nature methods
BRIEF COMMUNICATION
https://doi.org/10.1038/s41592-020-0959-9

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

Kumar Yugandhar[1,2], Ting-Yi Wang[1,2], Shayne D. Wierbowski[1,2], Elnur Elyar Shayhidin[1,2] and Haiyuan Yu[1,2] ✉

Thorough quality assessment of novel interactions identified by proteome-wide cross-linking mass spectrometry (XL-MS) studies is critical. Almost all current XL-MS studies have validated cross-links against known three-dimensional structures of representative protein complexes. Here, we provide theoretical and experimental evidence demonstrating that this approach can drastically underestimate error rates for proteome-wide XL-MS datasets, and propose a comprehensive set of four data-quality metrics to address this issue.

XL-MS is a powerful platform capable of unveiling protein interactions and capturing their structural dynamics[1]. The wealth of information from proteome-wide XL-MS approaches facilitates large-sale identification of protein–protein interactions[2,3], and high-throughput three-dimensional (3D) structural modeling of functional protein complexes[4–6]. With the increased throughput of these techniques, the number of false positive cross-links and incorrect interactions can quickly add up with just one large-scale XL-MS experiment, if one is not careful. Therefore, thorough quality assessment has become critically important.

It has been previously shown that the conventional false discovery rate (FDR) calculations for XL-MS can be susceptible to error propagation[7] (Supplementary Note 1). Currently, almost all proteome-wide XL-MS studies leverage available 3D structures of representative complexes for validation and quality assessment[8,9]. Here, we demonstrate fundamental flaws in this structure-based quality assessment approach that can drastically underestimate the error rates of large-scale XL-MS datasets.

In small-scale XL-MS studies, the fraction of cross-linked residue pairs that satisfy the maximum distance a given cross-linker can span (for example, 30 Å for disuccinimidyl sulfoxide (DSSO)[10]) provides meaningful insights into protein flexibility and the quality of the cross-links detected. In proteome-wide XL-MS studies, researchers extend this concept and use representative, highly abundant complexes such as the ribosome and the proteasome to estimate the quality of all cross-links reported. However, true positive and false positive cross-links in these large-scale studies are not equally likely to successfully map onto an existing 3D structure, leading to massive underestimation of false positives (Fig. 1a).

To theoretically demonstrate this, let us consider a reference protein complex structure consisting of 100 subunits. Because a false positive cross-link can be detected between any two random proteins within the proteome (~20,000 proteins for human proteome-wide experiments), for a given false positive with one of its ends mapped to the reference complex, the probability that the second end also maps to this complex by random chance is 5×10^{-3} (100/20,000). It should be noted that this probability would be even lower for the often-used ribosome (76 subunits: Protein Data Bank (PDB) ID 5T2C) or proteasome (34 subunits: PDB ID 5GJQ) complexes. However, these probabilities only hold for random peptide pairs (derived from false positive cross-links); true positive cross-links are much more likely to perfectly map to existing 3D structures. Conceptually, this is very similar to the fact that false positive cross-links are much more likely to be interprotein than intraprotein as shown by previous studies[11,12]. We expect that almost all false positive cross-links will have only one peptide mapped to the reference complex structure. The current structural-mapping approach explicitly considers only cross-links where both peptides map to the same complex structure, and, in doing so, it enriches for true positive cross-links and massively underestimates the error rates for proteome-wide XL-MS datasets. Consequentially, this validation approach may erroneously annotate artifacts as novel interactions, resulting in less-reliable experimental datasets for further studies.

To demonstrate our theory experimentally, we obtained a subset of 122 raw files from our recent proteome-wide human K562 XL-MS study[2]. Next, to generate three sets of cross-links with drastically different qualities, we ran the XlinkX search engine (Proteome Discoverer 2.2) using three criteria of increasing stringency ('10% FDR', '1% FDR' and '1% FDR with ΔXlinkX score ≥ 50'; see the Methods). As shown in Fig. 1b, at 10% FDR, a set of 35,561 interprotein cross-links were identified (we intentionally chose 10% FDR to obtain a low-quality set of cross-links with many false positives); 1% FDR yielded 16,591 interprotein cross-links; whereas '1% FDR with ΔXlinkX score ≥ 50' yielded 985 interprotein cross-links. We mapped the interprotein cross-link residue pairs from these three sets separately onto the 3D structure of the human proteasome following the conventional methodology. We then calculated the percentage of mapped residue pairs that satisfied DSSO's theoretical constraint (≤30 Å). We observed that there was no significant difference (all $P > 0.85$) among the three sets in terms of their percentage of residue pairs satisfying the distance constraint (Fig. 1c), even though the overall qualities of these three sets are drastically different by design. Additionally, we utilized our recently published search engine, MaXLinker[2], to repeat the analysis and observed similar results, confirming that these findings are software-independent (Extended Data Fig. 1a,b). We further re-analyzed raw files from two other publicly available studies representing different organisms (*Escherichia coli*[13] and mouse[10]) and

[1]Department of Computational Biology, Cornell University, Ithaca, NY, USA. [2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY, USA. ✉e-mail: haiyuan.yu@cornell.edu

cellular compartments (mitochondria[10]) (Fig. 2a,b and Extended Data Fig. 2a,b). These experimental results confirm that the current structure-mapping approach fails to capture the underlying error rate and indicate an urgent need for reliable metrics to estimate the quality of proteome-wide cross-linking datasets.

To address the pitfalls of the current validation approach, we propose the following comprehensive set of four measurements:

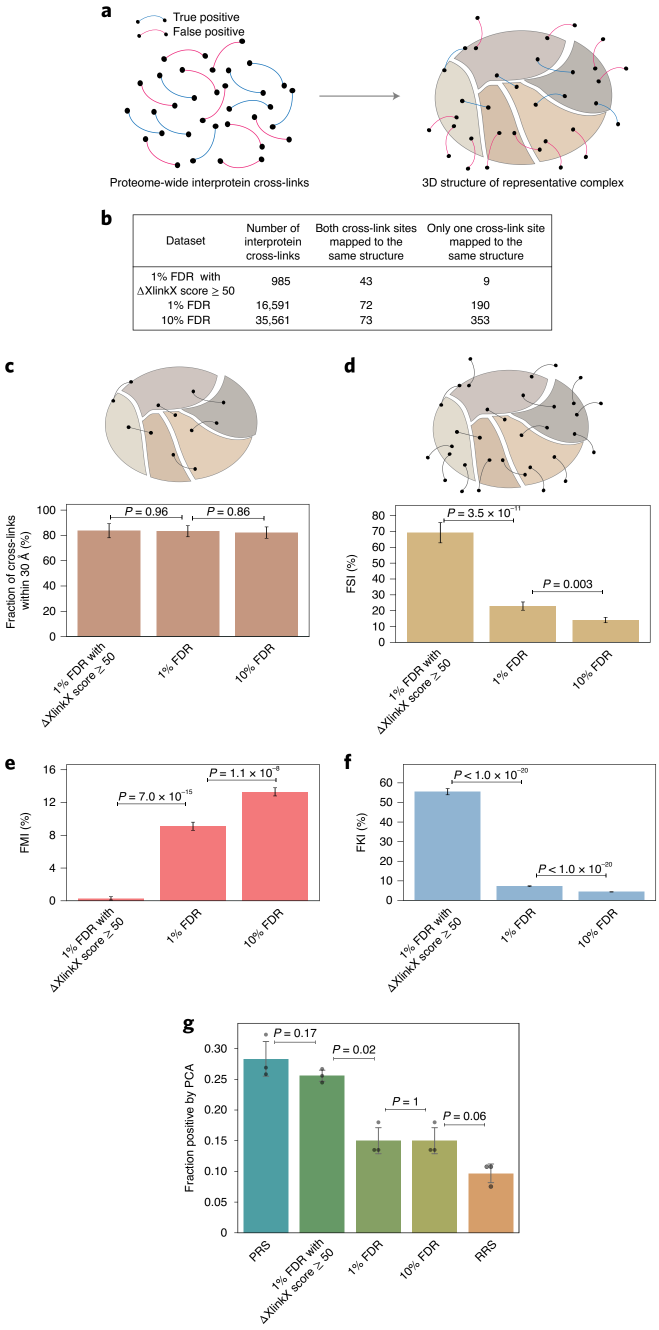
(1) *Fraction of structure-corroborating identifications (FSI)*: The current structure-based validation approach considers only those cross-links where both peptides mapped to the reference structure. Here, we propose FSI as an improved structure-based metric that uses the number of all interprotein cross-links with at least one peptide mapped to the reference structure, not just those with both peptides mapped, as the denominator (see the Methods).

(2) *Fraction of misidentifications (FMI)*: Including the proteome of an unrelated organism in the search database as an internal negative control can be an efficient way to independently assess the underlying error rate of the cross-link search algorithm[2,14,15] (see the Methods).

(3) *Fraction of interprotein cross-links from known interactions (FKI)*: Using previous knowledge of experimentally detected protein interactions to calculate the FKI provides a comparative quality estimate (see the Methods).

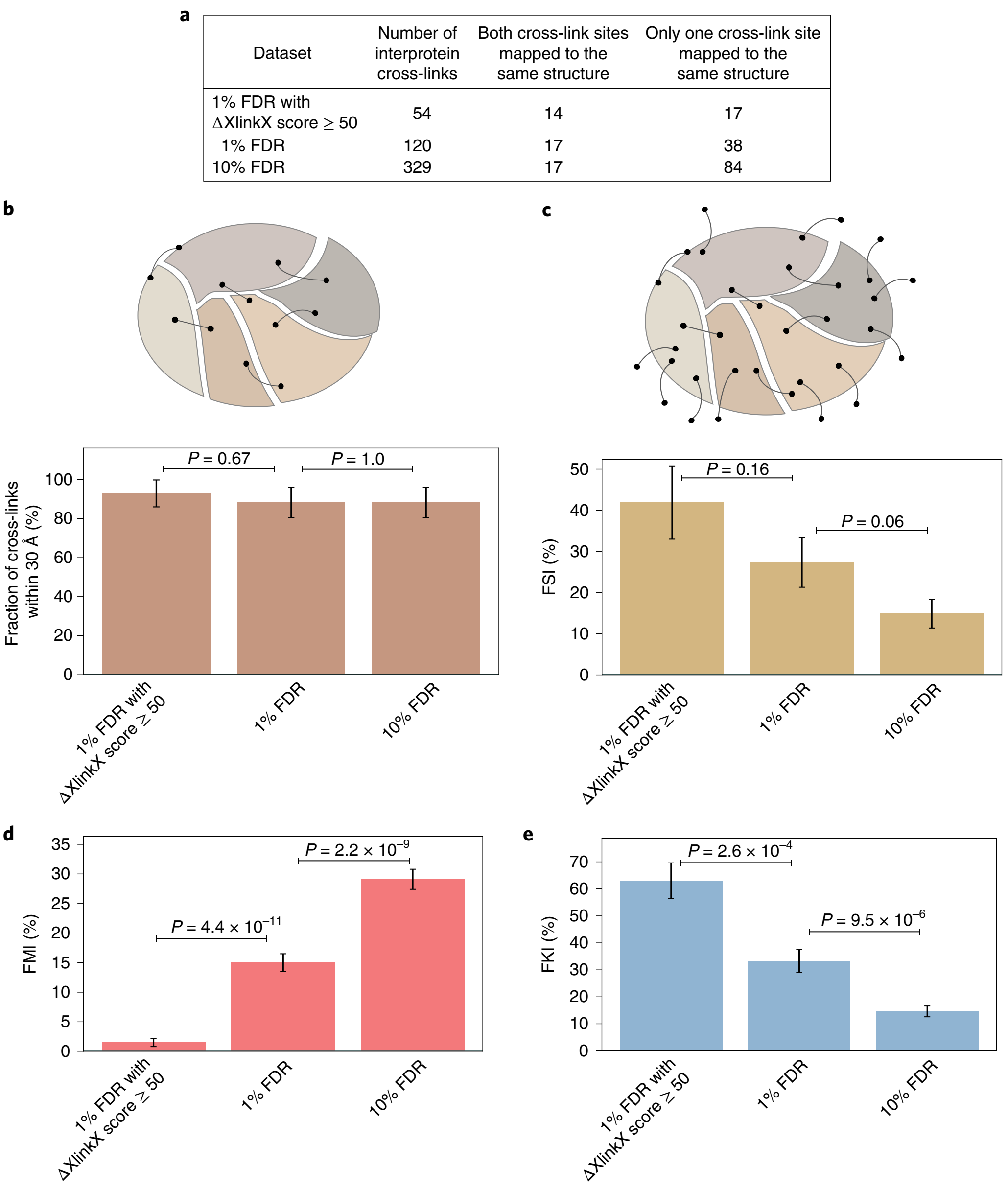
(4) *Fraction of validated novel interactions using orthogonal experimental assays*: It is essential to validate a representative set of novel interactions identified in proteome-wide XL-MS studies using an orthogonal experimental assay (for example, yeast two-hybrid (Y2H), protein complementation assay (PCA)), to ensure data quality and reproducibility (see the Methods). Furthermore, using a Bayesian framework[16,17] (Supplementary Note 2) and leveraging the validation rates among a positive reference set (PRS) of well-known interactions and a negative reference set (random reference set (RRS)) of random pairs, we can calculate the absolute precision of the novel interactions detected in an XL-MS study.

We next applied our proposed metrics on our human proteome-wide XL-MS results, and demonstrated how each of them efficiently captures the differences in data quality among the three filtered sets (Fig. 1d–g). Figure 1d shows that our improved structure-based metric, FSI, differentiates the three sets with statistical significance, which could not be achieved by the conventional structure-based approach (Fig. 1c). The results are consistent with our earlier theoretical expectation that applying more stringent quality filters would remove predominantly (likely false positive) cross-links with only one peptide mapped to the structure, and thereby result in higher FSI values (Fig. 1b,d). Furthermore, Fig. 1e



Dataset	Number of interprotein cross-links	Both cross-link sites mapped to the same structure	Only one cross-link site mapped to the same structure
1% FDR with ΔXlinkX score ≥ 50	985	43	9
1% FDR	16,591	72	190
10% FDR	35,561	73	353

Fig. 1 | Evaluation of the conventional 3D structure-based validation approach for proteome-wide XL-MS using human K562 DSSO XL-MS data[2]. **a**, In the current structure-mapping approach for validating cross-link identifications, most false positive cross-links only have one peptide mapped to the structure and are therefore ignored. **b**, Table showing the number of interprotein cross-links obtained at different filtering criteria, and upon mapping to a representative 3D structure of a human 26S proteasome. **c**, Fraction of cross-links satisfying the maximum distance constraint (≤30 Å) across the three sets, according to the conventional structure-based validation approach ($n = 43$ cross-links for '1% FDR with ΔXlinkX score ≥ 50'; $n = 72$ cross-links for '1% FDR'; $n = 73$ cross-links for '10% FDR'). **d**, FSI across the three sets ($n = 52$ cross-links for '1% FDR with ΔXlinkX score ≥ 50'; $n = 262$ cross-links for '1% FDR'; $n = 426$ cross-links for '10% FDR'). **e**, FMI across the three sets ($n = 668$ cross-links for '1% FDR with ΔXlinkX score ≥ 50'; $n = 3{,}029$ cross-links for '1% FDR'; $n = 4{,}957$ cross-links for '10% FDR'; see the Methods). **f**, FKI across the three sets ($n = 985$ cross-links for '1% FDR with ΔXlinkX score ≥ 50'; $n = 16{,}591$ cross-links for '1% FDR'; $n = 35{,}561$ cross-links for '10% FDR'). For **c**–**f**, *P* values were calculated using a two-sided *Z*-test. The error bars indicate ±s.e. of the proportion and the centers of the error bars indicate the proportion. **g**, Orthogonal experimental validation of a random subset of novel interactions from the three sets using PCA. PRS: mean fraction positive: 0.286; RRS: mean fraction positive: 0.098; '10% FDR': mean fraction positive: 0.152; '1% FDR': mean fraction positive: 0.152; '1% FDR with ΔXlinkX score ≥ 50': mean fraction positive: 0.258. The error bars indicate ±s.d. and the centers of the error bars indicate mean fraction positive; *P* values were calculated using a two-sided *t*-test on the log-transformed measurements ($n = 3$ independent experiments; see the Methods); 95% confidence interval; *t*-statistic 4.04 for '10% FDR' versus RRS, 7.20 for '1% FDR with ΔXlinkX score ≥ 50' versus '1% FDR', 2.13 for PRS versus '1% FDR with ΔXlinkX score ≥ 50'; 2 degrees of freedom.



Dataset	Number of interprotein cross-links	Both cross-link sites mapped to the same structure	Only one cross-link site mapped to the same structure
1% FDR with ΔXlinkX score ≥ 50	54	14	17
1% FDR	120	17	38
10% FDR	329	17	84

Fig. 2 | Demonstration of our set of validation metrics on a publicly available *E. coli* proteome-wide XL-MS dataset[13]**. a**, Table showing the number of interprotein cross-links obtained at different filtering criteria, and upon mapping to representative 3D structures. **b**, Fraction of cross-links satisfying the maximum distance constraint (≤30 Å) across the three sets, according to the conventional structure-based validation approach ($n=14$ cross-links for '1% FDR with ΔXlinkX score ≥ 50'; $n=17$ cross-links for '1% FDR'; $n=17$ cross-links for '10% FDR'). **c**, FSI ($n=31$ cross-links for '1% FDR with ΔXlinkX score ≥ 50'; $n=55$ cross-links for '1% FDR'; $n=101$ cross-links for '10% FDR'). **d**, FMI ($n=340$ cross-links for '1% FDR with ΔXlinkX score ≥ 50'; $n=553$ cross-links for '1% FDR'; $n=755$ cross-links for '10% FDR'). **e**, FKI ($n=54$ cross-links for '1% FDR with ΔXlinkX score ≥ 50'; $n=120$ cross-links for '1% FDR'; $n=329$ cross-links for '10% FDR'). For **b**–**e**, the *P* values were calculated using a two-sided *Z*-test and the error bars indicate ±s.e. of proportion.

reveals the exact same trend: FMI is significantly lower for the '1% FDR with ΔXlinkX score ≥ 50' set compared to the other two sets. Moreover, as shown in Fig. 1f, FKI exhibits great agreement with the expected data quality of different datasets (at '1% FDR with ΔXlinkX score ≥ 50', FKI is 55.5%; but at '10% FDR', FKI is merely 4.4%; $P < 1 \times 10^{-20}$).

Finally, we performed a thorough orthogonal experimental validation of randomly chosen novel interactions from the three sets using PCA[18,19]. The fraction of PCA-positive novel interactions from the '1% FDR with ΔXlinkX score ≥ 50' set (the highest-quality set) is distinctively higher compared with the other two sets and indistinguishable from that of PRS ($P=0.17$; Fig. 1g). Notably, the fractions of PCA-positive interactions for '1% FDR' and '10% FDR' are indistinguishable from that of RRS. Furthermore, using the Bayesian framework[16,17] (Supplementary Note 2), we calculated the absolute precision of the novel interactions detected in our human XL-MS study (Extended Data Fig. 3). Especially since the true FDR at the protein pair level can be substantially higher than the estimated FDR at the peptide pair level[7,15], absolute precision will be critically important for confirming the quality of novel protein–protein interactions identified in a large-scale cross-linking study. Finally, we confirmed the usefulness and robustness of the three computational metrics (namely, FSI, FMI and FKI) on the re-analyzed *E. coli* (Fig. 2c–e) and mouse mitochondrial (Extended Data Fig. 2c–e) XL-MS datasets, and using the additional search engine MaXLinker[2] (Extended Data Fig. 1c–e).

Taken together, our four metrics constitute a comprehensive framework to facilitate both relative comparison across different datasets and absolute estimation of error rates. Moreover, because these metrics stem from different principles, they provide complementary insights to various aspects of the data quality. FMI provides an orthogonal estimation of FDR and serves as an absolute measure of error rate. In fact, other methods[11,14,20] have been reported to provide complementary error estimates for XL-MS studies, and show good agreement with FMI in terms of the relative data quality across different datasets (Supplementary Note 3). Since FSI typically leverages thoroughly studied complexes, in theory, it should provide an absolute estimate of quality. Nonetheless, we do note that it may only provide relative comparison especially in cases where limited or incomplete 3D reference structures are available (Fig. 2c and Extended Data Fig. 2c). FKI and 'fraction of validated novel interactions using orthogonal experimental assays' specifically address the quality of detected interactions inferred from interprotein cross-links. Because a large fraction of true protein interactions is yet to be discovered, FKI only provides relative estimates of quality among comparable datasets. Finally, even if high-throughput orthogonal assays are not available, we recommend that low-throughput validation assays (such as coimmunoprecipitation[21]) be performed on a meaningful subset of the interactions identified (Supplementary Note 4).

In conclusion, we theoretically and experimentally illustrated the limitation of the current structure-based validation approach for evaluating proteome-wide XL-MS results. Furthermore, we proposed a comprehensive set of four metrics, and demonstrated their ability to distinguish datasets with varying qualities. Moreover, we acknowledge that this drastic underestimation of the error rate by the conventional structure-based approach is unlikely to pose a serious issue for XL-MS studies focused on specific proteins and individual complexes as long as the cross-link search is performed against only proteins that are included in the experiment. Importantly, this issue is highly relevant for the increasingly popular proteome-wide XL-MS experiments[8,9] and cross-linking immunoprecipitation–mass spectrometry studies[22]. Going forward, a comprehensive and accurate quality assessment framework such as the one proposed in this work needs to be adapted to aid in the advancement of XL-MS technologies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0959-9.

Received: 3 April 2019; Accepted: 20 August 2020;
Published online: 29 September 2020

References

1. Yu, C. & Huang, L. Cross-linking mass spectrometry: an emerging technology for interactomics and structural biology. *Anal. Chem.* **90**, 144–165 (2018).
2. Yugandhar, K. et al. MaXLinker: proteome-wide cross-link identifications with high specificity and sensitivity. *Mol. Cell. Proteomics* **19**, 554–568 (2020).
3. Iacobucci, C., Götze, M. & Sinz, A. Cross-linking/mass spectrometry to get a closer view on protein interaction networks. *Curr. Opin. Biotechnol.* **63**, 48–53 (2020).
4. Ferber, M. et al. Automated structure modeling of large protein assemblies using crosslinks as distance restraints. *Nat. Methods* **13**, 515–520 (2016).
5. 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. *Nat. Methods* **14**, 897–902 (2017).
6. Hauri, S. et al. Rapid determination of quaternary protein structures in complex biological samples. *Nat. Commun.* **10**, 192 (2019).
7. Fischer, L. & Rappsilber, J. Quirks of error estimation in cross-linking/mass spectrometry. *Anal. Chem.* **89**, 3829–3833 (2017).
8. O'Reilly, F. J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat. Struct. Mol. Biol.* **25**, 1000–1008 (2018).
9. Klykov, O. et al. Efficient and robust proteome-wide approaches for cross-linking mass spectrometry. *Nat. Protoc.* **13**, 2964–2990 (2018).
10. Liu, F., Lössl, P., Rabbitts, B. M., Balaban, R. S. & Heck, A. J. R. The interactome of intact mitochondria by cross-linking mass spectrometry provides evidence for coexisting respiratory supercomplexes. *Mol. Cell. Proteomics* **17**, 216–232 (2018).
11. Keller, A., Chavez, J. D., Felt, K. C. & Bruce, J. E. Prediction of an upper limit for the fraction of interprotein cross-links in large-scale in vivo cross-linking studies. *J. Proteome Res.* **18**, 3077–3085 (2019).
12. Bartolec, T. K. et al. Cross-linking mass spectrometry analysis of the yeast nucleus reveals extensive protein–protein interactions not detected by systematic two-hybrid or affinity purification-mass spectrometry. *Anal. Chem.* **92**, 1874–1882 (2020).
13. Liu, F., Lössl, P., Scheltema, R., Viner, R. & Heck, A. J. R. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat. Commun.* **8**, 15473 (2017).
14. Chen, Z.-L. et al. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* **10**, 3404 (2019).
15. Götze, M., Iacobucci, C., Ihling, C. H. & Sinz, A. A simple cross-linking/mass spectrometry workflow for studying system-wide protein interactions. *Anal. Chem.* **91**, 10236–10244 (2019).
16. Yu, H. et al. High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
17. Vo, TommyV. et al. A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human. *Cell* **164**, 310–323 (2016).
18. Nyfeler, B., Michnick, S. W. & Hauri, H.-P. Capturing protein interactions in the secretory pathway of living cells. *Proc. Natl Acad. Sci. USA* **102**, 6350–6355 (2005).
19. Braun, P. et al. An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* **6**, 91–97 (2008).
20. Beveridge, R., Stadlmann, J., Penninger, J. M. & Mechtler, K. A synthetic peptide library for benchmarking crosslinking-mass spectrometry search engines for proteins and protein complexes. *Nat. Commun.* **11**, 742 (2020).
21. Rual, J.-F. et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178 (2005).
22. 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. *Mol. Cell. Proteomics* **15**, 854–865 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Data processing. Cross-links were identified using XlinkX software (Proteome discoverer 2.2). Proteome Discoverer (PD) templates for different XlinkX search methodologies were obtained from Rosa Viner (Thermo Fisher Scientific). The raw files for the *E. coli* XL-MS dataset (MS2–MS3 aquisition) were obtained through e-mail request to Dr. Fan Liu. In addition to filtering cross-links at '10% FDR' and '1% FDR', we further filtered the '1% FDR' set using 'ΔXlinkX score' cut off ≥50. ΔXlinkX score is a cross-link spectrum match level scoring parameter in XlinkX software that indicates confidence in identifying a peptide pair over the next best competing peptide pair for a given precursor mass (higher score implies better quality). In addition to the three sets, we also filtered cross-links at '20% FDR' and carried out the structure-based mapping analyses to verify that the trend observed in Figs. 1c and 2b and Extended Data Fig. 2b holds at this much higher FDR threshold (Extended Data Fig. 4). During generation of the 20% FDR set using MaXLinker software, the FDR was estimated at the cross-link spectrum match level.

Target protein sequences were downloaded from the Uniprot database[23] (with filter 'reviewed'): (1) *E. coli*: 5,268 sequences; downloaded on 28 October 2017; (2) *Saccharomyces cerevisiae:* 7,904 sequences; downloaded on 28 September 2017 ('reviewed: yes'); (3) human (*Homo sapiens*): 42,202 sequences (20,206 canonical; 21,996 isoforms); downloaded on 23 June 2017); and (4) mouse (*Mus musculus*): 17,019 sequences; downloaded on 8 July 2019. More specifically, the human database consists of 21,996 isoform sequences in addition to the 20,206 canonical sequences. The mouse database consists of the canonical sequences for 17,019 proteins. The *E. coli* database contains of 5,268 sequences in total, consisting of 4,436 sequences from the K12 strain (4,436; most common) and the remaining 832 sequences from other less common strains. Similarly, for *S. cerevisiae*, the fasta database consists of 6,721 sequences from the common strain 'ATCC 204508', and the remaining sequences come from other less common strains such as 'YJM789', 'RM11-1a' and 'JAY291'. We utilized the full list of protein entries (did not rely on the protein grouping) to classify each cross-link as 'interprotein' or 'intraprotein', to avoid any inconsistencies that might occur due to potential protein grouping artifacts. When performing searches for Fig. 1e and Extended Data Fig. 5a, XlinkX crashed multiple times given the huge number of raw files (122 files) and the enormous search space (*H. sapiens* + *S. cerevisiae*). Hence, we ran the searches on a smaller set of raw files (25 files) to generate Fig. 1e and Extended Data Fig. 5a.

Mapping of cross-links to existing PDB structures. Cross-links from our human K562 proteome-wide XL-MS dataset were mapped to the 3D structure of the human 26S proteasome (PDB ID 5GJQ) utilizing residue-level mappings between Uniprot and PDB entries obtained from the SIFTS[24] database. In cases where multiple positions within the PDB structure were valid, the mapping with the shortest distance was prioritized. For the re-analyzed mouse mitochondrial XL-MS dataset[10], the cross-links were mapped to homologous complexes (PDB IDs 1EUC, 1T9G, 5LNK, 1ZOY, 1NTM, 1V54) as shown previously[10]. In brief, the protein sequences for all proteins involved in detected cross-links were aligned against a reference database containing PDB sequences of interest using BLAST[25]. All BLAST matches with significant *E* value and percentage identity greater than 70% were retained. Exact positions for each cross-link were mapped against homologous PDB structures using a pairwise alignment, and cross-links were only considered successfully mapped if the cross-linked lysine was conserved in the structure. In cases where multiple positions within the PDB structure were valid, the mapping with the shortest distance was prioritized. Any cross-links where the exact position of the cross-linked lysine was not structurally resolved in a homologous PDB structure were considered partially mapped. Because SIFT residue-level mapping for most of the representative structures (PDB IDs 2VRH, 1DKG, 1PCQ, 3JCD, 4PC1 and 2LRX) was unavailable for the *E. coli* dataset[13], we utilized the above-mentioned homology-based approach and the closest homologous complexes (PDB IDs 5MY1, 5ADY, 5ME0, 2RDO, 2VRH, 4JK2, 4YLN, 4YLO, 4XO2, 4YFH, 4YF0).

FSI. FSI can be calculated using the following equation:

$$\mathrm{FSI}(\%) = \frac{\begin{array}{c}\text{Number of interprotein XLs within the}\\ \text{Euclidean distance constraint of the linker}\end{array}}{\begin{array}{c}\text{Number of interprotein XLs with at least}\\ \text{one of the two residues mapped to structure}\end{array}} \times 100 \qquad (1)$$

In this work, we used 30 Å as the maximum distance constraint for DSSO.

FMI. FMI is the fraction of cross-link identifications from a false search space (from an unrelated organism) among all of the identified cross-links. It can be calculated using the following equation:

$$\mathrm{FMI}(\%) = \frac{\text{Number of mis-identifications}}{\text{Total number of identifications}} \times 100 \qquad (2)$$

In the current work, all of the raw files were re-analyzed against a sequence database containing all of the sequences from the target organism's proteome and all of the sequences from the *S. cerevisiae* proteome. Then the FMI, that is, cross-links with at least one of the two linked residues unambiguously mapped to proteins from *S. cerevisiae*, is calculated (if any cross-link had a peptide shared between homologous proteins from the target organism and *S. cerevisiae*, it was considered a true identification). Importantly, when choosing an unrelated organism, it is critical to make sure that there is no potential experimental contamination with proteins from that organism. It should be noted that another decoy database (reverse sequences of proteomes from both organisms) is generated for the FDR calculation by Proteome Discoverer. It is also noteworthy that FMI is estimated after the cross-link results are filtered at a conventional FDR threshold ('1% FDR' in the current study). Additionally, it should be pointed out that similar to the conventional FDR calculations[26], FMI calculations can also be sensitive to drastic differences in size of the proteome database of the unrelated organism. We utilized the following equation adapted from Fischer and Rappsilber[7] to account for differences in database size and observed a similar trend to that of uncorrected FMI across all three datasets analyzed in the current study (Extended Data Fig. 5).

$$\mathrm{FMI}_{\mathrm{corrected}}(\%) = \frac{\mathrm{TD} + \mathrm{DD}\left(1 - \frac{\mathrm{TD_{DB}}}{\mathrm{DD_{DB}}}\right)}{\mathrm{TT}} \times 100 \qquad (3)$$

where TT is the number of target–target matches, DD is the number of decoy–decoy matches and TD is number of target–decoy and decoy–target matches. $\mathrm{TD_{DB}}$ is the number of all possible unique target–decoy and decoy–target peptide pairs and $\mathrm{DD_{DB}}$ is the number of all possible unique decoy–decoy peptide pairs.

FKI. FKI for proteome-wide XL-MS studies can be defined as the fraction of the identified interprotein cross-links from previously known protein–protein interactions. It can be derived using the following equation:

$$\mathrm{FKI}(\%) = \frac{\text{Number of true positives}}{\text{Total number of postives}} \text{ X } 100 \qquad (4)$$

where, 'positives' refers to all of the identified interprotein cross-links, and 'true positives' refers to interprotein cross-links from known protein–protein interactions. If a given interprotein cross-link represents multiple potential interactions and at least one of those potential interactions was mapped to the list of known protein–protein interactions, it was counted as a 'true positive'. We compiled the known protein–protein interactions for *E. coli* (24,745), mouse (40,527) and human (336,033) from seven primary interaction databases. These databases include IMEx[27] partners IntAct[28], MINT[29] and DIP[30]; IMEx observer BioGRID[31]; and additional sources HPRD[32], MIPS[33] and iRefWeb[34]. Furthermore, iRefWeb combines interaction data from CORUM[35], BIND[36], MPPI[33] and OPHID[37]. We converted all gene identifiers in each database to Entrez gene IDs and then mapped to Uniprot IDs.

We would like to point out that FSI and FKI are calculated using similar denominators, conceptually. For FSI, the dominator consists of all interprotein cross-links with at least one of the two peptides mapped to the reference structure. In the case of FKI, the denominator consists of all of the interprotein cross-links in the dataset. Even though FKI's equation does not explicitly require all of the cross-links to have at least one of the two proteins to be present in the reference interactome database, we expect that almost all interprotein cross-links satisfy this criterion. Moreover, we analyzed all of the datasets from the current study and noted that all of the datasets have more than 97% of all of their interprotein cross-links with at least one of the proteins in the reference interactome database. We acknowledge that someone who has a smaller reference database might not note the same observation. However, we argue that such a case would lead to underestimation of FKI (that is, overestimation of error rate), thereby making FKI more stringent.

Fraction of validated novel interactions using orthogonal experiment, namely PCA. The open reading frames of novel protein–protein interactions in pDONR223 vector were inoculated from hORFeome v.8.1 library[38]. In each of the categories, namely '1% FDR with ΔXlinkX score ≥ 50', '1% FDR' and '10% FDR', 93 protein pairs were randomly picked without any overlaps between categories. The Gateway LR reactions were performed to clone the individual bait and prey proteins of each protein pair into the expression plasmids containing the complementation fragments of the fluorescent protein Venus. To perform the assay, the HEK293T cells were prepared in DMEM supplemented with 10% fetal bovine serum (ATCC) in black 96-well flat-bottom plates (Costar) with 5% CO_2 at 37 °C. Upon reaching 60–70% confluency, the cells were cotransfected with both plasmids containing the Venus fragments-tagged bait and prey open reading frames (100 ng for each) which were premixed and incubated with polyethylenimine (Polysciences) and OptiMEM (Gibco). For positive and negative controls, the sets containing the previously published 92 positive reference pairs and 92 negative reference pairs were simultaneously examined[19,39]. After 58 h, the fluorescence intensity of the transfected cells was measured and recorded using an Infinite M1000 microplate reader (Tecan) (excitation = 514 ± 5 nm/emission = 527 ± 5 nm). The PCA experiments were performed and analyzed in triplicate. We performed a statistical power analysis (using in-built R v.3.6.3 functions and Python 2.7) and confirmed that using 92 interactions would give us >97% power to detect the difference for the 'Positive Reference Set (PRS)' versus the 'Random Reference Set (RRS)', and the '1% FDR with ΔXlinkX score ≥ 50' versus the '1% FDR' and the '10% FDR' datasets. The effect sizes (Cohen's *d*) were calculated from the means and pooled standard deviations of two given groups under comparison (all effect sizes were large, that is, $d > 0.8$). The results are provided in Supplementary Table 1. Additionally, a short discussion on the utility of PCA to validate interactions from large-scale XL-MS studies on cell organelles and different organisms is provided in Supplementary Note 5.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The human K562 XL-MS raw files (122 raw files (97 HILIC and 25 SCX fractions) from our recent proteome-wide human K562 XL-MS study[2]) analyzed in this study have been deposited to the ProteomeXchange Consortium via the PRIDE[40] partner repository with the dataset identifier PXD018771. Raw data from our PCA experiments are available from the corresponding author upon request. Protein sequences were obtained from the Uniprot database (https://www.uniprot.org/). Residue-level mapping was performed using data from the SIFTS database (https://www.ebi.ac.uk/pdbe/docs/sifts/index.html). Protein three-dimensional structures utilized in this study were obtained from the PDB (accession codes: 5GJQ, 1EUC, 1T9G, 5LNK, 1ZOY, 1NTM, 1V54, 5MY1, 5ADY, 5ME0, 2RDO, 2VRH, 4JK2, 4YLN, 4YLO, 4XO2, 4YFH and 4YF0). Source data are provided with this paper.

References

23. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
24. Dana, J. M. et al. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* **47**, D482–D489 (2018).
25. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
26. Gupta, N., Bandeira, N., Keich, U. & Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* **22**, 1111–1120 (2011).
27. Orchard, S. et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* **9**, 345–350 (2012).
28. Kerrien, S. et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* **40**, D841–D846 (2012).
29. Licata, L. et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–D861 (2012).
30. Salwinski, L. et al. The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
31. Chatr-aryamontri, A. et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* **43**, D470–D478 (2015).
32. Keshava Prasad, T. S. et al. Human protein reference database—2009 update. *Nucleic Acids Res.* **37**, D767–D772 (2009).
33. Pagel, P. et al. The MIPS mammalian protein–protein interaction database. *Bioinformatics* **21**, 832–834 (2005).
34. Turner, B. et al. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* **2010**, baq023–baq023 (2010).
35. Ruepp, A. et al. CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* **38**, D497–D501 (2010).
36. Alfarano, C. et al. The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res.* **33**, D418–D424 (2005).
37. Brown, K. R. & Jurisica, I. Online predicted human interaction database. *Bioinformatics* **21**, 2076–2082 (2005).
38. Yang, X. et al. A public genome-scale lentiviral expression library of human ORFs. *Nat. Methods* **8**, 659–661 (2011).
39. Venkatesan, K. et al. An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2008).
40. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2018).

Acknowledgements

We thank R. Viner for support in data processing with XlinkX workflow in Proteome Discoverer. K.Y. thanks the Sam and Nancy Fleming Research Fellowship. This work was supported by grants from the National Institutes of Health (grant nos. GM124559 and GM125639) and the National Science Foundation (grant no. DBI-1661380) to H.Y.

Author contributions

H.Y. conceived and oversaw all aspects of the study. K.Y. performed the computational analyses with assistance from S.D.W. T.-Y.W. performed laboratory experiments with assistance from E.E.S. K.Y. and H.Y. wrote the manuscript with inputs from all of the authors.

Competing interests

The authors declare no competing interests.

Additional information

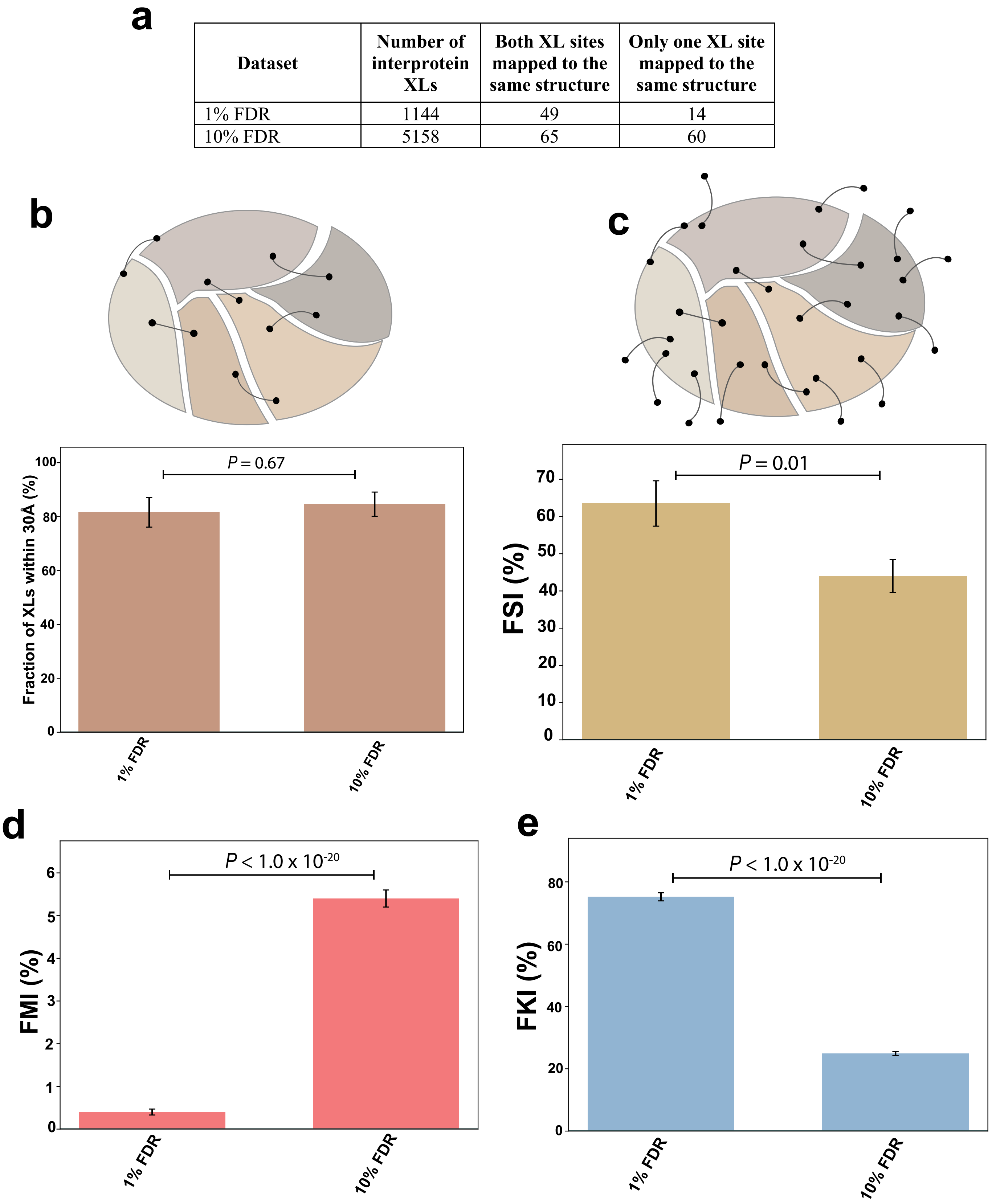
Extended data is available for this paper at https://doi.org/10.1038/s41592-020-0959-9.

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0959-9.

Correspondence and requests for materials should be addressed to H.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Editor recognition statement Allison Doerr was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.



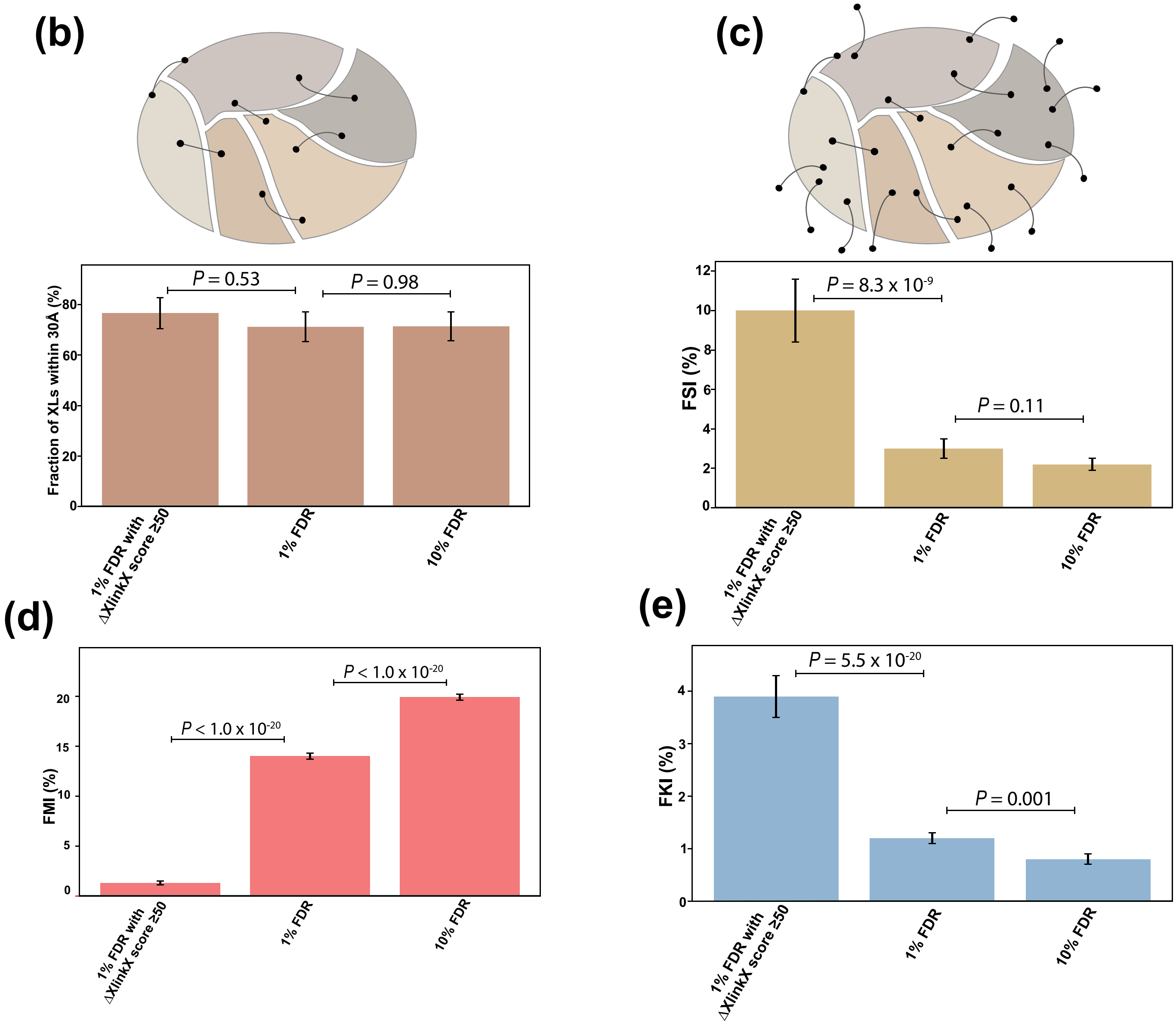
Dataset	Number of interprotein XLs	Both XL sites mapped to the same structure	Only one XL site mapped to the same structure
1% FDR	1144	49	14
10% FDR	5158	65	60

Extended Data Fig. 1 | See next page for caption.

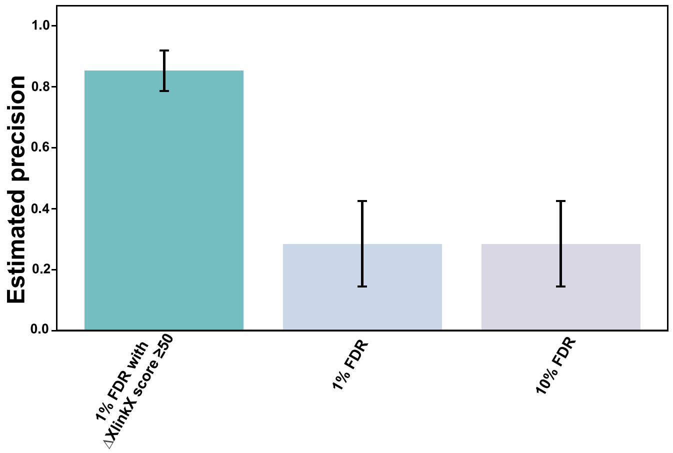
Extended Data Fig. 1 | Analysis of the human proteome-wide XL-MS dataset using MaXLinker software. (a) Table showing the number of interprotein cross-links obtained at different filtering criteria, and upon mapping to a representative 3D structure of a human 26S proteasome (PDB id: 5GJQ). **(b)** Comparison of the fraction of validated cross-links using the conventional structure-based approach (n = 49 XLs for '1% FDR'; n = 65 XLs for '10% FDR). **(c)** Comparison using the fraction of structure-corroborating identifications (FSI) (n = 63 XLs for '1% FDR'; n = 125 XLs for '10% FDR). **(d)** Comparison using the fraction of mis-identifications (FMI) (n = 8127 XLs for '1% FDR'; n = 15110 XLs for '10% FDR). **(e)** Comparison using the fraction of interprotein cross-links from known interactions (FKI) (n = 1144 XLs for '1% FDR'; n = 5158 XLs for '10% FDR). for **(b–e)**, the P values were calculated using a two-sided Z-test and the error bars indicate +/- SE of proportion.

(a)

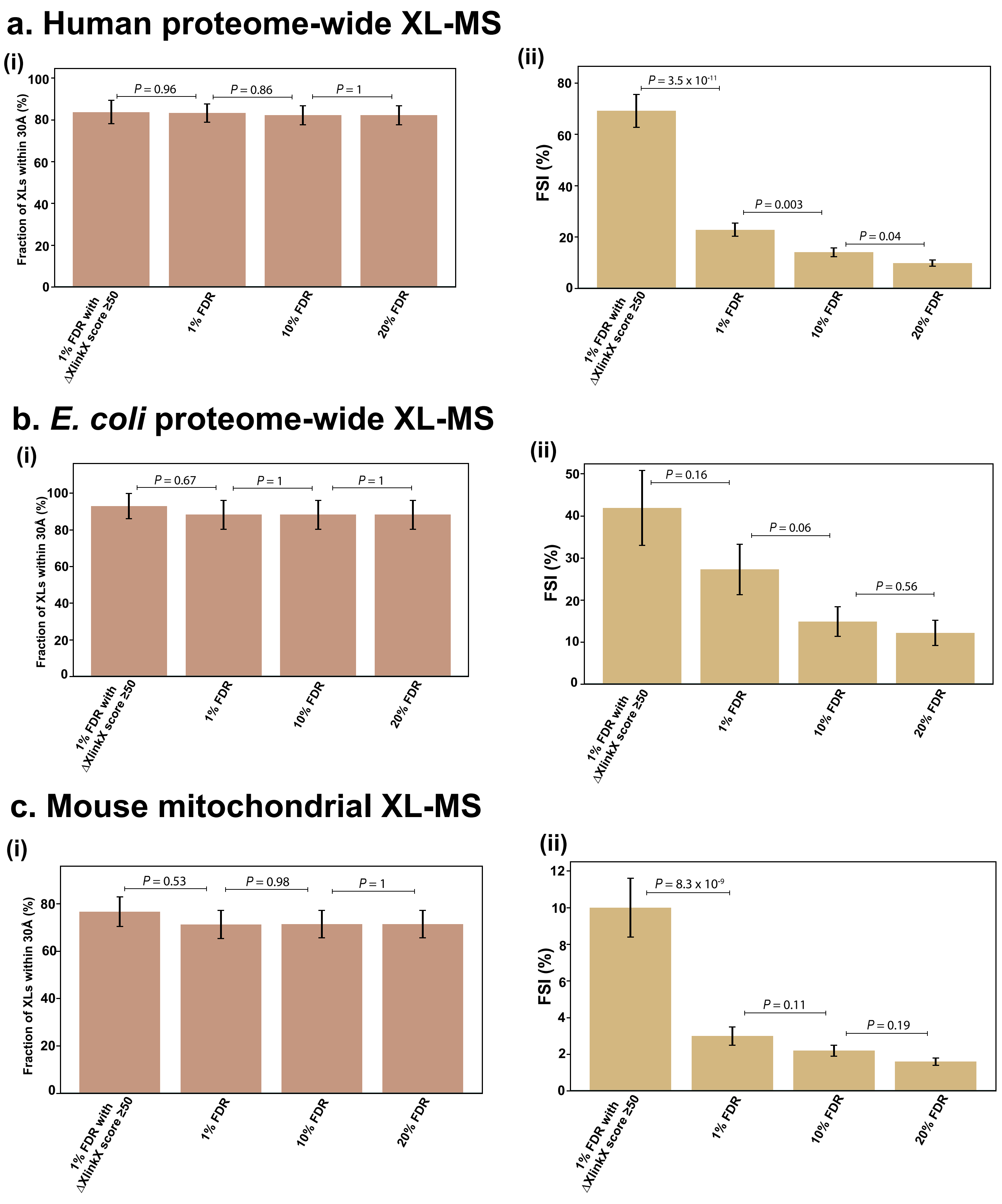
Dataset	Number of interprotein XLs	Both XL sites mapped to the same structure	Only one XL site mapped to the same structure
1% FDR with ΔXlinkXScore≥50	2368	47	313
1% FDR	11418	59	1343
10% FDR	19665	63	2034



Extended Data Fig. 2 | Demonstration of the utility of our comprehensive set of validation metrics on a publicly available mouse mitochondrial XL-MS dataset. (a) Table showing the number of interprotein cross-links obtained at different filtering criteria, and upon mapping to representative 3D structures. **(b)** Conventional structure-based validation (n = 47 XLs for '1% FDR with ΔXlinkX score≥50'; n = 59 XLs for '1% FDR'; n = 63 XLs for '10% FDR'). **(c)** Fraction of structure-corroborating identifications (FSI) (n = 360 XLs for '1% FDR with ΔXlinkX score≥50'; n = 1402 XLs for '1% FDR'; n = 2097 XLs for '10% FDR'). **(d)** Fraction of mis-identifications (FMI) (n = 4814 XLs for '1% FDR with ΔXlinkX score≥50'; n = 15323 XLs for '1% FDR'; n = 24317 XLs for '10% FDR'). **(e)** Fraction of interprotein cross-links from known interactions (FKI) (n = 2368 XLs for '1% FDR with ΔXlinkX score≥50'; n = 11418 XLs for '1% FDR'; n = 19665 XLs for '10% FDR'). P values in **(b-e)** were calculated using a two-sided Z-test and the error bars indicate +/- SE of proportion.

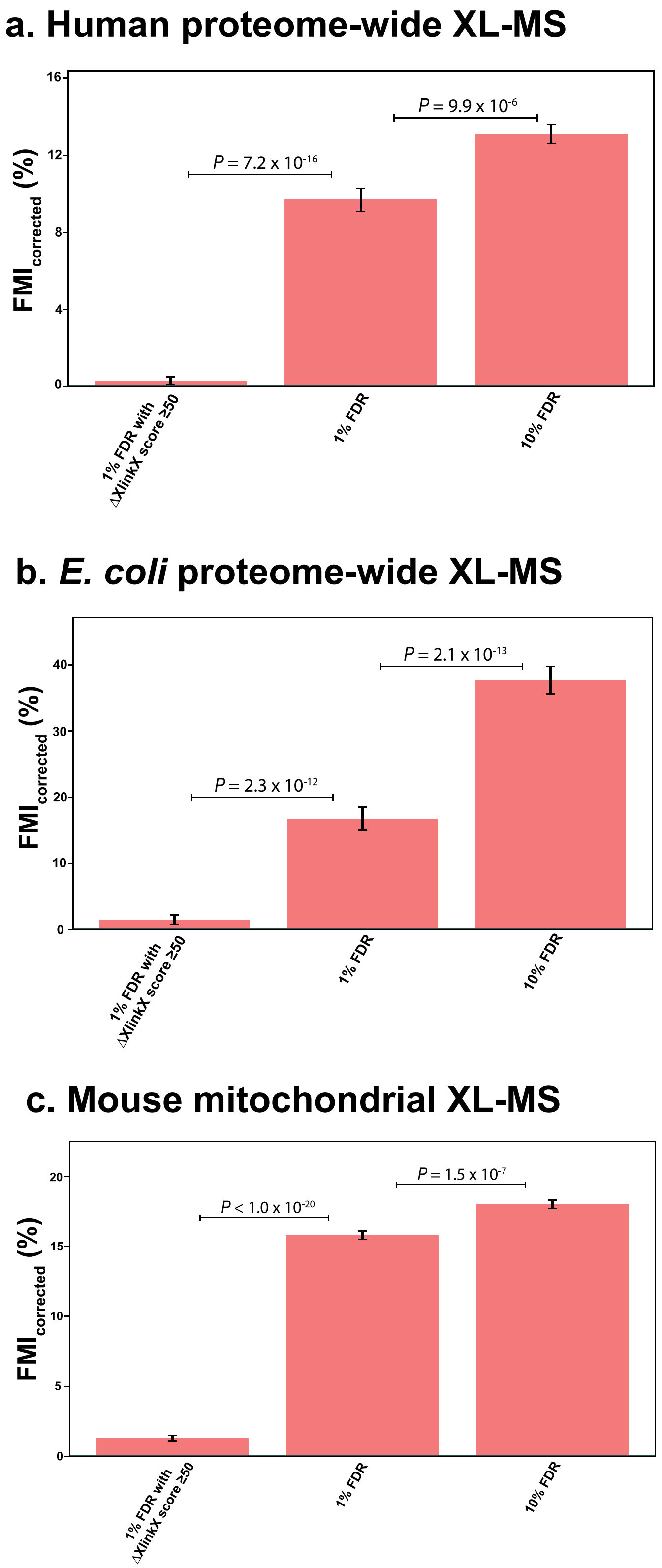


Extended Data Fig. 3 | Estimated precision using PCA experiments for the three datasets of different quality from our human K562 proteome-wide XL-MS study. Derived from Fig. 1g (n = 3 independent experiments; See Methods). The error bars indicate +/- SE of proportion (see Supplementary Note 2 for a detailed description of the methodology).



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Structure-based mapping analysis at 20% FDR, extension to the analysis shown in Fig. 1, Fig. 2, and Extended Data Fig. 2. a. Human proteome-wide XL-MS study: (i) Conventional structure-based validation (n = 43 XLs for '1% FDR with ΔXlinkX score≥50'; n = 72 XLs for '1% FDR'; n = 73 XLs for '10% FDR'; n = 73 XLs for '20% FDR'). (ii) Fraction of structure-corroborating identifications (FSI) (n = 52 XLs for '1% FDR with ΔXlinkX score≥50'; n = 262 XLs for '1% FDR'; n = 426 XLs for '10% FDR'; n = 605 XLs for '20% FDR'). **b**. *E. coli* proteome-wide XL-MS study: (i) Conventional structure-based validation (n = 14 XLs for '1% FDR with ΔXlinkX score≥50'; n = 17 XLs for '1% FDR'; n = 17 XLs for '10% FDR'; n = 17 XLs for '20% FDR'). (ii) Fraction of structure-corroborating identifications (FSI) (n = 31 XLs for '1% FDR with ΔXlinkX score≥50'; n = 55 XLs for '1% FDR'; n = 101 XLs for '10% FDR'; n = 123 XLs for '20% FDR'). **c**. Mouse mitochondrial XL-MS study: (i) Conventional structure-based validation (n = 47 XLs for '1% FDR with ΔXlinkX score≥50'; n = 59 XLs for '1% FDR'; n = 63 XLs for '10% FDR'; n = 63 XLs for '20% FDR'). (ii) Fraction of structure-corroborating identifications (FSI) (n = 360 XLs for '1% FDR with ΔXlinkX score≥50'; n = 1402 XLs for '1% FDR'; n = 2097 XLs for '10% FDR'; n = 2751 XLs for '20% FDR'). P values in all the panels were calculated using a two-sided Z-test and the error bars indicate +/- SE of proportion.



a. Human proteome-wide XL-MS

b. *E. coli* proteome-wide XL-MS

c. Mouse mitochondrial XL-MS

Extended Data Fig. 5 | Corrected FMI for the three datasets analyzed in the study (Utilizing Equation 3 from Methods section). (a) Human proteome-wide XL-MS (n = 668 XLs for '1% FDR with ΔXlinkX score≥50'; n = 3029 XLs for '1% FDR'; n = 4957 XLs for '10% FDR). **(b)** *E. coli* proteome-wide XL-MS (n = 340 XLs for '1% FDR with ΔXlinkX score≥50'; n = 553 XLs for '1% FDR'; n = 755 XLs for '10% FDR). **(c)** Mouse mitochondrial XL-MS (n = 4814 XLs for '1% FDR with ΔXlinkX score≥50'; n = 15323 XLs for '1% FDR'; n = 24317 XLs for '10% FDR). P values in all the panels were calculated using a two-sided Z-test and the error bars indicate +/- SE of proportion.

nature research

Corresponding author(s): Haiyuan Yu

Last updated by author(s): Jul 2, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

No software was used for data collection

Data analysis

XlinkX (from Proteome Discover 2.2) and MaXLinker version 1.0 (Yugandhar et al., Mol. Cell. Proteomics, 2020) were used for cross-link identification from different datasets analyzed in the study. Statistical tests were performed using models built in R (3.6.3) and Python (2.7)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The human K562 XL-MS raw files (97 HILIC fractions and 25 sex fractions) analyzed in this study have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD018771. Protein sequences were obtained from Uniprot database (https://www.uniprot.org/). Residue level mapping was performed using data from SIFTS database (https://www.ebi.ac.uk/pdbe/docs/sifts/index.html). Protein three dimensional structures utilized in this study were obtained from PDB (https://www.rcsb.org/ ; Accession codes: 5GJQ, 1EUC, 1T9G, 5LNK, 1ZOY, 1NTM, 1V54, 5MY1, 5ADY, 5ME0, 2RDO, 2VRH, 4JK2, 4YLN, 4YLO, 4XO2, 4YFH and 4YF0).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences　□ Behavioural & social sciences　□ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to predetermine sample size. Sample sizes in high-thoughput experiments were determined by maximizing the total number of samples that could be analyzed in a particular experiment
Data exclusions	No data were excluded from the analyses.
Replication	The PCA experiments were performed and analyzed in triplicate and all attempts at replication were successful.
Randomization	The identified cross-linked protein pairs were selected in an unbiased manner as well as under the availability of the ORF clones in our library collections. For PCA test, the samples were mixed with the positive and negative sets in positioning, in order to avoid the potential position bias.
Blinding	Data blinding present through all computational and experimental measurements.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
☒	□	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	K562 and HEK293T cells were obtained from ATCC.
Authentication	Cell lines have been thoroughly tested and authenticated by ATCC.
Mycoplasma contamination	Both K562 and HEK293T cells were tested negative for Mycoplasma contaminations.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.