Cell Systems

CellPress

Voices

What is the current bottleneck in mapping molecular interaction networks?

Michael A. Skinnider
Princeton University

Integration by parts

Network biologists today have access to a rich assortment of interaction networks produced by assays such as affinity purification-mass spectrometry (AP-MS), yeast two-hybrid (Y2H) screening, co-fractionation mass spectrometry (CF-MS), or thermal proximity co-aggregation (TPCA), to name just a few. But high-throughput interaction data are notoriously noisy, such that similar experiments performed in different laboratories can produce very different networks. A long-standing challenge is distinguishing genuine interactions from experimental artifacts.

Data integration offers a path forward. Meta-analysis of similar experiments performed in different laboratories can reveal interactions that are reproducible across dozens or even hundreds of datasets. A more ambitious goal is integrating datasets that span diverse assay types or that connect distinct classes of biomolecules, since it is unlikely that a single interaction assay can capture the totality of biologically relevant interactions.

However, data integration presents both practical and conceptual challenges. First, relevant datasets must be identified and prepared for integration, which may involve reprocessing of raw experimental data to minimize unwanted computational variation. To date, integration efforts have incorporated at most a handful of distinct assay types, highlighting an untapped opportunity to develop more comprehensive network models. Moreover, computational best practices for integration itself are still evolving. Historically, data integration has been primarily achieved using supervised machine learning, but this approach requires the definition of a set of "gold standard" interactions that the model should reproduce, and this definition in turn dictates the content and utility of the integrated network. Lastly, integration of biomolecular interaction datasets that span multiple modalities (for instance, protein-metabolite interactions) remains largely uncharted territory.

Katja Luck
Institute of Molecular Biology (IMB)

The rugged path to completion

Most protein-protein interaction (PPI) mapping efforts are done under nonphysiological conditions and thus result in biophysically possible PPIs lacking information about when a PPI might exert a function. While this can be seen as a major limitation, I would like to argue that this can also be seen as an opportunity. Determining protein interactomes across physiological conditions at acceptable completion is unrealistic. A more achievable goal is mapping all biophysically possible PPIs within a given organism and determining the relevant "sub" protein interactome for a given cellular context via *in silico* filtering using protein expression, localization, and imaging data.

To map the biophysically possible protein interactome, we need to understand methodological biases as well as how to translate them into complementarities and how to define completeness. If binding affinity is infinite, is the interactome, too? Should we consider a contact between two proteins as a PPI if this contact has a functional effect? Practically, this is not measurable. Interactome completeness will have to be defined based on methodological considerations. Another important angle to interactome mapping is proteoform resolution. While experimental methods are improving in their ability to distinguish proteoforms, computational efforts generally collapse PPI data to the gene level, a limitation that is resolvable. Obtaining structural information on PPIs will tremendously help in making sense of PPI data and connecting networks with function.

With the advent of artificial intelligence (AI), we are closer to structurally resolved interactomes than ever. However, experimental data used to train AI tools are biased, and it will be of utmost importance to understand how biases in training lead to biases in prediction—i.e., with respect to conservation, interaction stability, and chain flexibility. As a community, we should come together to formulate realistic goals and milestones toward the completion of the human protein interactome. The field still awaits its Human Cell Atlas moment.

M. Shahid Mukhtar
Clemson University

Context-specific interaction networks

Mapping molecular interaction networks in specific biological contexts—such as distinct tissues, cell types, developmental stages, or disease states—is crucial for understanding cellular functions. Omics technologies such as single-cell RNA sequencing and ATAC sequencing reveal cellular heterogeneity but face limitations like data sparsity, technical noise, and context-specific biases, while spatial transcriptomics offers spatial insights yet remains limited by cost and scalability. Proteomics and metabolomics offer condition-specific insights but are complex to generate and interpret. Bulk interactomics further obscures cell-type-specific interactions, resulting in incomplete or biased datasets that limit precise network mapping.

Building meaningful context-specific networks requires integrating diverse omics layers—transcriptomics, epigenomics, proteomics, and metabolomics—each with different data structures and resolutions. Inconsistent metadata, like tissue identity and developmental stage, add complexity, requiring robust statistical modeling, biological annotation, and scalable computation. Emerging computational approaches, including graph-based models and probabilistic frameworks, are being developed to address these challenges.

However, issues with scalability, interpretability, and generalizability remain. AI, particularly machine and deep learning, holds promise for overcoming these limitations. Tools like PINNACLE use geometric deep learning trained on multiorgan single-cell atlases to generate context-aware protein representations, improving annotation and therapeutic predictions. Similarly, frameworks like scNET are advancing the resolution of regulatory network inference at the single-cell level. Together, these innovations are bringing researchers closer to constructing accurate, high-resolution molecular interaction networks that reflect the true complexity of biological systems.

Martin Garrido-Rodriguez and Julio Saez-Rodriguez
EMBL, EMBL-EBI, and Heidelberg University

Balancing knowledge and data-driven approaches

We distinguish two main strategies for constructing molecular networks: knowledge-driven and data-driven approaches. Knowledge-driven methods rely on established biological information as a scaffold, integrating context-specific data onto this foundation. By contrast, data-driven approaches build networks directly from experimental observations, offering potentially greater dynamism and specificity. For instance, gene regulatory networks can be inferred from transcriptomic and chromatin accessibility data, while PPIs are mapped using techniques such as size exclusion chromatography, co-melting profiling, or cross-linking proteomics. These methods differ in the directness of the evidence they provide, influencing the reliability of the resulting networks.

As high-throughput technologies continue to advance, data-driven approaches are becoming increasingly prominent. However, prior biological knowledge remains essential for validating network inferences and guiding model regularization. Key challenges persist, including variable data quality, lack of universal benchmarks to assess network robustness and identifiability, and limited transferability of networks across different biological contexts. Understanding how interactions identified under one condition, such as a tissue type or drug treatment, translate to others remains a major obstacle.

Looking ahead, integrating structural predictions and expanded large-scale perturbation datasets, as well as developing deep learning methods that embed biological knowledge, offer promising pathways forward, enabling network biology to deliver increasingly powerful models for understanding and predicting complex biological systems.

Jolanda van Leeuwen
UMass Chan Medical School and University of Lausanne

Scaling up molecular insights

A major challenge in the mapping of interaction networks is the scale at which we can perform experiments. We know that molecular interactions are often context dependent. Ideally, we would thus like to measure the effect of all variants in every gene on the binding of the encoded protein to all other biological molecules in every genetic background and environmental condition. Obviously, this is not feasible. We are therefore often forced to limit the scope of our experiments, which causes problems for data interpretation, integration, and translatability. For example, it can be difficult to integrate PPI network data mapped under nutrient-rich conditions in epithelial cells with genetic interaction networks mapped under starvation conditions in cancer cell lines and then use the combined data to learn something biologically relevant about humans.

Despite these challenges, the field has made phenomenal progress in increasing the scale of experiments and developing approaches to integrate diverse biological networks. A decade ago, we were using libraries of ~100,000 mutants in a single experiment. Today, our experiments are two orders of magnitude larger. And as sequencing costs continue to drop, even larger experiments will become affordable. With these continuing developments, I am excited to see where the next decade of network mapping will take us!

Pedro Beltrao
Institute of Molecular Systems Biology ETH Zürich and Swiss Institute of Bioinformatics

No method to rule them all

Molecular interactions are at the heart of the function of biomolecules and have been instrumental in studying protein function and linking genetic variants to disease. The major bottleneck in mapping molecular interaction networks is the lack of a single scalable approach that captures the full condition-specific spectrum of different types of interactions. Indeed, no single method is available to determine all interaction partners even for physical PPIs, in part due to differences in interactions for stable complexes and the transient interactions of signaling enzymes. This challenge compounds when adding other interactions for metabolites, DNA, RNA, and enzymes with substrates: each layer requires its own detection technology and expert research groups. Additionally, mapping a molecular interactome has no natural finish line because—unlike the genome—networks can rewire across cell types or developmental stages.

Nevertheless, exciting progress in this field with improvements in both experimental and computational methods enables increasingly scalable mapping efforts of interaction networks. A concerted initiative that applies complementary experimental and computational tools to a handful of model systems would be a timely initiative that could transform sparse interaction lists into dynamic, context-resolved networks and unlock new insights into cell-type identity, disease mechanisms, and the evolution of multicellularity.

Anne-Ruxandra Carvunis
University of Pittsburgh School of Medicine

We are missing so many nodes!

Network biologists know this: our interaction maps are biased toward well-studied genes—the ones with curated functions, available reagents, previously identified interactions . . . But the scope of this problem is bigger than we thought.

Thousands of newly discovered translated elements are not yet integrated into molecular network maps—and, in fact, are still largely absent from reference annotations. These elements are short and go by many names: small open reading frames (sORFs or smORFs), noncanonical open reading frames (nORFs or ncORFs), novel unannotated open reading frames (nuORFs), proto-genes, proto-ORFs, microproteins, miniproteins, micropeptides, sORF-encoded polypeptides (SEPs), small proteins, and more.

Some are deeply conserved; others are species specific. Some map to transcripts previously thought to be noncoding; others map to known mRNAs. They can encode functional proteins, be processed as cell-surface epitopes, regulate the translation of genes, and contribute to disease. Very few have been characterized to date: everything is possible, very little is known. If interaction networks are to truly reflect the biology of

the cell, we must broaden our definition of a node and systematically incorporate these small but mighty elements into our mapping efforts.

Mikko Taipale
University of Toronto

Maybe less is more?

When identifying bottlenecks in interaction network mapping, we often highlight technical limitations. Having faster screens, better reagents, or higher sensitivity can give us more interactions—and more is good. But I'm going to resist the temptation and ask: what if we actually need fewer interactions? Let's take the guardian of the genome, p53. It currently has about 2,500 physical interactions in interaction databases. Or how about EGFR, with over 3,000 interactions? Although both proteins are important and associate with many other proteins and molecules in the cell, I wager that most of these interactions do not actually happen in cells—or, if they do, they are biologically irrelevant.

There are two explanations: the (understandable) pursuit of finding ever more interactions and the limits of natural selection. With new methods that are more scalable and sensitive, we have both the incentive and the opportunity to discover and report ever more interactions. New methods are, by design, evaluated by their ability to find new interactions. Moreover, interaction databases grow asymmetrically: old, likely spurious interactions are rarely if ever removed, even when better data emerge. On the other hand, just like our genome is dominated by junk DNA, our cells are full of meaningless molecular encounters. Natural selection is simply too weak to weed out spurious interactions that have only mildly deleterious effects on fitness. If we continue to chase quantity, we risk building molecular interaction networks that are increasingly disconnected from biological reality.

Perhaps we need *more* bottlenecks?

Andrew Emili
Oregon Health and Science University

Cracking the cell interactome code

Despite transformative advances in single-cell and spatial transcriptomics, our ability to map dynamic protein interaction networks at single-cell resolution in complex, heterogeneous tissues like the tumor microenvironment remains severely limited. PPIs are the functional scaffolds of cell state and function, yet capturing their context-specific, transient, and multivalent nature—especially across diverse cell types and changing states within evolving pathological landscapes—is a daunting task. This challenge is magnified when studying tissues with profound spatial heterogeneity, fluctuating microenvironments, and clinically divergent disease trajectories such as therapy response, drug resistance, or progression.

Current technologies fall short in sensitivity, scalability, or spatial precision. Mass-spectrometry-based interactomics lacks single-cell resolution, proximity labeling and imaging-based methods struggle to scale to networks, and computational inference fails to account for dynamic cell state transitions, the complexity of tissues, or patient variability. To overcome this, we urgently need an entirely new generation of tools—ultrasensitive, spatially precise, and temporally resolved. Advances in single-molecule *in situ* protein sequencing, AI-driven structural modeling, and ultrahigh-sensitivity spatial proteomics promise to break these barriers. But the path forward requires more than technology—it demands integrated frameworks that can unify these data to infer actionable, disease-relevant networks. Only then can we truly decode and leverage the molecular choreography underpinning health and disease.

Martha L. Bulyk
Brigham & Women's Hospital and Harvard Medical School

Going beyond gene-level molecular interaction networks

Most molecular interaction networks in biology are dynamic, with the network nodes and edges varying depending on the cell type or cell state. What is considered a node has a significant impact on interpretation of the network model. PPI networks and gene regulatory networks are often based on a gene-level view of proteins as the nodes in the networks, whereas alternative isoforms can differ dramatically in their sequences and functions from those of the reference isoform or other alternative isoforms.

For example, our recent study—in collaboration with the Vidal, Fuxman Bass, Salomonis, and other labs—of hundreds of transcription factor (TF) isoforms revealed that two-thirds of alternative TF isoforms differed from the reference isoform in at least one molecular function. Identification of the different node isoforms expressed in different cell types or states, together with data on how the interactions of alternative isoforms differ from those of their cognate reference isoforms, is essential for building molecular interaction networks that accurately model biological functions in a cell-type/state-dependent manner.

Going beyond a basic nodes/edges framework of networks, quantitative data on the nodes, the strengths of their interactions, and their functional consequences on cellular outputs are needed to understand how biological states are altered by mutations or genetic variation and how to best target pathological states therapeutically. High-throughput approaches to generate such datasets across cell types/states as well as computational methods to predict isoforms' functions from their primary sequences are needed to fill existing gaps in our ability to build such predictive models for any given cell type of interest.

Nevan J. Krogan
University of California, San Francisco, and Gladstone Institute of Data Science & Biotechnology

People power in network mapping

For me, the current bottleneck in mapping molecular interaction networks spans three key areas. First is *resolution*—we must move beyond simply detecting PPIs; we need to know *how* these molecules interact at a domain and residue level in quantitative and structural detail. This demands the integration of advanced structural technologies like cryoelectron microscopy and cryoelectron tomography, combined with functional genetics to validate biological relevance.

Second is the *temporal and spatial dimension*: interactions are dynamic, context-dependent processes that shift across time, cellular location, and conditions such as mutations, environmental stress, or drug treatments. Capturing these changes requires tools that can monitor molecular interactions in real time and *in situ*.

Finally, neither of these problems can be solved in isolation—*people-people interactions* are a key, though often overlooked, bottleneck. The complexity of the task demands that structural biologists, geneticists, systems biologists, and computational scientists work together seamlessly, yet coordination remains difficult due to institutional, cultural, and communicative barriers. However, bringing together experts from diverse fields is essential, as even the most advanced tools risk operating in silos. Only through more effective interdisciplinary collaboration can we build a truly comprehensive and functional map of molecular interactions.

DECLARATION OF INTERESTS

A.-R.C. is a member of the scientific advisory board for ProFound Therapeutics. J.S.-R. reports in the last 3 years funding from GSK and Pfizer and fees/honoraria from Travere Therapeutics, STADApharm, Astex, Pfizer, Owkin, Moderna, and Grunenthal. M.A.S. is a member of the Rutgers Cancer Institute of New Jersey (RCINJ). The Krogan laboratory has received research support from Vir Biotechnology, F. Hoffmann-La Roche, and Rezo Therapeutics. N.J.K. has a financially compensated consulting agreement with Maze Therapeutics. He is the president and is on the board of directors of Rezo Therapeutics, and he is a shareholder in Tenaya Therapeutics, Maze Therapeutics, Rezo Therapeutics, and GEn1E Lifesciences.