Supplementary information for "Integration of heterogeneous experimental data improves global map of human protein complexes"

Jose Lugo-Martinez
Computational Biology Department
Carnegie Mellon University
Pittsburgh, Pennsylvania, U.S.A.
jlugomar@cs.cmu.edu

Jörn Dengjel
Department of Biology
Université de Fribourg
1700 Fribourg, Switzerland
joern.dengjel@unifr.ch

Ziv Bar-Joseph*
Computational Biology Department
Carnegie Mellon University
Pittsburgh, Pennsylvania, U.S.A.
zivbj@cs.cmu.edu

Robert F. Murphy*
Computational Biology Department
Department of Biological Sciences
Carnegie Mellon University
Pittsburgh, Pennsylvania, U.S.A.
murphy@cmu.edu

*To whom correspondence should be addressed.

This file includes:

Supplementary figures

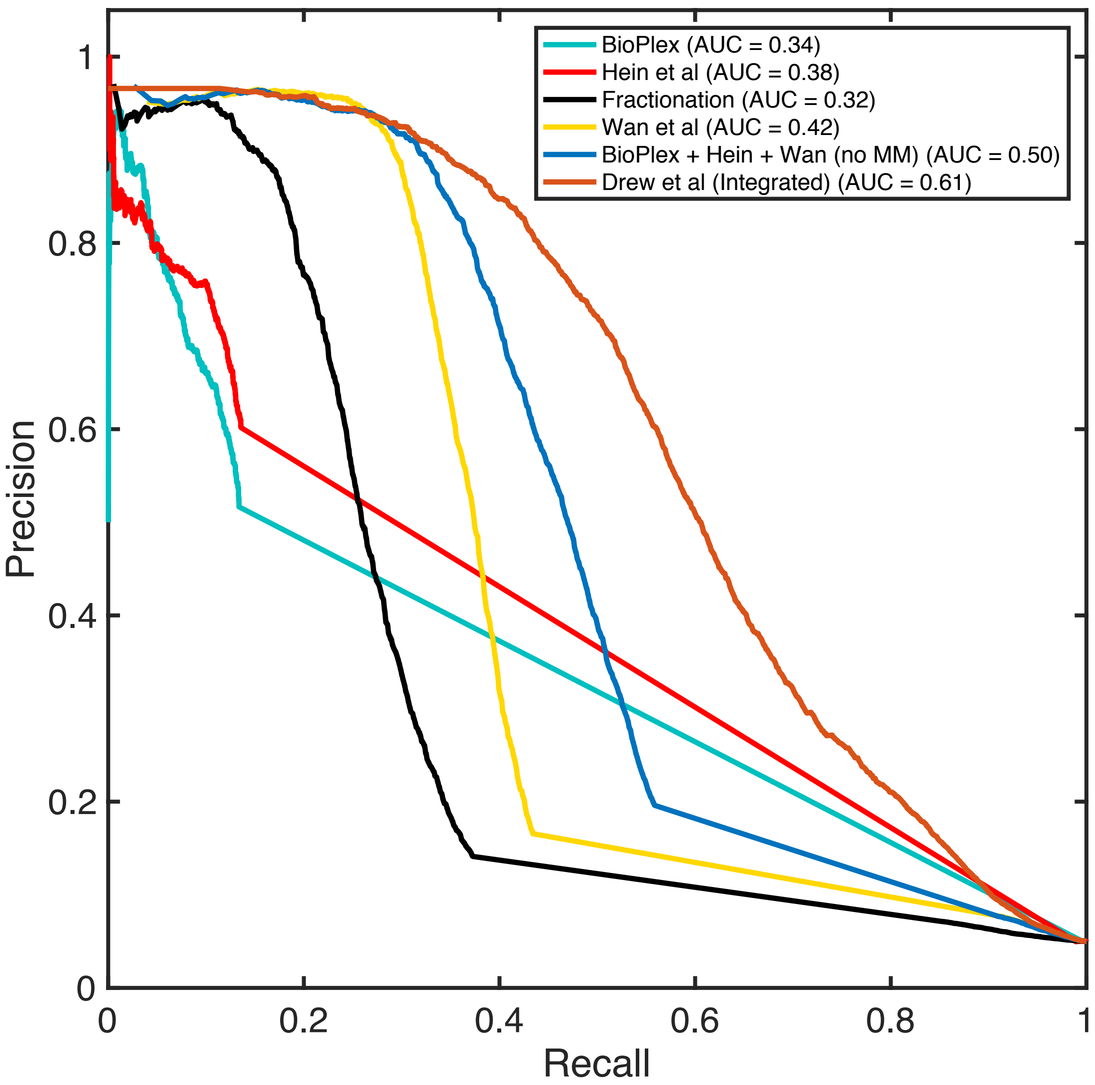

Figure S1: **Replication of performance comparison of pairwise protein interactions prediction as originally reported by Drew *et al* [1].** For each method, we show precision-recall curve and area under the curve (AUC) using our in-house implementation.

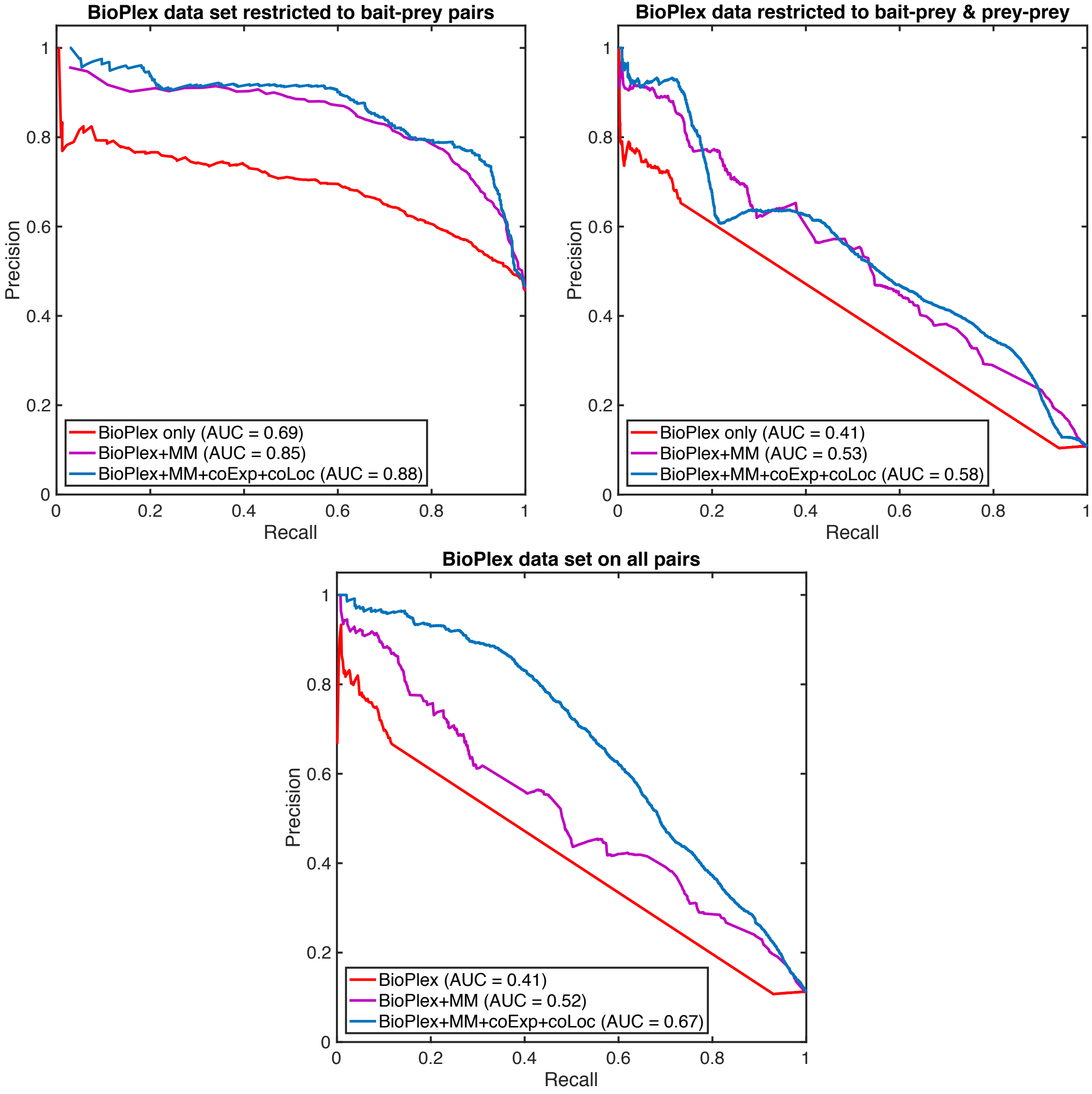

Figure S2: **Performance comparison of pairwise protein interactions prediction on BioPlex [3] data.** Figure shows precision-recall curve and area under the curve (AUC) under three different protein pair models of our proposed method (blue) compared against a baseline approach which only uses Bioplex specific features (red) and a previously proposed approach which adds weighted matrix model (MM) features (magenta). (A) Figure shows the comparison results restricted to bait-prey pairs. (B) Figure shows the comparison results restricted to bait-prey and prey-prey pairs. (C) Figure shows the comparison results on all possible pairs in BioPlex for which we had data.

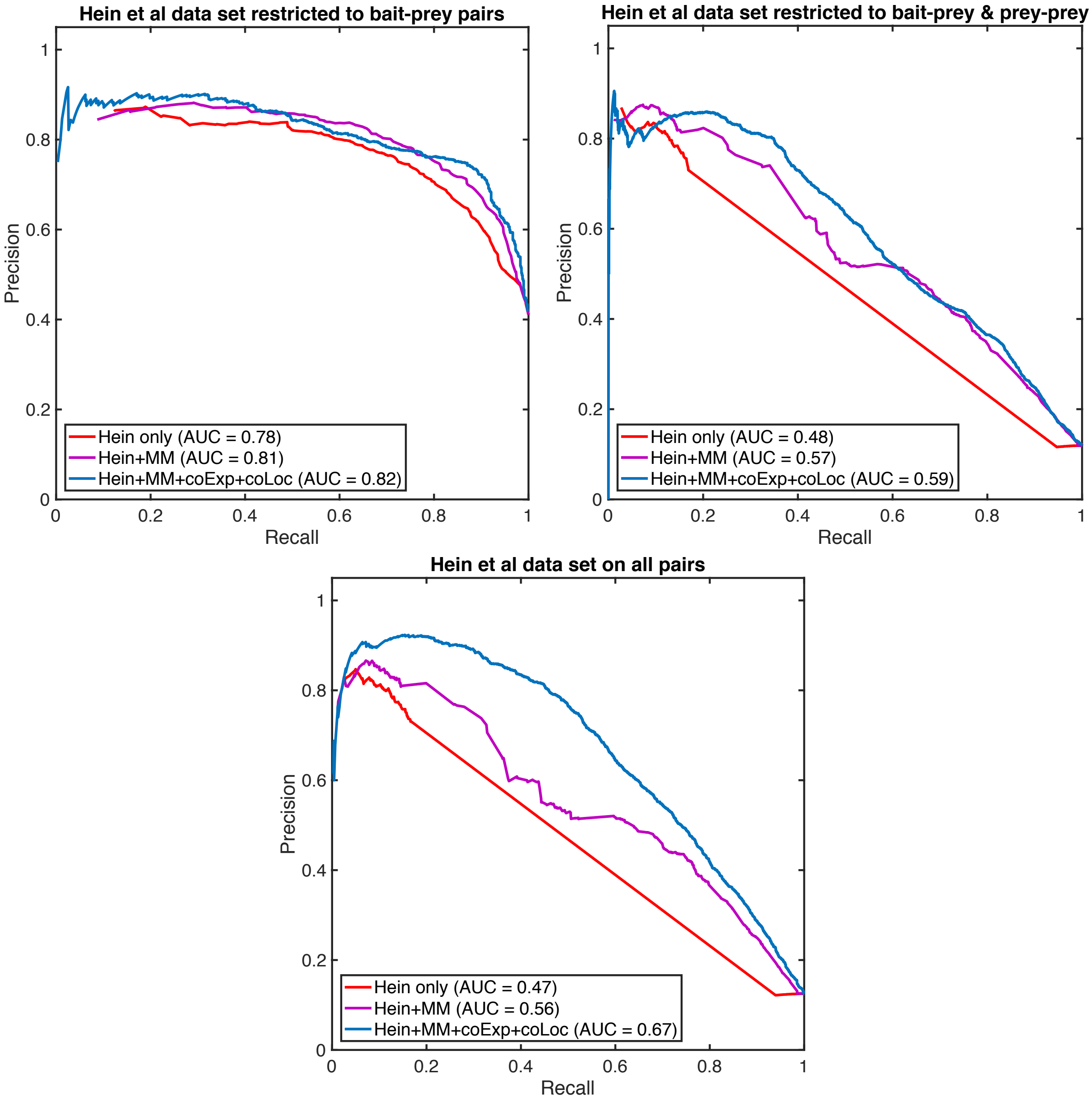

Figure S3: **Performance comparison of pairwise protein interactions prediction on Hein *et al* [2] data.** Figure shows precision-recall curve and area under the curve (AUC) under three different protein pair models of our proposed method (blue) compared against a baseline approach which only uses Hein *et al* specific features (red) and a previously proposed approach which adds weighted matrix model (MM) features (magenta). (A) Figure shows the comparison results restricted to bait-prey pairs. (B) Figure shows the comparison results restricted to bait-prey and prey-prey pairs. (C) Figure shows the comparison results on all possible pairs in Hein *et al* for which we had data.

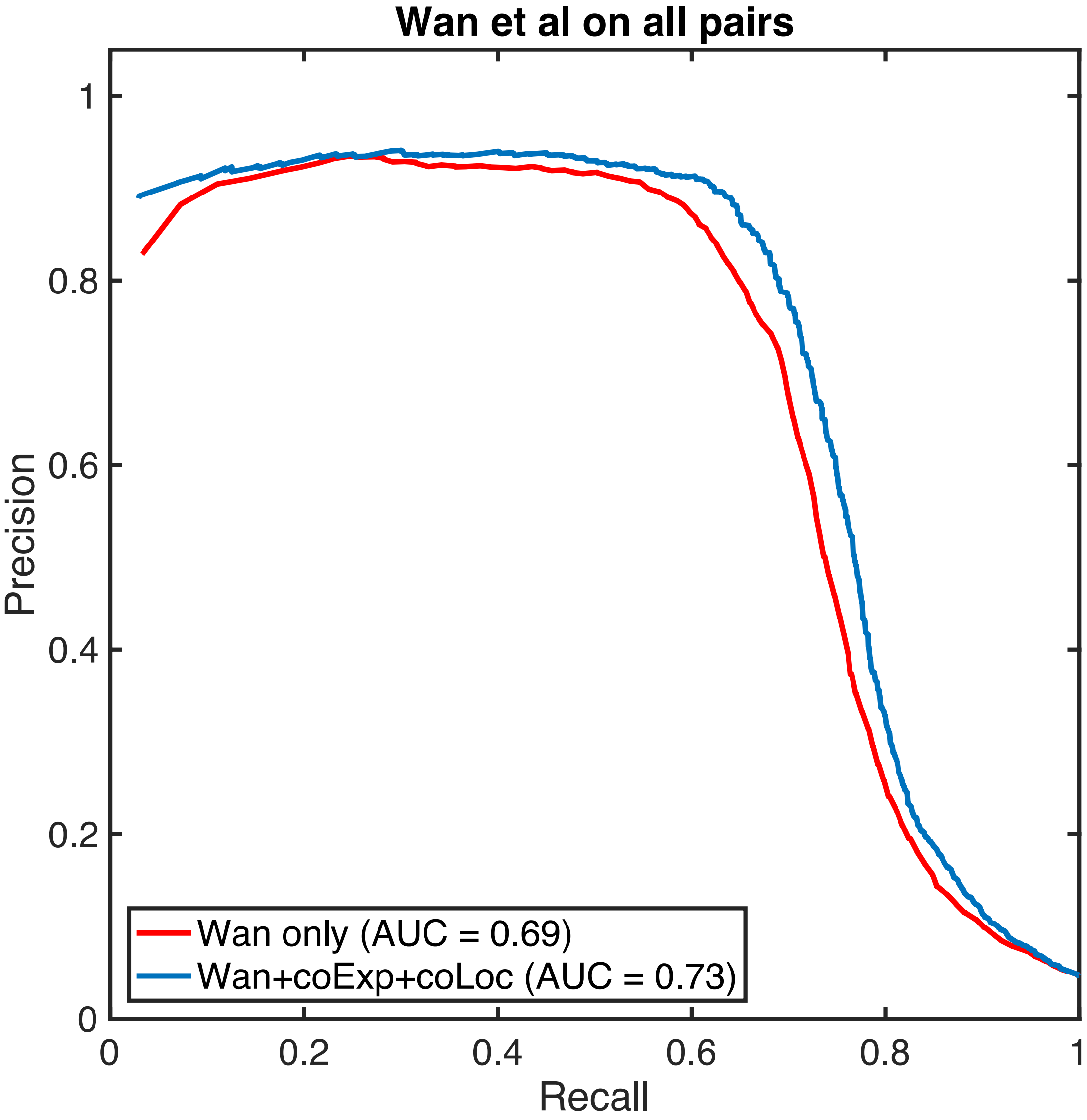

Figure S4: **Performance comparison of pairwise protein interactions prediction on Wan *et al* [4] data.** Figure shows precision-recall curve and area under the curve (AUC) of our proposed method (blue) compared against a baseline approach which only uses Wan *et al* specific features (red) over all protein pairs.

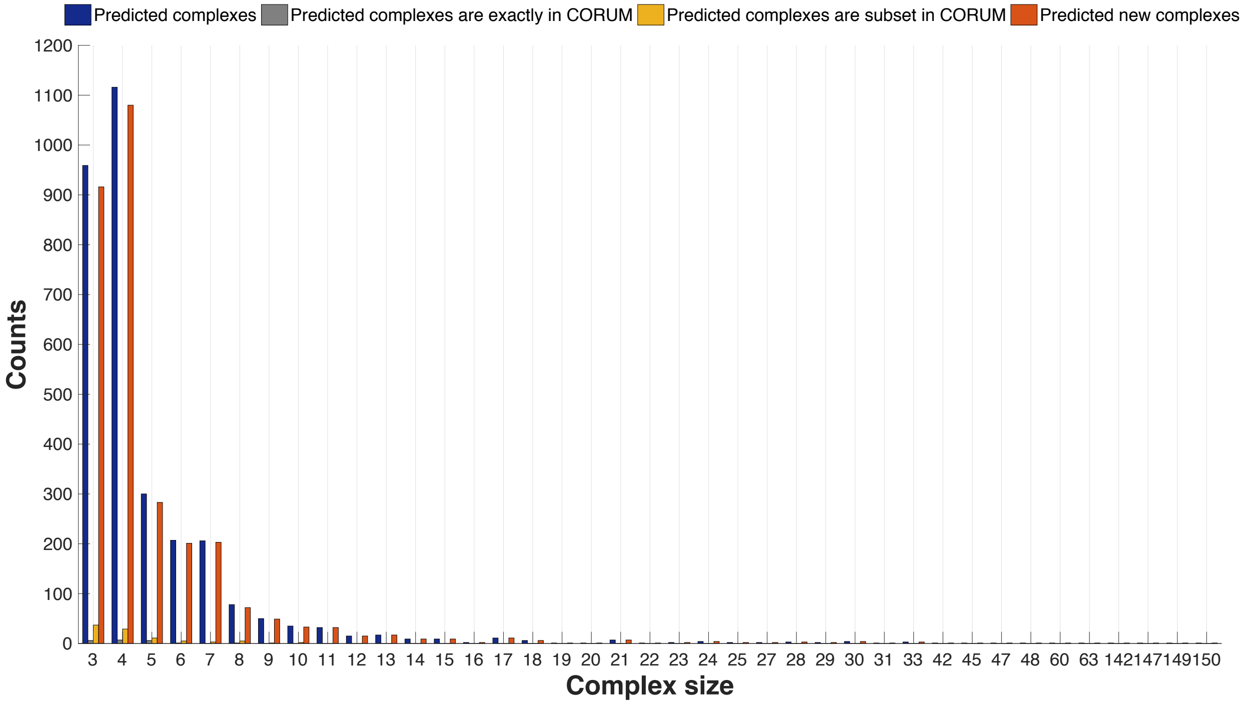

Figure S5: **Comparison of hu.MAP complexes against gold standard CORUM.** Figure shows hu.MAP complexes (blue) across three different categories as a function of complex size. The categories are (1) identical match to complex in CORUM (gray), (2) strict subset to a complex in CORUM (yellow), and (3) potentially novel complex (orange).

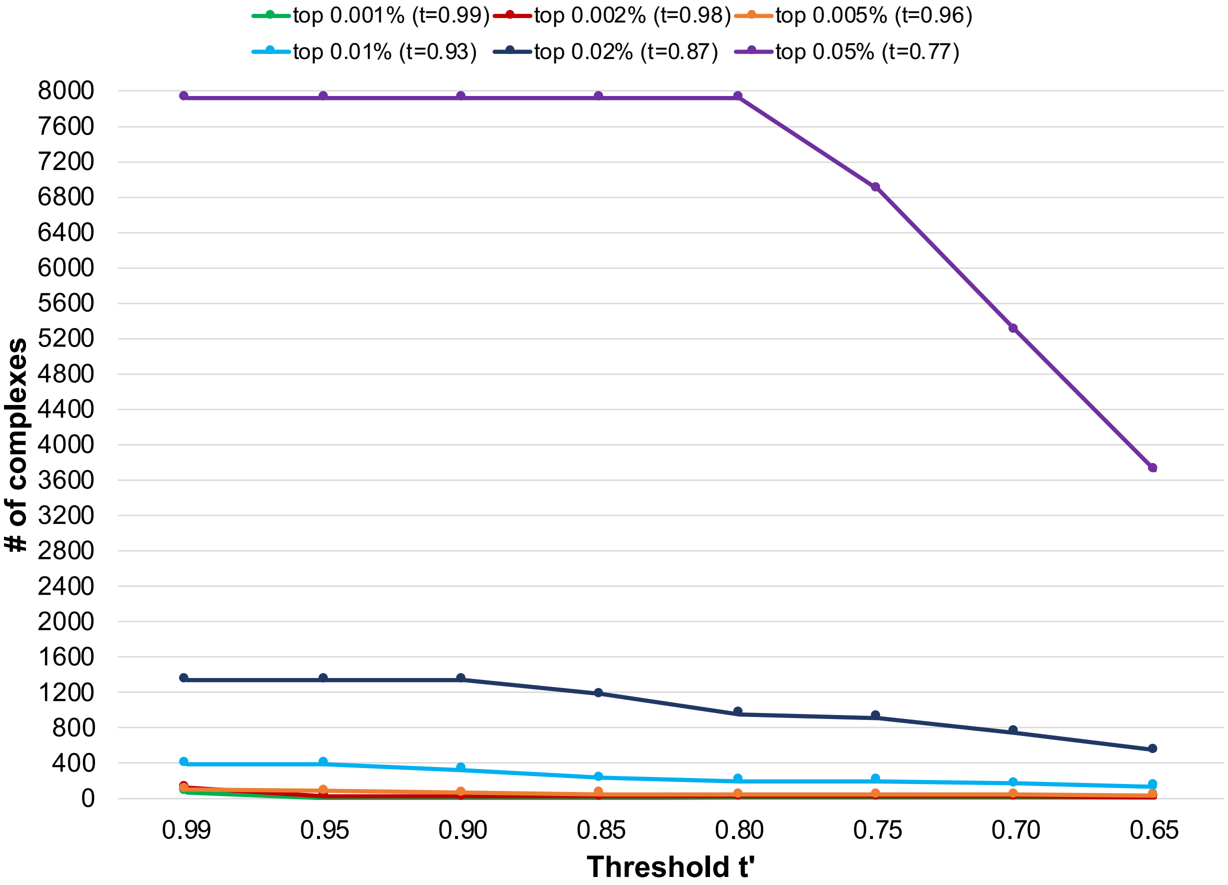

Figure S6: **Effect of protein complex refinement.** Figure shows the number of predicted complexes as a function of parameter t' using Algorithm 2. For each threshold, we show parameter t in parenthesis.

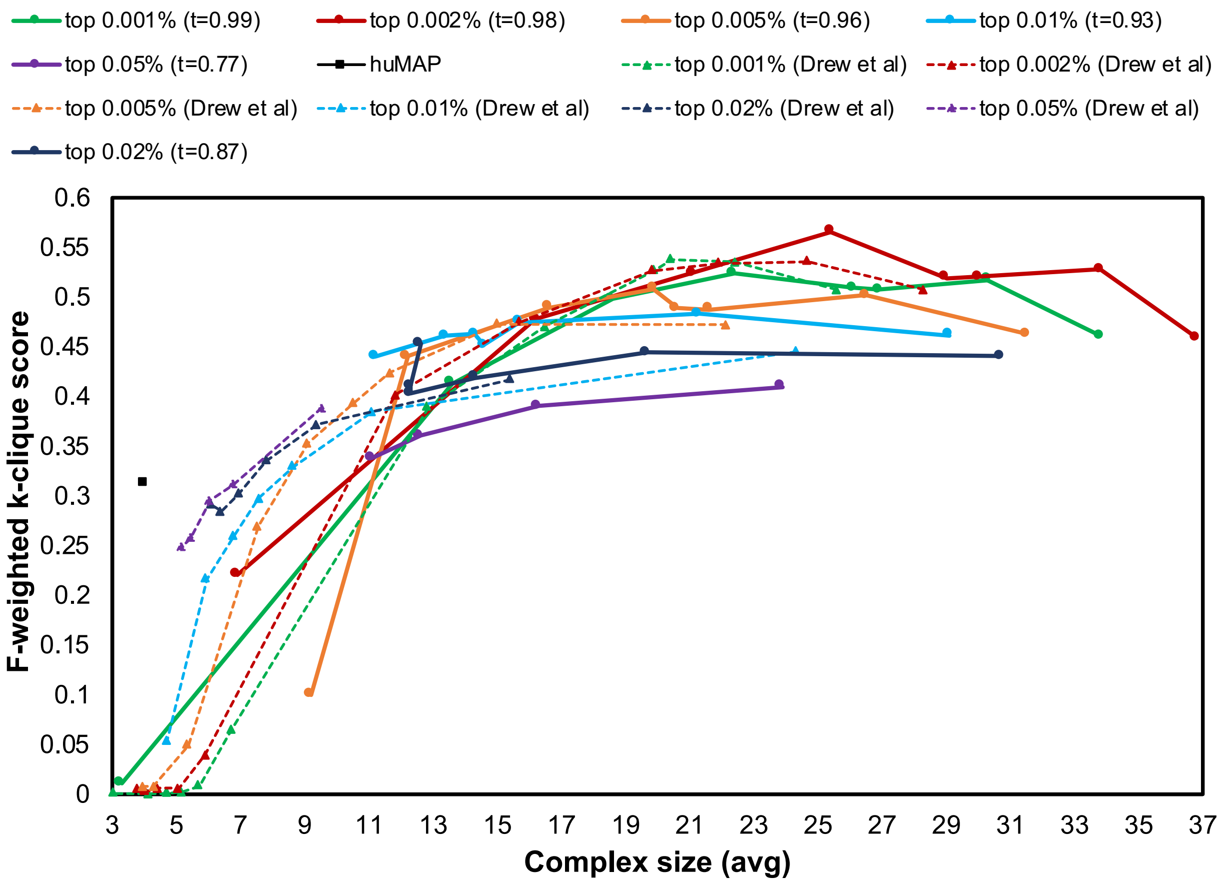

Figure S7: **Comparison between predicted and hu.MAP complexes.** Figure shows F-weighted k-clique score [1] of our method (solid lines with circles) as a function of average complex size for each t' threshold. Figure also shows corresponding scores for hu.MAP (square) and an in-house implementation that uses hu.MAP pairwise scores as input to Algorithms 1 and 2 (dotted lines with triangles).

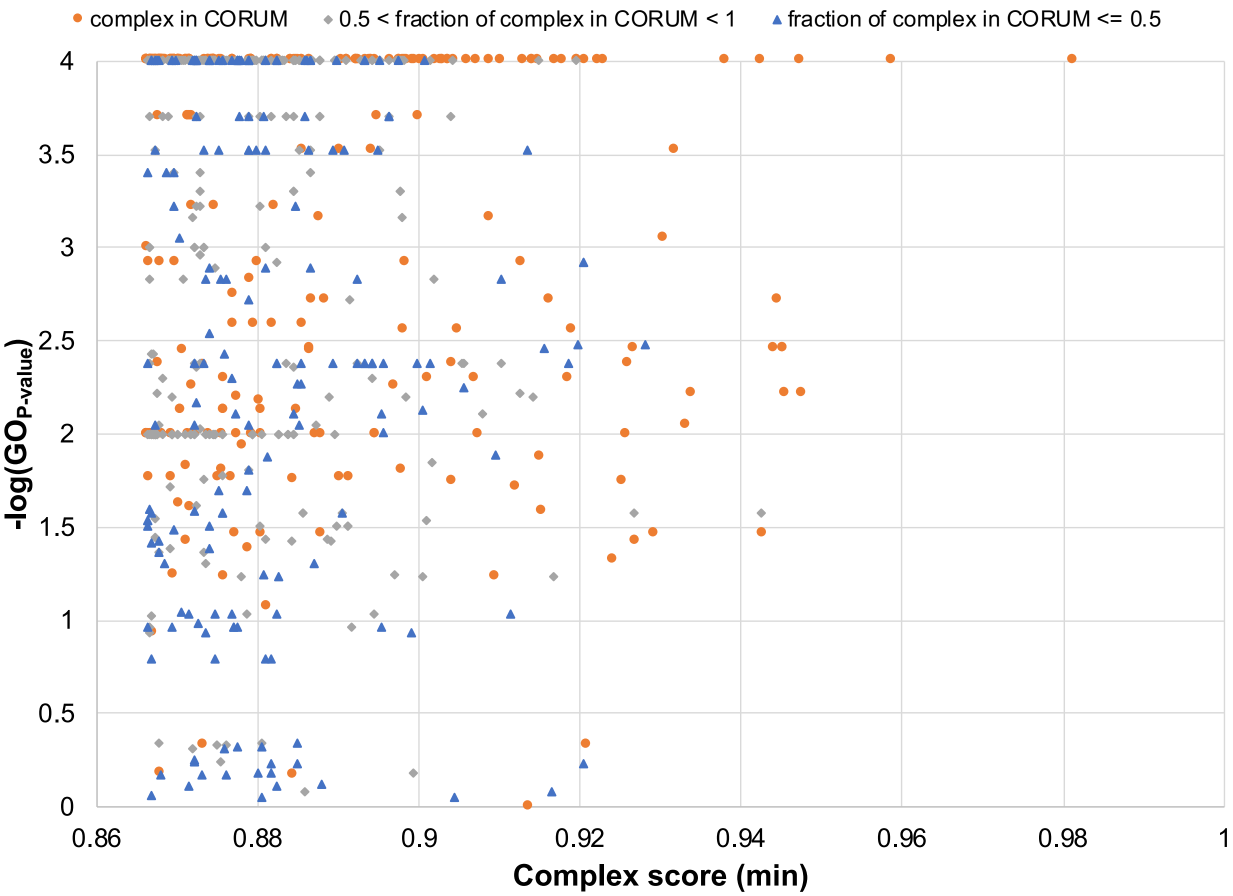

Figure S8: **Distribution of enriched functional annotation as a function of complex score.** For each complex, we show the distribution of the largest p-value from enriched functional annotations (plotted as $-\log(\text{GO}_{\text{P-value}})$) computed using g:Profiler and further adjusted to the estimated occurrence of significant enrichment from $10,000$ random complexes of the same size as a function of complex score. Additionally, each complex is assigned to one of the following three classes based on varying degrees of overlap with CORUM complexes: (i) full overlap with CORUM complex (orange circles), (ii) at least half the member proteins overlap with CORUM complex (grey diamonds), and less than half of co-member proteins overlap with CORUM complex (blue triangles).

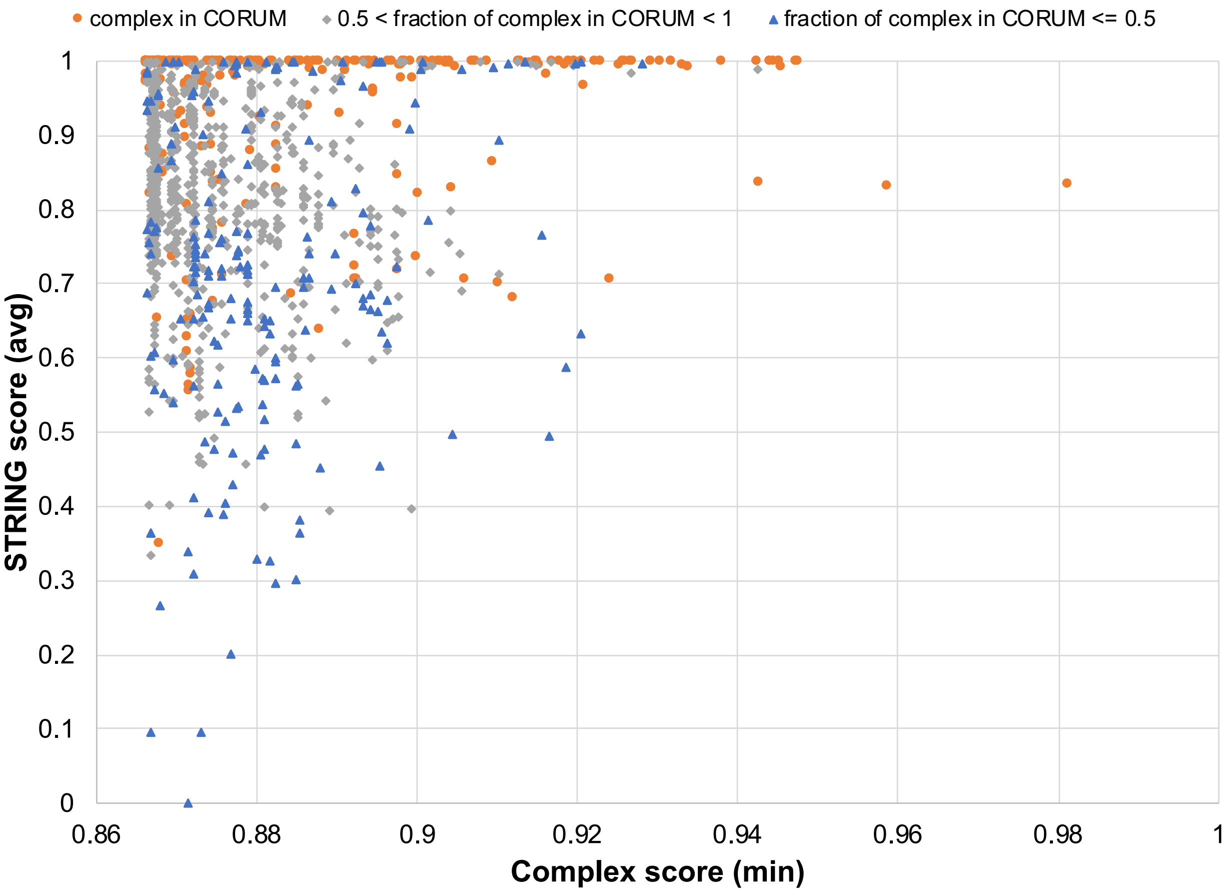

Figure S9: **Distribution of average STRING score as a function of complex score.** For each complex, we show the distribution of the average STRING score as a function of complex score. Additionally, each complex is assigned to one of the following three classes based on varying degrees of overlap with CORUM complexes: (i) full overlap with CORUM complex (orange circles), (ii) at least half the member proteins overlap with CORUM complex (grey diamonds), and less than half of co-member proteins overlap with CORUM complex (blue triangles).

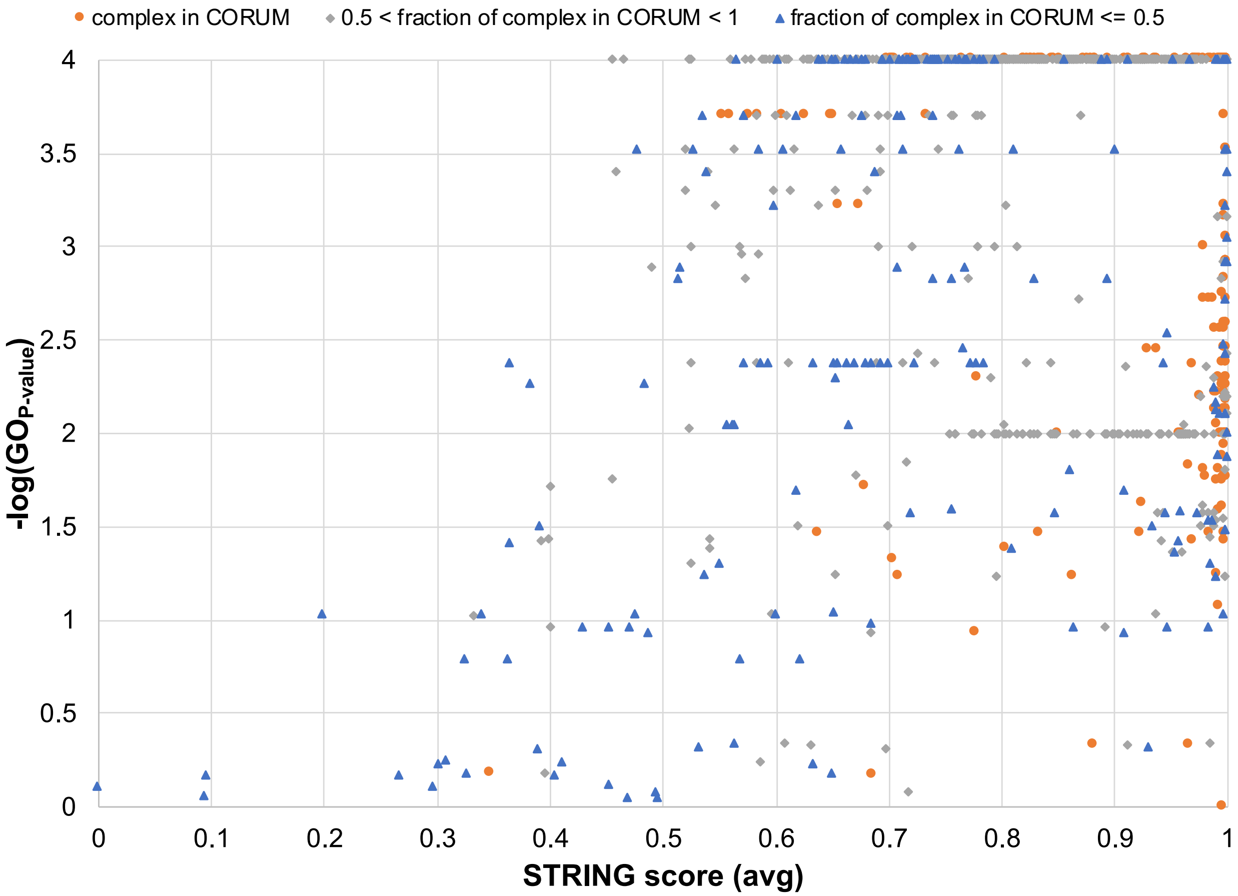

Figure S10: **Distribution of enriched functional annotation as a function of average STRING score.** For each complex, we show the distribution of the largest p-value from enriched functional annotations (plotted as -log($\mathrm{GO}_{\text{P-value}}$)) computed using g:Profiler and further adjusted to the estimated occurrence of significant enrichment from 10, 000 random complexes of the same size as a function of average STRING score. Additionally, each complex is assigned to one of the following three classes of based on varying degrees of overlap with CORUM complexes: (i) full overlap with CORUM complex (orange circles), (ii) at least half the member proteins overlap with CORUM complex (gray diamonds), and less than half of co-member proteins overlap with CORUM complex (blue triangles).

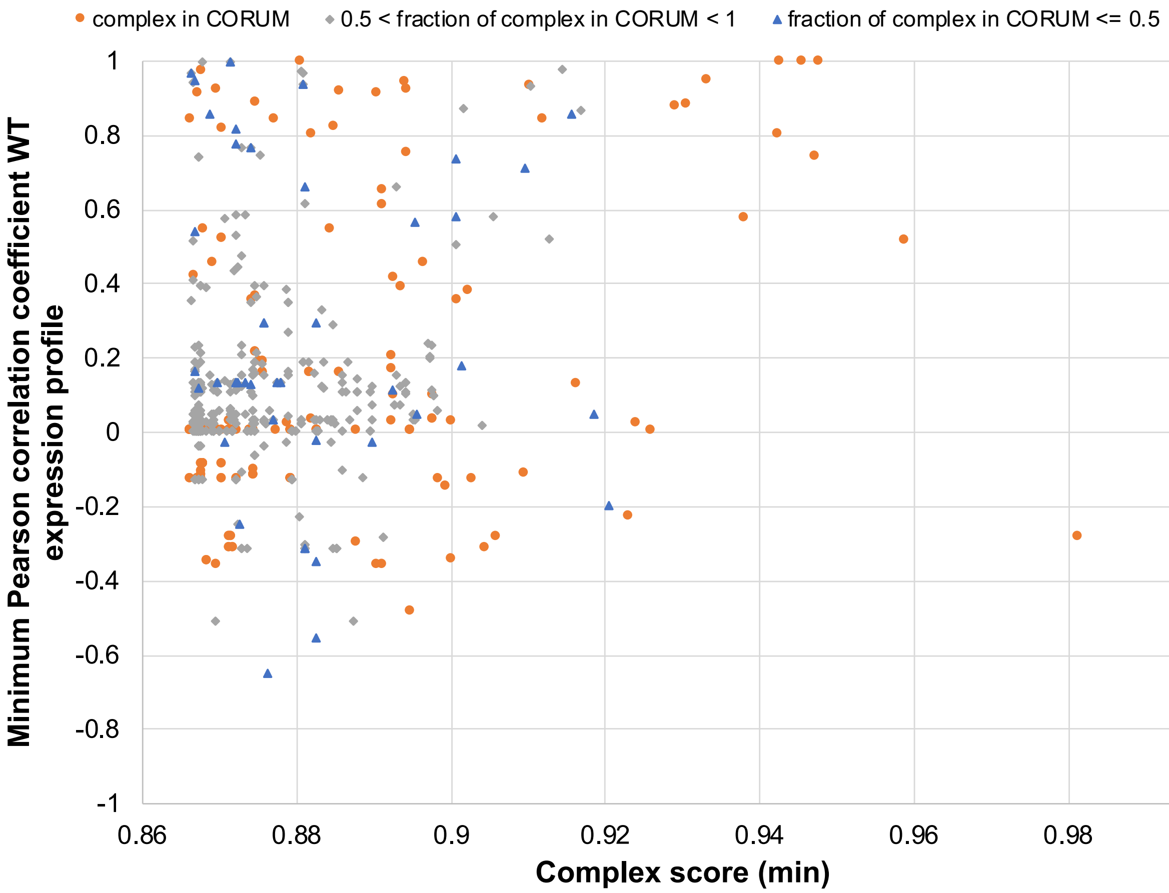

Figure S11: **Distribution of minimum Pearson's correlation coefficient as a function of complex score.** For each complex, we show the distribution of the minimum Pearson's correlation coefficient as a function of complex score. We further restrict the number of complexes to those with at least 50% of protein members with an entry in the expression profiles. Additionally, each complex is assigned to one of the following three classes based on varying degrees of overlap with CORUM complexes: (i) full overlap with CORUM complex (orange circles), (ii) at least half the member proteins overlap with CORUM complex (grey diamonds), and less than half of co-member proteins overlap with CORUM complex (blue triangles).

Supplementary files

Supplementary file 1 Full list of predicted proteins complexes available in the following link:
`http://murphylab.cbd.cmu.edu/software/2019_PPI/Uncharacterized_protein_complexes.xlsx`

References

[1] K. Drew, C. Lee, R. L. Huizar, F. Tu, B. Borgeson, C. D. McWhite, Y. Ma, J. B. Wallingford, and E. M. Marcotte. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol Syst Biol*, 13(6):932, 2017.

[2] M. Y. Hein, N. C. Hubner, I. Poser, J. Cox, N. Nagaraj, Y. Toyoda, I. A. Gak, I. Weisswange, J. Mansfeld, F. Buchholz, A. A. Hyman, and M. Mann. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*, 163(3):712–723, 2015.

[3] E. L. Huttlin, L. Ting, R. J. Bruckner, F. Gebreab, M. P. Gygi, J. Szpyt, S. Tam, G. Zarraga, G. Colby, K. Baltier, R. Dong, V. Guarani, L. P. Vaites, A. Ordureau, R. Rad, B. K. Erickson, M. Wühr, J. Chick, B. Zhai, D. Kolippakkam, J. Mintseris, R. A. Obar, T. Harris, S. Artavanis-Tsakonas, M. E. Sowa, P. De Camilli, J. A. Paulo, J. W. Harper, and S. P. Gygi. The BioPlex network: A systematic exploration of the human interactome. *Cell*, 162(2):425–440, 2015.

[4] C. Wan, B. Borgeson, S. Phanse, F. Tu, K. Drew, G. Clark, X. Xiong, O. Kagan, J. Kwan, A. Bezginov, K. Chessman, S. Pal, G. Cromar, O. Papoulas, Z. Ni, D. R. Boutz, S. Stoilova, P. C. Havugimana, X. Guo, R. H. Malty, M. Sarov, J. Greenblatt, M. Babu, W. B. Derry, E. R. Tillier, J. B. Wallingford, J. Parkinson, E. M. Marcotte, and A. Emili. Panorama of ancient metazoan macromolecular complexes. *Nature*, 525(7569):339–344, 2015.