diagonal) than our method, arising from those straight-line paths penetrating the protein. Similar results are obtained when measuring $SLD_{C\beta}$ (Fig. S2H, second column). Measuring $SASD_{C\beta}$ instead of a straight line yields a completely different distribution (Fig. S2H, third column), whereby the vast majority of distances are longer. $SASD_{N\zeta}$ values (Fig. S2H, fourth column) are inevitably equal or longer than those obtained with $\{SASD_{N\zeta}\}$. This is because the position of each linking atom, as identified in the provided protein structure, is included in the ensemble of alternative conformations considered by the latter measuring method. $\{SLD\}_{N\zeta}$ distances are instead bound to be always equal or shorter than $\{SASD\}_{N\zeta}$ as, between the same ensembles of alternative atom arrangements, no distance can be shorter than a straight line.

In order to quantify the relative performance of these different distance metrics, we calculated the Pearson cross-correlation coefficient of each alternative method against the result from $\{SASD\}_{N\zeta}$. The correlation coefficient can take values between -1 (anti-correlation) and 1 (correlation). Both $SLD_{C\alpha}$ and $SLD_{C\beta}$, which employ the coarsest approximations and are widely used, correlate poorly (0.70 and 0.74, respectively, where a value of 1 represents perfect correlation). Interestingly, although $SASD_{N\zeta}$ measures curved paths, it also yields poor performance (0.73). This is because the increase in precision gained by measuring a SASD instead of an SLD is counterbalanced by a large uncertainty about Nζ location (*i.e.* the position of Nζ in the crystal structure correlates poorly with its best possible linking position). Much better performance is obtained using $\{SLD\}_{N\zeta}$ (0.82), while the best correlation is obtained by the $SASD_{C\beta}$ metric (0.86) that benefits from the slower but more accurate SASD measure.

Measurement of execution times

For each distance metric, we measured the execution times of DynamXL while analysing all proteins in our PiQSi database benchmark. All calculations were performed on a workstation equipped with Intel i7-3770 3.40 GHz cores. Only lysine pairs at a distance <24 Å were assessed, and we expect longer execution times when measuring longer SASDs. For metrics accounting for side chain flexibility, i.e. $\{SLD\}_{N\zeta}$ and $\{SASD\}_{N\zeta}$, σ = 4 $Å^2$ was used (see Fig. S2D for a benchmark reporting on the effect of σ on execution

times). All SASD metrics require a preprocessing phase to identify accessible regions around the protein taking on average 27.66 s (depending of protein size and shape, see above). Average execution times for the proteins in the PiQSi database, excluding this preprocessing time but including the time required to identify side chains alternate conformations, are given in Table S3.

DATA AND SOFTWARE AVAILABILITY

DynamXL is available for download together with a user manual and example files at http://dynamXL.chem.ox.ac.uk.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
DynamXL software and manual	http://dynamXL.chem.ox.ac.uk	DynamXL

Figure 1

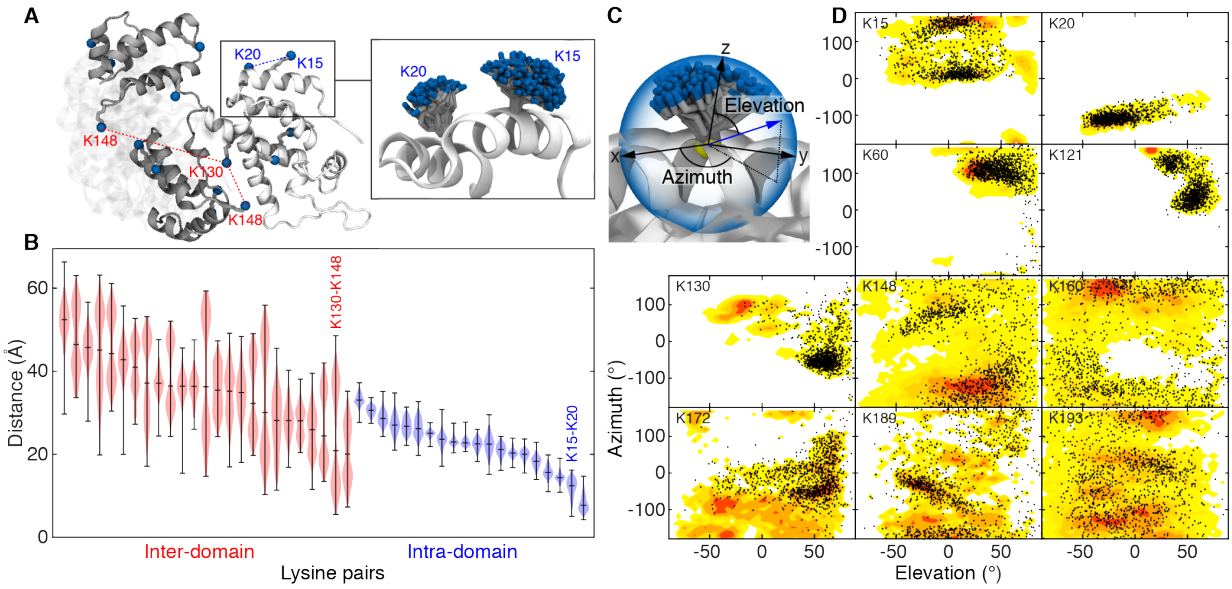

A
K20
K15
K148
K130
K148
K20
K15
B
Distance (Å)
60
40
20
0
K130-K148
K15-K20
Inter-domain
Intra-domain
Lysine pairs
C
z
Elevation
x
y
Azimuth
D
K15
K20
K60
K121
K130
K148
K160
K172
K189
K193
100
0
-100
Azimuth (°)
-50
0
50
Elevation (°)

Figure 2

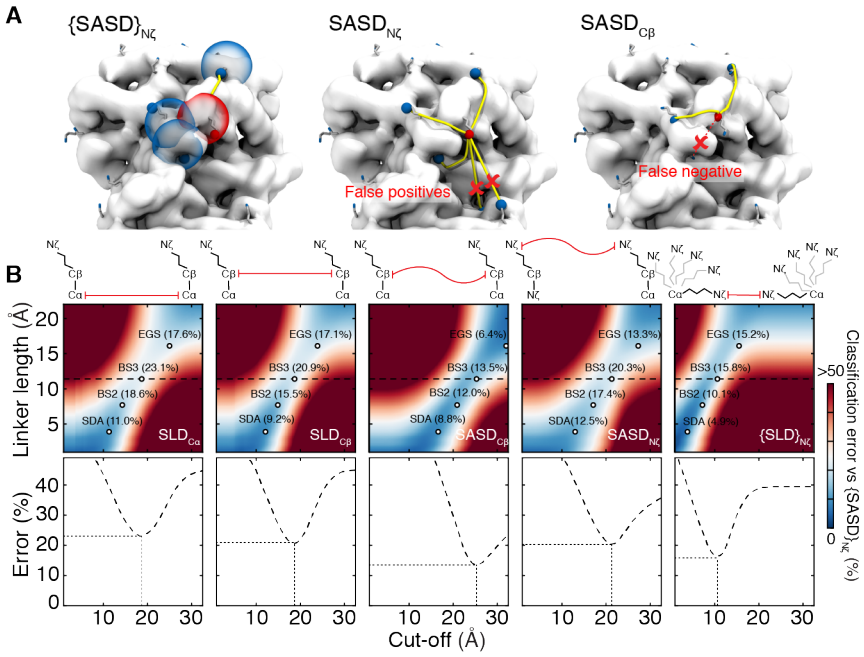

A
{SASD}Nζ
SASDNζ
SASDCβ
False positives
False negative
B
Linker length (Å)
Error (%)
Cut-off (Å)
EGS (17.6%)
BS3 (23.1%)
BS2 (18.6%)
SDA (11.0%)
SLDCα
EGS (17.1%)
BS3 (20.9%)
BS2 (15.5%)
SDA (9.2%)
SLDCβ
EGS (6.4%)
BS3 (13.5%)
BS2 (12.0%)
SDA (8.8%)
SASDCβ
EGS (13.3%)
BS3 (20.3%)
BS2 (17.4%)
SDA(12.5%)
SASDNζ
EGS (15.2%)
BS3 (15.8%)
BS2 (10.1%)
SDA (4.9%)
{SLD}Nζ
>50
0
Classification error vs {SASD}Nζ (%)

Figure 3

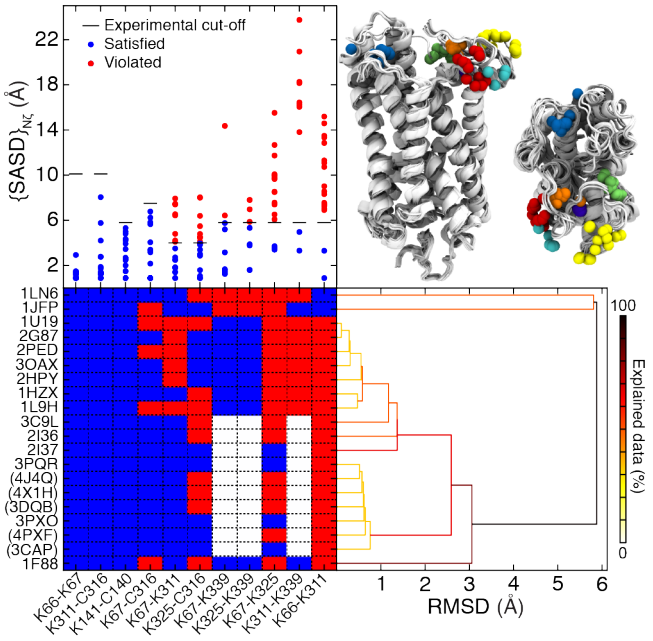
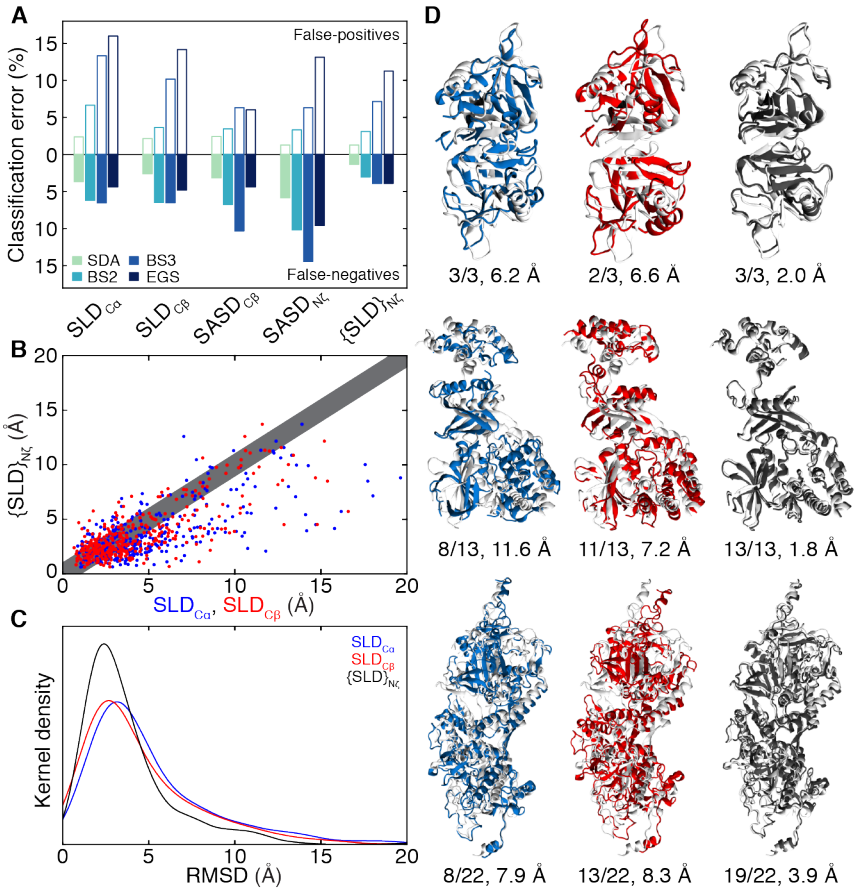

Experimental cut-off
Satisfied
Violated
{SASD}$_{N\zeta}$ (Å)
22
18
14
10
6
2
1LN6
1JFP
1U19
2G87
2PED
3OAX
2HPY
1HZX
1L9H
3C9L
2I36
2I37
3PQR
(4J4Q)
(4X1H)
(3DQB)
3PXO
(4PXF)
(3CAP)
1F88
K66-K67
K311-C316
K141-C140
K67-C316
K67-K311
K325-C316
K67-K339
K325-K339
K67-K325
K311-K339
K66-K311
1
2
3
4
5
6
RMSD (Å)
100
0
Explained data (%)

Figure 4


A
Classification error (%)
False-positives
False-negatives
SDA
BS2
BS3
EGS
SLD Cα
SLD Cβ
SASD Cβ
SASD Nζ
{SLD}Nζ
B
{SLD}Nζ (Å)
SLDCα, SLDCβ (Å)
C
Kernel density
RMSD (Å)
SLDCα
SLDCβ
{SLD}Nζ
D
3/3, 6.2 Å
2/3, 6.6 Å
3/3, 2.0 Å
8/13, 11.6 Å
11/13, 7.2 Å
13/13, 1.8 Å
8/22, 7.9 Å
13/22, 8.3 Å
19/22, 3.9 Å

- SUPPLEMENTAL INFORMATION -

Accommodating protein dynamics in the modelling of chemical cross-links

Matteo T. Degiacomi*[#], Carla Schmidt[†], Andrew J. Baldwin, Justin L.P. Benesch*

Department of Chemistry, Physical and Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford, OX1 3QZ, U.K.

* Correspondence to: matteo.t.degiacomi@durham.ac.uk and justin.benesch@chem.ox.ac.uk

[#] Present address: Chemistry Department, Durham University, South Road, Durham, DH1 3LE, UK

[†] Present address: Interdisciplinary research center HALOmem, Martin Luther University Halle-Wittenberg, Kurt-Mothes-Str. 3, 06120 Halle, Germany

Supplemental Tables

Cross-linker	Spacer length	$SLD_{C\alpha}$	$SLD_{C\beta}$	$SASD_{C\beta}$	$SASD_{N\zeta}$	$\{SLD\}_{N\zeta}$
SDA	3.9	11.0 (9.2%)	12.1 (9.2%)	16.6 (8.8%)	13.0 (12.5%)	3.8 (4.9%)
BS2G	7.7	14.3 (18.6%)	14.9 (15.5%)	20.9 (12.0%)	17.1 (17.4%)	7.2 (10.1%)
DMP	9.2	15.8 (21.5%)	16.1 (18.6%)	22.3 (12.9%)	18.8 (19.3%)	8.6 (12.7%)
DMS	11.0	18.4 (22.9%)	18.1 (20.7%)	24.9 (13.6%)	21.2 (20.4%)	10.1 (15.3%)
BS3, DSS	11.4	18.7 (23.1%)	18.7 (20.9%)	25.3 (13.5%)	21.3 (20.3%)	10.6 (15.7%)
EGS	16.1	25.0 (17.6%)	23.9 (17.1%)	32.0 (6.4%)	27.3 (13.3%)	15.5 (15.2%)

Table S1, related to Figure 2. Suggested cut-off distance for several commonly used lysine-specific or heterobifunctional cross-linkers and different measurement methods. All distances are reported in Å, and the error with respect to a classification obtained by using the $\{SASD\}_{N\zeta}$ is reported in parentheses. The cross-linkers investigated are: SDA (succinimidyl 4,4'-azipentanoate); BS2G (bis(sulfosuccinimidyl) 2,2,4,4-glutarate); DMP (Dimethyl pimelimidate•2HCl); DMS (Dimethyl suberimidate•2HCl); BS3 (bis(sulfosuccinimidyl)suberate); DSS (Disuccinimidyl suberate); and EGS (Ethylene glycol bis (succinimidylsuccinate)). All measurements include computational error (see STAR Methods). As such, classification by the reference $\{SASD\}_{N\zeta}$ was obtained by accounting for a 11.8% computational error in distance measurement (i.e. including shortest path algorithm, discretization at σ = 4 $Å^2$, and grid location errors). The $SASD_{C\alpha}$ and $SASD_{C\beta}$ measures include an error of 5.7% (shortest path algorithm and grid location errors), and the $\{SLD\}_{N\zeta}$ metric included an error of 6.1% (discretization error using $\sigma = 4$ $Å^2$).

Metric	Average execution time per protein (s)	Average execution time per distance (s)
$SLD_{C\alpha}$	$(8.74 \pm 6.50)\times10^{-5}$	$(9.15 \pm 63.2)\times10^{-7}$
$SLD_{C\beta}$	$(2.31 \pm 3.14)\times10^{-4}$	$(9.29 \pm 18.8)\times10^{-7}$
$SASD_{C\beta}$	18.99 ± 30.84	0.17 ± 0.18
$SASD_{N\zeta}$	22.35 ± 37.146	0.16 ± 0.13
$\{SLD\}_{N\zeta}$	3.32 ± 5.09	0.012 ± 0.01
$\{SASD\}_{N\zeta}$	3505.05 ± 6226.71	11.68 ± 15.04

Table S2, related to Figure 2. Execution time for different distance metrics in DynamXL. Average execution times for the proteins in the PiQSi database on an Intel i7-3770 3.40 GHz core for each distance metric. The average execution time for the measurement of a single distance was obtained by normalizing the total execution time by the number of lysine pairs at a distance <24 Å for each protein. Means ± one standard deviation are reported.

Cross-linker	**Measurement method**	**Best models' rank** (a)	**Best models' % of satisfied restraints** (b)	**% Cases with all restraints satisfied** (c)	**% Cases with high quality model** (d)	**Best models' RMSD** (e)
SDA	$SLD_{C\alpha}$	31.1	97.0 ± 11.6	52.4	31.9	5.0 ± 3.7
	$SLD_{C\beta}$	24.0	98.0 ± 9.5	70.0	41.8	4.7 ± 4.0
	$\{SLD\}_{N\zeta}$	10.6	99.2 ± 5.1	88.9	51.5	3.9 ± 2.9
BS3	$SLD_{C\alpha}$	22.5	69.7 ± 28.25	34.3	36.6	4.7 ± 3.1
	$SLD_{C\beta}$	19.6	72.0 ± 26.9	34.9	39.3	4.5 ± 3.1
	$\{SLD\}_{N\zeta}$	14.6	90.9 ± 17.5	68.7	44.5	4.0 ± 2.8

Table S3, related to Figure 4. Docking benchmark results. Measuring distances with $\{SLD\}_{N\zeta}$ leads to better models, compared to alternative SLD metrics. It is also noticeable that in the case of $\{SLD\}_{N\zeta}$, using the shorter SDA cross-linker leads to better performances than using BS3. For $SLD_{C\alpha}$ and $SLD_{C\beta}$, on the other hand, docking performances deteriorate. This is because the shorter the cross-linker, the more relevant the relative side-chain orientation becomes. Note that all values reported in this table display a dependence on the number of imposed distance restraints (Fig. S4).

(a) Average rank for the model having the smallest RMSD against the known dimeric structure.

(b) Average percentage of satisfied restraints by the model having the smallest RMSD.

(c) Percentage of models in the benchmark dataset satisfying all restraints.

(d) Percentage of best models having an RMSD < 3 Å.

(e) Average RMSD of best models.

Supplemental Figures

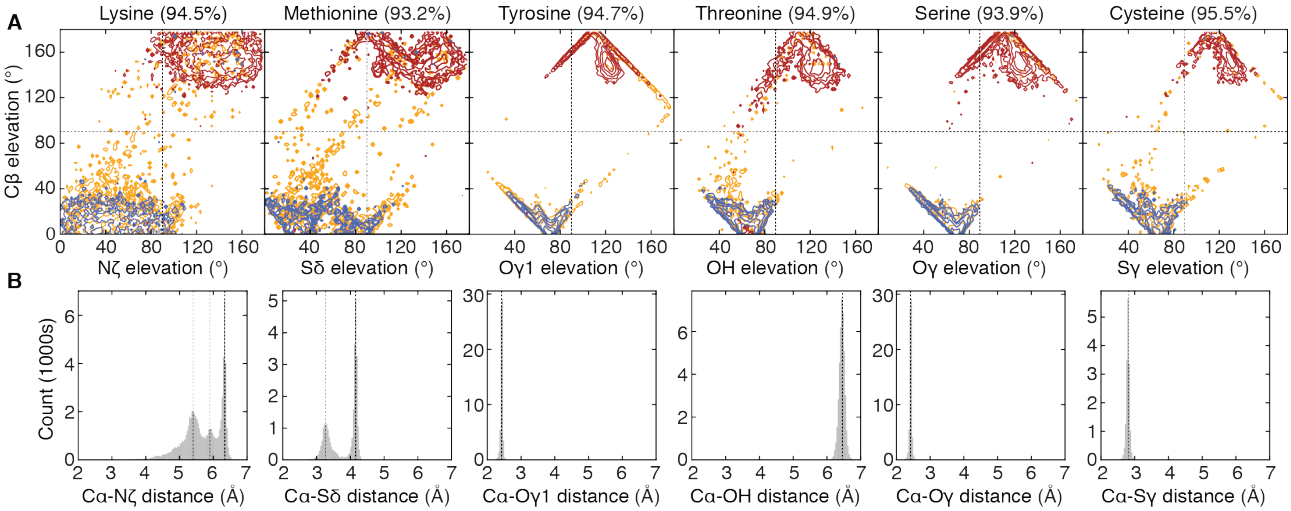


Figure S1, related to Figure 1. Arrangement of amino acid side chains. **A** Alignment with respect to the backbone plane for commonly cross-linked amino acids. For each amino acid in all the structures in the PiQSi database, we calculated the orientation with respect of the normal to the backbone plane (O, C and N atoms) and reported the resulting elevation (Fig. 1C). Elevation distributions for different secondary structure elements are represented separately (helix in red, coil in yellow and sheet in blue). In >93% of the cases (percentages indicated in parentheses), the linking atom is found on the same side of the backbone as its associated Cβ. This indicates that approximating the side-chain accessible space as a hemisphere incorporates the vast majority of the linking atom's accessible space. **B** Distribution of distances between the Cα and linking atoms due to rotations about the side-chain bonds. In the case of lysines, the distance distribution features three distinct peaks at 5.4, 5.9 and 6.3 Å (with a maximal length of 6.5 Å). These distances were selected as radii for our concentric spheres to sample the space accessible to lysine residues.

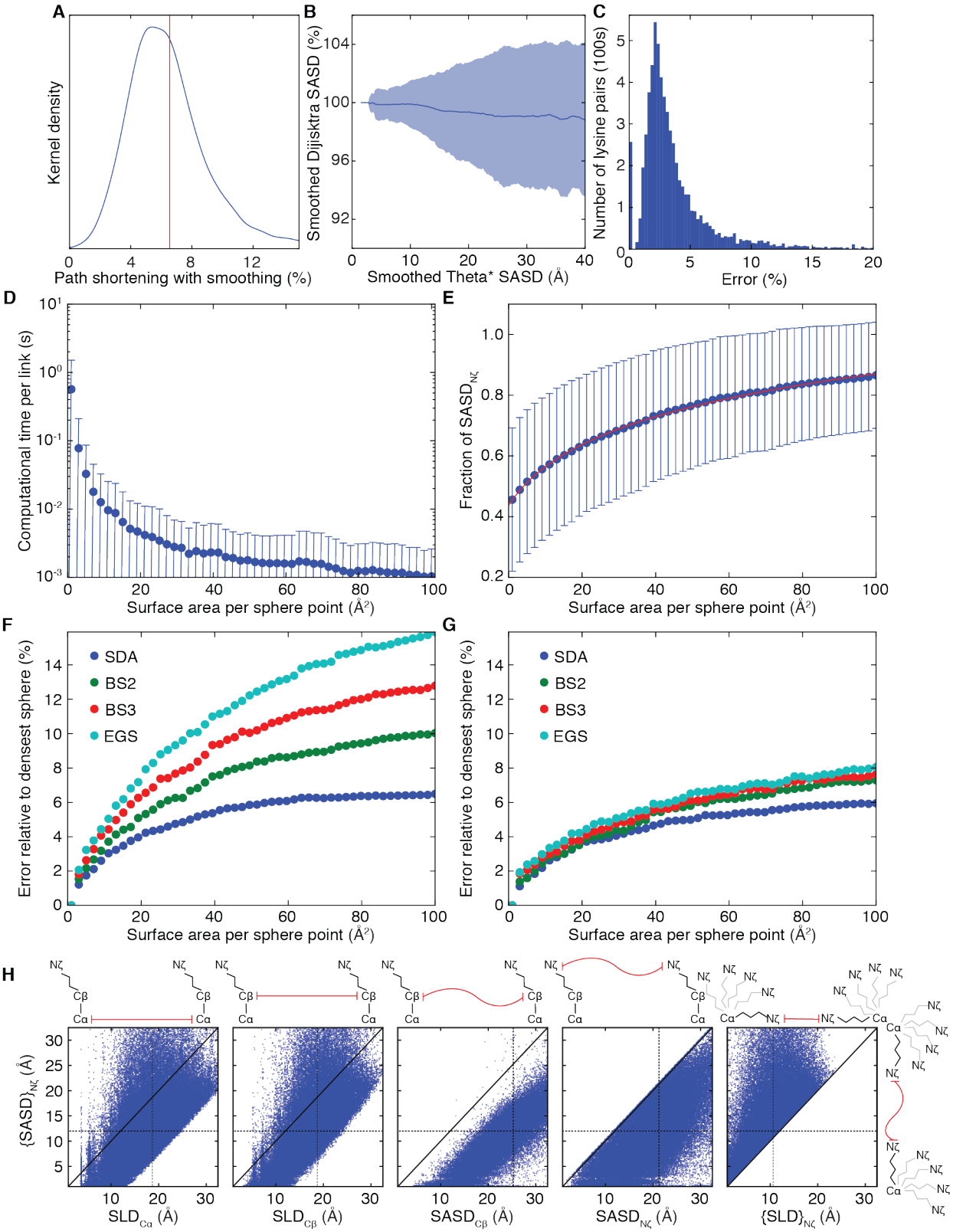


Figure S2, related to Figure 2. Assessment of errors in SASD measurement. A Analysis of path smoothing on SASD. The Theta* algorithm is not restricted to travelling through graph nodes if there is line-of-sight between two distant nodes. As such, shorter paths than those given by traditional shortest-path algorithms can be returned. Still, changes in direction are restricted to nodes, resulting therefore in unnaturally "angular" paths. These are therefore smoothed, obtaining distances on average 6.5% shorter (red

line) than the original path returned by Theta*. **B** Comparison of smoothed Theta* lengths against smoothed Dijkstra ones. On average, distances are comparable, although for long distances (>20 Å), Theta* returns on average slightly (<1%) shorter distances. This is because this algorithm exploits the available search space more effectively, as it is not limited to a search within the mesh grid. The shaded area corresponds to ± one standard deviation, is dependent on SASD, and is on average 2.1% in our region of interest (<24 Å). **C** The error caused by grid location with respect to the protein. Measurements are performed multiple times for every lysine pair, perturbing the position of our 1-Å mesh grid by 0.25 Å in the *x*, *y* and *z* directions. On average, an error equal to 3.6% is expected. **D** Increasing the number of hemisphere sampling points (i.e. reducing the surface-per-point σ) leads to an increase in computational times. The error bars correspond to ± one standard deviation. **E** Relationship between $SASD_{N\zeta}$ and $\{SASD\}_{N\zeta}$ using a range of σ values. The more detailed the spheres, the shorter the resulting path. The relationship between σ and path-shortening can be fitted with a shifted logarithmic function (red line). The error bars correspond to ± one standard deviation. **F** Using $\sigma = 1$ Å^2 as reference, classification errors for four lysine cross linkers can be estimated. Classification error is related to both the cross-linker length used (larger error for longer linkers) and σ. **G** By applying a correction derived from the shifted logarithmic fitting function (from **E**) to all measured distances, classification errors can be significantly reduced, especially in the case of the longer cross-linkers. **H** For each selected lysine-pair in PiQSi database, scatter plots report a comparison of their distances measured with a range of alternative distance metrics against the most accurate, $\{SASD\}_{N\zeta}$. On each plot, dashed lines indicate the most suitable cut-off for a BS3 cross-linker. In the case of $\{SASD\}_{N\zeta}$, the cut-off is set at 13 Å, i.e. the length of the cross-linker plus computational error (horizontal dashed line). For all other metrics, the cut-off distance (vertical dashed lines) is selected so that the classification error, the sum of false-positives (points in top left quadrant) and false-negatives (points in bottom right quadrant), is minimized.

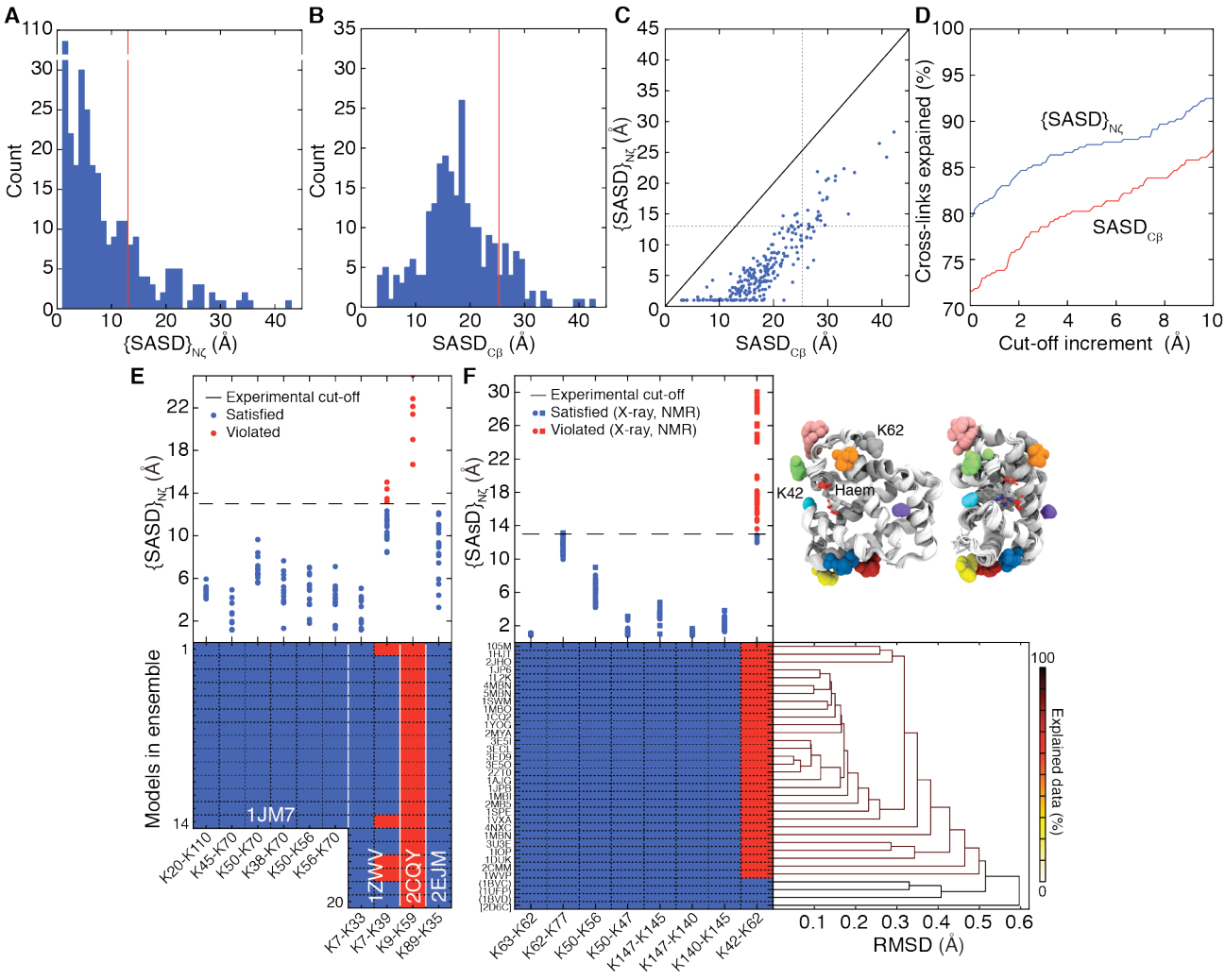


Figure S3, related to Figure 3. Comparison of SASD measurements against experimentally determined cross-links. A A histogram of distances measured for all the BS3 cross-links deposited in the XLdb, calculated using $\{SASD\}_{N\zeta}$. In this test, distances have been measured on a single protein structure per cross-link, even if multiple instances exist in the PDB. The criterion used to accept or reject a cross-link is indicated in red. A $\{SASD\}_{N\zeta}$ measure would confirm the cross-link if <13 Å (i.e. BS3 length equal to 11.4 Å, plus theoretical error, see Table S1). 79.7% of experimentally determined cross-links are confirmed as a result. **B** $SASD_{C\beta}$, by contrast, would accept the link if <25.3 Å (the length found to minimize classification error, see Table S1). Only 69.1% links are confirmed in this case. **C** Comparison between $SASD_{C\beta}$ and $\{SASD\}_{N\zeta}$. In nearly all cases, if a distance is accepted by the former method, it will be accepted by the latter too (top left quadrant). Rejected cross-links can only be explained by either protein backbone flexibility, or by cross-links wrongly assigned in the experiment. **D** When testing our distance metrics against XLdb data (panels **A**-**C**), $\{SASD\}_{N\zeta}$ accounted explicitly for side-chain flexibility in the given protein structures, while

for $SASD_{C\beta}$ we selected the cut-off minimizing the classification error caused by implicitly accounting for such flexibility. In the absence of alternative protein conformations, the distance cut-off of every metric can be increased in order to accommodate for backbone and domain movements. We tested the performance of $\{SASD\}_{N\zeta}$ and $SASD_{C\beta}$ in explaining XLdb data when incrementing their "ideal cut-offs" by further amounts. We observe that, at all increments, the $\{SASD\}_{N\zeta}$ metric always explains more cross-links than $SASD_{C\beta}$, even when extremely large cut-off increments are used. **E** Analysis of cross-links against NMR ensembles. The representations are analogous to those in Fig. 3. **F** Analysis of myoglobin cross-links against an ensemble of different crystal structures. The representations are analogous to those in Fig. 3. Only apo-myoglobin (indicated in parentheses) structures and a structure featuring a structural isomer of Haem, iron propicene (square brackets), can fully explain the experimental data. Clustering reveals that the structures are highly similar, the most different ones being structures of apo-myoglobin, and myoglobin bound to iron propicene.

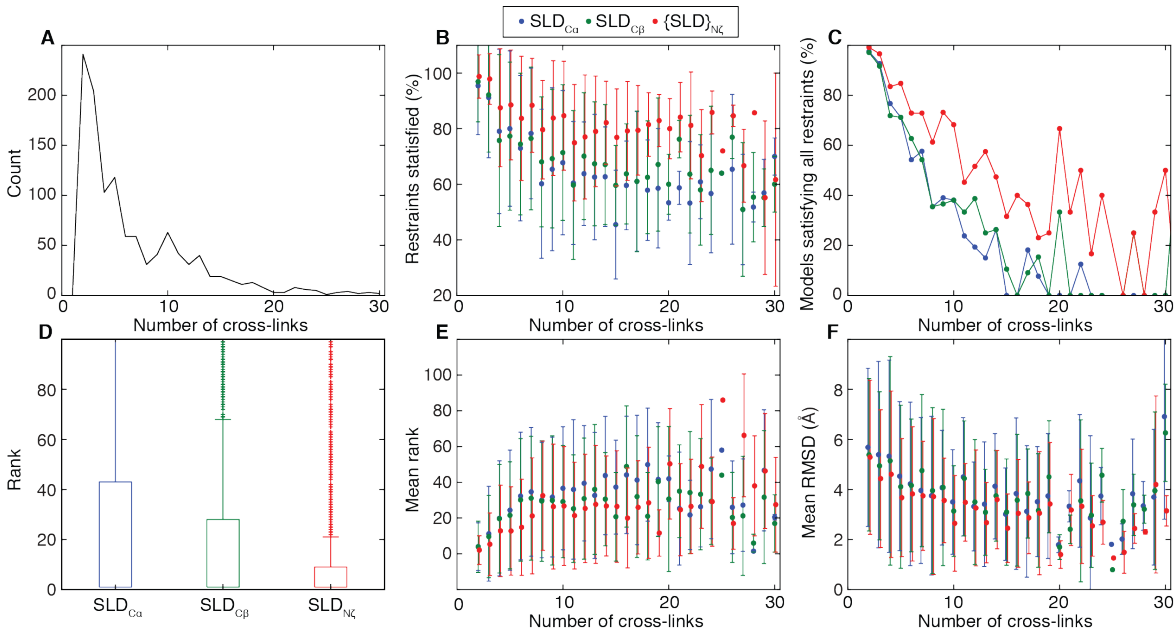


Figure S4, related to Figure 4. Comparing the performance of different distance metrics for docking using BS3 cross-links as restraints. **A** We prepared a dataset of 1180 simulated BS3 cross-linking cases, from 245 different dimers. The plot shows the number of docking cases as a function of the number of cross-links they employ as distance restraints. **B** Number of restraints satisfied by the best models, as a function of the total numbers of cross-links provided. $\{SLD\}_{N\zeta}$ identifies models satisfying a higher number of restraints. The error bars correspond to ± one standard deviation. **C** Percentage of best models satisfying all restraints, per total number of cross-links. The best models produced using $\{SLD\}_{N\zeta}$ satisfy all distance restraints more often. The decreasing trend indicates that the more cross-links employed, the higher the likelihood that at least one will be impossible to satisfy. **D** Boxplots indicating the ranking of the solution with lowest RMSD, for the three tested distance metrics. Using $\{SLD\}_{N\zeta}$ reduces the likelihood of having the model with smallest RMSD ranking poorly. **E** Decomposition of the ranking of the best model in each case, as a function of the total number of cross-links. The error bars correspond to ± one standard deviation. **F** RMSD of best model against known structure, as a function of the total number of cross-links. A small improvement is observed when using $\{SLD\}_{N\zeta}$. The error bars correspond to ± one standard deviation. See also Table S3.