

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Modeling direct protein interaction networks from mass spectrometry data

Permalink

<https://escholarship.org/uc/item/4t37f9dn>

Author

Palar, Aji

Publication Date

2024

Peer reviewed|Thesis/dissertation

Modeling direct protein interaction networks from mass spectrometry data

by
Aji Palar

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

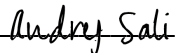
Biophysics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

Signed by:



C0A58494B03449B...

Andrej Sali

Chair

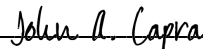
Signed by:



D0A58494B03449B...

Tanja Kortemme

Signed by:



CBF388B27CEB4AC...

John A. Capra

Committee Members

Copyright 2024

by

Aji Palar

To Max

Acknowledgements

First and foremost, I would like to thank my committee chair and advisor Andrej Sali who provided me with the freedom and support to pursue many opportunities at UCSF. Andrej, you taught me to sculpt the world. I would like to thank my committee members: Tanja Kortemme for her mentorship over the many years at UCSF and for her leadership as former director of the Biophysics program; Tony Capra for bringing a fresh perspective, you made this dissertation more accessible. Next, I would like to thank my former committee member Mike Keiser for providing a direct and complimentary perspective, thank you for taking the time and effort to support graduate students.

This dissertation would not have been possible without the help of several scientists. I am particularly grateful to my mentor and co-author Ignacia Echeverria – chapter 2 would not exist in its current form without our discussions and your input, thank you for helping me navigate both science and UCSF. Next, I would like to thank my co-author and collaborator Zhi Lin (Lindsey), our collaboration was one of the highlights of my PhD. Next, I'd like to thank Ben Webb, you seem to keep everything running as smoothly as possible. A big thanks to my former lab mates Ilan Chemmama, Seth Axen, and Sai Ganesan for welcoming me to the Sali lab. I would like to thank everyone I shared the lab space with: Daniel, Vipul, Tracy, Sree, Abantika, Andrew, Ken, Surabhi, Neelesh, Thomas, Kala, Dibyendu, Matthew, Cate, Eli, Shruthi, Elliot, Jeremey, Barak, Atreya, Arthur, Jared, Rakesh, Leah, and Vikas – I learned so much from all of you. Thank you to my friends and classmates: Christina, Calla, Elizabeth, Garrett, Maru, Matt, Wren, and Calla. Next, I would like to thank my dear Bay Area friends Aaron, Bryce, Christy, Daria, Emily, and Isaac. Our adventures are my best memories outside graduate school, see you soon...

Going further back I am indebted to many mentors and scientists. I would like to thank Richard Baker who welcomed me into Professor Andres Leschziner's research group at UC San Diego, I'll never see a model of a nucleosome without thinking of your density map. Andres, thank you for setting me on my path to UCSF, you taught me a method is a means to answer a question. I would especially like to thank my early

scientific teacher Jason Camara at Cabrillo College whose influence cannot be overemphasized, you taught me what complexity is.

Going further back to my childhood there are simply too many people to thank, I would like to thank my wilderness awareness mentors who sparked my curiosity of the natural world. Thank you, Alan, Greg, Mark, Neill, Tyler, and Paul. Next, I would like to thank my family for their love and support: my mother Valerie who taught me resilience, my sister Kartika who taught me trust, my brother Adhi our love of games, my Uncle Eric for prioritizing family, my grandmother Daisy for her thoughtful gifts and encouragement, and my grandfather Hank for our wide-ranging conversations. Finally, thank you Aish for all your kindness and support – this dissertation wouldn't exist without you.

Contributions

Chapter 2

Palar A, Echeverria I, Sali A. Integrative modeling of direct protein interaction networks based on affinity purification mass spectrometry data. *bioRxiv*. 2024.

Chapter 3

Lin Z, Schaefer K, Lui I, Yao Z, Fossati A, Swaney D, Palar A, Sali A, Wells J. Multi-scale photocatalytic proximity labeling reveals cell surface neighbors on and between cells.” *Science* 2024;385:ead15763.

Modeling direct protein interaction networks from mass spectrometry data

Aji Palar

Abstract

A complex network of molecular interactions underpins cellular physiology, with each interaction contributing to the cell's overall function. In normal physiological states, these networks are tightly regulated, but in disease, their structure and dynamics can shift, leading to dysregulations and pathogenesis. Predicting the structure of disease-relevant networks has the potential to enhance therapeutic target identifications, improve disease prognosis predictions, and refine models of complex molecular systems. In the first half of this work we develop, implement, benchmark, and illustrate Integrative Network Modeling, an algorithm for modeling disease relevant protein interaction networks based on affinity purification mass spectrometry (AP-MS) data. We find AP-MS experiments contain more information about a protein's direct protein interactions than previously thought. In the second half of this work, we predict the presence of protein interactions in the epidermal growth factor receptor (EGFR) molecular neighborhood using proximity labeling mass spectrometry. We apply a deep-learning model to predict the three-dimensional structure of EGFR binary complexes; we identify multiple proteins in complex with EGFR. The computational methods developed and applied in these studies are aimed at modeling complex molecular systems based on the integration of information from mass spectrometry and protein structure. Together, they are a step towards bridging the gap between structural and systems biology.

Table of Contents

Introduction.....	1
1.1 Models of complex biological systems based on mass spectrometry data.....	1
1.2 Limitations of existing methods for constructing models based on mass spectrometry data.....	3
1.3 Modeling in structural and systems biology contexts.....	5
1.4 Network modeling as an optimization problem.....	6
1.5 References.....	7
Integrative modeling of direct protein interaction networks based on affinity purification mass spectrometry data.....	11
2.1 Abstract.....	11
2.2 Introduction.....	11
2.3 Methods.....	14
2.4 Results.....	21
2.5 Discussion.....	24
2.6 Availability of software and data.....	30
2.7 Acknowledgements.....	30
2.8 Figures.....	31
2.9 References.....	39
Multi-scale photocatalytic proximity labeling reveals cell surface neighbors on and between cells.....	42
3.1 Abstract.....	42
3.2 Introduction.....	42

3.3 Results.....	44
3.4 Discussion.....	54
3.5 Acknowledgements.....	56
3.6 Declaration of interests.....	57
3.7 Data and materials availability.....	57
3.8 Figures.....	58
3.9 References.....	83
Conclusion.....	91
4.1 Improvements to integrative network modeling.....	91
4.2 Bridging the gap between structural and systems biology.....	93
4.3 References.....	95

Chapter 1

Introduction

A complex network of molecular interactions underpins cellular physiology, with each interaction contributing to the cell's overall function (Kuchina et al. 2022; Robinson, Sali, and Baumeister 2007). In normal physiological states, these networks are tightly regulated, but in disease, their structure and dynamics can shift, leading to dysregulations and pathogenesis. Predicting the structure of disease-relevant networks has the potential to enhance therapeutic target identifications, improve disease prognosis predictions, and refine models of complex molecular systems (Gordon, Jang, et al. 2020; Singh et al. 2024; Krogan et al. 2006; Jäger et al. 2011; Gordon, Watson, et al. 2020). In systems biology, determining these networks is crucial for describing how molecular interactions drive cellular behavior and for generating unbiased hypotheses for mechanistic studies and as input for other models (Raveh et al. 2021).

1.1 Models of complex biological systems based on mass spectrometry data

Protein interaction networks have traditionally been based on yeast two-hybrid (Y2H) (Gavin et al. 2002; Ito et al. 2001; Bader and Hogue 2002) or affinity purification mass spectrometry (AP-MS) data (Krogan et al. 2006; Ho et al. 2002). Networks based solely on Y2H experiments are “binary” in that each experiment only detects the presence of a single direct protein interaction through the activation of a reporter gene. In contrast, an AP-MS experiment can identify and quantify multiple types of proteins functionally linked to an affinity tagged “bait” protein. Typically, an AP-MS network is based on the integration of multiple AP-MS experiments that include both replicate experiments (using the same bait and condition) and experiments using different baits (including replicates). Recent AP-MS studies map protein interaction networks in specific disease contexts such as HIV infection (Jäger et al. 2011; Hüttenhain et al. 2019), SARS-CoV-2 infection (Gordon, Jang, et al. 2020), autism spectrum disorder (Wang et al. 2024), and certain cancers (Swaney et al. 2021). The networks obtained from such studies represent “snapshots” of a protein interaction network for a certain cell type under specific experimental conditions. These networks have led to the identification of disease-relevant protein complexes and functional modules. Chapter 2 deals with

the development, implementation, and assessment of a statistical inference framework for modeling direct protein interaction networks from AP-MS data.

Proximity labeling proteomics (PLP) is a method capable of detecting physically proximal proteins that may interact transiently. Compared to AP-MS and Y2H experiments, proximity labeling may detect proteins that are on average, further away (on the order of 1-10nm). Chapter 3 deals with the development of a photoactivatable proximity labeling strategy and an application of proximity labeling the epidermal growth factor receptor (EGFR) molecular neighborhood. Here, I orthogonally validate candidate EGFR neighbors by: (i) predicting the quaternary structure of EGFR-neighbor binary complexes, and (ii) assessment of the structure quality using various metrics.

Efforts to integrate multiple types of proteomics experiments into a single “map” of protein interactions have been made. One standout example is the human protein complex map 2.0 (hu.MAP 2.0), which is based on AP-MS, proximity labeling, and other data. The confidence scores obtained for hu.MAP 2.0 are highly accurate for predicting the presence of protein interactions (Drew, Wallingford, and Marcotte 2021). More recently, proteomics data has been used in combination with imaging data to produce other representations of the spatial localization of protein molecules in a cell. Standout examples include the OpenCell project (Cho et al. 2022) and the Multi-Scale Integrated Cell (MuSIC) map (Qin et al. 2021).

Finally, AP-MS and chemical cross-linking mass spectrometry has been used to model the macromolecular architecture of multi-protein complexes – a prototypical example is the integrative modeling of the yeast nuclear pore complex (NPC) (Alber, Dokudovskaya, Veenhoff, Zhang, Kipper, Devos, Suprpto, Karni-Schmidt, Williams, Chait, Rout, et al. 2007; Alber, Dokudovskaya, Veenhoff, Zhang, Kipper, Devos, Suprpto, Karni-Schmidt, Williams, Chait, Sali, et al. 2007).

A note on the inconsistent use of the term “protein-interaction”

The term “protein interaction” is defined differently in different domains of science. In a structural biology context, “protein interaction” refers to the direct binding of two protein molecules. In systems biology, a

“protein interaction” may refer to the observation of one type of protein under an experiment with a particular “bait” protein. For AP-MS this “prey” is a member of one of the complexes the bait participates in. In this sense, the prey may interact directly with the bait or not. In the introduction we use the term “protein interaction” in the looser systems biology sense.

1.2 Limitations of existing methods for constructing models based on mass spectrometry data

Protein interaction networks based on AP-MS data have been critical for identifying disease-relevant complexes, functional modules, and for generating mechanistic hypotheses. Despite their utility, the methods of constructing them are limited by: (i) the physical meaning of an edge may be undefined, (ii) uncertainties in the input information are not properly propagated to the output model, (iii) construction of a network through pairwise feature calculation may limit accuracy, (iv) networks are modeled as a single state but may exist in multiple states, and (v) mass spectrometry data likely contain inaccuracies.

The physical meaning of an edge may be undefined

A network model consists of a graph where nodes represent a type of molecule. In general, a node stands for many copies of a protein molecule in the sample(s) used to construct the model. In some cases, an edge is rigorously defined by a probabilistic model. For example, in the Significance Analysis of INteractome (SAINT) a weighted edge is the posterior probability a “prey” co-purifies with a “bait” under a mixture model (Choi et al. 2011). In chapter 2, we define an edge as the presence of at least one protein interaction between two molecules represented by the respective nodes. For anything to have “physical meaning” it must be formally defined in a theoretical framework (e.g., the physical meaning of an atom in molecular dynamics is different from the physical meaning of an atom in quantum mechanics). Often, an edge is the result of a series of information processing steps. For example, a BioGRID “physical interaction” may be based on AP-MS bait-prey pairs or co-crystal structures (Stark et al. 2006). As such a “physical interaction” is a result of the BioGRID curation process. Similar definitions exist for other databases such as STRING (Szklarczyk et al. 2023) and CORUM (Tsitsiridis et al. 2023). The curation efforts may not be easy to

understand. In the case of manual curation, they may not be reproducible. Despite these critiques, the above databases are often useful. For example, new experimental techniques may be assessed based on their enrichment of certain types of “physical interactions.”

Uncertainties in input information are not propagated to the output model

All methods of constructing a protein interaction network are limited by uncertainties in the input data. Such uncertainties should be explicitly and accurately represented when possible. It is common to represent uncertainty in the presence of a protein interaction with a confidence score with a value between 0 (no interaction) and 1 (protein interaction) (e.g., the hu.MAP 2.0 support vector machine (SVM) score and the SAINT score). It is less common to represent uncertainty in the entire network. This may be because: (i) many networks are constructed with each edge constructed independently of every other edge – thus representing the uncertainty of every edge independently is equivalent to representing the uncertainty of the network, (ii) protein interaction networks tend to be high-dimensional (thousands of nodes, hundreds of thousands to millions of edges, tens of millions of possible edges), analysis of such high-dimensional objects may be more challenging or impractical, (iii) the construction of protein interaction networks have not been formalized as an inference problem.

Construction of networks by pairwise feature calculation may limit accuracy

As mentioned above, many protein interaction networks are constructed by taking the union of pairwise protein interactions. For AP-MS a bait-prey pairwise interaction may be constructed based on mass spectrometry features (e.g., reproducibility of prey identification across replicate experiments). Real protein interaction networks likely have statistical properties that may be informative when trying to construct new protein interaction networks – for example they may be scale-free or have an average degree (Krogan et al. 2006). Pairwise construction of a protein interaction network cannot accommodate the integration of such statistical preferences. More often, these preferences are applied during post-hoc analysis (e.g., degree filtering) (Shannon et al. 2003).

Networks are represented as a single state but may exist in multiple states

Protein interaction networks typically depict a single “state”. Real protein interaction networks likely vary in time, as a function of experimental perturbation, or disease. Comparison of two protein interaction networks (two states A and B) is common (i.e., differential network analysis), the simultaneous modeling of multiple network states is not. Simultaneous modeling of multiple network states could allow for more principled incorporation of information, for example a hierarchical Bayesian model could be used where a prior network informs both network states A and B.

Mass spectrometry data likely contain inaccuracies

Mass spectrometry will likely be inaccurate or biased towards certain proteins. For example, MS may favor proteins that produce many peptides, highly abundant proteins, or proteins in particular cellular compartments (e.g., non-membrane proteins). As such, the construction of a protein interaction network would ideally be based on multiple types of orthogonal information.

To motivate the development of integrative network modeling in chapter 2 and our application of protein-complex structure prediction in chapter 3, we next introduce modeling in two domains of science.

1.3 Modeling in structural and systems biology contexts

A “model” is a depiction of a system that is useful for rationalizing existing information and for making predictions about outcomes of future experiments (Rout and Sali 2019). Different types of models are characteristic of different domains of science. For example, in traditional structural biology, an archetypal model is a model of the three-dimensional structure of a protein molecule in a low energy conformation. In systems biology, an archetypal model is a protein interaction network where nodes represent types of proteins and edges represent some physical ‘link’ between two protein types (Ho et al. 2002; Robinson, Sali, and Baumeister 2007). In both domains, other model representations occur (e.g., a coarse-grained structure, a gene regulatory network). Modeling is the process of converting the input information about the system into a model of the system. A protein interaction network is a model of how different types of

protein molecules are “wired” together in some cell-type that is useful for identifying functional complexes and modules.

1.4 Network modeling as an optimization problem

To address some of these limitations, we were motivated to develop integrative network modeling (INM) – a computational method to model direct protein interaction networks based on primarily AP-MS data.

Modeling proceeds by (i) gathering input information that informs the output network model, (ii) defining a network model representation where a node represents a protein type and an edge represents a direct physical interaction, (iii) constructing a scoring function that quantifies an agreement of a network model with the input information, (iv) sampling alternative network models guided by the scoring function, and (v) analyzing and assessing the network models.

The advantages of INM are: (i) its explicit representation of direct protein interaction networks, (ii) the accuracy and extensibility of its scoring function, (iii) its data efficiency, (iv) the flexibility and computational efficiency of sampling, and (v) its ability to characterize new systems under specific experimental conditions. These advantages are discussed in detail in chapter 2.

1.5 References

1. Alber, Frank, Svetlana Dokudovskaya, Liesbeth M. Veenhoff, Wenzhu Zhang, Julia Kipper, Damien Devos, Adisetyantari Suprpto, Orit Karni-Schmidt, Rosemary Williams, Brian T. Chait, Michael P. Rout, et al. 2007. “Determining the Architectures of Macromolecular Assemblies.” *Nature* 450 (7170): 683–94.
2. Alber, Frank, Svetlana Dokudovskaya, Liesbeth M. Veenhoff, Wenzhu Zhang, Julia Kipper, Damien Devos, Adisetyantari Suprpto, Orit Karni-Schmidt, Rosemary Williams, Brian T. Chait, Andrej Sali, et al. 2007. “The Molecular Architecture of the Nuclear Pore Complex.” *Nature* 450 (7170): 695–701.
3. Bader, Gary D., and Christopher W. V. Hogue. 2002. “Analyzing Yeast Protein-Protein Interaction Data Obtained from Different Sources.” *Nature Biotechnology* 20 (10): 991–97.
4. Choi, Hyungwon, Brett Larsen, Zhen-Yuan Lin, Ashton Breikreutz, Dattatreya Mellacheruvu, Damian Fermin, Zhaohui S. Qin, Mike Tyers, Anne-Claude Gingras, and Alexey I. Nesvizhskii. 2011. “SAINT: Probabilistic Scoring of Affinity Purification-Mass Spectrometry Data.” *Nature Methods* 8 (1): 70–73.
5. Cho, Nathan H., Keith C. Cheveralls, Andreas-David Brunner, Kibeom Kim, André C. Michaelis, Preethi Raghavan, Hirofumi Kobayashi, et al. 2022. “OpenCell: Endogenous Tagging for the Cartography of Human Cellular Organization.” *Science (New York, N.Y.)* 375 (6585): eabi6983.
6. Drew, Kevin, John B. Wallingford, and Edward M. Marcotte. 2021. “hu.MAP 2.0: Integration of over 15,000 Proteomic Experiments Builds a Global Compendium of Human Multiprotein Assemblies.” *Molecular Systems Biology* 17 (5): e10016.
7. Gavin, Anne-Claude, Markus Bötsche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jörg Schultz, et al. 2002. “Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes.” *Nature* 415 (6868): 141–47.
8. Gordon, David E., Gwendolyn M. Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M. White, Matthew J. O’Meara, et al. 2020. “A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug Repurposing.” *Nature* 583 (7816): 459–68.

9. Gordon, David E., Ariane Watson, Assen Roguev, Simin Zheng, Gwendolyn M. Jang, Joshua Kane, Jiewei Xu, et al. 2020. “A Quantitative Genetic Interaction Map of HIV Infection.” *Molecular Cell* 78 (2): 197–209.e7.
10. Ho, Yuen, Albrecht Gruhler, Adrian Heilbut, Gary D. Bader, Lynda Moore, Sally-Lin Adams, Anna Millar, et al. 2002. “Systematic Identification of Protein Complexes in *Saccharomyces Cerevisiae* by Mass Spectrometry.” *Nature* 415 (6868): 180–83.
11. Hüttenhain, Ruth, Jiewei Xu, Lily A. Burton, David E. Gordon, Judd F. Hultquist, Jeffrey R. Johnson, Laura Satkamp, et al. 2019. “ARIH2 Is a Vif-Dependent Regulator of CUL5-Mediated APOBEC3G Degradation in HIV Infection.” *Cell Host & Microbe* 26 (1): 86–99.e7.
12. Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. 2001. “A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (8): 4569–74.
13. Jäger, Stefanie, Peter Cimermancic, Natali Gulbahce, Jeffrey R. Johnson, Kathryn E. McGovern, Starlynn C. Clarke, Michael Shales, et al. 2011. “Global Landscape of HIV-Human Protein Complexes.” *Nature* 481 (7381): 365–70.
14. Krogan, Nevan J., Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, et al. 2006. “Global Landscape of Protein Complexes in the Yeast *Saccharomyces Cerevisiae*.” *Nature* 440 (7084): 637–43.
15. Kuchina, Anna, Jason Yang, Bree Aldridge, Kevin A. Janes, Naeha Subramanian, Nevan J. Krogan, Mehdi Bouhaddou, Shirit Einav, Jason Papin, and Ronald N. Germain. 2022. “How Can Systems Approaches Help Us Understand and Treat Infectious Disease?” *Cell Systems* 13 (12): 945–49.
16. Qin, Yue, Edward L. Huttlin, Casper F. Winsnes, Maya L. Gosztyla, Ludivine Wacheul, Marcus R. Kelly, Steven M. Blue, et al. 2021. “A Multi-Scale Map of Cell Structure Fusing Protein Images and Interactions.” *Nature* 600 (7889): 536–42.
17. Raveh, Barak, Liping Sun, Kate L. White, Tanmoy Sanyal, Jeremy Tempkin, Dongqing Zheng, Kala Bharath, et al. 2021. “Bayesian Metamodeling of Complex Biological Systems across Varying

- Representations.” Proceedings of the National Academy of Sciences of the United States of America 118 (35): e2104559118.
18. Robinson, Carol V., Andrej Sali, and Wolfgang Baumeister. 2007. “The Molecular Sociology of the Cell.” *Nature* 450 (7172): 973–82.
 19. Rout, Michael P., and Andrej Sali. 2019. “Principles for Integrative Structural Biology Studies.” *Cell* 177 (6): 1384–1403.
 20. Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11): 2498–2504.
 21. Singh, Digvijay, Neelesh Soni, Joshua Hutchings, Ignacia Echeverria, Farhaz Shaikh, Madeleine Duquette, Sergey Suslov, et al. 2024. “The Molecular Architecture of the Nuclear Basket.” *Cell* 187 (19): 5267–81.e13.
 22. Stark, Chris, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. “BioGRID: A General Repository for Interaction Datasets.” *Nucleic Acids Research* 34 (Database issue): D535–39.
 23. Swaney, Danielle L., Dana J. Ramms, Zhiyong Wang, Jisoo Park, Yusuke Goto, Margaret Soucheray, Neil Bhola, et al. 2021. “A Protein Network Map of Head and Neck Cancer Reveals PIK3CA Mutant Drug Sensitivity.” *Science (New York, N.Y.)* 374 (6563): eabf2911.
 24. Szklarczyk, Damian, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, et al. 2023. “The STRING Database in 2023: Protein-Protein Association Networks and Functional Enrichment Analyses for Any Sequenced Genome of Interest.” *Nucleic Acids Research* 51 (D1): D638–46.
 25. Tsitsiridis, George, Ralph Steinkamp, Madalina Giurgiu, Barbara Brauner, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. 2023. “CORUM: The Comprehensive Resource of Mammalian Protein Complexes-2022.” *Nucleic Acids Research* 51 (D1): D539–45.

26. Wang, Belinda, Rasika Vartak, Yefim Zaltsman, Zun Zar Chi Naing, Kelsey M. Hennick, Benjamin J. Polacco, Ali Bashir, et al. 2024. “A Foundational Atlas of Autism Protein Interactions Reveals Molecular Convergence.” bioRxiv.org: The Preprint Server for Biology, February. <https://doi.org/10.1101/2023.12.03.569805>.

Chapter 2

Integrative modeling of direct protein interaction networks based on affinity purification mass spectrometry data

2.1 Abstract

Accurate, precise, and complete modeling of a cellular protein interaction network requires the integration of large amounts of information. We have therefore developed integrative network modeling (INM) – a statistical inference framework extensible to multiple types of data that proceeds by (i) representing a network where nodes represent protein types and edges represent the presence of at least one direct interaction between proteins of the corresponding nodes, (ii) scoring the agreement between a network and input information, (iii) sampling alternative configurations of network models using a Hamiltonian Monte Carlo scheme, and (iv) analyzing the output sample of network models. To illustrate INM, we model the protein interaction network of a host-pathogen system based on affinity purification mass spectrometry (AP-MS) experiments. We benchmark our model on direct protein interactions extracted from the Protein Data Bank. Modeling predicts physical interactions with an AUC accuracy of 0.82. We compare INM networks to networks generated by state-of-the-art methods, provide recommendations for future models incorporating AP-MS data, and explore additional data types that can be incorporated into the INM framework. We illustrate the utility of INM on the HIV-1 CRL5 host-pathogen protein interaction network.

2.2 Introduction

A complex network of molecular interactions underpins cellular physiology, with each interaction contributing to the cell's overall function. These networks are tightly regulated in normal physiological states, but their structure and dynamics can shift in disease, leading to dysregulations and pathogenesis. Predicting the structure of disease-relevant networks might enable therapeutic target identifications, improve disease prognosis predictions, and refine models of complex molecular systems. In systems

biology, determining these networks is crucial for describing how molecular interactions drive cellular behavior, generating unbiased hypotheses for mechanistic studies, and as input for other models.

One of the most successful ways to model protein interaction networks is based on Affinity-Purification Mass Spectrometry (AP-MS) studies (Gavin et al. 2006; Krogan et al. 2006; Ewing et al. 2007; Jeronimo et al. 2007; Richards, Eckhardt and Krogan 2021). In these purifications, a tagged “bait” protein is expressed in some cell type, and co-purifying “prey” proteins are identified by mass spectrometry (MS). Statistical scoring models, such as SAINT, ComPASS, or MiST (Choi et al. 2011; Jäger et al. 2011; Wenger et al. 2011), are then applied to assign confidence scores to bait-prey interactions, distinguishing true, co-purifying preys from non-specific background contaminants. These confidence scores are subsequently used to generate a "spoke" model, where each interaction is represented as an edge between the bait node and each prey node, but not between the prey proteins themselves. The “spoke” model simplifies the representation of the interaction network, assuming a direct link between the bait and its identified preys, which may include direct or indirect interactions. Here, we define a direct interaction as those in which the proteins share a physical interface. Direct interactions may occur transiently or be stable. Although the bait-prey interactions predicted by these models do not always imply a direct protein interaction, many bait-prey pairs physically interact (Wang et al. 2024).

Knowing the structure of direct protein interaction networks is crucial to understanding the functional modules modulating biological processes. Further, these networks can serve as input for other modeling methods approaches, including integrative structural modeling (Webb et al. 2018) and machine learning models (Drew, Wallingford and Marcotte 2021; Qin et al. 2021; Cho et al. 2022), which might assist in identifying the functional protein complexes (Robinson, Sali and Baumeister 2007).

One approach to identifying the components of protein complexes is through matrix models (Hart, Lee and Marcotte 2007; Drew et al. 2017); in this framework, all proteins identified in an AP-MS experiment are included, regardless of their classification as prey or bait. The matrix representation enables the representation of relationships between proteins in a structured format, where rows and columns correspond

to individual protein types, and entries indicate interactions, typically marked by values of 1 for interaction and 0 for no interaction (Hart, Lee and Marcotte 2007). This model corresponds to a network, where each node corresponds to a row-column pair, and each edge corresponds to an entry in the matrix. This method facilitates the visualization and analysis of complex interaction networks, enabling the identification of clusters and exploration of network topology (Krogan et al. 2006; Hart, Lee and Marcotte 2007). Another approach used to identify protein assemblies is to integrate large amounts of proteomics data of varying kinds using machine learning (ML) (Drew, Wallingford and Marcotte 2021). While these ‘big-data’ approaches are accurate in predicting certain protein interactions, they may be limited in predicting transient interactions or condition-specific interactions.

Here, we develop integrative network modeling (INM), a statistical inference framework to model direct protein interactions that is extensible to multiple types of data. INM can predict not only bait-prey interactions but also prey-prey interactions. INM requires relatively few AP-MS purifications – on the order of tens – comparable to ‘spoke’ type models. Critically, INM requires orders of magnitude less information than current ML models, increasing its scope of use. The efficiency of data usage allows for studies of PPI networks under specific conditions, including the use of chemical perturbations such as small molecules, and facilitates the characterization of cell-type specific protein interaction networks.

The inference framework offers several key advantages. First, it can easily incorporate new types of data that may provide insight into the protein interaction networks. For example, it can be extended to incorporate data from proximity labeling MS, in situ/in vivo cross-linking MS, large-scale functional genetic interaction or CRISPRi/a experiments, or AI-derived pairwise protein structure predictions (Larson et al. 2013; Jumper et al. 2021; Braberg et al. 2022; Lin et al. 2024). Second, it can account for higher-order network properties (e.g., degree, centrality, path-length), using information from known biological networks. Third, the framework by design includes a process that captures the uncertainty in the input information, allowing us to evaluate the confidence in predicted interaction. Fourth, in some cases, it may be possible to obtain multiple network ‘states’ out of a computed network model; enabling us to describe

PPI networks under different experimental conditions or with different mutations in the components. Fifth, the network models can be used to prioritize virtual or experimental PPI screening. Finally, the output of network models could be useful as an input for other models. For example, the predicted physical interaction could be incorporated into integrative structure modeling calculations or as an input feature in a deep-learning model.

This paper is organized as follows. First, we describe the INM modeling framework (**Fig. 1**) consisting of (i) gathering the input information, (ii) representing the network model, (iii) building an extensible scoring function, (iv) Monte Carlo sampling, and (v) assessing the output models. Second, we describe how the model was benchmarked using a PDB derived gold-standard dataset. Third, we compare our computed network to other state-of-the-art methods and previously computed networks. Finally, we illustrate the INM approach by predicting the HIV-1 CRL5 protein interaction network.

2.3 Methods

We aim to model a network of direct physical interactions between pairs of protein types based on AP-MS data; a pair of protein types interact directly when their molecules are in physical contact with each other. Our network modeling consists of: (i) gathering input information that informs the output network model, (ii) defining a network model representation where a node represents a protein type and an edge represents a direct physical interaction, (iii) constructing a scoring function that quantifies an agreement of a network model with the input information, (iv) sampling alternative network models guided by the scoring function, and (v) analyzing and assessing the network models. We now discuss each stage in turn.

Integrative network modeling

Gathering input information

Input information for our network modeling consists of an input set of K AP-MS purifications. Each AP-MS purification identifies the protein interactions involving a specific 'bait' protein, which captures its directly or indirectly interacting partners, referred to as 'preys'. Both 'prey' and 'bait' proteins are quantified

by their spectral counts. If a protein is not identified in a purification, the spectral counts are set to 0. The input set of AP-MS purifications contains replicas obtained from multiple experiments using the same bait and experimental conditions as well as AP-MS purifications with different baits and/or experimental conditions.

Additionally, we use the SAINT (Significance Analysis of INTeractome) confidence scores. The SAINT score is a probabilistic score that quantifies the confidence of each bait-prey interaction based on the spectral count data, adjusting for experimental noise and non-specific background interactions (Choi et al. 2011).

Representation of network model

The network model representation defines the variables whose values are determined by modeling. Given the N proteins identified in the input set of AP-MS purifications, each represented as a network node, the model variables include edge variables A for all pairs of nodes and additional nuisance variables γ , whose values are computed by modeling but are generally not of primary interest when using a model (Rieping, Habeck and Nilges 2005). The edge variables indicate the presence (value of 1) or absence (value of 0) of an edge between two nodes; there are $N(N-1)/2$ such variables for N nodes in the network. In general, a node stands for many copies of a protein molecule in the sample(s) used for AP-MS. As mentioned above, an edge between two nodes represents a direct interaction between any pair of protein molecules represented by the two nodes. Hence, our network model representation includes a matrix model (Hart, Lee and Marcotte 2007) representation and the nuisance variables.

As a consequence of these definitions, a set of nodes in which each node is connected to at least one other node does not necessarily correspond to a single physical complex of the proteins represented by the nodes in the set. The connected nodes could also represent a mixture of complexes. In contrast, proteins in a complex will always form a set of connected nodes. For instance, a protein may be present in different cellular environments, each with a distinct set of interacting partners. In this scenario, the network model

will include edges connecting the protein to both sets of partners, even if no single protein molecule interacts with all the partners simultaneously.

Scoring of network models

The purpose of a scoring function is to rank alternative network models according to their agreement with the input information. Our scoring function is:

$$S(M) = S_R(M) + S_E(M) + S_S(M) + S_D(M), \quad (1)$$

where $S_R(M)$ reflects the deviation between the edges in M and the expected edges based on the AP-MS data; $S_E(M)$ reflects the deviation between the number of edges in M and the expected number of edges, which in turn depends on the number of nodes; $S_S(M)$ reflects the deviation between the edges in M and the expected edges based on a SAINT score (Choi et al. 2011); and $S_D(M)$ reflects the deviation between the degree of each node and its expected degree. Next, we define each score in turn.

$S_R(M)$ scores the value of each edge variable A_{ij} , based on all AP-MS purifications. We first construct the AP-MS profile for each protein: the AP-MS profile of a protein, whether a bait or prey, is a vector of its spectral counts from all K AP-MS purifications. We noticed that protein type pairs interacting directly or indirectly (*i.e.*, they co-occur in the same protein complex) tend to have a high AP-MS profile correlation, while protein type pairs that are not members of the same complex tend to have low profile correlations (**Fig. 2**). Thus, we quantify the observation of an edge between two nodes by calculating the Pearson correlation coefficient (R) of their associated AP-MS profiles. Based on these observations, we score edge variables as follows:

$$S_R(M) = -\sum_{i<j} \log[N(Z_{ij} | R_{ij} - 0.5, R_{ij}^2 + \epsilon)], \quad (2)$$

where N is a Gaussian distribution, Z_{ij} is a latent representation of the edge variable A_{ij} , R_{ij} is the AP-MS profile correlation coefficient for the protein type pair identified by i and j , and ϵ is a small numerical constant to ensure the distribution is defined if R_{ij} equals 0. Values of Z_{ij} less than 0.5 map to an edge

variable value A_{ij} of 0 while values of Z_{ij} equal to or greater than 0.5 map to edge variable values of A_{ij} equal to 1. The mapping is performed using a logistic function. In this formulation, for node pairs with low AP-MS profile correlation coefficients, the density around Z_{ij} is centered near -0.5 with small variance, effectively restraining A_{ij} to 0. In contrast, for node pairs with high AP-MS profile correlation coefficients, the density is centered near 0.5 with a wide variance, allowing A_{ij} to take on values of either 0 or 1. S_S scores the value of each edge variable based on the SAINT pair score:

$$S_S(M) = -\sum_{i<j} \log[N(Z_{ij} | p_{ij} - 0.5, p_{ij}^2 + \epsilon)], \quad (3)$$

where the SAINT pair score p_{ij} is the multiplication over the maximal SAINT score of each prey over all input purifications.

S_E scores the number of edges N_E in M based on the expected number of edges:

$$S_E(M) = -\log[B(N_E | u, n)U(u | \alpha, \beta)], \quad (4)$$

where B is a Binomial distribution, u is the number of edges in the network relative to the total number of possible edges n , U is a uniform distribution between α and β . α and β were manually set to 0.02 and 0.1.

S_D scores the node degree (number of edges connected to a given node) based on the expected node degree:

$$S_D(M) = -\sum_{i<j} \log[N(\text{deg}(i) | \mu_i, \sigma_i)], \quad (5)$$

where N is a Gaussian distribution, $\text{deg}(i)$ is the degree of node i , μ_i is the expected degree, and σ_i is the expected standard deviation. Both μ_i and σ_i were set to 3 for all nodes. A statistical sample of direct protein interaction networks could be used to place more informative prior parameters μ_i and σ_i .

Sampling of network models

Alternative network models were sampled using the No-U-Turn (NUTS) adaptive Hamiltonian Monte Carlo (HMC) sampler (Neal 2012; Mathew D Hoffman 2014). Our scoring function corresponds to the Hamiltonian potential energy, with auxiliary kinetic energy momentum variables represented as

uncorrelated multivariate Gaussians. In our definition, the edge variables are discrete while HMC requires them to be continuously differentiable. Therefore, we represent edge variables in a latent space Z and map from Z to A using a logistic function with a midpoint at 0.5 and a slope of 1000. The step size and number of leapfrog integrator steps are tuned by the integrator during a warmup period of 1000 steps. A random initial network configuration was chosen by assigning each latent variable a random value in the $(-2, 2)$ range. Twenty HMC chains were run in parallel for 20,000 steps each.

Analyzing the network models

A model must be assessed before its interpretation. We followed an assessment protocol similar to the one developed for integrative structure modeling (Viswanath et al. 2017; Saltzberg et al. 2021). We began by assessing the convergence of sampling using two criteria as follows. First, we assessed whether the best model score continues to improve as more models are sampled. Random model score subsets of several sizes (*e.g.*, 20, 40, 60, 80%, and the complete set) are each created several times (replicates), simulating increasing amounts of sampling. The best score in each subset is averaged across the replicates. Plotting the average best score for each model subset size shows whether the best score converges as the number of sampled models increases. Second, we assessed the similarity of score distributions for two unrelated model subsets. Specifically, we used a two-sided nonparametric Kolmogorov-Smirnov test (Siegal and Castellan 1988) to compare the two samples of scores, estimating the odds of the two score samples originating from the same parent distribution; we used a P -value threshold of 0.01 and a test-statistic threshold D of 0.03 to assess significance (**Fig. 3**).

Construction of an average network model

We constructed an average network model by averaging the sampled edge values over all the sampled network models. We used this average network model to compute a receiver operator characteristic (ROC) curve on our benchmark dataset. We define accuracy as the area under the ROC curve (AUC). Similarly to our comparison of the score distributions, we constructed two average network models for the two model

subsets and assessed their AUC accuracy. We also assessed AUC accuracy of three randomly selected network models. Finally, we determined the AUC accuracy as a function of the number of network models N used to calculate the average network model.

Modeling the HIV-CRL5 network

To benchmark and illustrate INM, we modeled the HIV-1 Cullin RING-E3-ligase (HIV-CRL5) network based on 64 AP-MS purifications performed in Jurkat cells under 3 viral infection conditions (Hüttenhain et al. 2019); a wild-type infection (WT), a Vif deletion (Δ VIF), and a mock infection. The protein level spectral count data were obtained exactly as published. The baits are CUL5, ELOB, CBF β , and LRR1. We selected 236 of 3,062 of the identified proteins to represent as nodes, requiring each protein to co-purify at a SAINT score of at least 0.5 in at least one AP-MS purification. Next, we applied INM for each of the three infection conditions independently and for all purifications combined.

Extracting reference edges from the PDB for benchmarking

To benchmark our method we compared the average edge values with direct protein interactions obtained from the structures of protein complexes. Sequence pairs were used to query the Protein Data Bank clustered at 70% sequence identity (PDB70) for matching chain pairs present in the same PDB file. The PDB70 was queried using all 4,686,391 pairwise combinations of sequences from the input set of 3,062 identified proteins on 9/12/2024. A pair of chains is defined as interacting if they bury at least 500 Å² of solvent-accessible surface area (SASA), according to the Shrake and Rupley algorithm (Shrake and Rupley 1973; Kunzmann et al. 2023) with a probe size of 1.4 Å. A chain is considered matching if it shares at least 30% sequence identity with the query sequence, is at least 88 amino acids long, and has an alignment e-value of at least 1e-7. Sequence alignments were performed by comparing multiple sequence alignment profiles represented as hidden Markov models using the program HHblits (Remmert et al. 2011). This procedure resulted in 2,985 direct edges (*i.e.*, corresponding to a physical interaction), and 15,666 edges

corresponding to chain pairs co-occurring in a PDB file. Benchmarking was performed using all 64 AP-MS purifications as input information.

Comparison to other methods

We compared INM to other methods of scoring protein interactions from AP-MS data, including SAINT and Mass Spectrometry interaction Prediction (MSiP, version 1.3.7) (Rahmatbakhsh 2023) using our AP-MS data as input. Comparisons were made using an ROC framework, where the edge values of each method were used to classify each pair as either interacting or not over a range of threshold values. MSiP implements other prey-prey scoring methods including the Dice coefficient, the overlap score, the Jaccard coefficient, and the log hypergeometric score (Guruharsha et al. 2011). Additionally, MSiP implements a support vector machine (SVM) classifier using the input scores as features. We trained 2,000 classifiers using our reference of direct protein interactions as positive training labels and assessed the performance of both a representative classifier and the input scores.

Finally, we compared the INM network to the Human Protein Complex Map 2.0 (hu.MAP 2.0, 3/7/2024) (Drew, Wallingford and Marcotte 2021) using the hu.MAP 2.0 SVM weights as a binary classifier of our benchmark protein interactions.

Construction of a synthetic benchmark

To benchmark our method on a reference where the ground truth is known by construction, we create a reference network and synthetic data. The reference network consists of 236 nodes and 55 edges selected at random. The synthetic data consists of AP-MS profile correlations for all edge variables. For edge variables equal to 0, AP-MS profile correlations were generated by randomly sampling values from the null distribution (**Fig. 8**). For the 55 edge variables equal to 1, AP-MS profile correlations were generated by uniformly sampling values between 0.3 and 1. The synthetic benchmark scoring function omits the SS term because we are unable to generate SAINT scores.

2.4 Results

To assess our modeling approach, we computed the HIV-1 CRL5 protein interaction network using all input purifications and compared the average network model to a reference of direct protein interactions obtained from the PDB. Our analysis includes, (i) demonstrating that that direct protein interactions are enriched for high values of AP-MS profile similarities, (ii) assessing the convergence and exhaustiveness of sampling – a necessary step to construct an average network model, (iii) modeling the HIV-CRL5 network using alternative variations of the scoring function and determining that the AP-MS profile similarity term is the most informative, (iv) comparing INM to other state-of-the-art methods using our benchmark dataset, (v) comparing INM to previously published networks, and (vi) illustrating the utility of our method by predicting a subset of Vif-dependent interactions.

AP-MS profile similarities are informative of protein's physical interactions

Proteins may be identified in one or more AP-MS purifications by their spectral counts, which provides a semi-quantitative measure of protein abundance (**Fig. 2**). The AP-MS profiles showcase how the spectral counts vary among conditions or replicas (**Fig. 2**). It seemed likely that two proteins with similar AP-MS profiles are physically proximal to one another in the mixture of complexes (Braberg et al. 2020). Using the HIV-1 CRL5 data (Hüttenhain et al. 2019) we demonstrate that the distribution of AP-MS profile similarities has multiple modes (**Fig. 2**). To confirm that the different modes represent information encoded in the AP-MS profiles, we shuffled the spectral counts within each AP-MS profile (without replacement) prior to calculating the AP-MS profile similarities (**Fig. 8**). The obtained distribution of AP-MS profile similarities has a single mode near 0 (null distribution) and total ablation of the rightmost mode. Next, we plotted the AP-MS profile similarities for those protein pairs known to interact in our benchmark dataset and compared that distribution to all other pairs of proteins. We find the interacting protein pairs are highly enriched towards high AP-MS profile values (**Fig. 2**). We also observe a population of AP-MS profiles with negative profile similarities that fall outside the null distribution.

We interpret these results as follows: (i) The rightmost peak corresponding to pairs of proteins with high AP-MS profile similarities may be explained by directly interacting proteins and by pairs of proteins that participate in one or more complexes, (ii) low values of the AP-MS profile similarity are indicative of two proteins not interacting directly, (iii) negative AP-MS profile correlations may correspond to pairs of proteins that are not simultaneously present in any AP-MS purification. They may therefore represent competitive binding or some other mechanism that prevents the presence of both preys in a single purification. These observations justify converting AP-MS profile similarities into a probabilistic scoring term to assess protein interactions (**Eq. 2**).

Assessment of sampling and scoring

We followed an assessment protocol similar to the one developed for integrative structure modeling (Viswanath et al. 2017; Saltzberg et al. 2021). The score convergence test shows that the best score does not continue to improve significantly with an increased number of models sampled (**Fig. 3**). The two score distributions drawn from subsets of models are similar to each other (KS P -value=0.33, D =0.004) (**Fig. 3**).

We assessed the scoring function based on the AUC accuracy on a reference of 55 direct protein interactions obtained from the PDB. While the AUC accuracy of any individual network model is low (**Fig. 3**), the AUC of the average of N network models is increasingly higher, plateauing near 0.8 (**Fig. 3**). The AUCs standard deviation (s.d.) across independent modeling runs as a function of N is low (approximately 0.01). Compared to 5 references of 55 randomly chosen protein interactions, the AUC accuracy remains near 0.5 and the s.d. is high near 0.05 (**Fig. 3**).

Next, we assessed the performance of the different terms in the scoring function by computing the AUC accuracy of scoring function variations. Such variations include different combinations of the scoring terms. We find that the AP-MS profile similarity term (SR) is the most informative of direct protein interactions (AUC 0.82, **Fig. 4**). The degree term (SD) is only informative at high threshold values, as indicated by the constant slope of the ROC curve for false positive rates (FPR) between 0 and 0.6. The term controlling the

base rate of protein interactions (SE) is weakly informative (AUC 0.58). The SAINT term (SS) has is accurate (AUC of 0.78) but is uninformative at low FPR values; however, it becomes more informative than the profile similarity term (SR) at an FPR near 0.2. These results suggest that the accuracy of the scoring function is driven mostly by the AP-MS profile similarity term.

Comparison of INM to other methods

We compared the expected edge value from INM to edge weights from other AP-MS matrix models (**Fig. 4**), focusing on the MSiP software package. MSiP implements multiple methods of scoring protein interactions based on AP-MS data (Rahmatbakhsh 2023), as well as an SVM that provides a score between 0 and 1 for each interaction. The performance of the SVM is a stochastic function of the training. We compared both the individual scores and the SVM scores to INM using ROC curves. To assess the performance of a representative SVM classifier we independently trained 2,000 classifiers on our benchmark data. The AUC accuracies of the Dice coefficient and the negative log hypergeometric score are lower than that of the INM method. The AUC accuracies of the SVM classifiers range from 0.4 to 0.8, with a mean of 0.64 (**Fig. 5**).

Spoke models based on SAINT scores link prey proteins to baits, without providing information about bait-bait and prey-prey interactions. To facilitate the comparison of the HIV-CRL5 SAINT-derived spoke model to our model, we constructed an all-pairs matrix that includes inferred SAINT scores for prey-prey interactions. These SAINT scores were obtained by multiplying the maximal SAINT score over all purifications of prey i and prey j . Thus, if both proteins i and j have a high SAINT score in at least one input purification, their SAINT pair score is high. If either protein i or j have a low maximal SAINT score, the corresponding pair score is low. The AUC accuracy of the maximal pairwise SAINT score is 0.78 (**Fig. 4**).

We compare the average HIV-CRL5 network generated by INM to the human protein complex map 2.0 (hu.MAP 2.0) network using our benchmark dataset (**Fig. 4**). The hu.MAP 2.0 network was created by integrating over 15,000 proteomics experiments using a ML model. This integration includes various types

of proteomics data, including AP-MS, yeast-two-hybrid, and RNA hairpin. The hu.MAP 2.0 SVM scores have an AUC accuracy of 0.86 on our benchmark.

Vif dependent protein interactions

To assess whether our approach can identify condition-specific protein interactions, we modeled protein interaction networks under WT (containing Vif), Δ VIF, and mock conditions. By examining a subset of Vif-dependent interactions corresponding to proteins identified in the WT condition and not the Δ VIF conditions, we observe several features of the average network model. DDB1-and-CUL4-associated factor 11 (DCA11) is correctly predicted to interact with both CUL4A and CUL4B, supporting its putative role as a substrate adaptor (Angers et al. 2006). Similarly to CRL5, CRL4 complexes are hijacked during viral infection to target host machinery for proteolytic degradation (Dobransky et al. 2024). Interestingly, Programmed cell death 6-interacting protein (PDC6I, AIP1, ALIX) is predicted to interact with CUL4A, CUL4B, and DCA11, suggesting that PDC6I may engage with both CUL4A-DCA11 and a CUL4B-DCA11 complexes. PDC6I binds to a short linear motif in the HIV-1 viral protein Gag, playing a role in viral budding through the ESCRT (endosomal sorting complex required for transport) pathway (Zhai et al. 2008). Ubiquitination of Gag is required for ESCRT-mediated HIV-1 budding (Sette et al. 2013). CRL4 may serve as the ubiquitin-conjugating enzyme for the ubiquitination of Gag mediated by PDC6I. We do not represent We did not include Gag in our network because it did not meet our SAINT score cutoff due to low spectral counts.

2.5 Discussion

We developed a statistical inference framework (INM) to model direct protein interaction networks based on AP-MS spectral count data and SAINT scoring. We benchmarked INM against a reference of direct protein interactions obtained from the PDB. We compared the networks from INM to the hu.MAP 2.0 network and a SAINT network using our PDB reference. Here we discuss (i) the relationship between our method and others; (ii) the benefits of the INM modeling framework (iii) the limitations of INM arising

from inaccurate scoring and imperfect sampling; and (iv) the scope of adding additional information into INM.

Comparison with other methods

Recently, over 15,000 proteomics experiments were integrated into the human protein complex map 2.0 (hu.MAP 2.0) (Drew, Wallingford and Marcotte 2021). The map spans 15,373 human protein types and over 17 million protein interaction scores derived from different experiments. Additionally, the SAINT scoring was previously applied to score thousands of bait-prey pairs identified in 64 HIV-CRL5 AP-MS purifications (Hüttenhain et al. 2019). We compared the networks generated from each method by binary classification of the PDB-derived direct protein interactions. The hu.MAP 2.0 scores are highly predictive of direct protein interactions, achieving an AUC accuracy of 0.86 (**Fig. 4**). In contrast, the AUC accuracies of INM and SAINT have AUC were 0.82 and 0.78, respectively. The AUC accuracy of hu.MAP 2.0 may be attributed to the diverse datasets (*e.g.*, AP-MS, yeast-two-hybrid) used in network modeling, the total amount of input information (*i.e.*, tens of thousands of experiments), or the inclusion of additional features from the AP-MS data (*e.g.*, MS1 intensity and spectral counts). Although hu.MAP 2.0 is accurate, it may not be applicable to uncharacterized systems for which multiple data types are not available. In such cases, methods relying primarily on AP-MS (*i.e.*, INM,SAINT, MSiP) may be advantageous. Furthermore, there might be some overlap between the hu.MAP 2.0 training labels and our PDB benchmark dataset, which may contribute to the accuracy of hu.MAP 2.0 in the CRL5 system. Notably, the slope of the hu.MAP 2.0 ROC curve greatly decreases for the last ~25% of reference interactions. We see similar features for INM,SAINT, and MSiP. INM outperforms the MSiP classifier and the MSiP scores (Jaccard, Dice, Overlap, negative log HGScore) on our benchmark. Since no method recovers these interactions, we conclude that the evidence for these interactions by mass spectrometry is weak or absent.

Unlike hu.MAP 2.0, INM, MSiP, and SAINT rely on the same set of input purifications, yet INM achieves a higher AUC accuracy. This difference can be explained in part by the fact that SAINT was not designed to model prey-prey interactions. We view INM and SAINT as complementary. For example, predicting all

prey-prey interactions for the 3,062 identified proteins would require constructing an impractically large pairwise matrix—something beyond the scope of INM. However, SAINT scoring allowed us to select a subset of highly confident protein interactions. INM captures information from the AP-MS purifications that is predictive of direct protein interactions that INM, which SAINT does not retrieve (**Fig. 5**).

The benefits of the INM modeling framework

The advantages of INM are: (i) its explicit representation of direct protein interaction networks, (ii) the accuracy and extensibility of its scoring function, (iii) its data efficiency, (iv) the flexibility and computational efficiency of sampling, and (v) its ability to characterize new systems under specific experimental conditions. We now discuss each in turn.

Bridging the gap between structural and systems biology

Different types of models are characteristic of different domains of science. For instance, in system biology, an archetypal model is a molecular network where nodes represent protein types, and edges indicate direct or indirect interaction between some proteins of those types. In contrast, in structural biology, an archetypal model depicts the spatial coordinates of the individual components of the system (not their types) such that the proximity between them indicates a physical interaction. These representations facilitate using input information characteristic of the corresponding domain of science. Here, we attempt to bridge the gap between systems biology and structural biology by using a depiction that is intermediate to the two archetypal depictions. Namely, the nodes represent protein types (*i.e.*, systems biology nodes) and the edges represent the presence of at least one protein interaction between molecules of the corresponding types. This representation facilitates using input information from both system biology and structural biology experiments, thus maximizing the accuracy, precision, and completeness of the model.

The network model representation (nodes as protein types and edges as the presence of at least one protein interaction corresponding to the nodes) may be well suited to bridge the gap between systems and structural biology. The predicted edges may be used as input for spatial modeling of macromolecular complexes.

Likewise, the network output network models may be directly interpreted or used as input features in AI/ML models. Moreover, combining structural and network degrees of freedom in a single unified modeling approach may enable the modeling of molecular neighborhoods.

Accuracy of the INM scoring function

The scoring function is accurate in predicting direct protein interactions, generating an average network model with an AUC accuracy of 0.82. Remarkably, INM is nearly as accurate as hu.MAP 2.0, despite using 3 orders of magnitude less data. The accuracy of the scoring function could be further improved by incorporating additional new scoring terms representing additional information. Ideally, the scoring function would adopt a fully Bayesian formulation, where each additional term represents the likelihood of the data given a network model.

Sampling is both efficient and exhaustive, typically running at 1-3 seconds per Monte Carlo step. If fewer compute resources are available, the sampling efficiency may be tuned during the warmup phase. Alternatively, the amount of sampling could be decreased to obtain initial approximations of the network structure.

INM is particularly advantageous for characterizing new systems that may be inaccessible to other methods, as it relies on a limited number of purifications collected under specific conditions. For instance, INM can effectively model host-pathogen interactions or human systems under specific perturbations, provided that corresponding AP-MS data is available.

Limitations of the INM modeling framework

In general, modeling can fail to produce accurate and precise models because of insufficient input information or inferior modeling, which in turn can be caused by either an inappropriate representation, inaccurate scoring function, or insufficient sampling.

Input information

One concern is that incomplete input information would result in missing proteins (nodes) or protein interactions (edges); however, this issue can be mitigated by integrating diverse types of datasets, as it is unlikely that any single protein or interaction would be systematically absent across all datasets. When necessary, gene ontology (GO) annotations can be utilized to reduce the risk of missing protein components involved in larger complexes or components associated with a given biological process.

During benchmarking, we observed that all models are weakly informative or uninformative of roughly a quarter of the benchmark protein-protein interactions, as indicated by the low slope of the ROC curve in the FPR range 0.6-1.0 (**Fig. 4**), in contrast to the steep slope in the 0.0-0.2 range. This effect is present both for hu.MAP 2.0 and INM. One explanation is that the proteomics data may not support a subset of the benchmark interactions. Alternatively, it is possible that none of the methods accurately model these interactions, even though the information is present. In a benchmark using synthetic data where the ground truth is known by construction, INM recovers interactions with high accuracy (**Fig. 5**) The difference in accuracy between the real and synthetic case cannot be explained by insufficient sampling as sampling is exhaustive in both cases. Of the two explanations, we find the first more plausible.

Model representation

In INM, the network model representation is implemented as an all-pairs dense matrix model, which limits scalability to several hundred to a few thousand nodes due to the quadratic increase in memory usage. While sparse data structures would alleviate this limitation, the data-to-parameter ratio would still decrease linearly with increasing network sizes – potentially requiring additional regularizing priors to prevent overfitting. Additionally, the current network representation has limitations in capturing certain types of protein interactions, such as protein interactions occurring in homomers.

Scoring

The accuracy of the scoring function could be increased by adding additional likelihoods for orthogonal types of data such as functional genomics screens or proximity labeling MS. A more accurate prior could be constructed from a sample of network models derived from the PDB. Finally, a fully Bayesian data likelihood for AP-MS could replace both the SR and SS terms.

Sampling

For the HIV-CRL5 system sampling has converged and is exhaustive, additional network modeling studies on diverse systems are required to determine the amount of sampling required on each system.

Analysis

Currently we construct our average network model by taking an average over all edge values. Real biological protein interaction networks may exist in multiple states. With accurate scoring and sufficient sampling, INM can in principle recover such states by clustering network models rather than taking an average.

Opportunities for adding additional information to INM

The scope of adding additional information to the scoring function is large as many types of experiments or other approaches may be informative of direct protein interaction. It is of particular interest to develop scoring terms based on data that may inform indirect or transient protein interactions such as proximity labeling MS, in situ/in vivo cross-linking MS, large-scale functional genetic interaction or CRISPRi/a experiments, or AI-derived pairwise protein structure predictions (Larson et al. 2013; Jumper et al. 2021; Braberg et al. 2022; Lin et al. 2024).

Recently, AP-MS data have been integrated with imaging data to produce models of protein localization (Qin et al. 2021; Cho et al. 2022). Critically, the AP-MS data is represented as the output of a spoke-type model – a list of differentially abundant proteins. While powerful, this approach may underfit the AP-MS

data. The difference in accuracy between SAINT scores and AP-MS profile similarities suggests that using mass spectrometry features (such as spectral count or MS1 intensity) directly as input to a machine-learning model encodes more information about protein interactions than downstream representations of the data (e.g., bait-prey networks).

We have developed and benchmarked integrative network modeling (INM), a statistical inference method to infer the network of all-pairs of direct protein interactions, based on AP-MS data, which may be extended to other types of information. Our method may be used to model protein interaction networks under specific experimental conditions (e.g., mutation or viral infection).

2.6 Availability of software and data

All software and data are freely accessible. Integrative Network Modeling is implemented as a module of our open-source Integrative Modeling Platform (IMP) software (<https://integrativemodeling.org>) (Webb et al. 2018).

2.7 Acknowledgements

This work has been supported by NIH NIGMS grant GM083960 to AS, NIH NIAID grants U54AI170792 and 2U19AI135990 to AS and IE, and NIH NIGMS grant 1R35GM151256 to IE.

2.8 Figures

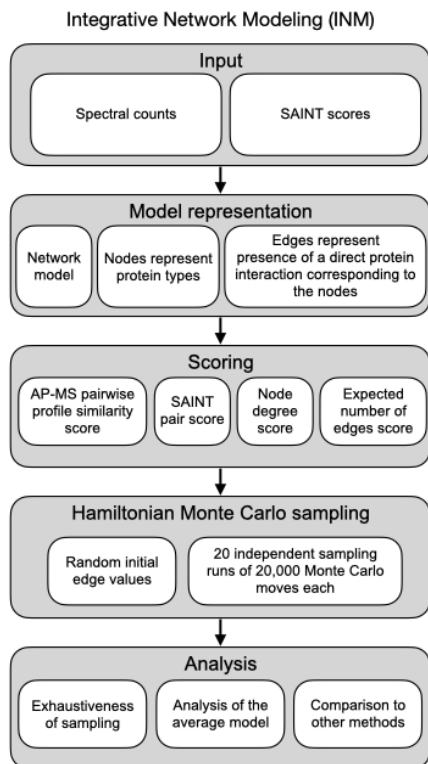


Figure 1. The flowchart of integrative network modeling (INM).

The five stages of Integrative Network Modeling (INM): (i) the input information are spectral counts Significance Analysis of INTeractome (SAINT) scores from K affinity purification mass spectrometry (AP-MS) purifications, (ii) the network model is represented as an undirected unweighted graph of N nodes and an edge variable for all $N(N-1)/2$ possible node pairs; edge variables values may be 0 or 1 representing the absence or presence of an edge, (iii) the agreement between any network model (set of edges for a set of nodes) and the input information is scored using a scoring function. The scoring function is a sum of terms, each one scores the agreement between the model and a particular type of input information (*e.g.*, AP-MS profiles, SAINT scores). (iv) alternative network models are sampled using a Monte Carlo scheme. A Monte Carlo move is accepted using the Metropolis criteria. Moves are proposed using Hamiltonian dynamics. The output is a set of t network models distributed according to e-S(M), (v) for the output set of network models, the exhaustiveness and convergence is assessed, an average network model is constructed by taking the average value of the edges over the t sampled network models. The average network model is compared to networks generated by other methods using a Protein Data Bank (PDB) derived benchmark of direct protein interactions.

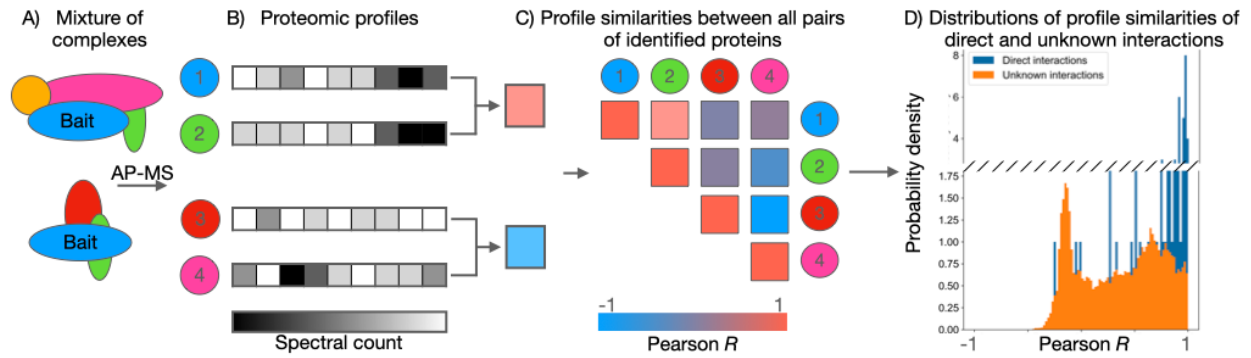


Figure 2. AP-MS proteomic profiles inform direct protein interactions.

A) AP-MS purifications are performed with one or more bait over multiple conditions. The identified prey may be in a single complex or a mixture of complexes. B) The spectral counts of each identified protein (prey) in each condition are compared to one another using Pearson correlation. C) a correlation matrix is constructed for such comparisons. D) The distribution of correlation coefficients is multimodal. The distribution of AP-MS profile correlation coefficients corresponding to direct interactions is skewed towards high-profile similarities.

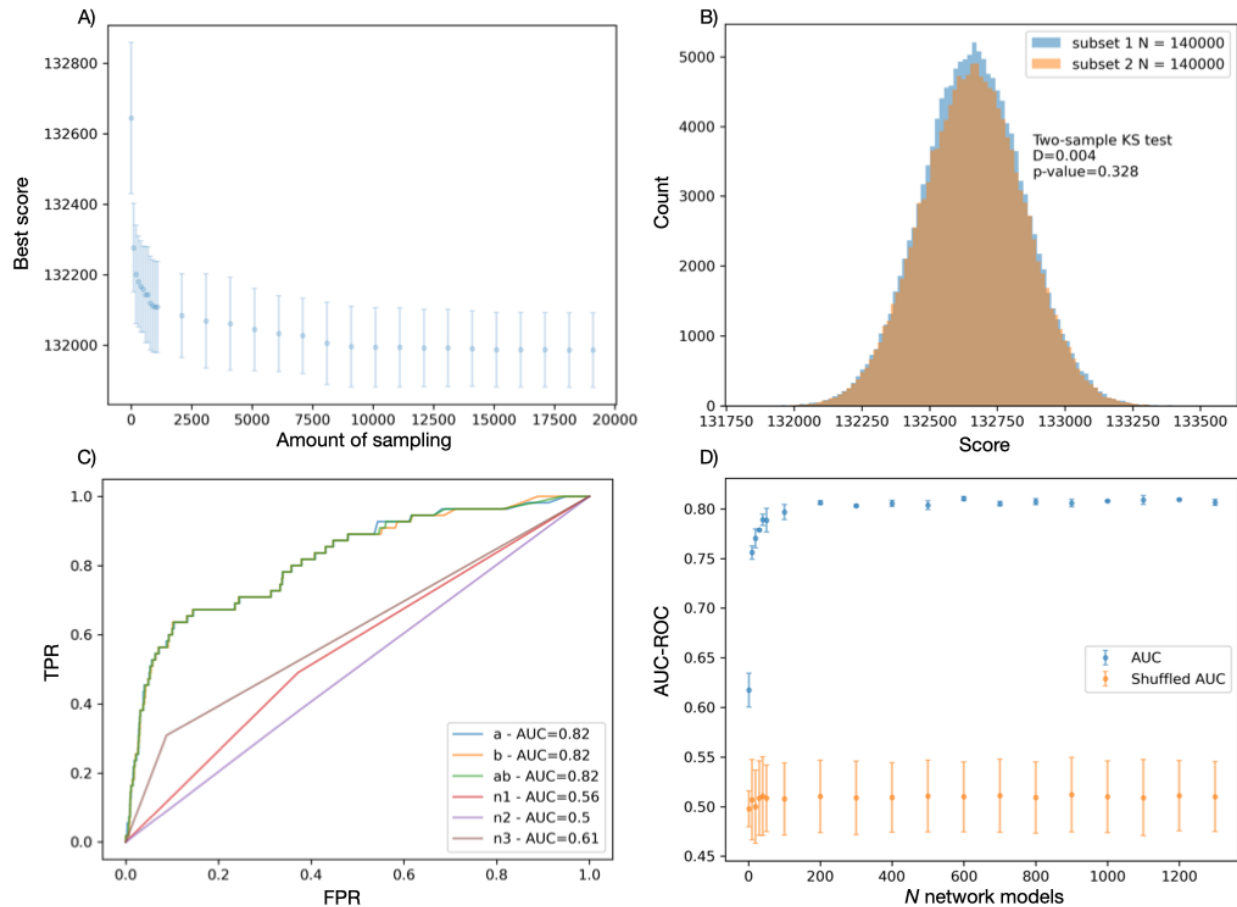


Figure 3. Convergence and exhaustiveness of sampling.

A) The best scoring model per modeling run converges as the number of sampled Monte Carlo steps increases. The average (dots) and standard deviation (bars) are plotted over for one modeling run. B) Two random subsets of scores from 20 independent network modeling runs are identically distributed. C) the ROC curves of the two random subsets of network models (A & B) are identical to the ROC curve of their union (AB). Individual network models are not accurate (n1, n2, n3). D) The AUC accuracy of the average model is plotted as a function of increasing N, where N is the number of models used to compute the average. Uncertainty is estimated by bootstrapping three times (blue). The network models are not accurate when compared to five random reference networks (orange).

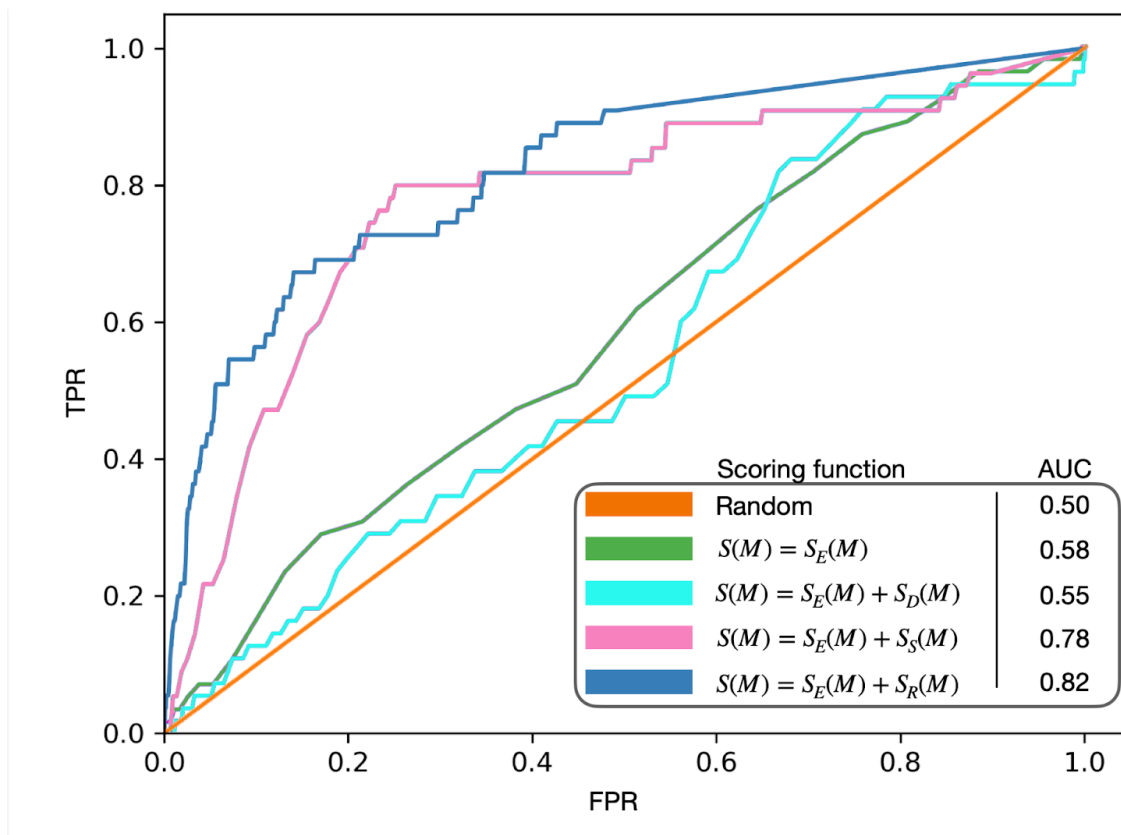


Figure 4. Accuracy of scoring function terms

We assessed four variations of the scoring function using 55 direct protein interactions obtained from the Protein Data Bank (PDB) as a reference. The scoring function variations are: (i) SE, the number of edges term, (ii) SE + SD, the addition of the degree term, (iii) SE + SS, the addition of the SAINT term, (iv) SE + SR, the addition of the AP-MS profile similarity term.

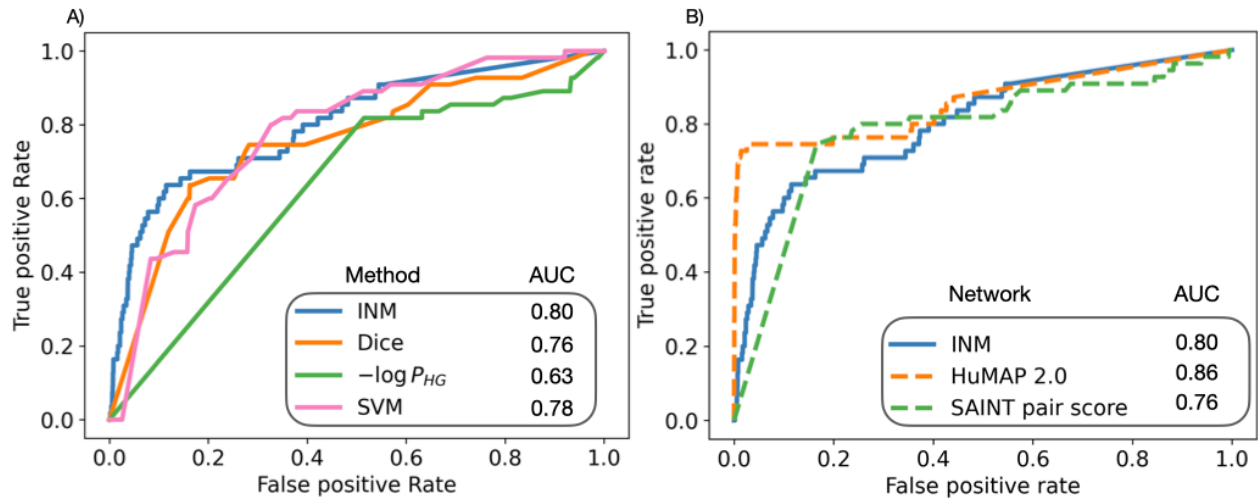


Figure 5. Comparison to other methods and networks

The AUC accuracy of each method is compared to the PDB-derived reference of 49 direct protein interactions. A) INM is compared to state-of-the-art methods implemented in the MSiP software. B) INM is compared to the hu.MAP 2.0 network and an all-pairs network derived from SAINT bait-prey scores.

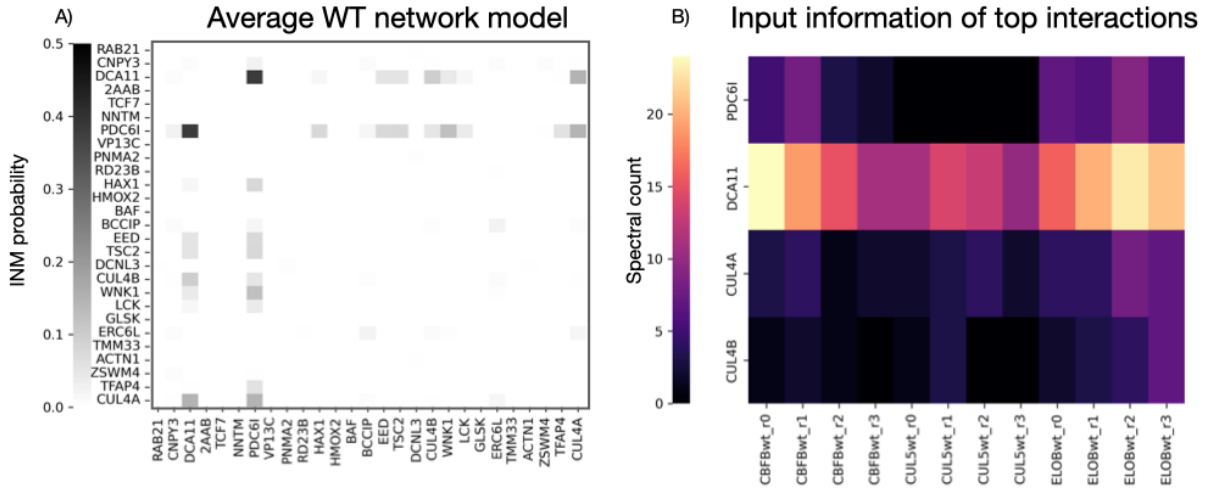


Figure 6. Illustration of INM on the HIV-1 CRL5 system

A) The average network model corresponding to proteins identified only in the WT condition and not in the Δ Vif condition. Labels are UniProt entry prefixes. DDB1-and-CUL4-associated factor 11 (DCA11) is correctly predicted to interact with both CUL4A and CUL4B, consistent with its putative role as a substrate adaptor. CUL4 and CUL4 are not predicted to interact with each other, consistent with their respective roles as catalytic units in CRL ubiquitin ligases. Programmed cell death 6-interacting protein (PDC6I) is predicted to interact with CUL4A, CUL4B, and DCA11. PDC6I has a known role in HIV viral budding. B) The predicted interactions are non-obvious from observation of the spectral count data. These predictions occur even though DCA11 has higher overall spectral counts than either CUL4A, CUL4B, or PDC6I.

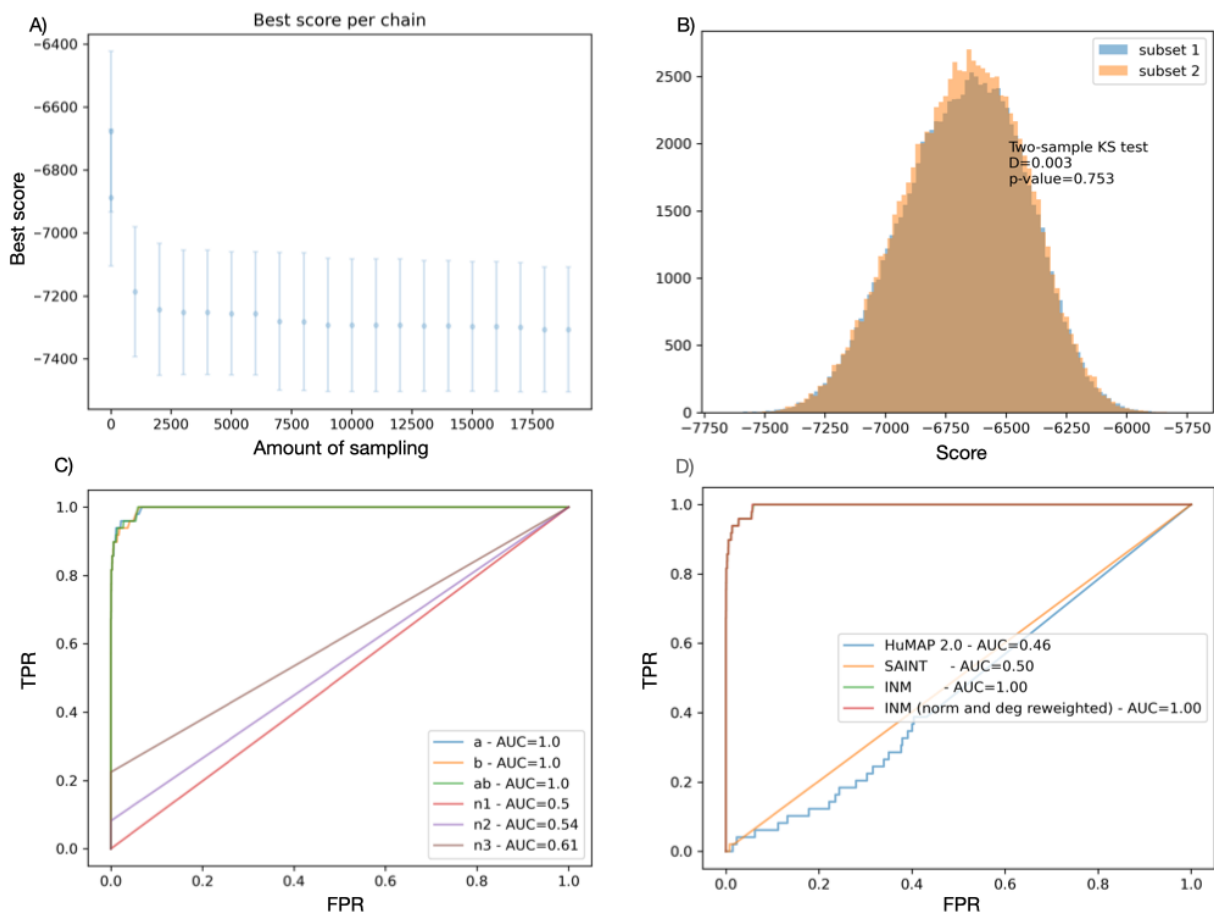


Figure 7. Benchmark for synthetic data

A) The best scoring model per modeling run converges as the number of sampled Monte Carlo steps increases. The average (dots) and standard deviation (bars) are plotted over for one modeling run. B) Two random subsets of scores from 20 independent network modeling runs are identically distributed. C) the ROC curves of the two random subsets of network models (A & B) are identical to the ROC curve of their union (AB). Individual network models are not accurate (n1, n2, n3). D) The AUC accuracy of the average model is plotted as a function of increasing N, where N is the number of models used to compute the average. Uncertainty is estimated by bootstrapping three times (blue). The network models are not accurate when compared to five random reference networks (orange). Normalization and degree reweighting were applied to the INM probability, this scheme ranks solutions identically.

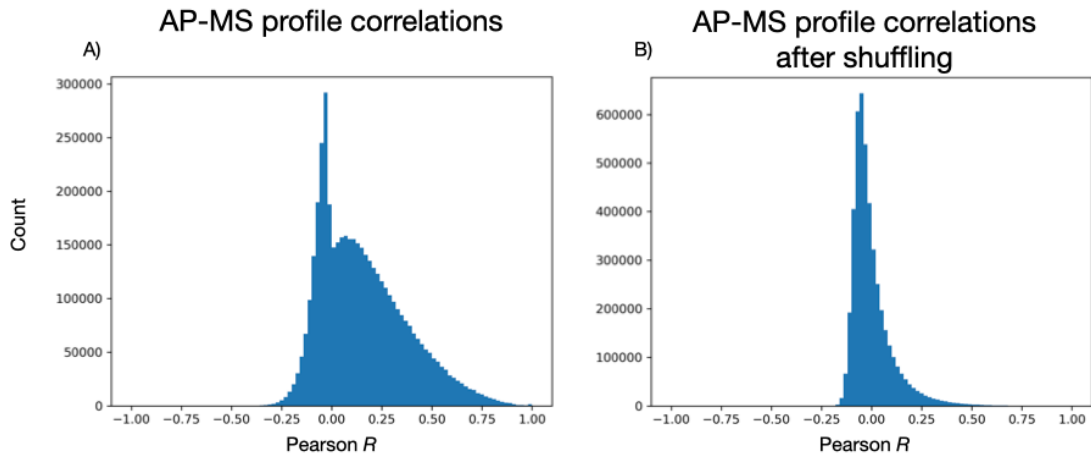


Figure 8. Shuffling of spectral counts ablates the rightmost mode

A) The observed distribution of affinity purification mass spectrometry (AP-MS) profile correlations (Pearson R). The distribution has two modes. B) The distribution of AP-MS profile correlations after shuffling the spectral counts within each proteomic profile without replacement (*i.e.*, null distribution). The distribution is asymmetric with a single peak near 0.

2.9 References

1. Angers S, Li T, Yi X et al. Molecular architecture and assembly of the DDB1-CUL4A ubiquitin ligase machinery. *Nature* 2006;443:590–3.
2. Braberg H, Echeverria I, Bohn S et al. Genetic interaction mapping informs integrative structure determination of protein complexes. *Science* 2020;370, DOI: 10.1126/science.aaz4910.
3. Braberg H, Echeverria I, Kaake RM et al. From systems to structure - using genetic data to model protein structures. *Nat Rev Genet* 2022;23:342–54.
4. Choi H, Larsen B, Lin Z-Y et al. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat Methods* 2011;8:70–3.
5. Cho NH, Cheveralls KC, Brunner A-D et al. OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science* 2022;375:eabi6983.
6. Dobransky A, Root M, Hafner N et al. CRL4-DCAF1 ubiquitin ligase dependent functions of HIV Viral Protein R and Viral Protein X. *Viruses* 2024;16:1313.
7. Drew K, Lee C, Huizar RL et al. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol Syst Biol* 2017;13:932.
8. Drew K, Wallingford JB, Marcotte EM. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol Syst Biol* 2021;17:e10016.
9. Ewing RM, Chu P, Elisma F et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* 2007;3:89.
10. Gavin A-C, Aloy P, Grandi P et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006;440:631–6.
11. Guruharsha KG, Rual J-F, Zhai B et al. A protein complex network of *Drosophila melanogaster*. *Cell* 2011;147:690–703.
12. Hart GT, Lee I, Marcotte ER. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* 2007;8:236.

13. Hüttenhain R, Xu J, Burton LA et al. ARIH2 is a Vif-dependent regulator of CUL5-mediated APOBEC3G degradation in HIV infection. *Cell Host Microbe* 2019;26:86–99.e7.
14. Jäger S, Cimermancic P, Gulbahce N et al. Global landscape of HIV-human protein complexes. *Nature* 2011;481:365–70.
15. Jeronimo C, Forget D, Bouchard A et al. Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. *Mol Cell* 2007;27:262–74.
16. Jumper J, Evans R, Pritzel A et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583–9.
17. Krogan NJ, Cagney G, Yu H et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006;440:637–43.
18. Kunzmann P, Müller TD, Greil M et al. Biotite: new tools for a versatile Python bioinformatics library. *BMC Bioinformatics* 2023;24:236.
19. Larson MH, Gilbert LA, Wang X et al. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc* 2013;8:2180–96.
20. Lin Z, Schaefer K, Lui I et al. Multiscale photocatalytic proximity labeling reveals cell surface neighbors on and between cells. *Science* 2024;385:ead15763.
21. Mathew D Hoffman AG. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J Mach Learn Res* 2014.
22. Neal RM. MCMC using Hamiltonian dynamics. *arXiv [statCO]* 2012.
23. Qin Y, Huttlin EL, Winsnes CF et al. A multi-scale map of cell structure fusing protein images and interactions. *Nature* 2021;600:536–42.
24. Rahmatbakhsh M. Systems Biology of Host-Pathogen Protein-Protein Interactions. Dr thesis 2023.
25. Remmert M, Biegert A, Hauser A et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2011;9:173–5.

26. Richards AL, Eckhardt M, Krogan NJ. Mass spectrometry-based protein-protein interaction networks for the study of human diseases. *Mol Syst Biol* 2021;17:e8792.
27. Rieping W, Habeck M, Nilges M. Inferential structure determination. *Science* 2005;309:303–6.
28. Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. *Nature* 2007;450:973–82.
29. Saltzberg DJ, Viswanath S, Echeverria I et al. Using Integrative Modeling Platform to compute, validate, and archive a model of a protein complex structure. *Protein Sci* 2021;30:250–61.
30. Sette P, Nagashima K, Piper RC et al. Ubiquitin conjugation to Gag is essential for ESCRT-mediated HIV-1 budding. *Retrovirology* 2013;10:79.
31. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol* 1973;79:351–71.
32. Siegal S, Castellan N. Nonparametric statistics for the behavioral sciences. McGraw-hill: Boston. Medio Aevo 1988.
33. Viswanath S, Chemmama IE, Cimermancic P et al. Assessing exhaustiveness of stochastic sampling for integrative modeling of macromolecular structures. *Biophys J* 2017;113:2344–53.
34. Wang B, Vartak R, Zaltsman Y et al. A foundational atlas of autism protein interactions reveals molecular convergence. *bioRxivorg* 2024, DOI: 10.1101/2023.12.03.569805.
35. Webb B, Viswanath S, Bonomi M et al. Integrative structure modeling with the Integrative Modeling Platform. *Protein Sci* 2018;27:245–58.
36. Wenger CD, Phanstiel DH, Lee MV et al. COMPASS: a suite of pre- and post-search proteomics software tools for OMSSA. *Proteomics* 2011;11:1064–74.
37. Zhai Q, Fisher RD, Chung H-Y et al. Structural and functional studies of ALIX interactions with YPX(n)L late domains of HIV-1 and EIAV. *Nat Struct Mol Biol* 2008;15:43–9.

Chapter 3

Multi-scale photocatalytic proximity labeling reveals cell surface neighbors on and between cells

3.1 Abstract

The cell membrane proteome is the primary biohub for cell communication, yet we are only beginning to understand the dynamic protein neighborhoods that form on the cell surface and between cells. Proximity labeling proteomics (PLP) strategies using chemically reactive probes are powerful approaches to yield snapshots of protein neighborhoods but are currently limited to one single resolution based on the probe labeling radius. Here, we describe a multi-scale PLP method with tunable resolution using a commercially available histological dye, Eosin Y, which upon visible light illumination, activates three different photo-probes with labeling radii ranging from ~ 100 to 3000 \AA . We applied this platform to profile neighborhoods of the oncogenic epidermal growth factor receptor (EGFR) and orthogonally validated >20 neighbors using immuno-assays and AlphaFold-Multimer prediction that generated plausible binary interaction models. We further profiled the protein neighborhoods of cell-cell synapses induced by bi-specific T-cell engagers (BiTEs) and chimeric antigen receptor (CAR)T cells at longer length scales. This integrated multi-scale PLP platform maps local and distal protein networks on cell surfaces and between cells. We believe this information will aid in the systematic construction of the cell surface interactome and reveal new opportunities for immunotherapeutics.

3.2 Introduction

Cell surface proteins are critical mediators of information, nutrients, and functions on cells and between them. The extracellular proteome, both secreted and membrane-bound, is encoded by more than 25% of the human genome (1, 2). Proteomics methods have made great strides in characterizing the composition of the surface proteome in health and disease models (3, 4). However, we know much less about the protein-protein interactions formed on the cell membrane, especially transient interactions that regulate cell signaling networks.

Proximity labeling proteomics (PLP) methods have enabled the identification of protein interactomes in complex cellular environments (5, 6). These methods typically generate a single reactive intermediate locally to label and profile nearby proteins using imaging or proteomics. The first generation of PLP methods used genetically encoded enzymes such as APEX (7), BioID (8) or TurboID (9, 10) to produce phenoxy radicals or activated AMP that have long reactive half-lives ($t_{1/2} > 100 \mu\text{sec}$). These methods are well-suited for characterizing cell-cell and organelle-specific interactomes given their long labeling range up to 3000 Å by labeling electron-rich amino acids (11). Singlet oxygen generators (SOG) triggers selective labeling on His (12) at a shorter range given the shorter half-life of singlet oxygen in water ($\sim 2\text{-}4 \mu\text{s}$) (13). However, proteins are estimated to be separated by only 60-70 Å on the crowded cell surface (14), thus making it challenging to identify the most proximal protein neighbors using long-range PLP approaches.

Most recently, PLP methods of very short range have emerged, enabling higher resolution mapping including μMap (15-17). These designs employ transition-metal or other photocatalysts attached to antibodies to trigger reactive intermediate with shorter half-lives such as carbenes or nitrenes ($t_{1/2} \sim 2$ and 10 ns, respectively) (11, 18). Activation of these probes enables labeling proteins at a significantly shorter range of $\sim 100\text{-}700 \text{ \AA}$ as well as broader amino acid coverage (11, 19), thus making it much more appropriate for nearest neighborhood analysis (20-23). Collectively, the suite of PLP methods can cover a broad length scale for labeling protein neighborhoods and synapses but require multiple photocatalysts for adjustable resolution.

Here, we report a multi-scale photocatalytic PLP technology, termed MultiMap (**Fig. 1**) that allows short-, intermediate-, and long-range labeling from a single photocatalyst, Eosin Y (EY). We discovered that EY, a fluorescent dye commonly used in food chemistry and biological staining (21) can efficiently trigger the generation of carbene, nitrene, singlet oxygen and phenoxy radicals from photo-probes with bio-compatible blue or green light. We applied MultiMap to profile high-resolution neighborhoods of the oncogenic epidermal growth factor receptor (EGFR) in different cellular contexts. We identified >20 neighbors and further validated their interactions via immunoprecipitation and in silico prediction models

using AlphaFold-Multimer (24). We demonstrated that MultiMap can capture long-range intercellular engagements between cancer cells and T lymphocytes induced by bi-specific T-cell engagers (BiTEs) and engineered chimeric antigen receptors (CARs). Our data show that MultiMap is an effective multi-scale PLP technology that can characterize local and distal cellular interactomes from a single photocatalyst. We believe that with artificial intelligence-assisted structural prediction methods integrated, the MultiMap workflow will be an important approach in the broad quest to define the spatial organization of the cell surface proteome and to reveal new drug discovery opportunities.

3.3 Results

Eosin Y (EY) activates a panel of photo-probes for protein labeling.

We explored Eosin Y (EY) for PLP (**Fig. 1**) based on its photocatalytic ability broadly used in polymer synthesis (25) and easy commercial access (26). In contrast, the current transition- metal photocatalysts used in μ Map, such as $\{\text{Ir}[\text{dF}(\text{CF}_3)\text{ppy}]_2(\text{dtbbpy})\}\text{PF}_6$, requires lengthy synthetic routes (15, 27). We first compared EY to the Ir-catalyst for photo-induced hydrolysis of the diazirine via LC-MS and observed 100% quantitative hydrolysis with 10 min blue LED illumination ($\lambda=450$ nm, **Fig. S1**). The absorption peak for EY ($\lambda_{\text{max}}=517$ nm, **Fig. S1**) is significantly red-shifted compared with the Ir-catalyst ($\lambda_{\text{max}}=420$ nm) (15), which could make EY more bio-compatible (28). Indeed, green LED illumination ($\lambda=525$ nm) of EY induced 100% diazirine hydrolysis, whereas the iridium catalyst did not induce any detectable conversion (**Fig. S1**). Erythrosin B ($\lambda_{\text{max}}=535$ nm), a dye structurally similar to EY, photo-catalyzed 31% hydrolysis of the diazirine with blue LED illumination and 100% with green LED illumination, indicating that the EY scaffold can be modified for specific photochemical properties. We next tested the ability of EY to label bovine serum albumin (BSA) using a diazirine-biotin probe in the presence of light (**Fig. 1**). We observed time- and light-dependent accumulation of biotinylated BSA via Western blot (WB) analysis; labeling plateaued within 6 min of blue LED illumination (**Fig. 1**). A pulse-light experiment (**Fig. 1**) demonstrated that the catalytic function of EY is light-dependent.

EY is structurally similar to other photocatalysts that can trigger aryl-azide and phenol probes (26, 29, 30). EY fully converted the aryl-azide probe to the aniline upon blue LED illumination (**Fig. S1**). We then tested the photocatalytic labeling on BSA by WB analysis and found that EY efficiently catalyzed biotin labeling in the presence of either aryl-azide-biotin or phenol-biotin (**Fig. 1**). Additionally, we evaluated the ability of EY to induce singlet-oxygen-based labeling using biocytin-hydrazide (12) and observed efficient light-dependent labeling. The extents of labeling of BSA among the four biotin-containing reactive probes, which we refer to as photo-probes, ranged in the following order: aryl-azide-biotin (>95%), biocytin-hydrazide (~80%), phenol-biotin (~15%) and diazirine-biotin (~2%) (**Fig. S1**). Under blue light illumination, the Ir catalyst could trigger labeling of BSA by the aryl-azide-biotin but not the phenol-biotin (**Fig. S1**). EY also efficiently catalyzed labeling of BSA with green LED while the Ir catalyst showed no labeling (**Fig. S1**). More than 80% labeling of BSA was achieved upon 3 min green LED exposure of EY with all four photo-probes. We also found that EY maintains its photocatalytic function above its pKa (pH=3.5) (**Fig. S1**) (26) and thus is compatible with labeling across a wide range of physiologic pH conditions.

Conjugation of EY onto proteins

We evaluated different conjugation methods for EY first onto BSA and then to antibodies (**Fig. 2** and **Fig. S2**). After synthesizing DBCO-PEG4-EY via an amine-isothiocyanate reaction (Scheme 1), we explored the conjugation efficiency and stoichiometry for attaching a click-compatible azido functionality specifically to Lys, Met or Cys using N-hydroxy succinimide (NHS) ester, oxaziridine, or maleimide/iodoacetamide warheads, respectively (**Fig. S2**). EY-conjugation via NHS-azide ligation produced the most efficient conjugation; conjugated EY also efficiently catalyzed BSA self-biotinylation with diazirine-biotin, aryl-azide-biotin and phenol-biotin (**Fig. S2**).

Next, we conjugated EY to cetuximab (Ctx), an FDA-approved antibody that selectively binds EGFR and competes for epidermal growth factor (EGF) binding, thus turning-off EGFR signaling and cell proliferation in cancer (**Fig. 2B**) (31, 32). Ctx does not have Lys, Met or Cys residues in the CDRs or in the contact epitope with the EGFR ectodomain (ECD, aa 1-645, PDB:1YY9) (31), suggesting all

bioconjugation methods are viable without impairing binding. We tested the same panel of bioconjugation warheads on Ctx, generating similar levels of conjugation as seen for BSA (**Fig. S2**). Quantification of the levels of conjugation by WB analysis or EY absorption indicated that a stoichiometry of eight and two EY catalysts were installed per Ctx-NHS-EY and Ctx-Ox-EY, respectively (**Fig. S2**).

We then tested the intra- and inter-molecular labeling of the Ctx-EY conjugates with equimolar amounts of recombinant human EGFR ectodomain (ECD, aa 1-645) and in competition with EGF (**Fig. 2**). 1 μ M Ctx-NHS-EY or Ctx-Ox-EY was mixed with 1 μ M EGFR-ECD with or without pre-incubation with 1 μ M EGF. 100 μ M diazirine-biotin was added and the mixture was illuminated with blue LED for 10 minutes. Proteins were purified via acetone precipitation and immunoblotted to evaluate the labeling efficiency (**Fig. 2**). Both Ctx-NHS-EY and Ctx-Ox-EY conjugates demonstrated self-labeling in a light-dependent manner. Not surprisingly, there was a higher degree of biotinylation with Ctx-NHS-EY which contains ~4-fold more conjugated EY than Ctx-Ox-EY (**Fig. S2**). Intermolecular EGFR labeling with both EY-conjugated constructs occurred in a light-dependent manner (**Fig. 2**), indicating that the conjugation of EY did not interfere with Ctx binding to EGFR as expected. Pre-incubation of EGF prevented labeling, demonstrating that direct binding is necessary for target labeling (**Fig. 2**). To explore the generality of the workflow, we performed the same NHS and oxaziridine bioconjugation and labeling using a trastuzumab (Trz) Fab that binds the extracellular domain of the HER2 receptor (**Fig. S3**). Similar intermolecular labeling of HER2 was observed with Trz-NHS-EY or Trz-Ox-EY in a light-dependent manner. The demonstration of EGFR and HER2 labeling in vitro supports the broad applicability of the bioconjugation strategy and photo-probe labeling workflow.

We chose to focus on the NHS-azide conjugate (abbreviated to Ctx-EY) given its higher bioconjugation and photo-probe labeling efficiency. We evaluated the EGFR labeling efficiencies with diazirine-, aryl-azide-, and phenol-biotin photo-probes in parallel (**Fig. 2**). All three probes labeled the EGFR ECD to increasing levels: aryl-azide-biotin > phenol-biotin > diazirine-biotin. The differing yields likely result from the combined effects of the reactive radical intermediates: half-lives (phenol >> aryl-azide > diazirine), yield

of reaction with protein (phenol>aryl- azide>diazirine), and broad amino-acid preference observed (diazirine~aryl-azide>>phenol) (**Fig. S3**) (33-35).

We analyzed the specific sites of biotinylation for self-labeling of BSA and binary complex labeling of Ctx and EGFR with different photo-probes using MS analysis (**Table S1-8** and **Fig. S3**). For diazirine-biotin, we found 17 biotinylated sites on BSA (**Table S1**), and 30 sites on the Ctx-EGFR complex (**Table S5**), with good coverage of modified peptides over the light and heavy chains of Ctx, as well as EGFR ECD (**Fig. 2** and **Fig. S3**). We further characterized the modification sites on BSA and Ctx-EGFR systems for the other probes (**Table S1-8** and **Fig. S3**). As expected, phenol-biotin mostly labeled Tyr/Trp, while labeling with biocytin-hydrazide was found exclusively on His. Diazirine-biotin and aryl-azide-biotin showed very broad amino acid preference, consistent with previous reports (19, 34).

Ctx-EY catalyzes targeted labeling of EGFR on cells.

We next evaluated the ability of Ctx-EY to bind EGFR and label live cells (**Fig. 3**). First, we incubated the Ctx-EY conjugate with an epithelial skin cancer cell line, A431 cells, that endogenously expresses very high levels of wild-type EGFR (nTPM=2978) (36). On-cell binding for the Ctx-conjugates, both Ctx-EY and Ctx-Ir, was confirmed via flow cytometry showing that the conjugation of EY or the Ir-catalyst did not affect binding (**Fig 3**). Detailed titration from 1 nM to 10 μ M of Ctx and Ctx-EY analyzed via flow cytometry further confirmed conjugation did not detectably affect cell binding (**Fig. 3** and **Fig. S4**). We also tested A549 cells with more typical levels of EGFR (37) (nTPM=59.7, **Fig. S4**) as well as NCI-H441 cells with very low EGFR expression (nTPM=29.8, **Fig. S4**), both of which showed proportionally reduced binding of Ctx-EY and was similar to Ctx.

We next performed on-cell proximity labeling with diazirine-, aryl-azide- and phenol- biotin photo-probes upon blue LED illumination (**Fig. 3** and **Fig S5**). We tested a range of Ctx-EY concentrations and observed efficient biotinylation on cells at 100 nM (**Fig. 3** and **Fig. S5**). The diazirine-biotin, aryl-azide biotin, and phenol-biotin labeling caused a major shift of biotinylation in the flow cytometry profile of 64%, 98%, and

94%, respectively. This is consistent with the order of labeling efficiencies observed in vitro. The Ctx-Ir only activated cell biotinylation with diazirine-biotin and aryl-azide-biotin and not phenol-biotin (**Fig. S5**). We further visualized cell biotinylation induced by Ctx-EY via confocal microscopy (**Fig. 3** and **Fig. S5**). The labeled A431, A549 and NCI-H441 cells were co-stained with both anti-human IgG-AlexaFluor488 and streptavidin-AlexaFluor647 to visualize Ctx and biotinylation, respectively. We confirmed that the Ctx-EY conjugate was located on the cell membrane. Biotinylation using the diazirine- and aryl-azide-biotin photo-probes were observed mainly on the cell membrane, whereas the phenol-biotin labeling was more diffuse, consistent with the longer half-life and labeling range of the phenoxy radical.

We next developed a proteomics workflow to label the EGFR neighborhood (**Fig. 4**), focusing first on A431 cells with highest levels of EGFR and using the most reactive diazirine-biotin photo-probe. We incubated A431 cells with or without EGF competition first and then performed the on-cell biotinylation workflow using Ctx-EY, followed by biotin enrichment using neutravidin beads. WB analysis confirmed selective biotinylation of EGFR which was ablated in the presence of EGF (**Fig. 4**). We also observed dose-dependent EGFR labeling over a wide range of Ctx-EY concentrations of 1-1000 nM, which was competed off by either EGF or unlabeled Ctx (**Fig. S6**).

Cells were treated with Ctx-EY in the presence or absence of EGF competition and biotinylated proteins were captured on neutravidin beads and digested on-bead with trypsin. Samples were prepared in biological triplicate for MS analysis using label-free quantitation (volcano plot shown in **Fig. 4**, heatmap shown in **Fig. S6**). We identified a total of 536 proteins with 41 proteins enriched by more than two-fold with Ctx-EY relative to EGF competition ($\log_2(\text{ratio}) \geq 1$, $p\text{-value} < 0.05$, $\text{unique peptide} \geq 2$; **Table S9** and **Fig. S6**). EGFR was among the highly enriched as expected. Gene Ontology (GO) analysis showed a significant representation of biological processes that include regulation of phosphatase activity as well as molecular function entities such as phosphatase activator activity (**Fig. S6**). These features are consistent with the functional roles of EGFR signaling and suggest that the enriched EGFR interactors are accurately represented.

We orthogonally confirmed that six top hits were biotinylated by biotin-IP, where streptavidin pull-down samples were analyzed by WB using specific antibodies following proximity labeling (**Fig. 4**). Among them, five were observed to co-IP with EGFR (**Fig. 4**). All six proteins are known to either functionally interact with EGFR or found in immunoprecipitation experiments (38-40). These include ITB1, which is critical for stable maintenance for EGFR on the cell membrane (41, 42), as well as macrophage migration inhibitory factor (MIF), an immunostimulatory cytokine regulated by matrix metalloproteinase 13 (MMP13) known to be inhibitory for EGFR activation (43). Others include substrates of EGFR such as glutathione S-transferase P1 GSTP1 (44) and tight junction protein ZO1 (45), both of which are known to be activated upon phosphorylation by EGFR. One target membrane-associated progesterone receptor component 1, PGRC1, was not observed in EGFR co-IP experiment, and we would expect that some interactions may not be strong enough to survive the co-IP workup in these cells.

Multi-scale EGFR interactome profiled via MultiMap

Having demonstrated the proteomic workflow of Ctx-EY triggered biotinylation on cells expressing high levels of EGFR, we expanded to cells expressing modest levels of EGFR. Lung cancer cell line, A549, for example, express lower amounts of EGFR (nTPM=59.7), which is more typical of native membrane proteins (37). We applied all photo-probes and showed EGFR was selectively biotinylated with each probe (**Fig. S7**). Applying the proteomics workflow, we then identified EGFR neighbors enriched with diazirine-biotin, aryl-azide-biotin and phenol-biotin by comparing labeling with Ctx-EY in the absence and presence of EGF (**Fig. 5**). We found that EGFR is one of the most enriched proteins from all three datasets (**Table S10-12**). Enriched proteins were identified with the same statistical thresholds [$\log_2(\text{ratio}) \geq 1$, $p\text{-value} < 0.05$, unique peptide ≥ 2], allowing us to compare protein identities across reactions with different photo-probes. We identified 72 proteins using diazirine-biotin, 188 using aryl-azide-biotin, and 188 using phenol-biotin (Tables S10-12 and **Fig. S7**).

As represented in a Venn diagram (**Fig. 5**), there were a total of 322 unique proteins enriched over the controls in at least one of the three photo-probes. As represented in a Venn diagram (**Fig. 5**), there were a

total of 322 unique proteins enriched over the controls in at least one of the three photo-probes. The aryl-azide-biotin and phenol-biotin labeled more proteins than diazirine-biotin reflecting their higher yields and their relatively long labeling radii. We found that >80% of the enriched proteins were annotated as plasma membrane proteins in UniProt (plasma membrane, GO:0005886) for all three photo-probes. GO enrichment analysis suggested molecular functions such as EGFR activity and EGF binding were highly enriched (**Fig. S7**).

Sixteen candidate neighbors were identified in all three datasets of MultiMap (**Fig. 5**). While no direct structural evidence has been reported for EGFR with any of these, CD44 and Galectin-3 have been functionally associated with EGFR: CD44 regulates EGFR functions in the presence of CD147 and hyaluronan (46, 47); Galectin-3 regulates EGFR localization and its interactions suggested through genetic studies in pancreatic cancers (48). Both targets were further validated by biotin-IP and EGFR co-IP (**Fig. 5**), supporting that they are proximal neighbors of EGFR.

We next expanded our list to the 29 proteins that were in common for diazirine-biotin and aryl-azide-biotin, the photo-probes with high labeling resolutions (**Fig. 5**). Among them, we found a paraoxonase, PON2, as well as two proteins associated with RTK phosphorylation and activation: beta-adducin ADBB and MAP kinase pathway member BRAF (49, 50). Both PON2 and ADBB were detected by biotin-IP and EGFR co-IP. Interestingly, BRAF, a cytosolic protein was enriched by biotin-IP but not EGFR co-IP suggesting it is close but may not be in physical contact (**Fig. 5** and **Fig. S7**). Between the diazirine and aryl-azide datasets, we identified the known EGFR functional interactors such as Tid1 (51) and ITB1 also found in the A431 cell experiments (33, 34), as well as previously unreported interaction partners including CKAP4 and RAC1 (**Fig. S7**).

The aryl-azide-biotin and phenol-biotin experiments contributed more proteins (293 in total). Among the top hits were the tyrosine-protein phosphatase receptor, PTPRF, glutathione transferase GSTP1, small GTPase Rab11a, Rho-related GTP binding protein RHOC and ESCRT protein PDC6I (**Fig. 5** and **Fig. S7**).

Remarkably, all were detected by biotin-IP with ten out of eleven of these proteins co-IPed with EGFR, suggesting that they form relatively stable complexes.

To provide a structural level of analysis, we turned to AlphaFold-Multimer, an exciting extension of AlphaFold developed over the last few years that uses artificial intelligence to generate plausible models of binary protein complexes (24, 52, 53). This community has developed scoring metrics such a predicted DockQ score (pDockQ), where a threshold of >0.23 retrieves 51% of true-positive interacting proteins with a false-positive rate of $\sim 1\%$ in large test set models (54). Additional criteria can be applied including buried solvent accessible surface area (BSASA) $\geq 500 \text{ \AA}^2$ (55, 56), predicted local distance difference test (pLDDT) > 50 for the interface residues and minimum predicted alignment error (PAE) $< 15 \text{ \AA}$ as described previously (52). As a true positive example, we derived an AlphaFold-Multimer model of the EGF:EGFR complex (**Fig. S8** and **Table S13**) that closely overlaid that of the known structure of EGF:EGFR (PDB: 1IVO, RMSD between 469 atom pairs is 0.924 \AA) (57).

We applied AlphaFold-Multimer to candidate neighbors validated by biotin-IP and/or EGFR co-IP in A431 and A549 cells and generated a total of 29 models. As shown in waterfall plots, the average pDockQ score (0.298) and BSASA (1466 \AA^2) for the 29 EGFR-protein pairs were both above the established criteria suggesting direct interactions (**Fig. S8B** and **Table S13**). We next applied AlphaFold-Multimer to compute models of all potential heterodimeric complexes from **Fig. 4** and **Fig. 5** (**Table S13**). To increase the accuracy of the models for transmembrane proteins (58), we calculated separately the ECD (aa 1-646) and intracellular domain (ICD, aa 695- 1022) of EGFR and paired them with the corresponding ECDs or ICDs of transmembrane protein targets. As previously described, we retained only high-confidence AlphaFold-Multimer models [average pLDDT > 50 , minimum predicted Alignment Error (PAE) $< 15 \text{ \AA}$] (52) and performed further filtering using the aforementioned criteria [pDockQ score ≥ 0.23 , BSASA $\geq 500 \text{ \AA}^2$]. The final list of validated complexes included the binary complexes of EGFR ECD with CD44 ECD (aa 1-153, pDockQ=0.375), PON2 (pDockQ=0.372) and MIF (aa 1-115, pDockQ=0.375) (**Fig. 5** and **Fig. S8**). In addition, the ICD of EGFR is predicted to bind Rab11a (pDockQ=0.264), GSTP1 (pDockQ=0.535) and

RAC1 (pDockQ=0.387) (**Fig. 5** and **Fig. S9**). Most interestingly, one of the AlphaFold-Multimer complexes predicted with the highest confidence is a cell-surface phosphatase PTPRF, where PTPRF ECD binds EGFR ECD (pDockQ=0.429) and likewise, the PTPRF ICD binds the EGFR ICD (pDockQ=0.476) (**Fig. 5** down).

MultiMap can capture distal synaptic protein networks

Extracellular protein-protein interactions occur not only in cis on the cell membrane but also in trans between cell-cell junctions. To explore PLP of cell synapses using MultiMap at different labeling radii, we assembled a co-culture system where the cell-cell interaction was induced by a bispecific T cell engager (BiTE) (**Fig. 6**). This BiTE contained the Ctx Fab genetically fused to an α -CD3 scFv (OKT3) (59). Two different cells were utilized in the co-culture system: a HEK293T cell was engineered in house to overexpress a Flag-tagged-EGFR (HEK-Flag-EGFR) and well-established Jurkat cells expressing a NFAT-GFP reporter (60). In this design, the Flag tag served as an orthogonal ecto-epitope for an EY-conjugated α -Flag nanobody (α -Flag-EY), thus allowing selectively recognition of Flag-tagged EGFR. In order to separately characterize the labeling on HEK-Flag-EGFR and Jurkat NFAT-GFP cells, we used α -CD3-PE signal to allow facile separation of CD3+ Jurkat cells from CD3- HEK-Flag-EGFR. Levels of cis- and trans-labeling from α -Flag-EY were determined by flow cytometry. Proteins labeled with different photo-probes were enriched using streptavidin beads and analyzed by WB (**Fig. 6**).

We first monitored BiTE engagement between HEK-Flag-EGFR and Jurkat NFAT-GFP cells using the standard GFP reporter gene readout. As expected, we observed dose-dependent BiTE activation of cell-cell engagement, with the GFP signal shifted to 80.3% in the presence of 8 nM EGFR BiTE and 92.3% with 50 nM BiTE (**Fig. S10**). The GFP signal shift was not affected by the presence of α -Flag-EY, indicating that the Flag tag recognition did not interfere with the cell synapse engagement. We then performed the MultiMap workflow using four photo-probes of increasing labeling range: diazirine-biotin, aryl-azide-biotin, biocytin-hydrazide and phenol-biotin. We monitored biotinylation in cis for HEK-Flag-EGFR and in trans for Jurkat NFAT-GFP using a streptavidin-AlexaFluor647 signal (**Fig. 6** and **Fig. S10**). Cis-labeling of HEK-Flag-EGFR cells occurred for >60% of cells for the diazirin-biotin, aryl-azide-biotin and phenol-

biotin with ~29% for the biocytin hydrazide (**Fig. S10**). In sharp contrast, minimal shift (~3-4%) was observed on control HEK-EGFR cells without the Flag tag, suggesting that α -Flag-EY is selectively recognizing the Flag tag. Interestingly, the trans-labeling of the Jurkat cells using the shorter-range diazirine-biotin, aryl-azide-biotin was limited to 3-4% (**Fig. S10**), while the intermediate-range biocytin-hydrazide and long-range phenol-biotin labeled 9% and 22%, respectively) (**Fig. 6** and **Fig. S10**). This is consistent with the cell-cell synapse distance based on the length of the Fab-ScFv BiTE (61), plus the size of the EGFR ECD and the CD3 complex. By further analysis via WB, the cis-target EGFR was observed enriched by biotin-IP in the presence of the three photo-probes, whereas trans-target CD3 was only significantly enriched in the phenol-biotin sample, with moderate amount observed in the aryl-azide-biotin-labeled sample (**Fig. 6**). Thus, longer-range photo-probes are more efficient for trans-labeling.

To further expand the generality of MultiMap for cell-cell synapses, we tested the BiTE system to two other cancer targets, HER2 and CDCP1 (**Fig. 6** and **Fig. S11**). We fused α -HER2 Fab sequence (Trz Fab, **Fig. S10**) and a previously generated α -CDCP1 Fab (4A06) (**Fig. 6** and **Fig. S11**) (59) onto the CD3 scFv scaffold. Again, we observed dose-dependent cell-cell engagement in the presence of the engineered BiTEs and antigen-expressing cells (**Fig. S11** and **Fig. S11**). On-cell biotinylation results were similar to the EGFR BiTE system; cis-labeling was found with all three photo-probes and trans-labeling activated primarily with phenol-biotin (**Fig. S10** and **Fig. S11**). In particular, by separating HEK293T-CDCP1 and Jurkat-NFAT-GFP cells, we confirmed selective biotinylation of CDCP1 with all three photo-probes, and CD3 only with phenol-biotin (**Fig. 6**). We quantitatively profiled the proteins captured at the cell synapse of HEK-Flag-CDCP1 and Jurkat NFAT-GFP in biological triplicate (**Fig. 6**, **Table S14**). We discovered that CDCP1 was enriched in the cis-labeled samples. Proteins from the CD3 complex including CD3d and CD3e were highly enriched in the trans-labeled samples. This observation indicates that interactome in the cell-cell synapse can be captured by the MultiMap workflow.

Lastly, we evaluated MultiMap labeling at a (CAR)T cell-cell synapse (**Fig. 6** and **Fig. S11**). Jurkat cells expressing a Myc-tagged CAR construct (Jurkat-CAR) that targets CD19 were mixed with K562 cancer

cells expressing CD19 ectodomain (K562-CD19). With an EY- conjugated α -myc antibody (α -myc-EY), we performed our workflow by introducing cis-labeling on K562-CD19 and trans-labeling on Jurkat-CAR cells. We confirmed cell engagement by monitoring the CAR activation with or without K562 cells (**Fig. S11**). We observed cis-labeling on interacting CAR cells with all photo-probes (**Fig. S11**). On the other hand, both aryl-azide- biotin and phenol-biotin achieved trans-labeling, with much lower level of biotinylation using the short-range diazirine-biotin (**Fig. 6**). The same results were confirmed by WB analysis (**Fig. 6**). These results are in line with the estimate of cell-cell distance between CAR-induced synapse at ~ 120 Å according to AlphaFold prediction (62), which is shorter than BiTE-induced synapse. We finally sorted each cell type for proteomics analysis and found both CD19 and CD3 component enriched for trans-labeling and cis-labeling (**Fig. 6** and Tables S15). Taken together, our data suggests that MultiMap can label cells at the cell-cell synapses and map the proteins in proximity via PLP. Only minimal alternation to the existing workflow is needed for labeling in different cell- cell engagement scenarios. Looking forward, we anticipate that this platform will be a useful technology to identify key proteins in different synaptic environments.

3.4 Discussion

Here we demonstrate a multi-scale PLP technology, MultiMap, that enables proximity labeling and interactome profiling with adjustable resolution depending on the half-life of the photo-probe using a single photocatalyst EY. EY is a unique photocatalyst in its ability to trigger a broad range of photo-probes. It is commercially available, bio-compatible, and shown to be readily conjugated to seven different proteins and antibodies by commonly accessible methods. Simple targeting by EY-conjugated antibodies obviates the need for cell engineering. EY-mediated labeling is rapid and light-dependent, which potentially allows kinetic control of the labeling. Future structure-activity relationship studies on EY, as done for rhodamine- or fluorescein-based scaffolds (63), could enhance our understanding of the photochemical mechanisms of EY, as well as improve its spectral properties, activation efficiency, and use in tissues (20, 27).

Coupled with standard biochemical validation and the recently developed AlphaFold- Multimer algorithm for structural prediction (24), the MultiMap proximity labeling proteomics workflow provides three orthogonal and integrated pillars for high-resolution profiling of protein neighborhoods. In addition to identification of new potential neighbors, we detected many proteins known to functionally interact with EGFR that are reported to stabilize, modulate, or act as substrates for EGFR. One of the most striking targets identified was the phosphatase, PTPRF, which could be a functional off-switch for EGFR. Interestingly, AlphaFold-Multimer predicts the ECD of PTPRF binds the back side of the EGFR ECD away from the dimer interface, and that the ICD of the phosphatase binds to the intracellular kinase domain of EGFR. Despite the fact that EY-antibodies recognize extracellular targets, we found that some of the high-confidence hits were intracellular proteins. Some are known to functionally associate with EGFR, for which high- confidence AlphaFold-Multimer binary models were constructed. It is not impossible that the labeling is caused by cell penetrance of the activated photo-probe when triggered by EY. We believe future work could fine-tune the properties of photo-probes such as charge and hydrophobicity to achieve extracellular-only or organelle-specific labeling.

It is unlikely that all these identified neighbors bind simultaneously to EGFR. In fact, some proteins are predicted to bind over the same sites. These data suggest EGFR can be in multiple neighborhoods which are dynamic and may have multiple functions yet to be revealed. It is also important to note that the binder we used in this study, Ctx, is an inhibitor of EGFR function. Thus, candidates identified in our studies are specifically from the EGFR off-state neighborhood. We envision that MultiMap will be useful to study on-state, drug-bound, and resistance mutant neighborhoods, which will give a comprehensive map of the EGFR interactomes.

MultiMap was also effective for long-range labeling of cell-cell synapses. As shown for the ones activated by BiTE or (CAR)T, we found that spatial variability among synaptic junctions can be addressed by using photo-probes with different labeling radii. The unique advantage of MultiMap allowing multi-scale labeling, potentiates its application for interactome profiling of additional intercellular interaction networks.

In cases where antibodies are not available, one can use a genetically encoded tag on the target ECD, similar to the Flag and myc ecto-tags we introduced in our study. We can also envision proteome-wide interactome profiling for membrane proteins using these ecto-tags on par with the scale for the intracellular OpenCell system (64).

Lastly, we recognize that despite the confidence and accelerated process in target identification via MultiMap, information on candidate neighborhoods warrants further confirmation via more in-depth structural, mutational, and functional studies. Nonetheless, we believe that MultiMap proximity labeling proteomics integrated with in silico prediction can begin to provide plausible models for binary protein interactions with high structural precision. This would be crucial step forward to begin to construct a structural map of the protein-protein interactome on the cell surface. In addition, understanding how the surfaceome functions in concert with the external environment communication may suggest new neo-complexes to target for both small molecules and biologics.

3.5 Acknowledgements

We thank Jie Zhou, Kevin Leung, Thomas Bartholow, Johnathan Maza, Kaan Kumru, Corleone Delaveris, Paul Burroughs and James Byrnes for insightful discussions. We also thank Susanna Elledge for the Her2 Fab expression plasmids, Jhoely Duque-Jimenez and Xin Zhou from Harvard Medical School for CAR-T and K562-CD19 cells as well as Juntao Yu from Harvard Medical School for guidance in conducting the AlphaFold-Multimer analysis.

Funding: We are grateful to generous support from NIH (1R01CA248323-01). K.S. is a Merck fellow of the Helen Hay Whitney Foundation. Z.Y. is supported by an NIH National Institute of General Medical Sciences F32 grant (1F32GM149084-01). A.S. is supported by NIH (P41GM109824 and R01GM083960). A.P. is supported by NIH (U19AI135990). The HDFCC LCA for cell sorting is funded by NIH (P30CA082103).

3.6 Declaration of interests

The authors declare the following competing financial interest: J.A.W and Z.L filed a provisional patent on the multi-scale interactome profiling of membrane proteins using photocatalytic proximity labeling.

3.7 Data and materials availability

All data are available in the main text or the supplementary materials.

3.8 Figures

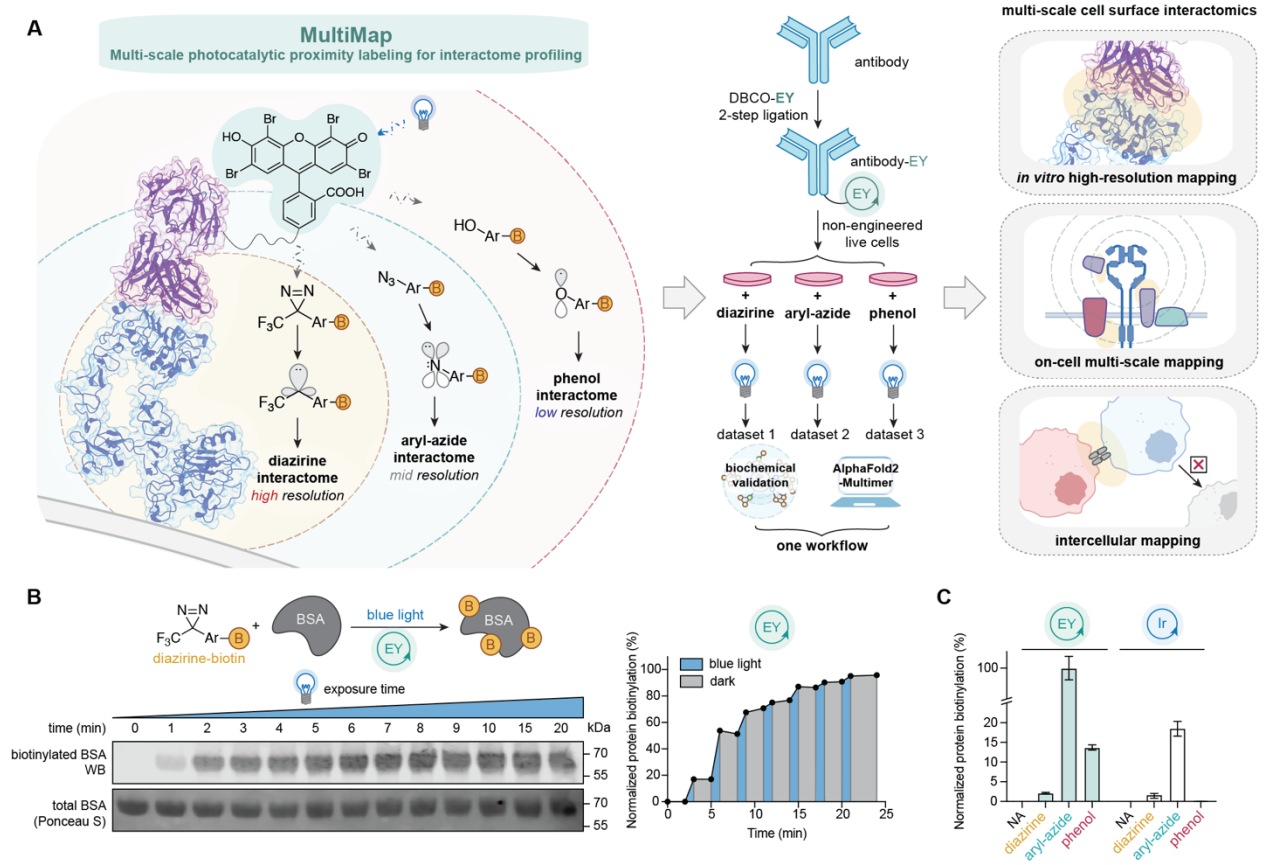


Fig. 1. MultiMap captures high-resolution snapshots of biological networks across a wide range of length scales. (A) A schematic of the MultiMap workflow. EY is conjugated to an antibody that binds the target of interest (*e.g.*, Fab arm of Ctx bound to the EGFR extracellular domain). Upon illumination, proteins are biotinylated, captured and digested for MS analysis. Proteomics hits are further examined by immunoprecipitation and predictive structural analysis via AlphaFold-Multimer. MultiMap is a useful platform for profiling local membrane protein interactomes both on live cells and between cell-cell synapses. (B) WB showing EY-mediated photocatalytic biotinylation of BSA by a diazirine-biotin probe upon blue LED illumination. Biotinylation can be controlled temporally by pulsed light. (C) EY triggers labeling of BSA with all three photo-probes (diazirine-biotin, aryl-azide-biotin and phenol-biotin), but Ir only activates the first two. All immunoblot images are representative of at least two biological replicates.

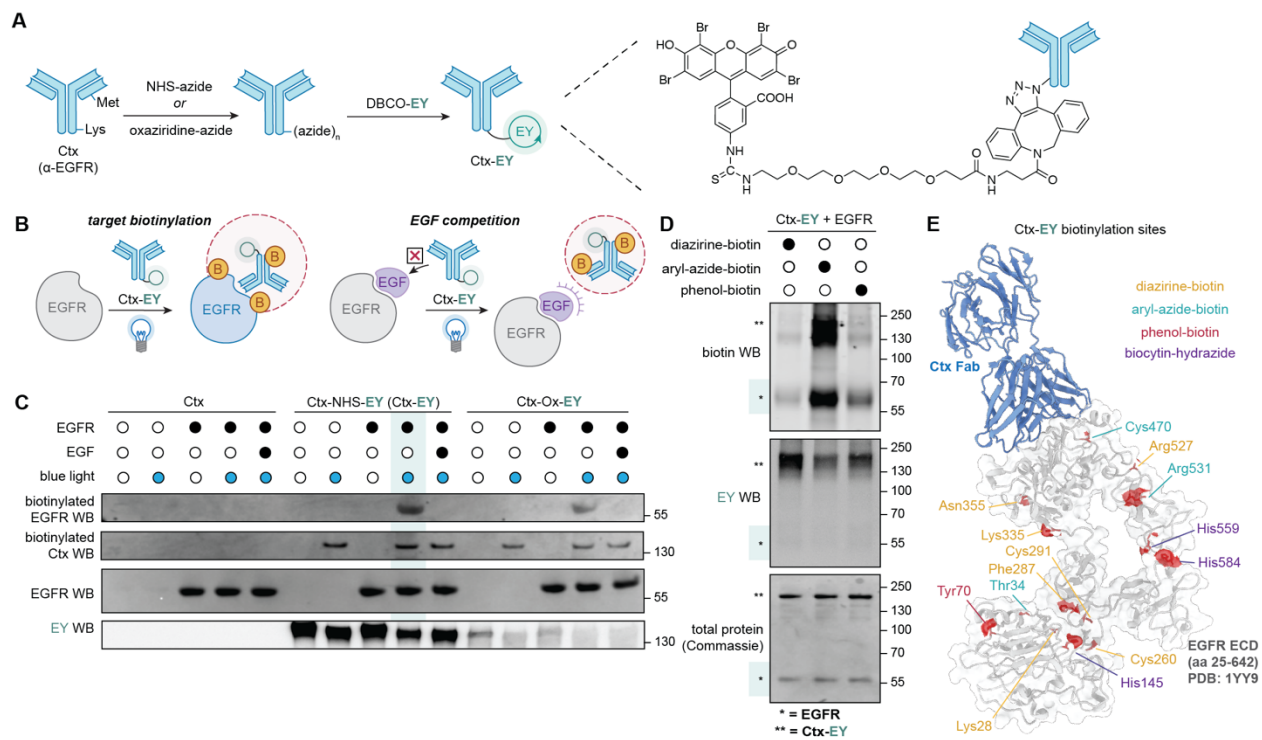


Fig. 2. Targeted labeling using antibody-EY conjugates in vitro. (A) Synthetic scheme of Ctx-EY via a two-step bioconjugation workflow. An azido functionality was first introduced onto either Lys or Met residues using NHS or oxaziridine chemistry, respectively, followed by bio-orthogonal click reaction to couple EY. (B) Schematic design to test intra- and inter-biotinylation of Ctx-EY and EGFR with or without EGF competition. (C) Targeted EGFR biotinylation with the diazirine-biotin photo-probe when triggered by either Ctx-NHS-EY (Ctx-EY) or Ctx-Ox-EY in vitro. Both conjugates selectively label EGFR in a light-dependent fashion, which is competed off by exogenous EGF. (D) EGFR is biotinylated by all three photo-probes: diazirine-biotin, aryl-azide-biotin or phenol-biotin using Ctx-EY. (E) Biotinylation sites of diazirine-biotin (yellow), aryl-azide-biotin (cyan), phenol-biotin (maroon) and biocytin-hydrazide (purple) highlighted on the crystal structure of the EGFR ECD (grey) in complex with Ctx Fab (blue) (PDB: 1YY9, data taken from **Table S5-8**). All immunoblot images are representative of at least two biological replicates.

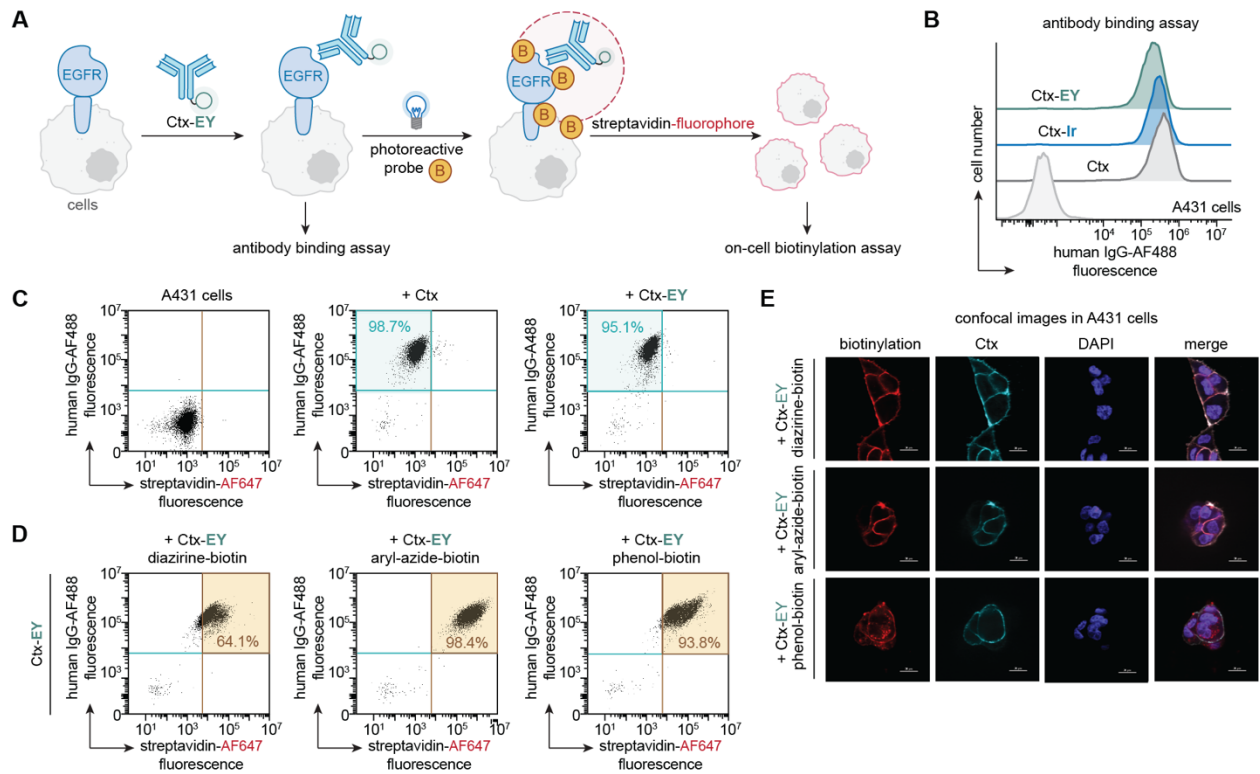


Fig. 3. Ctx-EY enables EGFR-dependent labeling on cells with different photo-probes. (A)

General on-cell labeling workflow using Ctx-EY conjugate and detection of biotin labeling using fluorescent streptavidin-AF647. (B) Cellular binding assay of 100 nM Ctx, Ctx-EY or Ctx-Ir conjugates on A431 cells via flow cytometry analysis shows similar on-cell binding. (C, D) Quantitative on-cell binding (C) and on-cell labeling (D) with diazirine-, aryl-azide- and phenol- biotin triggered by 100 nM Ctx-EY on A431 cells via flow cytometry analysis. (E) Confocal microscopy imaging of antibody binding and on-cell biotinylation of Ctx-EY on A431 cells shows labeling mostly confined to cell surface. Scale bar=20 μm .

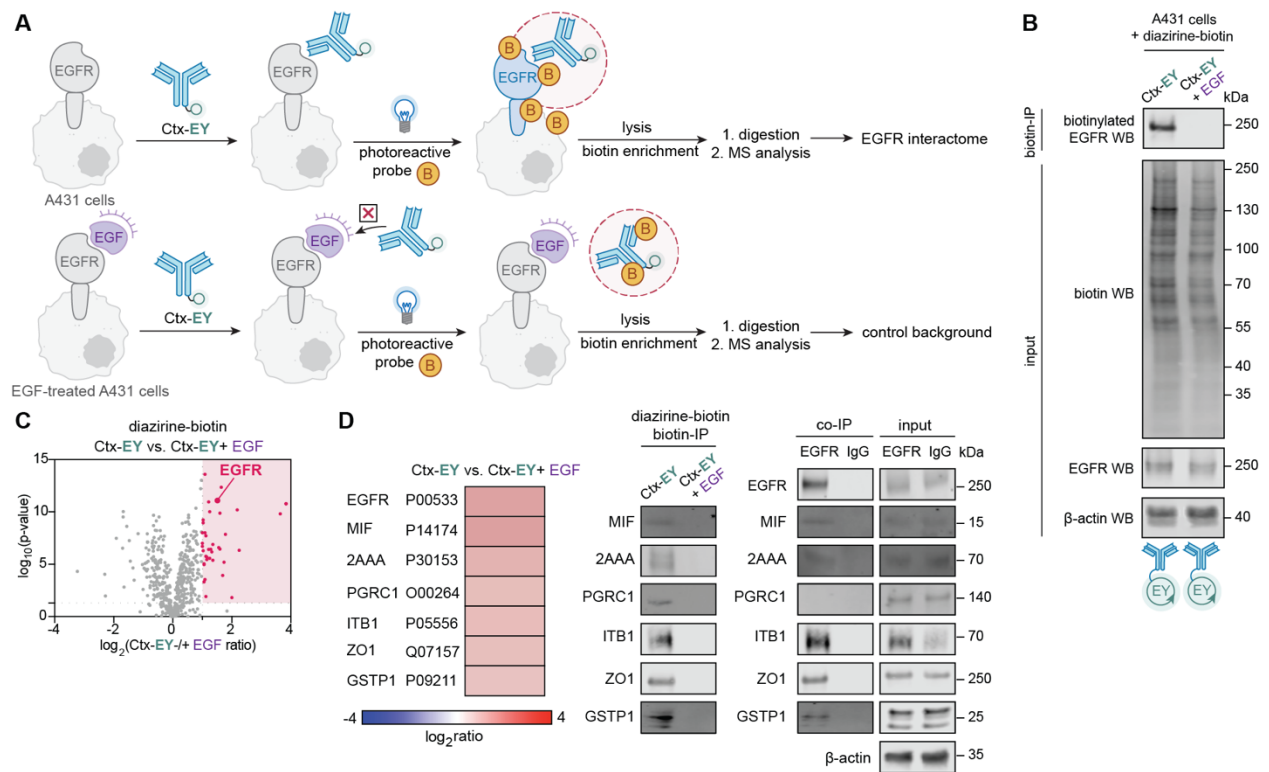


Fig. 4. High-resolution profiling of the EGFR neighborhood using MultiMap. (A) General proteomics workflow of interactome profiling using Ctx-EY conjugate with or without EGF competition. (B) WB showing biotinylation using the diazirine-biotin photo-probe on A431 cells using Ctx-EY in the absence and presence of EGF competition. (C) Volcano plot of Ctx-EY-mediated labeling of EGFR with or without EGF on A431 cells using diazirine-biotin. 41 significantly enriched proteins ($\log_2(\text{ratio}) \geq 1$, $p\text{-value} < 0.05$, unique peptide ≥ 2 , $n=3$) are highlighted in red and listed in **Table S9**. (D) Enrichment ratios for six top protein hits are displayed. Full heatmap is presented in **Fig. S6** and **Table S9**. All six proteins were confirmed to be biotinylated using biotin-IP blots, and five were selectively enriched in a separate EGFR co-IP experiment. All immunoblot images are representative of at least two biological replicates.

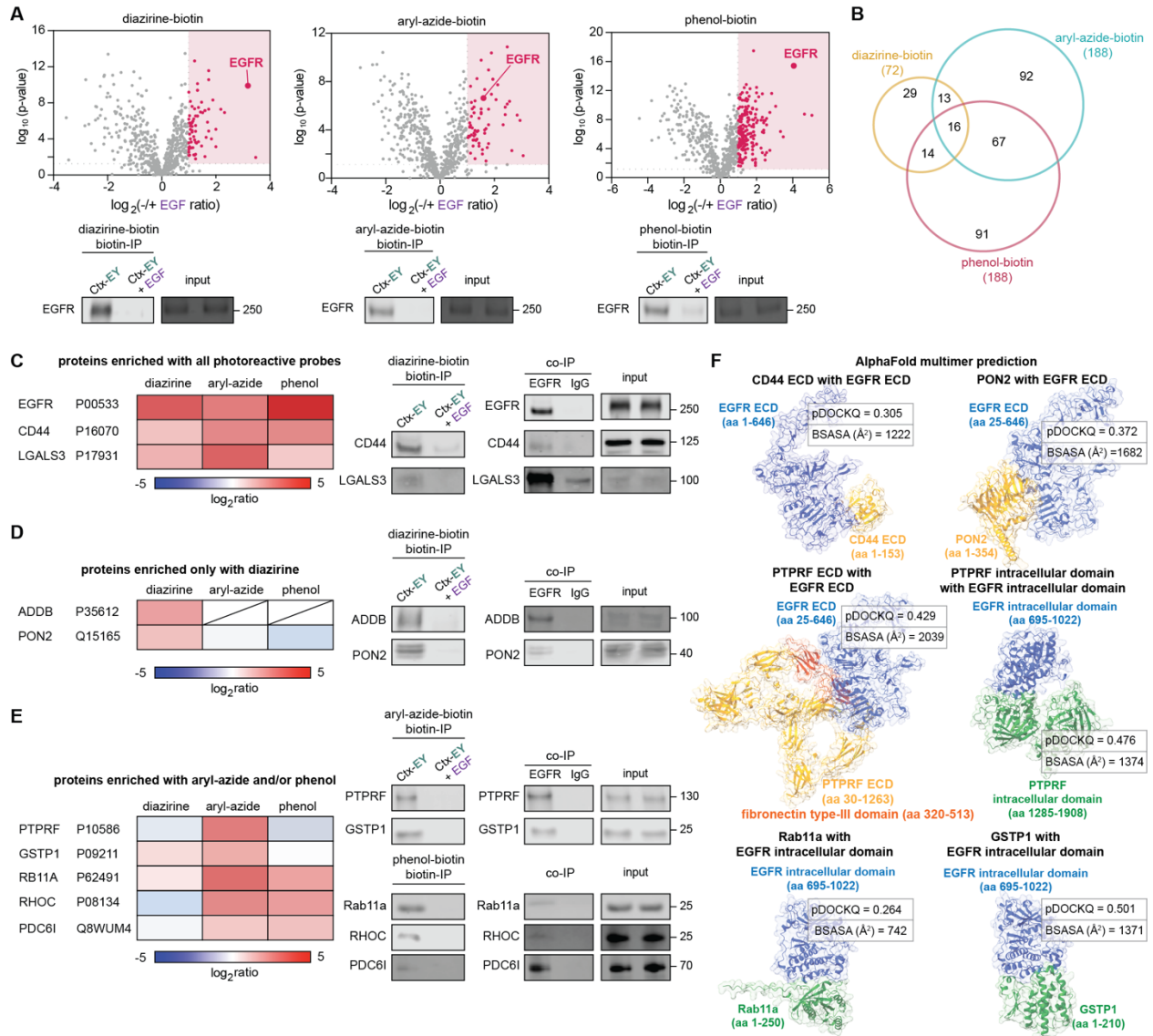


Fig. 5. MultiMap reveals a multi-scale EGFR interactome network. (A) Volcano plots of Ctx- EY mediated EGFR interactome profiling on A549 cells using three different photo-probes (biotin- diazirine, aryl-azide-biotin, or phenol-biotin, respectively, n=3). Significantly enriched proteins ($\log_2(\text{ratio}) \geq 1$, p-value < 0.05, unique peptide ≥ 2) are highlighted in red and listed in **Table S10-12**. (B) Venn diagram of EGFR interactome enriched from A431 cells using different photo-probes. (C) Enrichment ratios and validation of protein hits using all three photo-probes. (D) Enrichment ratios and validation of protein hits from only the diazirine-biotin dataset. (E) Enrichment ratios and validation of protein hits from both aryl-azide-biotin and/or phenol-biotin datasets. (F) AlphaFold-Multimer predictions of EGFR complexes confirmed the direct interactions of EGFR with interactors found via MultiMap. EGFR ECD or ICD (blue) is shown in complex with the corresponding interactor protein (yellow or orange-yellow for the ones interacting with EGFR ECD, green for the ones interacting with EGFR ICD) along with the pDockQ scores and BSASA. All immunoblot images are representative of at least two biological replicates.

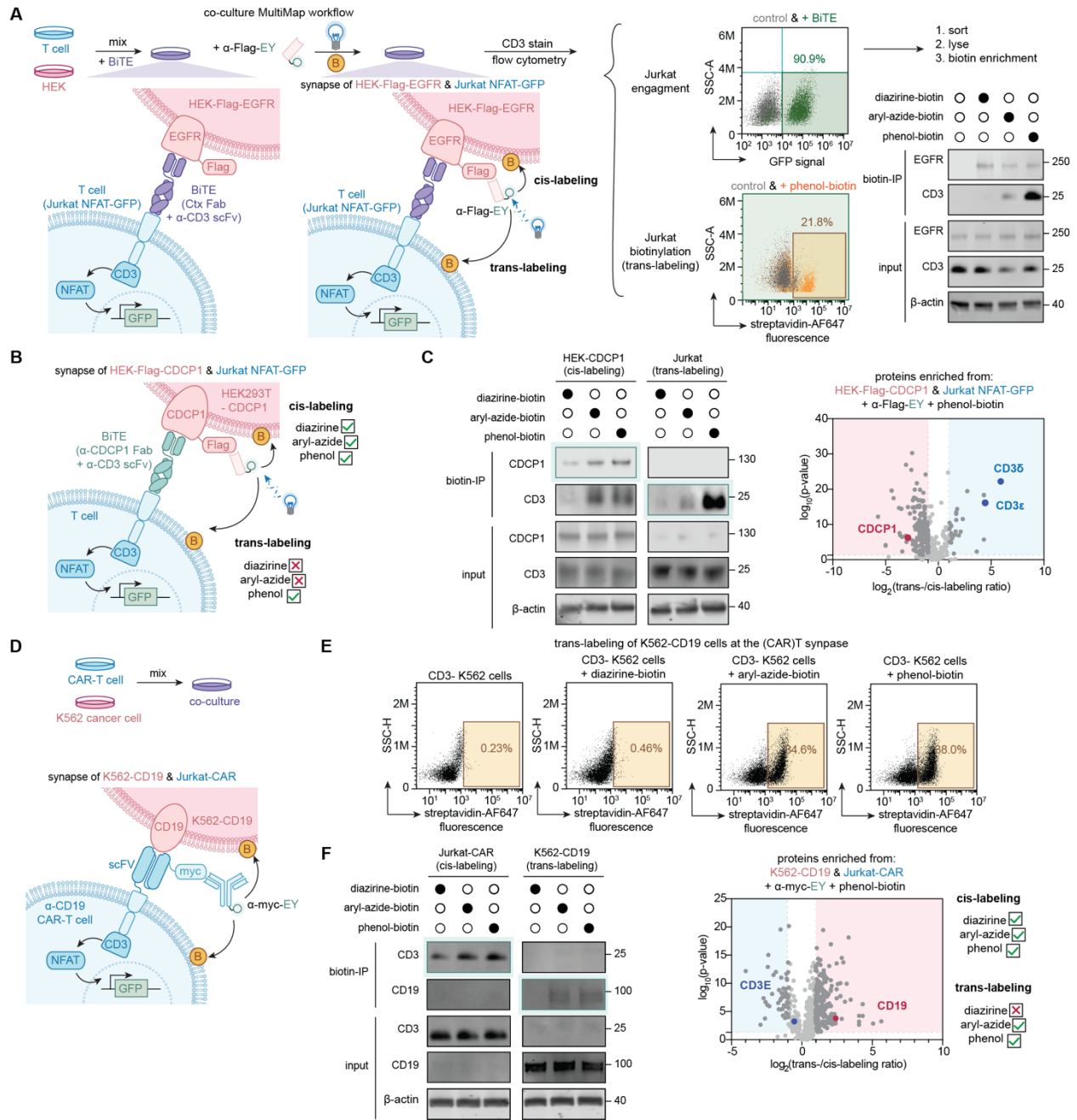


Fig. 6. MultiMap enables targeted snapshots of local cell-cell synapses. (A) On-cell labeling of the T-cell synapse using a bispecific T cell engager (BiTE) that recognizes EGFR. Jurkat NFAT- GFP and HEK293T-Flag-EGFR were co-cultured in the presence of the BiTE before MultiMap was performed using an EY-conjugated α -Flag nanobody (α -Flag-EY). Cell-cell engagement was monitored by NFAT-GFP reporter gene activation. Photocatalytic labeling was characterized by flow cytometry before biotin-enriched proteins were analyzed by WB. (B) Target biotinylation of CDCP1 and CD3 at the T cell synapse using a bispecific T cell engager (BiTE) (Figure caption continued on next page)

(Figure caption continued from previous page) that recognizes CDCP1. Longer labeling radius using phenol-biotin was necessary for trans-labeling on Jurkat NFAT-GFP. (C) Volcano plot of proteins biotinylated on HEK-Flag-CDCP1 (cis-labeling) and Jurkat NFAT-GFP (trans-labeling) using phenol-biotin (n=3). Significantly enriched proteins from HEK-Flag-CDCP1 ($\log_2(\text{ratio}) \leq -1$, p-value < 0.05, unique peptide ≥ 2) or Jurkat NFAT-GFP ($\log_2(\text{ratio}) \geq 1$, p-value < 0.05, unique peptide ≥ 2) are highlighted in blue and red, respectively. Full protein lists were shown in **Table S14**. (D) Scheme of on-cell labeling of anti-CD19 chimeric antigen receptor (CAR)T cell system. (E) (CAR)T cell-mediated trans-labeling of K562 cancer cells using all photo-probes. (F) Target biotinylation of CD3 and CD19 at the CAR-T synapses using WB and MS analysis. Cells were sorted to differentiate cis- and trans-labeling before biotinylated proteins were enriched for analysis. Both phenol-biotin and aryl-azide-biotin enabled trans-labeling. Volcano plot of proteins biotinylated on Jurkat-CAR (cis-labeling) and K562-CD19 (trans-labeling) using phenol-biotin (n=3). Significantly enriched proteins from Jurkat-CAR ($\log_2(\text{ratio}) \leq -1$, p-value < 0.05, unique peptide ≥ 2) or K562-CD19 ($\log_2(\text{ratio}) \geq 1$, p-value < 0.05, unique peptide ≥ 2) are highlighted in blue and red, respectively. Full protein lists were shown in **Table S15**. All immunoblot images are representative of at least two biological replicates.

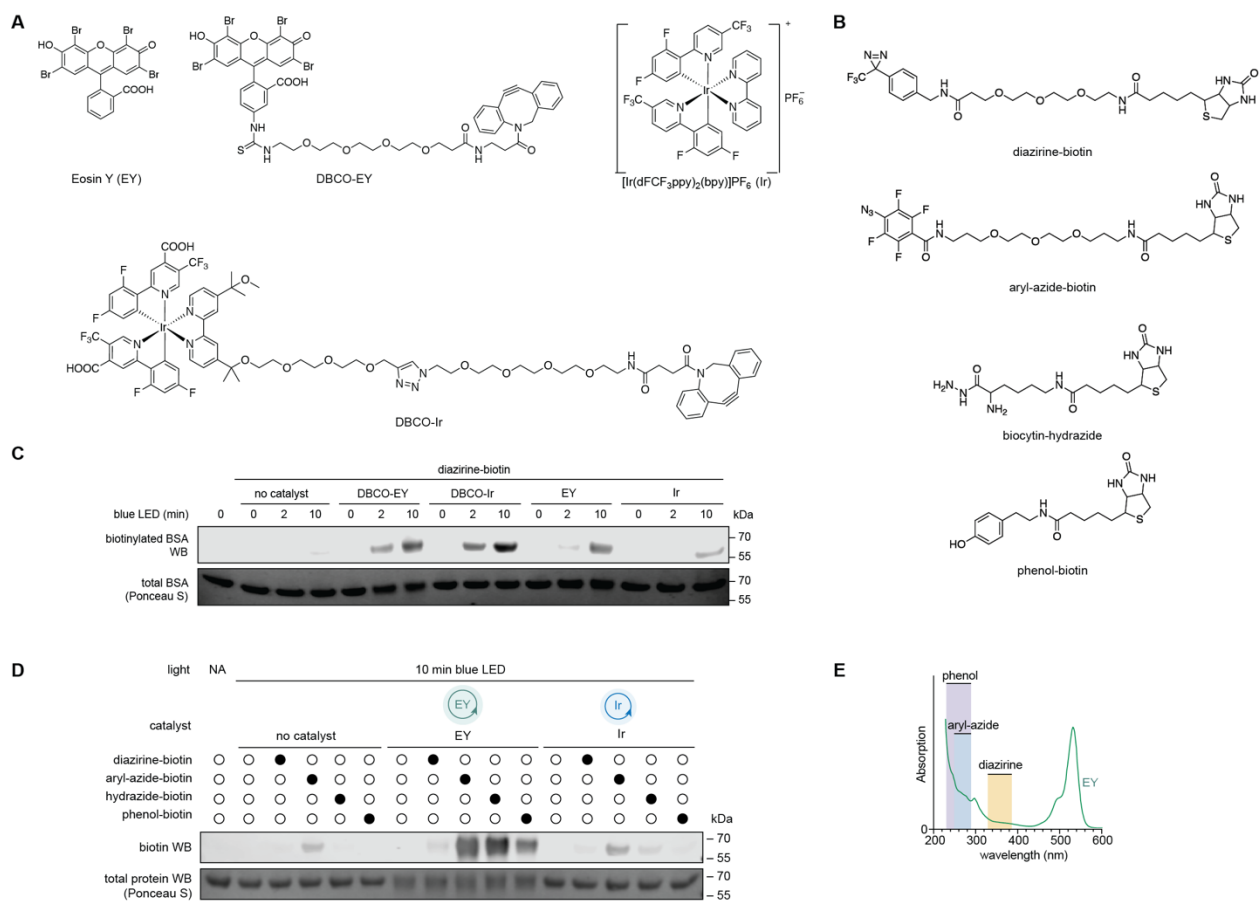


Fig. S1. Eosin Y (EY) as an organic photocatalyst that triggers the activation of photo-probes. (A) Chemical structures of photocatalysts used in this study. **(B)** Chemical structures of all photo-probes used in this study. **(C)** EY triggers the biotinylation of bovine serum albumin (BSA) using diazirine-biotin in a time-dependent manner. Significant amount of biotinylation was observed with both EY and Ir than the background labeling. **(D)** EY triggers biotinylation of BSA with all four photo-probes, diazirine-biotin, aryl-azide-biotin, biocytin-hydrazide and phenol-biotin upon blue LED illumination. Quantification of biotinylation is shown in **Fig. 1C**. **(E)** Absorption peaks of the photoreactive warheads (diazirine, aryl-azide and phenol, shown in colored boxes) and EY. Photoreactive warheads cannot be activated by blue or green LED and thus require EY for activation.

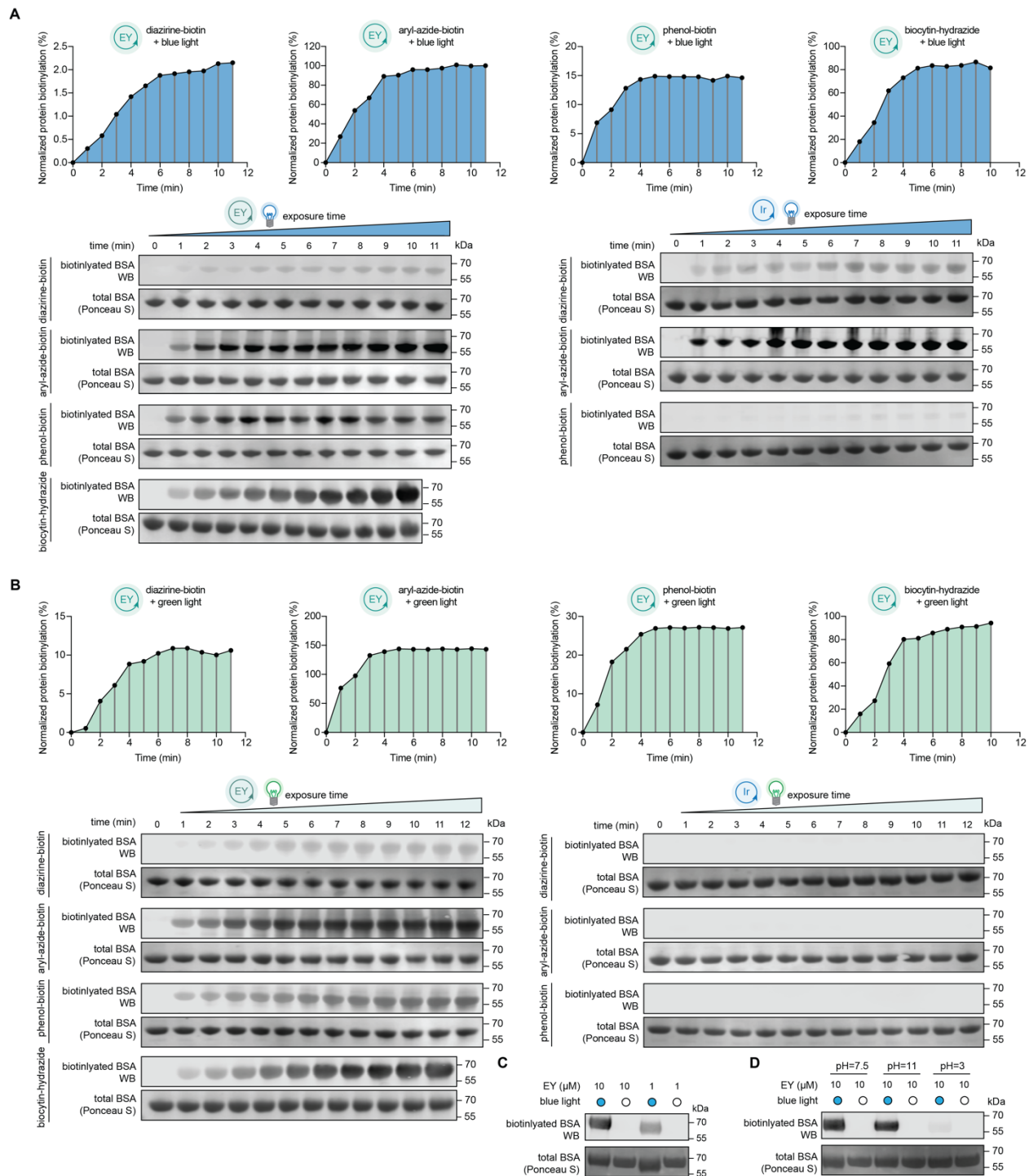


Fig. S2. EY triggers protein labeling in a light-dependent manner. (A) Time-dependent biotinylation on BSA demonstrated the rapid kinetics for labeling using EY with blue LED activation. Biotinylation levels of BSA in were tracked over time and quantified, indicating that EY can trigger >90% conversion with 3 min illumination with blue LED. **(B)** Time-dependent biotinylation on BSA demonstrated the rapid kinetics for labeling using EY with green LED activation. Biotinylation levels of BSA were tracked via WB analysis and quantified. (Figure caption continued on next page)

(Figure caption continued from previous page) **(C)** EY- induced biotinylation using diazirine-biotin is dose-dependent. **(D)** Photocatalytic ability of EY was not affected in basic condition (pH=11), but was significantly inhibited in acidic condition (pH=3).

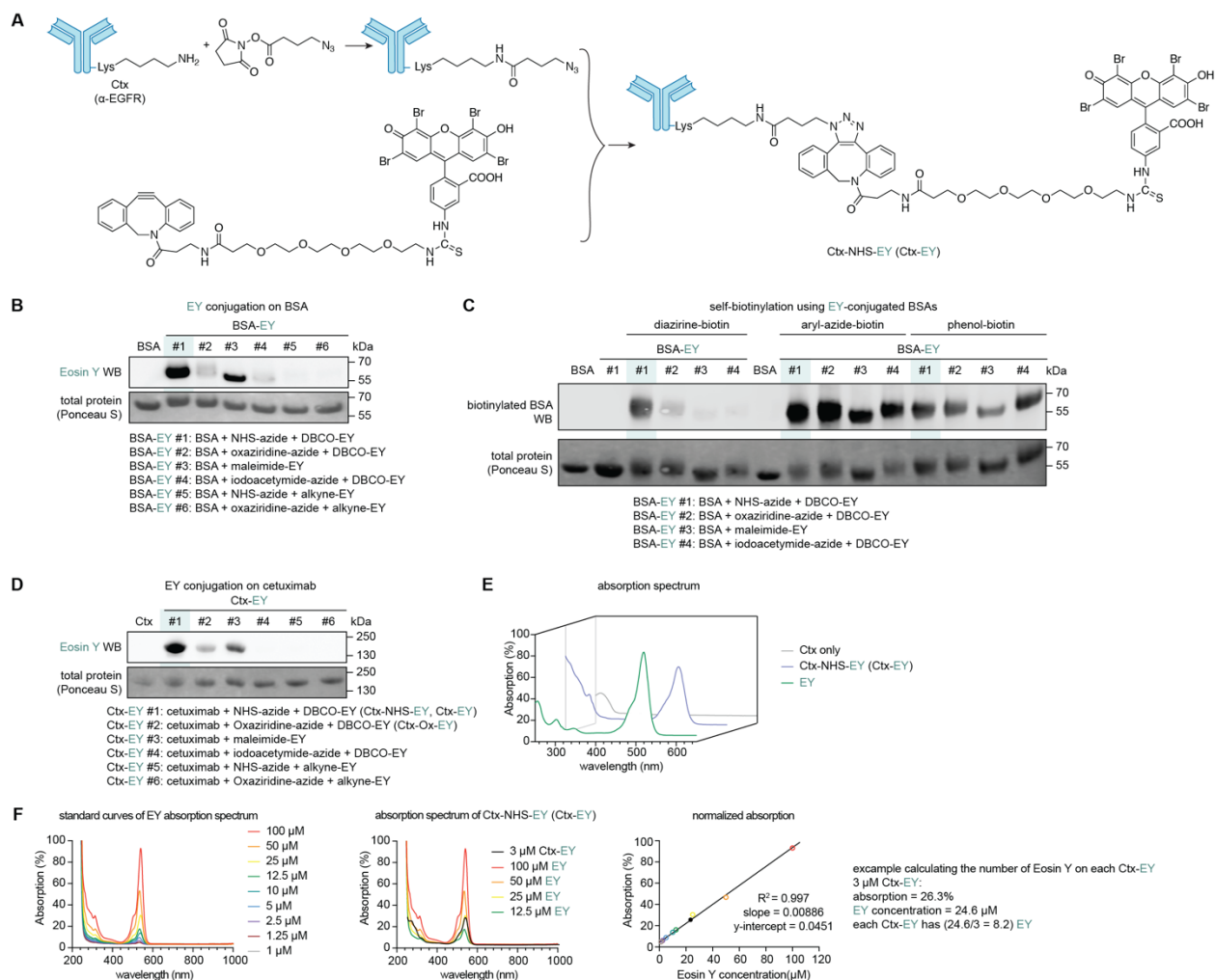


Fig. S3. Conjugation chemistry of EY onto BSA and Ctx for targeted protein biotinylation. (A) Lys-specific conjugation of EY using DBCO-PEG4-EY. **(B)** Screening of conjugation methods on BSA using different covalent warheads: NHS, oxaziridine, maleimide and iodoacetamide. The NHS-based amine coupling (highlighted in green) produced the highest extent of conjugation owed to presence of available Lys residues. **(C)** Evaluation of self-biotinylation using different EY-conjugated BSA constructs. Three photo-probes (diazirine-, aryl-azide- and phenol-biotin) were tested. **(D)** Screening of conjugation methods on Ctx with different residue-specific labeling methods mirrors results seen with BSA in Panel (B). **(E)** Overlaid absorption spectra of Ctx, Ctx- EY and EY. Conjugation of EY onto Ctx has minimal effect on the EY's absorption spectrum. **(F)** Calculation of conjugated EY stoichiometry using the photochemical property of EY as a dye. An example of Ctx-EY is shown, demonstrating that an average of eight EY molecules was conjugated per Ctx.

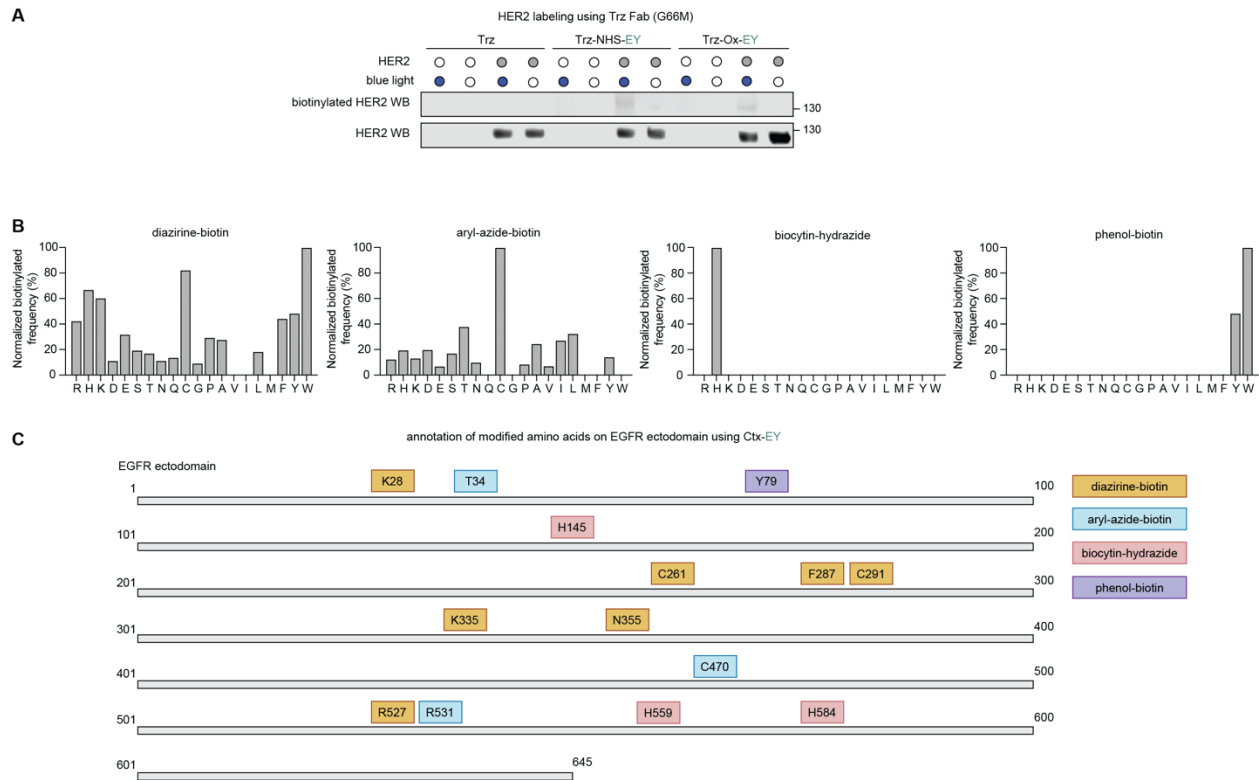


Fig. S4. Identification of specific sites and residues labeled with different biotin-photoprobes using EY-conjugated antibodies. (A) Selective labeling of the ecto-domain of HER2 using EY- conjugated Trz. A Trz Fab mutant (G68M) was conjugated using covalent warheads of NHS or oxaziridine, enabling targeted labeling of the purified HER2 ectodomain (amino acids 23 to 652) in vitro. **(B)** Statistics of amino acid labeling preference with four photo-probes from combined datasets of observed biotinylated peptides on BSA (45 peptides in total), Ctx and EGFR ECD (88 peptides in total). Spectrum assignments were shown in **Table S1-8**. **(C)** Identification of the EGFR ectodomain residues that were modified with different photo-probes. Amino acids labeled with diazirine-biotin (+ 616.25Da), aryl-azide-biotin (+ 620.23Da), phenol-biotin (+ 361.15Da) or biocytin-hydrazide (+ 384.50Da) are mapped on the sequence of EGFR or the crystal structure of EGFR ectodomain (PDB: 1YY9) shown in **Fig. 2E**. Spectrum assignments were shown in **Table S5-8**.

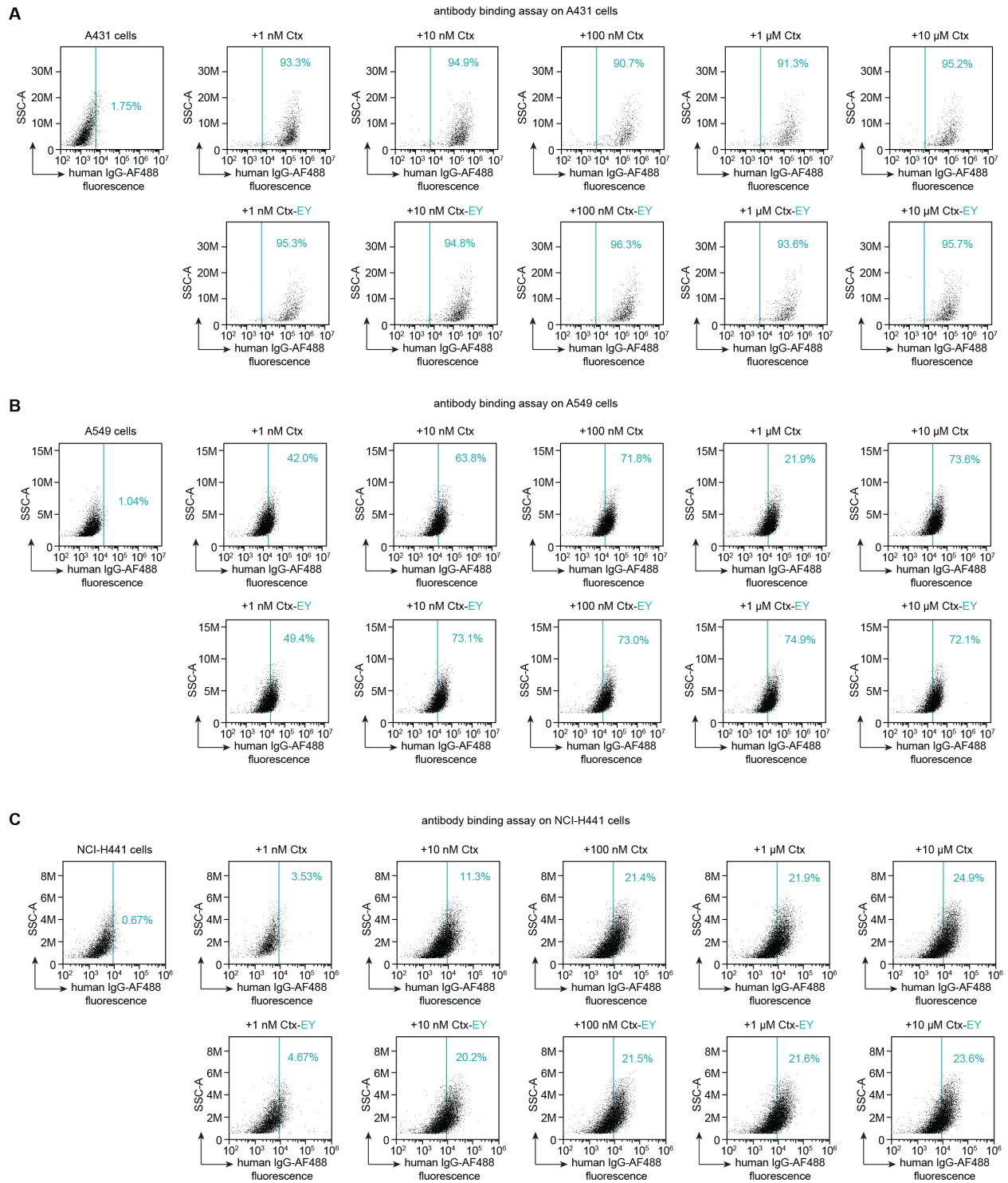


Fig. S5. Flow cytometry experiments with cells containing different levels of EGFR shows Ctx-EY binds to cells similarly to unconjugated Ctx. (A) Quantitative on-cell binding assay of Ctx and Ctx-EY on A431 cells with high EGFR expression levels (EGFR nTPM: 2978). (B) Quantitative on-cell binding assay of Ctx and Ctx-EY on A549 cells with low (Figure caption continued on next page)

(Figure caption continued from previous page) EGFR expression levels (EGFR nTPM: 59.7). (C) Quantitative on-cell binding assay of Ctx and Ctx-EY on NCI- H441 cells with very low EGFR expression levels (EGFR nTPM: 29.8).

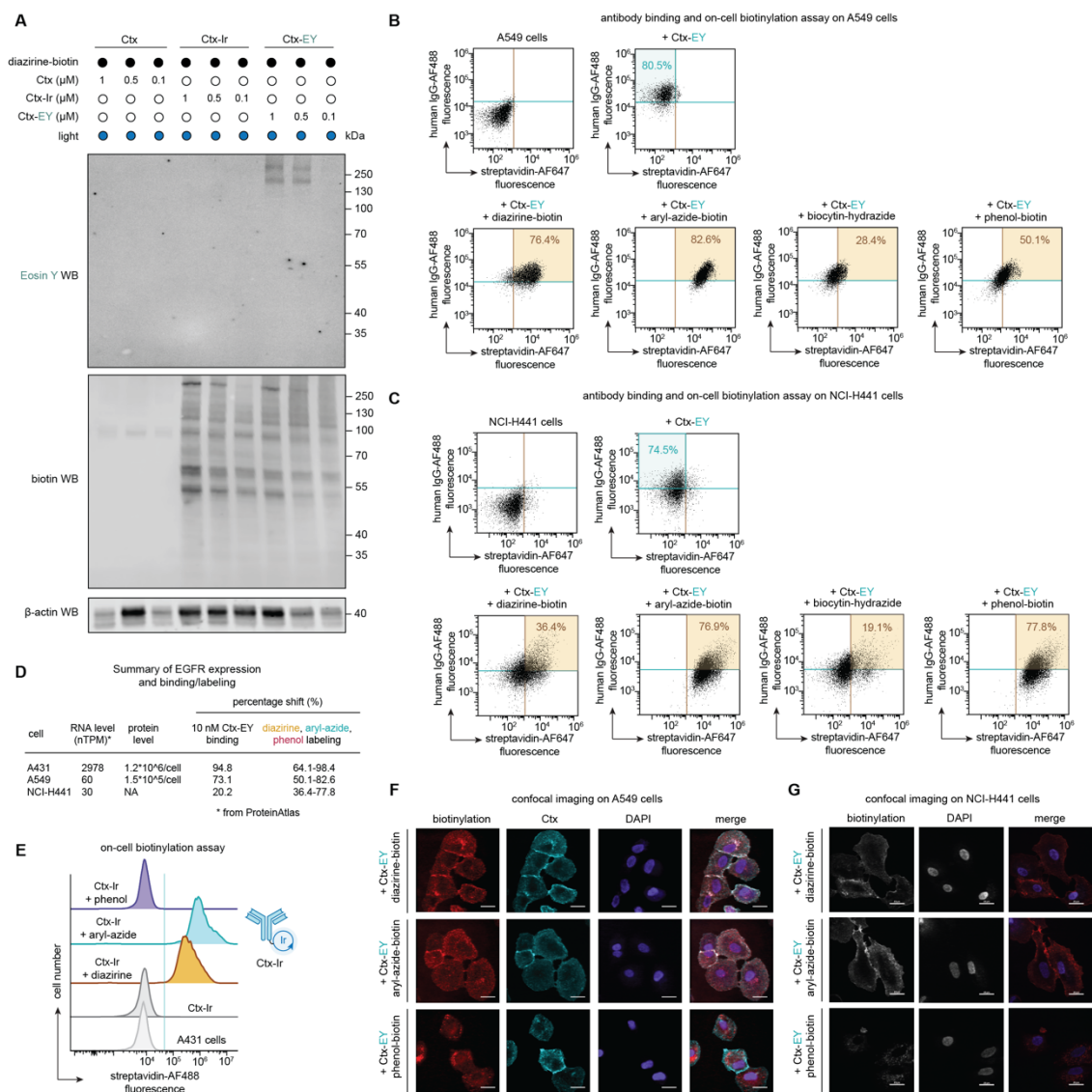


Fig. S6. Ctx-EY can activate cell-surface biotinylation on A431, A549 and NCI-H441 cells using different photo-probes. (A) On-cell biotinylation of A431 cells with diazirine-biotin from Fig. 3D confirmed by WB analysis. Dose-dependent labeling using Ctx-EY or Ctx-Ir are shown after 10 min blue LED illumination in the presence of 100 μM diazirine-biotin. **(B)** On-cell biotinylation with Ctx-EY on A549 cells expressing lower endogenous level of EGFR than A431 cells. **(C)** On-cell biotinylation with Ctx-EY on NCI-H441 cells expressing very low endogenous level of EGFR than either A549 or A431 cells. **(D)** Summary of Ctx-EY-mediated labeling and biotinylation results in cell lines with different expression level of EGFR. **(E)** Flow cytometry analysis of on-cell biotinylation of A431 cells using Ctx-Ir confirmed that the iridium catalyst can activate biotin-diazirine and aryl-azide-biotin, but not phenol-biotin. **(F)** Confocal microscopy imaging of cellular biotinylation and antibody binding with Ctx-EY on A549 cells expressing low endogenous level of EGFR. **(G)** Confocal microscopy imaging of cellular biotinylation with Ctx-EY on NCI-H441 cells expressing very low endogenous level of EGFR shows strong cell-surface labeling. Scale bar, 20 μm.

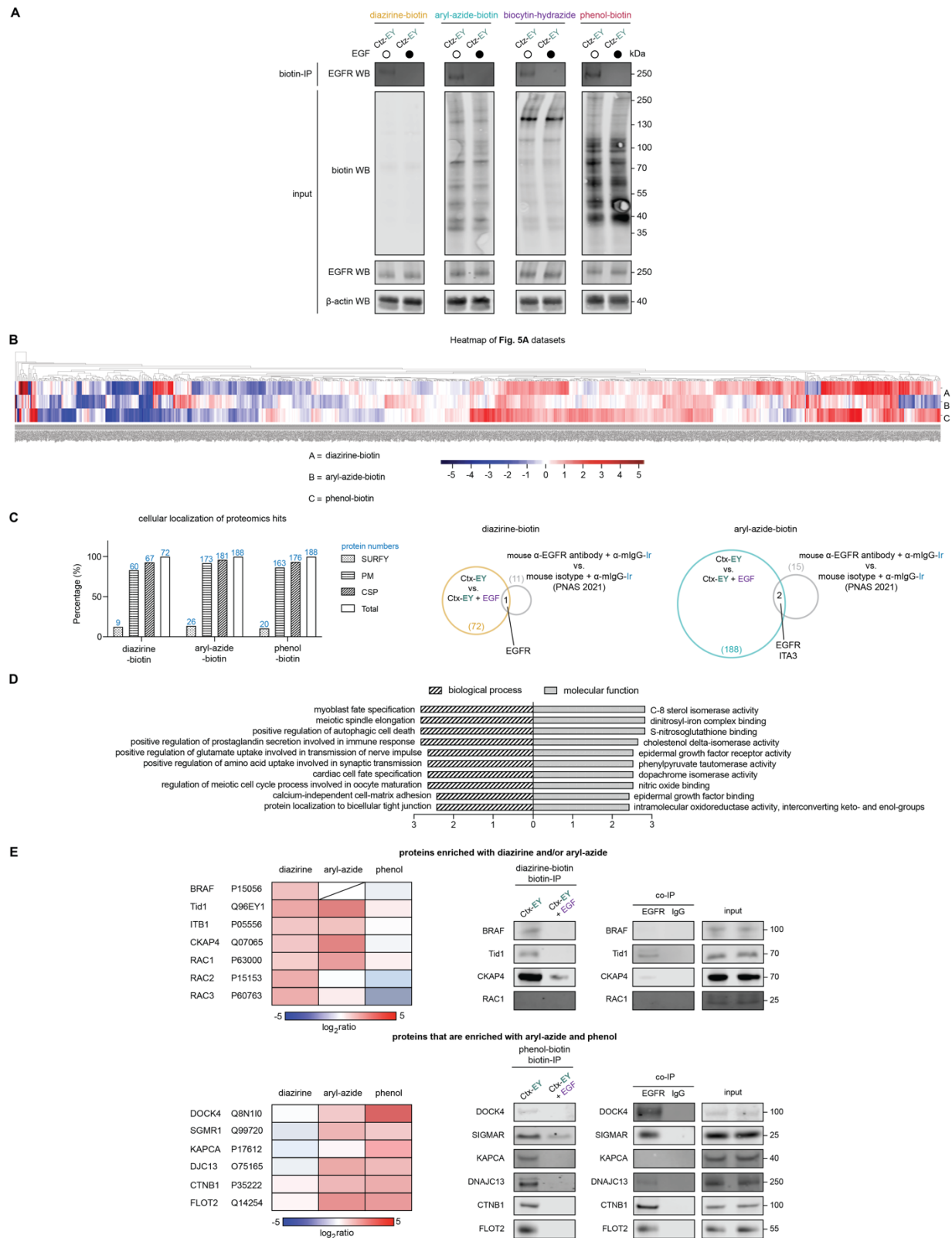


Fig. S8. EGFR interactome profiling using Ctx-EY on live A549 cells. (A) Full panel of WB analysis showing biotinylation of A549 cells using Ctx-EY showing it can activate biotinylation of A549 cells using biotin-diazirine, (Figure caption continued on next page)

(Figure caption continued from previous page) biotin-aryl-azide, biocytin-hydrazide and biotin-phenol. **(B)** Heatmap of the EGFR-interacting candidates identified in **Fig. 5A**. Significantly enriched proteins ($\log_2(\text{ratio}) \geq 1$, $p < 0.05$, unique peptide ≥ 2) are highlighted in red and listed in **Table S10-12**. **(C)** Categorized localization of protein hits discovered in **Fig. 5A** and direct comparison with published datasets using a tandem secondary antibody strategy. Analysis was performed as described in **fig. S7C**. **(D)** Gene Ontology (GO) analysis performed on the EGFR-interacting candidates in **Fig. 5A**. Analysis was performed as described in **fig. S7D**. Top 10 most significant biological process terms and molecular function terms were annotated with p-value. **(E)** Enrichment ratios and validation of protein hits from the diazirine-biotin, aryl-azide-biotin or phenol-biotin datasets.

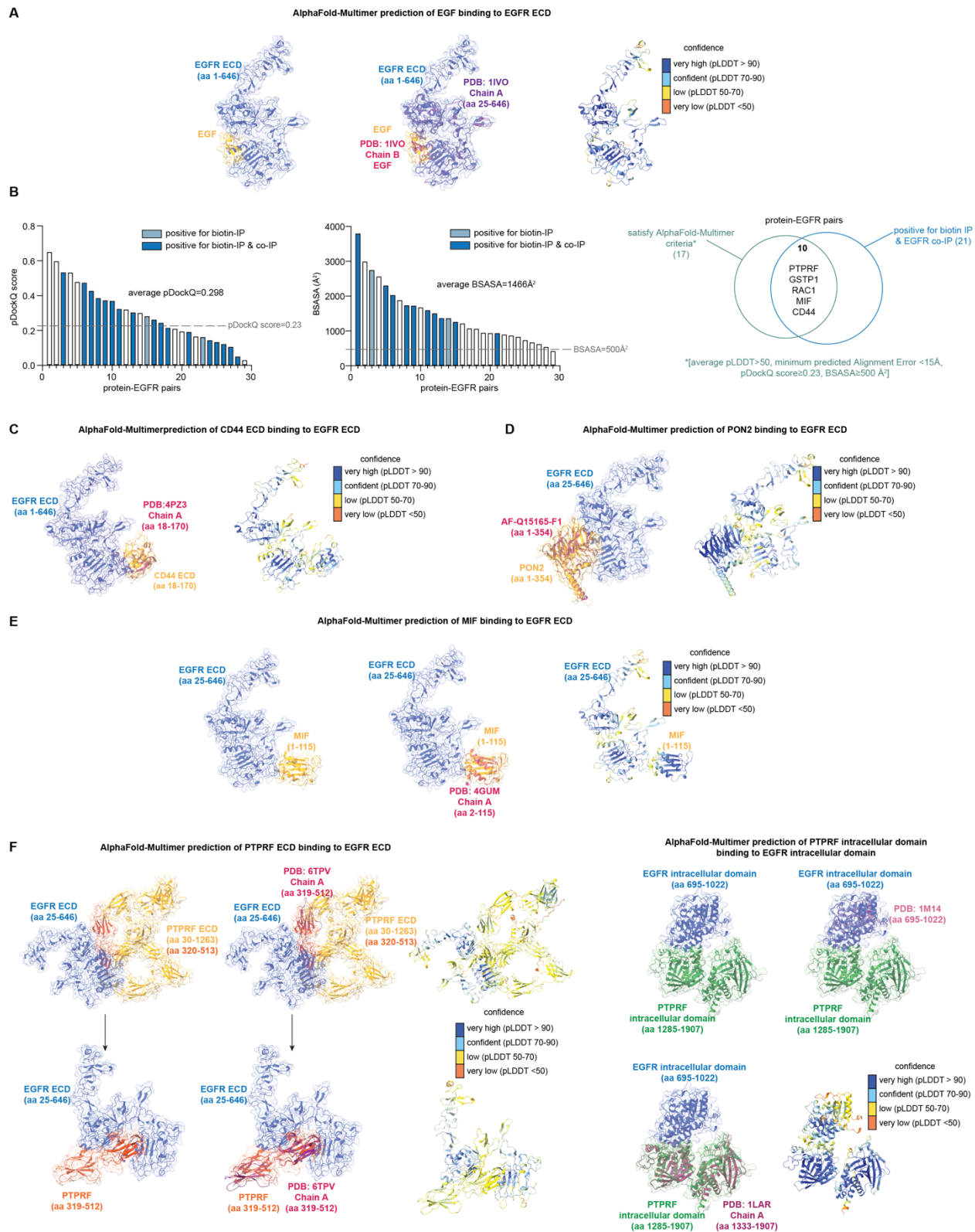


Fig. S9. AlphaFold-Multimer predictions of EGFR binary complexes with EGFR interactors from MultiMap datasets. (A) Predicted model of EGF bound to EGFR ECD via AlphaFold- Multimer compared with the crystal structure of (Figure caption continued on next page)

(Figure caption continued from previous page) EGF-bound EGFR (PDB: 1IVO, middle) and colored with pLDDT value (right). It is noteworthy that most residues were observed with high confidence (pLDDT>70). **(B)** Analysis of a selection of MultiMap hits using AlphaFold-Multimer and biochemical validation. An average pDockQ score of 0.298 and 1466Å² BSASA were observed for the 29 EGFR-protein pairs in the waterfall plots, both passing the criteria for highly confident heterodimeric AlphaFold-Multimer structures. **(C)** Predicted model of CD44 ECD bound to EGFR ECD via AlphaFold-Multimer shown in **Fig. 5F**. The predicted model was compared with the crystal structure of CD44 hyaluronan-binding domain in its ECD (PDB: 4PZ3, left) and colored with pLDDT value (right). **(D)** Predicted model of PON2 bound to EGFR ECD via AlphaFold-Multimer shown in **Fig. 5F**. Given that no crystal structures were available, the predicted structure was compared to the AlphaFold monomer prediction (AF-Q15165-F1, left) and colored with pLDDT value (right). **(E)** Predicted structure of MIF bound to the EGFR ECD via AlphaFold-Multimer. The predicted structure was compared with the crystal structure of MIF (PDB: 4GUM, middle) and colored with pLDDT value (right). **(F)** Predicted models of PTPRF ECD bound to EGFR ECD (left) and the PTPRF intracellular domain bound to the EGFR intracellular domain (right) via AlphaFold-Multimer shown in **Fig. 5F**. Predicted binding surface of PTPRF ECD (amino acids 319 to 512) was highlighted in orange-yellow and compared with the crystal structure of PTPRF ECD at the fibronectin type-III domain (PDB:6TPV). Both full PTPRF ECD and highlighted binding surface of PTPRF were colored with pLDDT value. Similarly, the predicted structure of the PTPRF intracellular domain bound to the EGFR intracellular domain was compared with the crystal structures of EGFR intracellular domain (PDB: 1M14) and PTPRF intracellular phosphatase domain (PDB: 1LAR), and colored with pLDDT value.

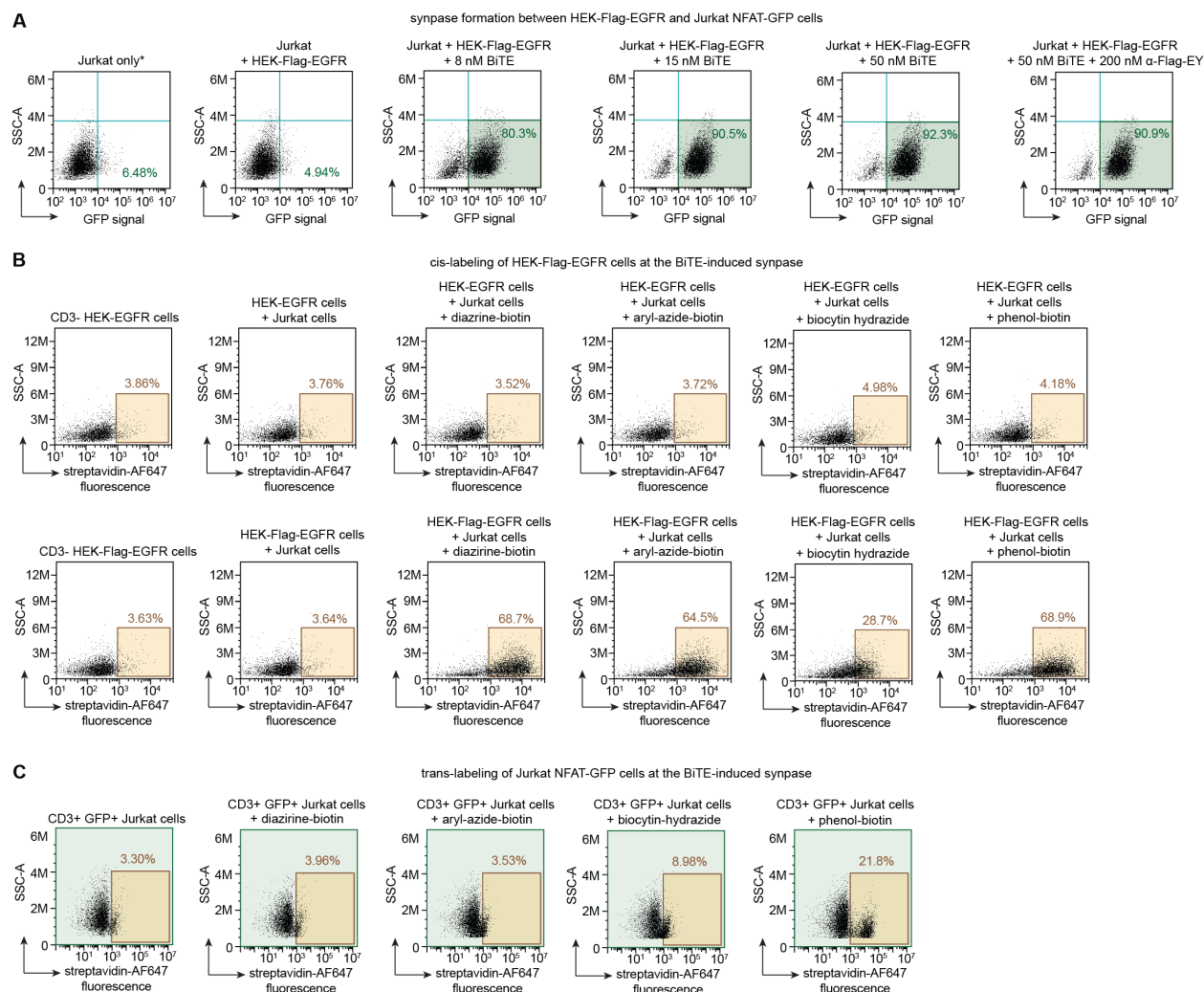


Fig. S11. Targeted labeling of BiTE-induced synapses between Jurkat NFAT-GFP and HEK-Flag-EGFR using α -Flag-EY. (A) Cell-cell engagement between Jurkat NFAT-GFP and HEK-Flag-EGFR induced by a bispecific T cell engager (BiTE) in a dose-dependent manner in Fig. 6A. Cell synapse formation was monitored by NFAT-GFP activation in Jurkat cells. Percentage of Jurkat cells engaging HEK293T-EGFR reached >90% and was not affected by the addition of EY-conjugated M1 α -Flag antibody (α -Flag-EY). (B) Cis-labeling of HEK-Flag-EGFR using α -Flag-EY and four different photo-probes. HEK293T cells transfected with untagged EGFR (HEK-EGFR) and Flag-tagged EGFR (HEK-Flag-EGFR) were compared in parallel. Significant labeling with all four photo-probes in HEK-Flag-EGFR. (C) Trans-labeling of Jurkat cells using α -Flag-EY and four different photo-probes. CD3⁺ Jurkat cells were gated for GFP⁺ cells before the biotinylation levels were quantified. Only biotin-phenol provided significant trans-labeling shift as shown in Fig. 6A.

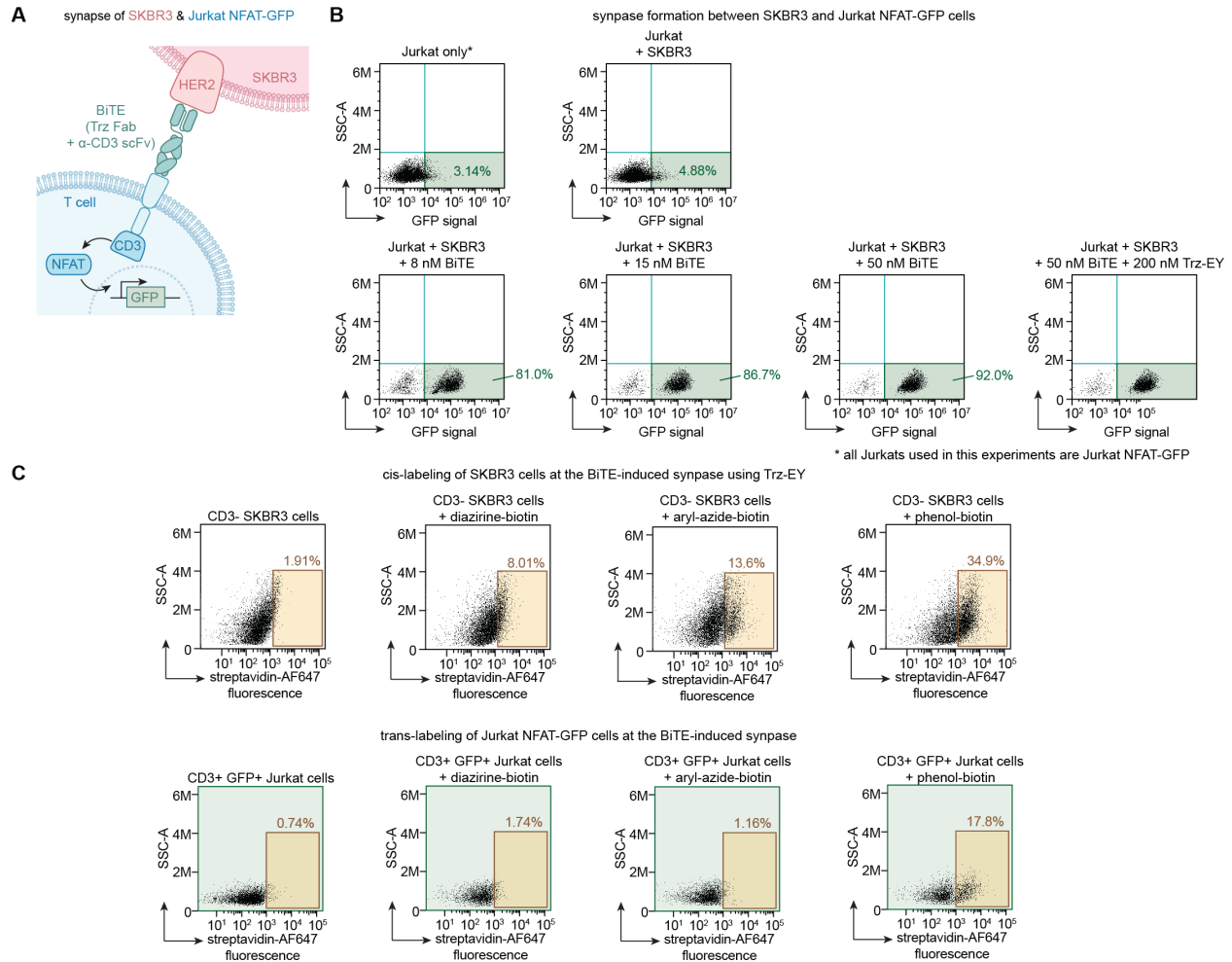


Fig. S12. Extended applications of targeted BiTE-induced cell synapse labeling. (A) Scheme of on-cell labeling of Jurkat NFAT-GFP and SKBR3 induced by a BiTE that recognizes HER2. (B) Cell-cell engagement between Jurkat NFAT-GFP and SKBR3 cells was monitored by NFAT- GFP activation. Percentage of Jurkat cells engaging SKBR3 reached >90% and the synapse was not affected by the addition of the EY-conjugated Trz (Trz-EY). (C) Cis- and trans-labeling SKBR3 and Jurkat NFAT-GFP between a BiTE-induced synapse using α -HER2-EY and three different photo-probes.

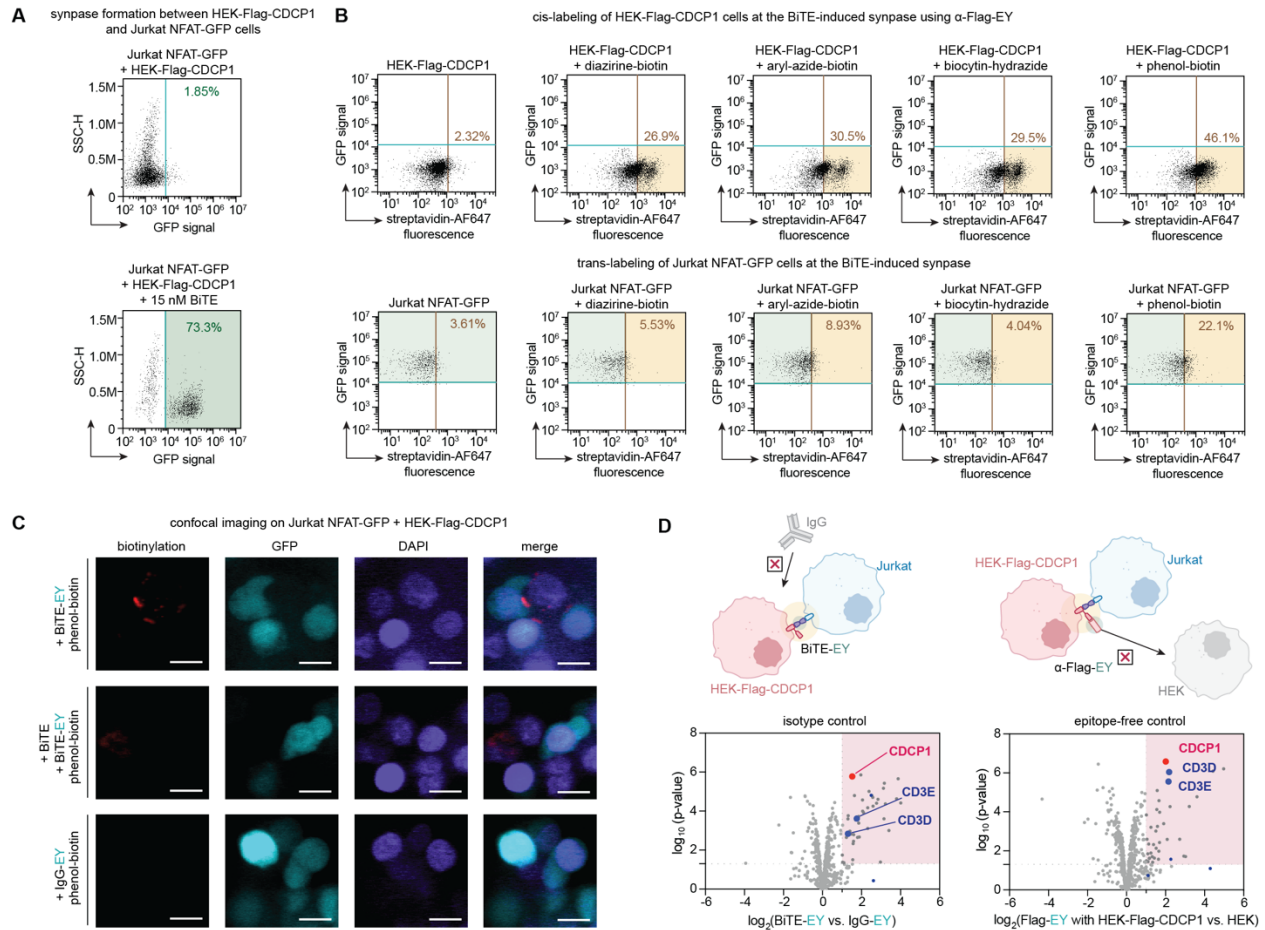


Fig. S13. Targeted labeling of BiTE-induced synapses between Jurkat NFAT-GFP and HEK-Flag-CDCP1 using α -Flag-EY and BiTE-EY. (A) Cell-cell engagement between HEK-Flag-CDCP1 and Jurkat NFAT-GFP monitored by NFAT-GFP activation. (B) Cis- and trans-labeling of HEK-Flag-CDCP1 and Jurkat NFAT-GFP between a BiTE-induced synapse using α -Flag-EY. (C) Confocal microscopy imaging of BiTE-EY-mediated biotinylation confirmed that phenol-biotin photo-probe labeling was primarily confined to cell-cell synapses. Scale bar, 10 μ m. (D) Proteins biotinylated on unsorted HEK-Flag-CDCP1 (cis-labeling) and Jurkat NFAT-GFP (trans-labeling) using phenol-biotin compared with isotype and epitope-free controls (n=3). Significantly enriched proteins ($\log_2(\text{ratio}) \geq 1$, $p < 0.05$, unique peptide ≥ 2) are shaded in red. CDCP1 from HEK-Flag-CDCP1, CD3 components from Jurkat NFAT-GFP and their associated proteins from STRING analysis are highlighted in red and blue, respectively. Full protein lists are shown in **Table S15-S16**.

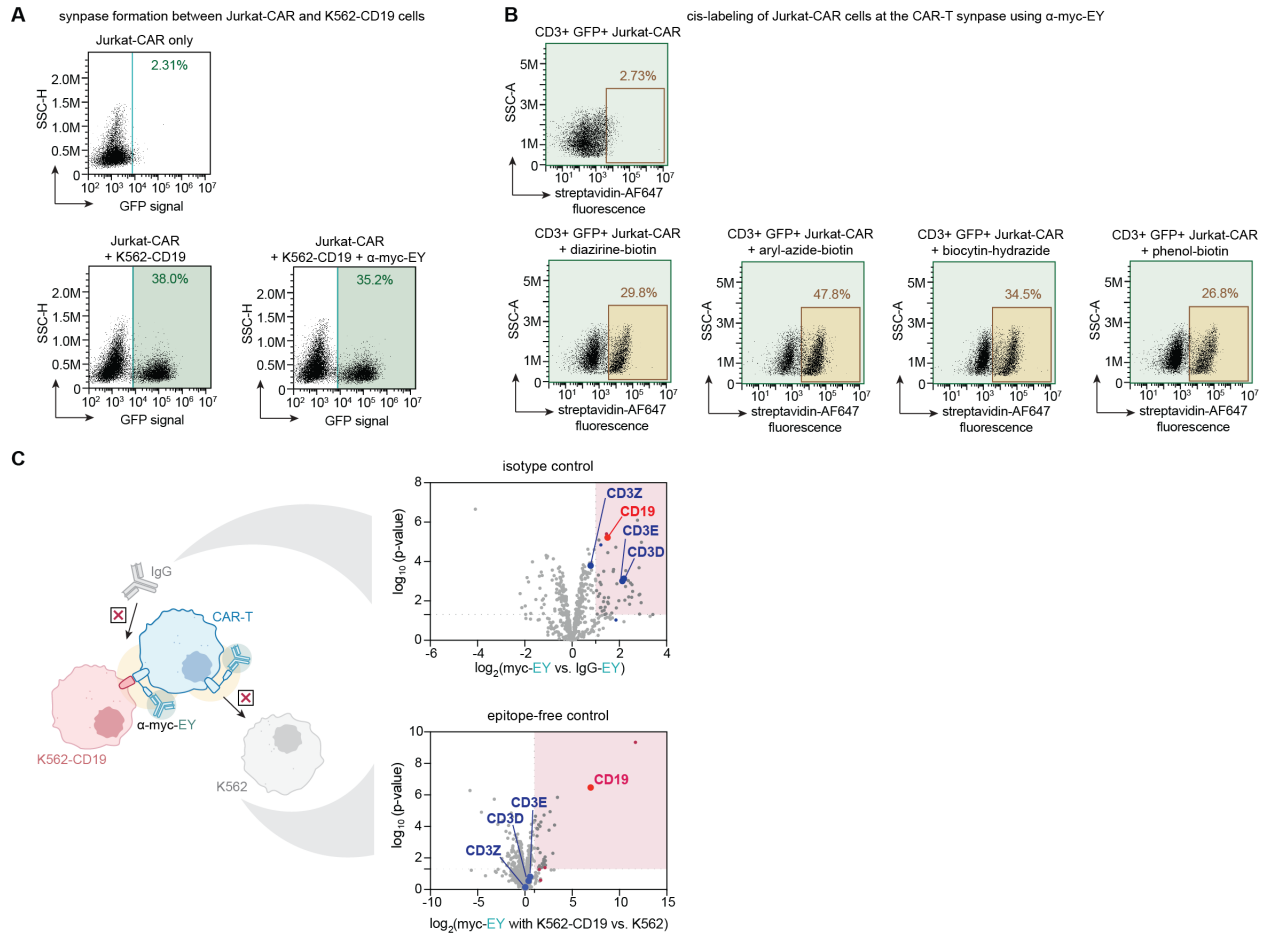


Fig. S14. Targeted labeling of CAR-induced synapses between Jurkat-CAR and K562-CD19 using α -Myc-EY. (A) Cell-cell engagement between Jurkat-CAR with K562-CD19 monitored by NFAT-GFP activation. (B) Cis-labeling of Jurkat-CAR with K562-CD19 shown in Fig. 6D. (C) Proteins biotinylated at the synapses of Jurkat-CAR (cis-labeling) and K562-CD19 (trans-labeling) using phenol-biotin compared with isotype and epitope-free controls (n=3). Significantly enriched proteins ($\log_2(\text{ratio}) \geq 1$, $p < 0.05$, unique peptide ≥ 2) are shaded in red. CD19 from K562-CD19, CAR components from Jurkat-CAR and their associated proteins from STRING analysis are highlighted in red and blue, respectively. Full protein lists are shown in Table S18-19.

3.9 References

1. M. Uhlén et al., The human secretome. *Sci. Signal.* 12, eaaz0274 (2019). doi: 10.1126/scisignal.aaz0274; pmid: 31772123
2. D. H. Siepe et al., Identification of orphan ligand-receptor relationships using a cell-based CRISPRa enrichment screening platform. *eLife* 11, e81398 (2022). doi: 10.7554/eLife.81398; pmid: 36178190
3. C. C. Wu, J. R. Yates 3rd, The application of mass spectrometry to membrane proteomics. *Nat. Biotechnol.* 21, 262–267 (2003). doi: 10.1038/nbt0303-262; pmid: 12610573
4. N. P. Barrera, C. V. Robinson, Advances in the mass spectrometry of membrane proteins: From individual proteins to intact complexes. *Annu. Rev. Biochem.* 80, 247–271 (2011).doi: 10.1146/annurev-biochem-062309-093307; pmid: 21548785
5. W. Qin, K. F. Cho, P. E. Cavanagh, A. Y. Ting, Deciphering molecular interactions by proximity labeling. *Nat. Methods* 18, 133–143 (2021). doi: 10.1038/s41592-020-01010-5; pmid: 33432242
6. C. P. Seath, A. D. Trowbridge, T. W. Muir, D. W. C. MacMillan, Reactive intermediates for interactome mapping. *Chem. Soc. Rev.* 50, 2911–2926 (2021). doi: 10.1039/D0CS01366H; pmid: 33458734
7. S. S. Lam et al., Directed evolution of APEX2 for electron microscopy and proximity labeling. *Nat. Methods* 12, 51–54 (2015). doi: 10.1038/nmeth.3179; pmid: 25419960
8. K. J. Roux, D. I. Kim, B. Burke, D. G. May, BioID: A Screen for Protein-Protein Interactions. *Curr. Protoc. Protein Sci.* 91, 23.1, 15 (2018). doi: 10.1002/cpps.51; pmid: 29516480
9. T. C. Branon et al., Efficient proximity labeling in living cells and organisms with TurboID. *Nat. Biotechnol.* 36, 880–887 (2018). doi: 10.1038/nbt.4201; pmid: 30125270
10. W. Qin et al., Dynamic mapping of proteome trafficking within and between living cells by TransitID. *Cell* 186, 3307–3324.e30 (2023). doi: 10.1016/j.cell.2023.05.044; pmid: 37385249
11. J. V. Oakley et al., Radius measurement via super-resolution microscopy enables the development of a variable radii proximity labeling platform. *Proc. Natl. Acad. Sci. U.S.A.* 119, e2203027119 (2022). doi: 10.1073/pnas.2203027119; pmid: 35914173

12. M. Müller et al., Light-mediated discovery of surfaceome nanoscale organization and intercellular receptor interaction networks. *Nat. Commun.* 12, 7036 (2021). doi: 10.1038/s41467-021-27280-x; pmid: 34857745
13. M. K. Kuimova, G. Yahioglu, P. R. Ogilby, Singlet oxygen in a cell: Spatially dependent lifetimes and quenching rate constants. *J. Am. Chem. Soc.* 131, 332–340 (2009). doi: 10.1021/ja807484b; pmid: 19128181
14. M. Lindén, P. Sens, R. Phillips, Entropic tension in crowded membranes. *PLOS Comput. Biol.* 8, e1002431 (2012). doi: 10.1371/journal.pcbi.1002431; pmid: 22438801
15. J. B. Geri et al., Microenvironment mapping via Dexter energy transfer on immune cells. *Science* 367, 1091–1097 (2020). doi: 10.1126/science.aay4106; pmid: 32139536
16. T. J. Bechtel, T. Reyes-Robles, O. O. Fadeyi, R. C. Oslund, Strategies for monitoring cell-cell interactions. *Nat. Chem. Biol.* 17, 641–652 (2021). doi: 10.1038/s41589-021-00790-x; pmid: 34035514
17. R. C. Oslund et al., Detection of cell-cell interactions via photocatalytic cell tagging. *Nat. Chem. Biol.* 18, 850–858 (2022). doi: 10.1038/s41589-022-01044-0; pmid: 35654846
18. D. C. Cabanero et al., Photocatalytic activation of aryl (trifluoromethyl) diazos to carbenes for high-resolution protein labeling with red light. *J. Am. Chem. Soc.* 146, 1337–1345 (2024). doi: 10.1021/jacs.3c09545; pmid: 38165744
19. B. A. G. DeGraff, D. W. Gillespie, R. J. Sundberg, Phenyl nitrene: A flash photolytic investigation of the reaction with secondary amines. *J. Am. Chem. Soc.* 95, 7491–7496 (1973).
20. T. G. Bartholow et al., Photoproximity labeling from single catalyst sites allows calibration and increased resolution for carbene labeling of protein partners in vitro and on cells. *ACS Cent. Sci.* 10, 199–208 (2023). doi: 10.1021/acscentsci.3c01473; pmid: 38292613
21. B. F. Buksh et al., mMap-Red: Proximity labeling by red light photocatalysis. *J. Am. Chem. Soc.* 144, 6154–6162 (2022). doi: 10.1021/jacs.2c01384; pmid: 35363468
22. N. E. S. Tay et al., Targeted activation in localized protein environments via deep red photoredox catalysis. *Nat. Chem.* 15, 101–109 (2023). pmid: 36216892

23. T. Reyes-Robles et al., Nanoscale mapping of EGFR and c-MET protein environments on lung cancer cell surfaces via therapeutic antibody photocatalyst conjugates. *ACS Chem. Biol.* 17, 2304–2314 (2022). doi: 10.1021/acscchembio.2c00409; pmid: 35939534
24. T. J. Bechtel et al., Proteomic mapping of intercellular synaptic environments via flavin-dependent photoredox catalysis. *Org. Biomol. Chem.* 21, 98–106 (2022). doi: 10.1039/D2OB02103J; pmid: 36477737
25. R. Evans et al., Protein complex prediction with AlphaFold- Multimer. *bioRxiv* 463034 [Preprint] (2022); [https://doi.org/ 10.1101/2021.10.04.463034](https://doi.org/10.1101/2021.10.04.463034).
26. A. J. Slezak et al., Tumor cell-surface binding of immune stimulating polymeric glyco-adjuvant via cysteine-reactive pyridyl disulfide promotes antitumor immunity. *ACS Cent. Sci.* 8, 1435–1446 (2022). doi: 10.1021/acscentsci.2c00704; pmid: 36313164
27. M. Majek, A. Jacobi von Wangelin, Mechanistic perspectives on organic photoredox catalysis for aromatic substitutions. *Acc. Chem. Res.* 49, 2316–2327 (2016). doi: 10.1021/acs.accounts.6b00293; pmid: 27669097
28. N. E. S. Tay et al., Targeted activation in localized protein environments via deep red photoredox catalysis. *Nat. Chem.* 15, 101–109 (2023). doi: 10.1038/s41557-022-01057-1; pmid: 36216892
29. Z. Huang et al., Bioorthogonal photocatalytic decaging-enabled mitochondrial proteomics. *J. Am. Chem. Soc.* 143, 18714–18720 (2021). doi: 10.1021/jacs.1c09171; pmid: 34709827
30. Y. Fang, P. Zou, Photocatalytic proximity labeling for profiling the subcellular organization of biomolecules. *ChemBioChem* 24, e202200745 (2023). doi: 10.1002/cbic.202200745; pmid: 36762434
31. K. Toh et al., Chemoproteomic identification of blue-light-damaged proteins. *J. Am. Chem. Soc.* 144, 20171–20176 (2022). doi: 10.1021/jacs.2c07180; pmid: 36306265
32. S. Lin et al., Redox-based reagents for chemoselective methionine bioconjugation. *Science* 355, 597–602 (2017). doi: 10.1126/science.aal3316; pmid: 28183972
33. S. Li et al., Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. *Cancer Cell* 7, 301–311 (2005). doi: 10.1016/j.ccr.2005.03.003; pmid: 15837620

34. K. Yonesaka et al., Activation of ERBB2 signaling causes resistance to the EGFR-directed therapeutic antibody cetuximab. *Sci. Transl. Med.* 3, 99ra86 (2011). doi: 10.1126/scitranslmed.3002442; pmid: 21900593
35. S. K. Elledge et al., Systematic identification of engineered methionines and oxaziridines for efficient, stable, and site-specific antibody bioconjugation. *Proc. Natl. Acad. Sci. U.S.A.* 117, 5733–5740 (2020). doi: 10.1073/pnas.1920561117; pmid: 32123103
36. D. I. Kim, K. J. Roux, Filling the Void: Proximity-based labeling of proteins in living cells. *Trends Cell Biol.* 26, 804–817 (2016). doi: 10.1016/j.tcb.2016.09.004; pmid: 27667171
37. A. V. West et al., Labeling preferences of diazirines with protein biomolecules. *J. Am. Chem. Soc.* 143, 6691–6700 (2021). doi: 10.1021/jacs.1c02509; pmid: 33876925
38. H. Wang et al., Selective mitochondrial protein labeling enabled by biocompatible photocatalytic reactions inside live cells. *JACS Au* 1, 1066–1075 (2021). doi: 10.1021/jacsau.1c00172; pmid: 34467350
39. H. Kimura et al., Antibody-dependent cellular cytotoxicity of cetuximab against tumor cells with wild-type or mutant epidermal growth factor receptor. *Cancer Sci.* 98, 1275–1280 (2007). doi: 10.1111/j.1349-7006.2007.00510.x; pmid: 17498200
40. H. Masui, L. Castro, J. Mendelsohn, Consumption of EGF by A431 cells: Evidence for receptor recycling. *J. Cell Biol.* 120, 85–93 (1993). doi: 10.1083/jcb.120.1.85; pmid: 8416997
41. J. R. Wiśniewski, M. Y. Hein, J. Cox, M. Mann, A “proteomic ruler” for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteomics* 13, 3497–3506 (2014). doi: 10.1074/mcp.M113.037309; pmid: 25225357
42. Z. Yao et al., A global analysis of the receptor tyrosine kinase-protein phosphatase interactome. *Mol. Cell* 65, 347–360 (2017). doi: 10.1016/j.molcel.2016.12.004; pmid: 28065597
43. S. Foerster et al., Characterization of the EGFR interactome reveals associated protein complex networks and intracellular receptor dynamics. *Proteomics* 13, 3131–3144 (2013). doi: 10.1002/pmic.201300154; pmid: 23956138

44. J. Li et al., Perturbation of the mutated EGFR interactome identifies vulnerabilities and resistance mechanisms. *Mol. Syst. Biol.* 9, 705 (2013). doi: 10.1038/msb.2013.61; pmid: 24189400
45. V. Morello et al., b1 integrin controls EGFR signaling and tumorigenic properties of lung cancer cells. *Oncogene* 30, 4087–4096 (2011). doi: 10.1038/onc.2011.107; pmid: 21478906
46. M. Petrás et al., Molecular interactions of ErbB1 (EGFR) and integrin-b1 in astrocytoma frozen sections predict clinical outcome and correlate with Akt-mediated in vitro radioresistance. *Neuro-oncol.* 15, 1027–1040 (2013). doi: 10.1093/neuonc/not046; pmid: 23595626
47. Y. Zheng et al., Secreted and O-GlcNAcylated MIF binds to the human EGF receptor and inhibits its activation. *Nat. Cell Biol.* 17, 1348–1355 (2015). doi: 10.1038/ncb3222; pmid: 26280537
48. T. Okamura et al., Tyrosine phosphorylation of the human glutathione S-transferase P1 by epidermal growth factor receptor. *J. Biol. Chem.* 284, 16979–16989 (2009). doi: 10.1074/jbc.M808153200; pmid: 19254954
49. T. Kaihara et al., Redifferentiation and ZO-1 reexpression in liver-metastasized colorectal cancer: Possible association with epidermal growth factor receptor-induced tyrosine phosphorylation of ZO-1. *Cancer Sci.* 94, 166–172 (2003). doi: 10.1111/j.1349-7006.2003.tb01414.x; pmid: 12708492
50. S. Meran et al., Hyaluronan facilitates transforming growth factor-b1-dependent proliferation via CD44 and epidermal growth factor receptor interaction. *J. Biol. Chem.* 286, 17618–17630 (2011). doi: 10.1074/jbc.M111.226563; pmid: 21454519
51. G. D. Grass, L. B. Tolliver, M. Bratoeva, B. P. Toole, CD147, CD44, and the epidermal growth factor receptor (EGFR) signaling pathway cooperate to regulate breast epithelial cell invasiveness. *J. Biol. Chem.* 288, 26089–26104 (2013).doi: 10.1074/jbc.M113.497685; pmid: 23888049
52. J. Merlin et al., Galectin-3 regulates MUC1 and EGFR cellular distribution and EGFR downstream pathways in pancreatic cancer cells. *Oncogene* 30, 2514–2525 (2011). doi: 10.1038/ onc.2010.631; pmid: 21258405

53. M. Ferrandi et al., Adducin- and ouabain-related gene variants predict the antihypertensive activity of rostaduroxin, part 1: Experimental studies. *Sci. Transl. Med.* 2, 59ra86 (2010). doi: 10.1126/scitranslmed.3001815; pmid: 21106940
54. E. C. Stites, The response of cancers to BRAF inhibition underscores the importance of cancer systems biology. *Sci. Signal.* 5, pe46 (2012). doi: 10.1126/scisignal.2003354; pmid: 23074264
55. C. Y. Chen et al., Tid1-L inhibits EGFR signaling in lung adenocarcinoma by enhancing EGFR Ubiquitinylation and degradation. *Cancer Res.* 73, 4009–4019 (2013). doi: 10.1158/0008-5472.CAN-12-4066; pmid: 23698466
56. Y. Lim et al., In silico protein interaction screening uncovers DONSON's role in replication initiation. *Science* 381, eadi3448 (2023). doi: 10.1126/science.adi3448; pmid: 37590370
57. X. Gu et al., The midnolin-proteasome pathway catches proteins for ubiquitination-independent degradation. *Science* 381, eadh5021 (2023). doi: 10.1126/science.adh5021; pmid: 37616343
58. P. Bryant, G. Pozzati, A. Elofsson, Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* 13, 1265 (2022). doi: 10.1038/s41467-022-28865-w; pmid: 35273146
59. A. Shrake, J. A. Rupley, Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* 79, 351–371 (1973). doi: 10.1016/0022-2836(73)90011-9; pmid: 4760134
60. F. P. Davis, A. Sali, PIBASE: A comprehensive database of structurally defined protein interfaces. *Bioinformatics* 21, 1901–1907 (2005). doi: 10.1093/bioinformatics/bti277; pmid: 15657096
61. H. Ogiso et al., Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. *Cell* 110, 775–787 (2002). doi: 10.1016/S0092-8674(02) 00963-7; pmid: 12297050
62. E. N. Banhos Danneskiold-Sams et al., Rapid and accurate deorphanization of ligand-receptor pairs using AlphaFold. *bioRxiv* 531341 [Preprint] (2023); <https://doi.org/10.1101/2023.03.16.531341>.
63. S. A. Lim et al., Targeting a proteolytic neoepitope on CUB domain containing protein 1 (CDCP1) for RAS-driven cancers. *J. Clin. Invest.* 132, e154604 (2022). doi: 10.1172/JCI154604; pmid: 35166238

64. F. Macian, NFAT proteins: Key regulators of T-cell development and function. *Nat. Rev. Immunol.* 5, 472–484 (2005). doi: 10.1038/nri1632; pmid: 15928679
65. J. S. Klein et al., Examination of the contributions of size and avidity to the neutralization mechanisms of the anti-HIV antibodies b12 and 4E10. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7385–7390 (2009). doi: 10.1073/pnas.0811427106; pmid: 19372381
66. Q. Xiao et al., Size-dependent activation of CAR-T cells. *Sci. Immunol.* 7, eabl3995 (2022). doi: 10.1126/sciimmunol.abl3995; pmid: 35930653
67. L. D. Lavis, Teaching old dyes new tricks: Biological probes built from fluoresceins and rhodamines. *Annu. Rev. Biochem.* 86, 825–843 (2017). doi: 10.1146/annurev-biochem-061516-044839; pmid: 28399656
68. D. P. Hari, B. König, Synthetic applications of eosin Y in photoredox catalysis. *Chem. Commun.* 50, 6688–6699 (2014). doi: 10.1039/C4CC00751D; pmid: 24699920
69. N. H. Cho et al., OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science* 375, eabi6983 (2022). doi: 10.1126/science.abi6983; pmid: 35271311
70. J. R. Klesmith et al., Retargeting CD19 chimeric antigen receptor T cells via engineered CD19-fusion proteins. *Mol. Pharm.* 16, 3544–3558 (2019). doi:10.1021/acs.molpharmaceut.9b00418; pmid: 31242389
71. Y. Ge et al., Target protein deglycosylation in living cells by a nanobody-fused split O-GlcNAcase. *Nat. Chem. Biol.* 17, 593–600 (2021). doi: 10.1038/s41589-021-00757-y; pmid: 3368629
72. Y. Li et al., Rapid enzyme-mediated biotinylation for cell surface proteome profiling. *Anal. Chem.* 93, 4542–4551 (2021). doi: 10.1021/acs.analchem.0c04970; pmid: 33660993
73. D. Bausch-Fluck et al., The in silico human surfaceome. *Proc. Natl. Acad. Sci. U.S.A.* 115, E10988–E10997 (2018). doi: 10.1073/pnas.1808790115; pmid: 30373828
74. Z. C. Drake, J. T. Seffernick, S. Lindert, Protein complex prediction using Rosetta, AlphaFold, and mass spectrometry covalent labeling. *Nat. Commun.* 13, 7846 (2022). doi: 10.1038/s41467-022-35593-8; pmid: 36543826

75. J. Jumper et al., Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). doi: 10.1038/s41586-021-03819-2; pmid: 34265844
76. M. Mirdita et al., ColabFold: Making protein folding accessible to all. *Nat. Methods* 19, 679–682 (2022). doi: 10.1038/s41592-022-01488-1; pmid: 35637307
77. P. Kunzmann, K. Hamacher, Biotite: A unifying open source computational biology framework in Python. *BMC Bioinformatics* 19, 346 (2018). doi: 10.1186/s12859-018-2367-z; pmid: 30285630
78. Z. Lin et al., Multiscale photocatalytic proximity labeling reveals cell surface neighbors on and between cells. *Dryad* (2024). <https://doi.org/10.5061/dryad.j6q573nmq>.

Chapter 4

Conclusion

This dissertation deals with the development and application of mass spectrometry data in combination with protein structure to model complex systems of interacting protein molecules. In chapter 2 we develop a statistical inference framework for modeling direct protein interaction networks. In chapter 3 we apply AlphaFold-Multimer virtual screening to orthogonally validate candidate epidermal growth factor receptor (EGFR) molecular neighbors. Now, I make suggestions for improving the INM method developed in chapter 2 with an emphasis on the types of data used in chapter 3.

4.1 Improvements to integrative network modeling

Recommendations for the development of an AlphaFold-Multimer term

Large scale virtual screening of protein interactions is now common-place using models such as AlphaFold-Multimer, ColabFold, and RoseTTAFold2 (Evans et al. 2021; Mirdita et al. 2022; Baek et al. 2023). Bacterial, yeast, and more recently human, proteomes have been screened for protein-protein interactions (Humphreys et al. 2021; Zhang et al. 2024; Humphreys et al. 2024). Such methods rely on co-evolutionary signatures obtained from the genomic sequences of many organisms. They may additionally rely on structural information obtained from the Protein Data Bank. As such, the predicted protein-protein interactions (PPI) may not be representative of the interactions under specific experimental conditions or disease contexts. Nonetheless, such information will likely be highly informative for network modeling. A straightforward approach to adding this information into the INM model (chapter 2) is to, (i) calculate a PPI score (e.g., pDOCKQ) – many such scores have been assessed (Akdal et al. 2022; Bryant, Pozzati, and Elofsson 2022), (ii) use the PPI score as input information for modeling, (iii) restrain the value of each edge variable using a distribution based on the PPI score (e.g., a normal distribution with mean equal to the PPI score), (iv) use the variance parameter of the distribution to encode some uncertainty in the interpretation of the confidence score.

Recommendations for the development of a proximity-labeling term

In chapter 3 we use proximity labeling proteomics (PLP) in combination with AlphaFold-Multimer virtual PPI screening to validate an EGFR molecular neighborhood. It would be useful to include proximity labeling information in the INM method developed in chapter 2. In general, PLP strategies may label more distant pairs of molecules in the 1-10nm range – a much larger range than chemical cross-linking or virtual PPI screening. As such, the interactors may be either direct or indirect (participates in the same complex but does not interact directly). One way to include such information would be to restrain a higher-order network feature based on the PLP data. One such feature could be the shortest-path-length between nodes i and j . We began to explore this direction by using a variation of the Floyd-Warshall algorithm to calculate shortest paths (Floyd 1962; Warshall 1962). A challenge for developing restraints based on higher order network features (e.g., degree, shortest path) is the construction of reference networks. A reference network of direct-protein interactions could be obtained from the PDB in roughly the following steps: (i) obtain all PDB structures, (ii) calculate pairwise contacts between chains, (iii) cluster sequence similar chains and assign nodes to each cluster, (iv) draw an edge between two nodes if a contact occurs between any corresponding chain pair, (v) bootstrap the network by selecting M random subgraphs consisting of all edges between N random nodes for the same N . This process would produce a statistical sample of direct networks from which restraints could be developed. A caveat to this approach is uncertainty in the reference. In general, many protein interactions may be missing. Such missing interactions may be reduced by adding interactions from sources such as AF-M screening. Nonetheless, these missing interactions will systematically bias the higher order features used for the restraints. In some cases, this problem may be avoided using an appropriate upper or lower bound. For example, if a proximity-labels are enriched at shorter path lengths - the maximal value of the shortest-path-length could be restrained. This is because the addition of missing protein interactions can only shorten the path-length and cannot make it longer. Conversely, if a labeled proteins are enriched at high-degree nodes, a lower bound can be placed on the degree of a node. This is because the true degree of the node can only increase with the addition of missing

edges. Such an analysis depends on the main source of inaccuracy being missing edges and not false edges. This is likely true for high-quality protein interaction datasets (e.g., PDB-bind, PDB) but may not be true for more broadly defined interaction datasets (e.g., STRING).

Recommendations for the development of a CRISPRi/a term

The addition of functional genomic information to INM is appealing because it is entirely orthogonal to proteomic approaches. The challenges with genomic information are (i) they cannot be applied to essential proteins, (ii) functional relationships may exist even if the protein molecules are physically distant, and (iii) functional relationships may exist through other, non-protein mediated mechanisms (e.g., RNA-mediated transcriptional control) (Henninger et al. 2021). These features make it generally more challenging to interpret genomic information in terms of direct protein interactions. An approach to this problem would be to (i) gather a large amount of CRISPRi/a data, (ii) obtain a reference of direct protein interactions (see above), (iii) calculate some pairwise score based on the CRISPRi/a data, (iv) obtain an empirical distribution for the likelihood of the score given a feature of the reference. This distribution is informative regardless of the non-protein mediated mechanisms. As a hypothetical example, the likelihood of a CRISPRi/a positive genetic interaction (GI) score may increase for pairs of nodes with a shortest-path-length of at least 2. Critically, a lower GI score does not imply that the path-length is longer – it is simply the lack of evidence.

4.2 Bridging the gap between structural and systems biology

Integrative modeling methods allow for the modeling of increasingly complex and dynamic systems (Rout and Sali 2019; Raveh et al. 2024; Latham et al. 2024). The advent of AlphaFold-Multimer (Evans et al. 2021; Bryant, Pozzati, and Elofsson 2022; Akdel et al. 2022) virtual PPI screening allows for structural information at proteome-wide scales for both bacterial and eukaryotic proteomes. Modeling methods that integrate this information are critical to answer questions relevant for understanding the fundamental workings of cells, engineering organisms, and developing therapeutic strategies Here, we develop INM

where a node represents many copies of protein molecules, and an edge represents the presence of at least one protein interaction between a pair of molecules corresponding to the two nodes. A model representation with “systems” biology type nodes and “structural” biology type edges may help bridge the gap between systems and structural biology.

4.3 References

1. Akdel, Mehmet, Douglas E. V. Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O. Zalevsky, Bálint Mészáros, Patrick Bryant, et al. 2022. “A Structural Biology Community Assessment of AlphaFold2 Applications.” *Nature Structural & Molecular Biology* 29 (11): 1056–67.
2. Baek, Minkyung, Ivan Anishchenko, Ian R. Humphreys, Qian Cong, David Baker, and Frank DiMaio. 2023. “Efficient and Accurate Prediction of Protein Structure Using RoseTTAFold2.” *bioRxiv*. <https://doi.org/10.1101/2023.05.24.542179>.
3. Bryant, Patrick, Gabriele Pozzati, and Arne Elofsson. 2022. “Improved Prediction of Protein-Protein Interactions Using AlphaFold2.” *Nature Communications* 13 (1): 1265.
4. Evans, Richard, Michael O’Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Židek, et al. 2021. “Protein Complex Prediction with AlphaFold-Multimer.” *bioRxiv*. <https://doi.org/10.1101/2021.10.04.463034>.
5. Floyd, Robert W. 1962. “Algorithm 97: Shortest Path.” *Communications of the ACM* 5 (6): 345.
6. Henninger, Jonathan E., Ozgur Oksuz, Krishna Shrinivas, Ido Sagi, Gary LeRoy, Ming M. Zheng, J. Owen Andrews, et al. 2021. “RNA-Mediated Feedback Control of Transcriptional Condensates.” *Cell* 184 (1): 207–25.e24.
7. Humphreys, Ian R., Jimin Pei, Minkyung Baek, Aditya Krishnakumar, Ivan Anishchenko, Sergey Ovchinnikov, Jing Zhang, et al. 2021. “Computed Structures of Core Eukaryotic Protein Complexes.” *Science (New York, N.Y.)* 374 (6573): eabm4805.
8. Humphreys, Ian R., Jing Zhang, Minkyung Baek, Yaxi Wang, Aditya Krishnakumar, Jimin Pei, Ivan Anishchenko, et al. 2024. “Essential and Virulence-Related Protein Interactions of Pathogens Revealed through Deep Learning.” *bioRxiv.org: The Preprint Server for Biology*, April. <https://doi.org/10.1101/2024.04.12.589144>.
9. Latham, Andrew P., Jeremy O. B. Tempkin, Shotaro Otsuka, Wanlu Zhang, Jan Ellenberg, and Andrej Sali. 2024. “Integrative Spatiotemporal Modeling of Biomolecular Processes: Application to the

Assembly of the Nuclear Pore Complex.” bioRxiv.org: The Preprint Server for Biology, August, 2024.08.06.606842.

10. Mirdita, Milot, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. 2022. “ColabFold: Making Protein Folding Accessible to All.” *Nature Methods* 19 (6): 679–82.
11. Raveh, Barak, Roi Eliasian, Shaked Rashkovits, Daniel Russel, Ryo Hayama, Samuel E. Sparks, Digvijay Singh, et al. 2024. “Integrative Spatiotemporal Map of Nucleocytoplasmic Transport.” bioRxiv.org: The Preprint Server for Biology, January. <https://doi.org/10.1101/2023.12.31.573409>.
12. Rout, Michael P., and Andrej Sali. 2019. “Principles for Integrative Structural Biology Studies.” *Cell* 177 (6): 1384–1403.
13. Warshall, Stephen. 1962. “A Theorem on Boolean Matrices.” *Journal of the ACM* 9 (1): 11–12.
14. Zhang, Jing, Ian R. Humphreys, Jimin Pei, Jinuk Kim, C. Choi, Rongqing Yuan, J. Durham, et al. 2024. “Computing the Human Interactome.” bioRxiv, October. <https://doi.org/10.1101/2024.10.01.615885>.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

Signed by:

Aji Palar

C36E9517E177410...

Author Signature

10/30/2024

Date