Structure-based validation can drastically under-estimate error rate in proteome-wide cross-linking mass spectrometry studies

Kumar Yugandhar[1,2], Ting-Yi Wang[1,2], Haiyuan Yu[1,2,*]

[1]Department of Computational Biology, Cornell University, Ithaca, New York,14853, USA

[2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, New York, 14853, USA

[*]To whom correspondence should be addressed. Tel: 607-255-0259; Fax: 607-255-5961; Email:haiyuan.yu@cornell.edu

Abstract

Cross-linking mass spectrometry (XL-MS) is an increasingly popular technique for capturing large-scale interactions and their structural dynamics. Almost all proteome-wide XL-MS studies to date validate their identified cross-links by mapping them onto the available 3D structures of representative complexes such as ribosome and proteasome. Here, we theoretically demonstrate and experimentally confirm that such a structure-based validation approach can drastically underestimate the underlying error rate. Given the broad use and interest in these XL-MS datasets, the unexpected high number of false positives will severely hinder the development of the field and the utility of these datasets. Thus, it is of utmost importance to bring this problem to the attention of the field.

Cross-linking mass spectrometry (XL-MS) is a powerful platform to unveil protein interactions and capture their structural dynamics[1]. Development of efficient MS-cleavable chemical cross-linkers such as disuccinimidyl sulfoxide (DSSO)[2], disuccinimidyl dibutyric urea (DSBU)[3] and protein interaction reporters (PIRs)[4] has expanded the applications of XL-MS from studying individual functional complexes to discovering proteome-wide interactions and their structural dynamics. In addition to using false discovery rate (FDR) for filtering the list of cross-link identifications at expected quality, available three-dimensional (3D) structures are utilized for their validation and quality assessment by almost all proteome-wide XL-MS studies[5, 6]. Here, we demonstrate fundamental flaws in such a structure-based quality assessment approach that drastically under-estimates the error rates of large-scale XL-MS datasets.

In order to assess the quality of proteome-wide XL-MS datasets, researchers map the cross-links onto available 3D structures of highly abundant and representative complexes such as ribosome and proteasome. Then, they use the fraction of cross-linked residue pairs that satisfy the theoretical distance constraint of the cross-linker (e.g., 30Å for DSSO) to evaluate the dataset. While this approach may provide a sense of quality about the intraprotein cross-links, it suffers from massive underestimation of false positives in case of the interprotein cross-links (Fig. 1a), which are key for inferring novel protein-protein interactions and modeling 3D structure for functional complexes. In other words, such approach would selectively pre-filter only highly likely true positives for the distance-based validation, ignoring the potential false positives.

Theoretically, the probability of both peptides of an interprotein cross-link from a human proteome-wide experiment (~20,000 proteins) being mapped to the same protein complex structure consisting of 100 subunits is $\sim 5\times10^{-3}$ (99/19999). The probability would be much lower for the currently used representative complexes such as ribosome (76 subunits: PDB ID 5T2C) and proteasome (34 subunits: PDB ID 5GJQ). Considering such a low probability, any interprotein cross-link mapped to different subunits of a single complex structure is highly likely to be a true positive identification. On the other hand, most false positives will have only one peptide mapped to the complex structure (Fig. 1a). Because the current structural-mapping approaches explicitly validate crosslinks that have both peptides mapped to the same complex structure, they tend to result in a massive underestimation of error rate for proteome-wide XL-MS datasets.

Consequentially, these methods may erroneously annotate artifacts as novel interactions, resulting in less reliable experimental datasets for further studies.

To validate our theory experimentally, we performed a proteome-wide human XL-MS experiment using MS-cleavable cross-linker DSSO on K562 cell lysate (25 fractions). Next, in order to generate three sets of cross-links with drastically different qualities, we ran XlinkX search engine (Proteome Discoverer 2.2) using three criteria of decreasing stringency (1% FDR with ΔXlinkX score ≥50, 1% FDR, and 10% FDR). As shown in Fig, 1b panel (i), at 1% FDR with ΔXlinkX score ≥50, only 70 interprotein crosslinks were identified; whereas at 10% FDR, 2,925 were identified (we intentionally chose 10% FDR to obtain a low-quality set of cross-links with many false positives). Further, we mapped the interprotein cross-link residue pairs from these three sets separately onto the 3D structures of human ribosome and proteasome. We then calculated the percentage of mapped residue pairs that satisfied DSSO's theoretical constraint (≤30Å). We observed that there was no significant difference (all $P > 0.59$) among the three datasets in terms of their percentage of satisfying residue pairs (Fig. 1b panel (ii)), even though the overall quality of these three sets are drastically different by design. Our experimental results confirm that the current structure-mapping approach overestimates the fraction of true positives and thereby fails to capture the underlying error rate. Furthermore, it also indicates an urgent need for additional measures to estimate the quality of proteome-wide cross-linking datasets.

To further illustrate the problem, we used a comparative quality metric called "Precision", which is often used in machine learning[7, 8]. Some of the previous studies utilized the known protein-protein interaction networks to visualize and infer biological insights from XL-MS datasets[9, 10]. Here we utilized the known set of interactions to calculate precision for comparative quality assessment among the three datasets:

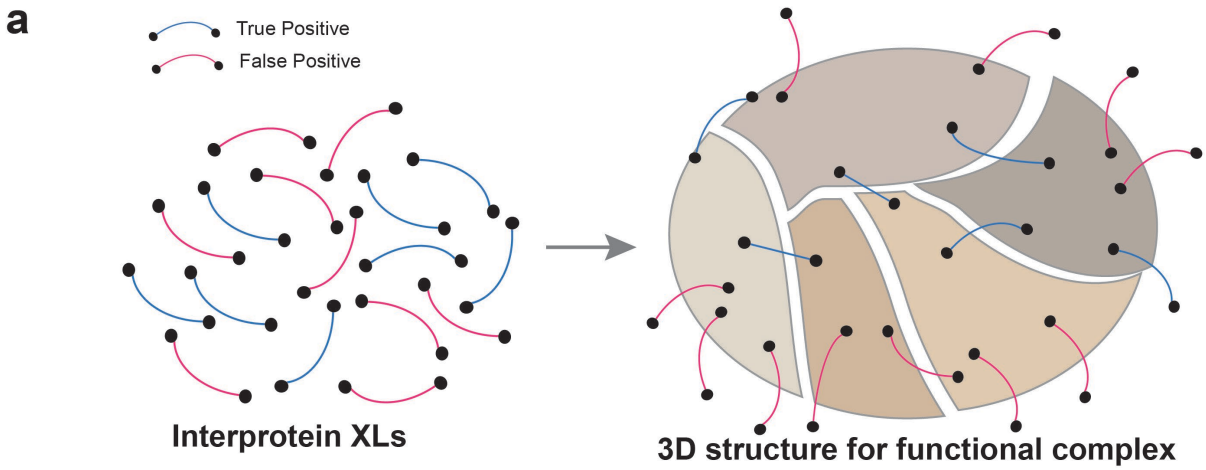
$$\text{Precision (\%)} = \frac{\text{Number of true positives}}{\text{Total number of postives}} X\ 100 \qquad (1)$$

where, 'positives' refer to all the identified interprotein cross-links, and 'true positives' refer to the interprotein cross-links between known interacting proteins. We calculated precision for the three datasets shown in Fig. 1b, utilizing a comprehensive set of 403,355 experimentally-confirmed human binary and co-complex protein interactions compiled from seven primary databases[11] (namely DIP[12], BioGRID[13], MINT[14], iRefWeb[15], IntAct[16], HPRD[17], and MIPS[18]). It should be

noted that, because a large fraction of true protein interactions is yet to be discovered, precision can only be used as a relative estimate and should not be used as an absolute measure for data quality.

As shown in Fig. 1b panel (i), precision exhibits great agreement with the expected data quality of different data sets (at 1% FDR with ΔXlinkX score ≥50, precision is 80.0%; but at 10% FDR, precision is merely 8.2%; $P < 1\times10^{-20}$). It should be noted that the cross-links in Fig. 1b panel (ii) would have a precision of 100% by definition. Importantly, the precisions for interprotein cross-links where only one peptide was mapped onto the ribosomal or proteasomal structures across all three sets are abysmal (≤2.4%).

We acknowledge that this issue of drastically under-estimating quality by structure-based measurements would not be applicable for XL-MS studies on specific proteins and individual complexes if the cross-link search is performed against only proteins that are included in the experiment; however, it will be applicable if the search is performed against the entire proteome. Importantly, it is highly relevant for the increasingly popular proteome-wide XL-MS experiments[5, 6] and cross-linking immunoprecipitation MS (xIP-MS) studies[19]. In these studies, structure-based evaluation (which is used in almost all studies) does not provide an adequate assessment of the data quality and will make even low-quality datasets seem good. Given the broad use and interest in these XL-MS datasets[20, 21], the unexpected high number of false positives will severely hinder the development of the field and the utility of these datasets. Thus, it is of utmost importance to bring this problem to the attention of the field. Going forward, a more comprehensive and accurate quality assessment framework needs to be developed to aid in the rapid advancement of XL-MS technologies.



b

	(i) Interprotein XLs	(ii) Both XL sites mapped to the same PDB ID	(iii) Only one XL site mapped to the same PDB ID
1% FDR	1371 (*Precision*: 11.5%)	41 (Fraction of XLs ≤ 30Å: 80.5%)	126 (*Precision*: 2.4%)
1% FDR with Δ XlinkX Score ≥ 50	70 (*Precision*: 80.0%)	18 (Fraction of XLs ≤ 30Å: 77.8%)	2 (*Precision*: 0.0%)
10% FDR	2925 (*Precision*: 8.2%)	58 (Fraction of XLs ≤ 30Å: 75.9%)	255 (*Precision*: 2.4%)

Figure 1. Evaluation of 3D structure-based validation approaches for proteome-wide cross-linking mass spectrometry (XL-MS) datasets. (a) Illustration of our theory about the limitation of structure-mapping approach in validating the false positive XL identifications. (b) Assessment of data quality for XL-MS datasets (filtered using different quality filters from a single experiment) by mapping them onto the known 3D structures of human ribosome (PDB ID: 5T2C) and proteasome (PDB ID: 5GJQ). Panel (i) shows the total number of interprotein XLs (*Precision* for '1% FDR with ΔXlinkX score ≥50' set is significantly higher than that of '1% FDR' set ($P < 1\times10^{-20}$) and that of '10% FDR' set ($P < 1\times10^{-20}$)). Panel (ii) shows the number of interprotein XLs where both the residues in the pair are mapped to the same structure (Fraction of XLs within 30Å for '1% FDR with ΔXlinkX score ≥50' set is statistically indistinguishable to that of '1% FDR' set ($P = 0.81$) and that of '10% FDR' set ($P = 0.87$)). All the P values were calculated using a two-sided Z-test. Panel (iii) shows the number of XLs and the corresponding *Precision*, where only one of the two residues mapped to the same structure.

References

1. Yu, C. & Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Analytical Chemistry* **90**, 144-165 (2018).
2. Kao, A. et al. Development of a Novel Cross-linking Strategy for Fast and Accurate Identification of Cross-linked Peptides of Protein Complexes. *Molecular & Cellular Proteomics* **10**, M110.002212 (2011).
3. Müller, M.Q., Dreiocker, F., Ihling, C.H., Schäfer, M. & Sinz, A. Cleavable Cross-Linker for Protein Structure Analysis: Reliable Identification of Cross-Linking Products by Tandem MS. *Analytical Chemistry* **82**, 6958-6968 (2010).
4. Tang, X. & Bruce, J.E. A new cross-linking strategy: protein interaction reporter (PIR) technology for protein–protein interaction studies. *Molecular BioSystems* **6**, 939-947 (2010).
5. O'Reilly, F.J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nature Structural & Molecular Biology* **25**, 1000-1008 (2018).
6. Klykov, O. et al. Efficient and robust proteome-wide approaches for cross-linking mass spectrometry. *Nature Protocols* **13**, 2964-2990 (2018).
7. Meyer, M.J. et al. Interactome INSIDER: a structural interactome browser for genomic studies. *Nature Methods* **15**, 107 (2018).
8. Kovács, I.A. et al. Network-based prediction of protein interactions. *Nature Communications* **10**, 1240 (2019).
9. Keller, A., Chavez, J.D., Eng, J.K., Thornton, Z. & Bruce, J.E. Tools for 3D Interactome Visualization. *Journal of Proteome Research* **18**, 753-758 (2019).
10. Chavez, J.D. et al. Quantitative interactome analysis reveals a chemoresistant edgotype. *Nature Communications* **6**, 7928 (2015).
11. Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology* **6**, 92 (2012).
12. Smith, A.J. et al. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research* **32**, D449-D451 (2004).
13. Chatr-aryamontri, A. et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Research* **43**, D470-D478 (2014).
14. Palma, A. et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research* **40**, D857-D861 (2011).
15. Turner, B. et al. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* **2010** (2010).
16. Bridge, A. et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Research* **40**, D841-D846 (2011).
17. Keshava Prasad, T.S. et al. Human Protein Reference Database--2009 update. *Nucleic acids research* **37**, D767-D772 (2009).
18. Ruepp, A. et al. The MIPS mammalian protein–protein interaction database. *Bioinformatics* **21**, 832-834 (2004).
19. Makowski, M.M., Willems, E., Jansen, P.W.T.C. & Vermeulen, M. Cross-linking immunoprecipitation-MS (xIP-MS): Topological Analysis of Chromatin-associated Protein Complexes Using Single Affinity Purification. *Molecular & Cellular Proteomics* **15**, 854 (2016).
20. Ferber, M. et al. Automated structure modeling of large protein assemblies using crosslinks as distance restraints. *Nature Methods* **13**, 515 (2016).
21. Karaca, E., Rodrigues, J.P.G.L.M., Graziadei, A., Bonvin, A.M.J.J. & Carlomagno, T. M3: an integrative framework for structure determination of molecular machines. *Nature Methods* **14**, 897 (2017).

Supplementary Methods

Cell culture and whole cell lysate preparation

The K562 cells (ATCC® CCL-243™) were purchased from American Type Culture Collection (ATCC). The cells were maintained in the Iscove's Modified Dulbecco's Medium (IMDM) (ATCC) supplemented with 10% fetal bovine serum (FBS) (ATCC) at 37°C with humidified ambient atmosphere containing 5% CO_2. The K562 cells were collected and washed three times with cold PBS. The cells were then resuspended in cold buffer composed of 50 mM HEPES, 150 mM NaCl, pH 7.5 supplemented with Protease Inhibitor Cocktail (Roche). The resuspended cells were lysed on ice by sonication (Amplitude 10% for 5 sec and repeat 6 times), followed by centrifugation at 15,000 *g* for 10 min at 4ºC. The supernatant was collected and measured the protein concentration using Bio-Rad Protein Assay Dye (Bio-Rad).

Cross-linking of human proteome

DSSO (Thermo Fisher Scientific) was freshly prepared as a 50 mM stock solution by dissolving in anhydrous DMSO. The 1 mg/mL lysate of K562 cells was reacted with 1 mM DSSO in 50 mM HEPES buffer, 150 mM NaCl, pH 7.5 for 1 hour at room temperature. The cross-linking reactions were terminated by 50 mM Tris-Cl buffer, pH 7.5.

Processing of cross-linked samples for analysis

The DSSO-treated cell lysate was processed by being denatured in 1% sodium dodecyl sulfate (SDS), reduced by dithiothreitol (DTT), and alkylated with iodoacetamide, followed by precipitated in cold acetone-ethanol solution (acetone:ethanol:acetic acid=50:49.9:0.1, v/v/v). The precipitates were dissolved in 50 mM Tris-Cl, 150 mM NaCl, 2 M urea, pH 8.0 and digested by TPCK-treated trypsin (Worthington Biochemical Corporation) at 37ºC overnight. After digestion, the sample was acidified by 2% trifluoroacetic acid-formic acid solution, desalted through Sep-Pak C18 cartridge (Waters), and dried using SpeedVac™ Concentrator (Thermo Fisher Scientific).

Fractionation of cross-linked peptides by Strong Cation Exchange (SCX)

The 1 mg of desalted trypsin-digested sample was dissolved in 25% acetonitrile/0.1% formic acid (v/v) and passed through a Spin-X centrifuge tube filters (Corning). The sample was fractionated by a PolySULFOETHYL A column (5 µm, 200 Å, 2.1 x 200 mm; PolyLC) operated by a Dionex UltiMate 3000 Series instrument (Thermo Fisher Scientific). Two following buffers were used: 10 mM potassium phosphate monobasic in 25% acetonitrile, pH 3.0 (Buffer A) and 10 mM potassium phosphate monobasic/500 mM potassium chloride in 25% acetonitrile, pH 3.0 (Buffer B). The fractionation was performed at a flow rate of 200 µL/min using a linear gradient from 5-60% of Buffer B in 40 min and 60-100% of Buffer B in an additional 10 min. The fractions were collected at 1-min intervals. The fractions from 23 to 60 min were desalted using SOLA HRP SPE cartridges (Thermo Scientific). The eluted peptides were dried by speed vacuum and stored at -20ºC until LC-MS/MS analysis.

NanoLC-MS^n analysis

The fractionated samples were analyzed using UltiMate3000 RSLCnano (Dionex) coupled to an Orbitrap Fusion (Thermo Fisher Scientific) mass spectrometer. Each sample was loaded onto an Acclaim PepMap 100 C18 trap column (5 µm, 100 µm x 20 mm, 100 Å, Thermo Fisher Scientific) and separated on an Acclaim PepMap C18 nano column (3 µm, 75 µm x 25 cm, Thermo Fisher Scientific) by 5-40% B at 300 nL/min in 120 min. For MS data acquisition, the CID-MS2-MS3 workflow was used. The Orbitrap Fusion was operating in positive ion mode with nano spray voltage set at 1.7 kV and source temperature at 275°C. The MS1 precursors were detected in Orbitrap mass analyzer (375-1575 m/z and resolution = 60,000). The precursor ions with the charge of 4+ to 10+ were selected for CID-MS2 acquisition in Orbitrap mass analyzer (resolution = 30,000, AGC target = 5 x 10^4, precursor isolation width = 1.6 m/z, and maximum injection time = 100 ms) with the collision energy of CID at 25%. The peaks with a mass difference (Δ = 31.9721) in the CID-MS2 spectrum triggered acquisition of CID-MS3 spectra in Ion Trap with CID collision energy of 35% and AGC target of 2 x 10^4. All spectra were recorded by Xcalibur 3.0 software and Orbitrap Fusion Tune Application v. 2.1 (Thermo Fisher Scientific).

Data processing

Cross-links were identified using XlinkX software (Proteome discoverer 2.2).

Statistics

Statistical analyses were performed using a two-sided *Z*-test as indicated in the figure legend. Exact *P* values are provided for all compared groups.