## Rapid Statistical Review Report for your manuscript

**Title:** State of the interactomes: an evaluation of molecular networks for generating biological insights.

Dear Author,

Thank you for giving us the opportunity to work with you!

For ease of understanding, this report is divided into the following sections:

| | |
|---|---|
| **Section 1** | **TECHNICAL CHECKS**<br>Details of the checks that we have undertaken as part of the review |
| **Section 2** | **OVERVIEW & NEXT STEPS**<br>Recommended next steps for you |
| **Appendix** | Frequently Asked Questions |

We will be happy to provide further clarifications or answer any queries you may have about this report.

## Section 1: TECHNICAL CHECKS

### ➢ Review of research design & methods

- The study evaluates 45 human interactomes to assess their utility in biological applications, specifically disease gene prioritization and interaction prediction.
- The inclusion of 45 diverse interactomes, categorized as Experimental, Curated, and Composite, allows for a broad evaluation across interaction types (e.g., protein-protein, regulatory, signaling).
- However, the selection of 45 interactomes, while expansive, is not exhaustive and may over-represent well-known databases (e.g., IntAct, BioGRID).
- All networks were standardized by mapping to NCBI Gene IDs, removing duplicates and self-interactions, and filtering non-human interactions, ensuring consistency and comparability across datasets.
- Interactomes and consensus networks (PCNets) are publicly available via NDEx (ndexbio.org), and the evaluation pipeline is accessible on GitHub, promoting reproducibility and transparency.
- The study employs two complementary approaches—disease gene prioritization (via network propagation) and interaction prediction (using L3 and MPS algorithms)—to assess network performance, capturing both functional and structural utility.
- Disease gene sets are primarily sourced from DisGeNET and GWAS Catalog, which may introduce literature bias. Experimental gene sets (n=17) are limited in number and scope.
- Interaction predictions are validated against external datasets (CORUM, PANTHER) and in silico using AlphaFold-Multimer, enhancing the robustness of findings.
- The study does not mention pilot testing the evaluation pipeline, which could refine metrics and ensure robustness across network types.
- The reliance on tools like AlphaFold-Multimer for in silico validation introduces potential biases, as it is optimized for physical interactions and may miss non-physical relationships.
- Consider complementing AlphaFold with other in silico tools (e.g., molecular docking, machine learning-based interaction predictors) to capture a wider range of interaction types.

### ➢ Data Analysis

- The use of non-parametric tests (e.g., Spearman correlation, Mann-Whitney U-test) for non-normal data (e.g., ipTM scores) and parametric tests (e.g., linear regression for size adjustment) where appropriate demonstrates rigorous statistical planning.
- The robust Z-statistic for gene set recovery, using median absolute deviation, accounts for outliers and non-normal distributions, enhancing reliability.
- The 10-fold cross-validation for interaction prediction ensures unbiased estimation of predictive performance, while external validation against CORUM and PANTHER strengthens generalizability.

- The study mentions linear regression for size adjustment but does not report model fit statistics (e.g., $R^2$, p-values) or variable selection methods, limiting interpretability.
- While p-values and correlations (e.g., rs=0.80 for network coverage vs. citation count) are reported, effect sizes (e.g., Cohen's d, Cramer's V) are absent, reducing the ability to assess practical significance.
- The study does not explore multivariate models to identify predictors of network performance (e.g., interaction type, data source, network topology), missing opportunities to uncover complex relationships.
- Multiple statistical tests (e.g., GO enrichment, Mann-Whitney U-tests) are conducted without correction (e.g., Bonferroni, FDR), increasing the risk of false positives.
- Consider including effect sizes for correlations (e.g., $rs^2$ for Spearman's) and group comparisons (e.g., Cohen's d for ipTM score differences). For example, report Cramer's V for GO enrichment to quantify functional bias strength.
- Implement generalized linear mixed models (GLMM) to account for clustering effects (e.g., interactome dependencies) and explore predictors of performance, such as interaction type or data source.
- Conducting a post-hoc power analysis will ensure adequate sample size for detecting meaningful differences in performance across interactomes, particularly for smaller gene sets.

## ➢ Critical appraisal of strengths/weaknesses

- The study assesses 45 diverse interactomes across multiple dimensions (gene coverage, functional representation, performance), providing a robust benchmark for network selection in biological research.
- The dual use of disease gene prioritization (network propagation) and interaction prediction (L3, MPS, AlphaFold-Multimer) offers a multifaceted evaluation, addressing both functional and predictive utility.
- The availability of standardized interactomes, PCNets, and the evaluation pipeline via NDEx and GitHub enhances transparency, reproducibility, and community engagement.
- The non-systematic selection of 45 interactomes may exclude emerging or specialized networks, limiting the study's comprehensiveness.
- The lack of detailed regression model statistics and effect sizes hampers interpretability and the assessment of practical significance.
- The focus on protein-coding genes and exclusion of non-coding RNAs/pseudogenes restricts applicability to broader genomic contexts.
- The absence of FDR or Bonferroni correction for multiple statistical tests increases the risk of false positives, potentially inflating reported associations.

## Section 2: OVERVIEW & NEXT STEPS

### SUMMARY

The research article "State of the interactomes: an evaluation of molecular networks for generating biological insights" provides a comprehensive benchmarking of 45 human interactomes, assessing their utility in disease gene prioritization and interaction prediction. Conducted as a computational study, it analyzed networks categorized as Experimental, Curated, and Composite, using standardized data from public sources (e.g., IntAct, BioGRID) and depositing results on NDEx (ndexbio.org).

Findings indicate that large composite networks (e.g., HumanNet, STRING, FunCoup) excel in prioritizing disease genes from Literature (DisGeNET, n=906) and Genetic (GWAS, n=699) datasets, driven by extensive gene and interaction coverage. Smaller networks (e.g., DIP, Reactome, SIGNOR) outperform in interaction prediction, achieving high precision for held-out (P@k=0.57 for DIP) and external interactions (CORUM, PANTHER). The study introduces PCNet2.0 (3.85M interactions), a consensus network balancing performance and parsimony, and validates 126 novel interactions using AlphaFold-Multimer, particularly from SIGNOR, enriching for receptor and regulatory functions. Biases toward highly cited/expressed genes and immune response functions were noted, with underrepresentation of non-coding RNAs and transporter activities.

Despite its scientific design, limitations include potential selection bias in interactome choice, incomplete statistical reporting (e.g., regression model fit), and reliance on computational validation without experimental confirmation. The lack of multiple testing correction and limited experimental gene sets (n=17) further constrain generalizability. Nevertheless, the study provides critical insights into interactome performance, offering publicly accessible tools and networks to guide network biology research and inform applications like GWAS interpretation and gene function discovery.

### RECOMMENDATIONS

We have listed focus areas that should be addressed to improve the robustness of your study.

**Major issues**:

| | Focus area | Recommendations |
|---|---|---|
| 1. | Selection bias | Systematically review available interactomes and define explicit inclusion criteria (e.g., data quality, update frequency). |
| 2. | Incomplete statistical reporting | Fully report regression models, including $R^2$, F-statistics, and coefficients. Use stepwise or lasso regression to identify key predictors. |
| 3. | Multiple tests without correction | Apply Bonferroni correction for all statistical tests and report adjusted q-values alongside p-values to control false positives. |

| | | |
|---|---|---|
| 4. | Limited experimental gene sets | Expand experimental gene sets and include functional sets from proteomics or single-cell RNA-seq to broaden applicability. |
| 5. | Dependence on AlphaFold | Complement AlphaFold with other in silico tools (e.g., molecular docking) and prioritize experimental validation. |

**Minor issues**:

| | **Focus area** | **Recommendations** |
|---|---|---|
| 1. | Lack of pilot testing | Conduct a pilot study on a subset of interactomes to optimize pipeline parameters and assess computational efficiency. |
| 2. | Effect size reporting | Report effect sizes for correlations (e.g., $rs^2$) and group comparisons (e.g., Cohen's d) to contextualize findings. |
| 3. | Subgroup analysis | Perform subgroup analyses by network type or biological context to provide tailored recommendations. |
| 4. | Power analysis | Conduct a post-hoc power analysis to confirm sufficient power for detecting meaningful performance differences. |

## Appendix: FREQUENTLY ASKED QUESTIONS

**Q: What is the technical experts' qualification?**
A: Our experts reviewers have a minimum qualification of a PhD in your relevant subject area and have extensive experience in publishing and peer-reviewing manuscripts. These experts also have experience of writing and publishing their own manuscripts in peer-reviewed journals. Many of our experts even serve as peer reviewers on journal editorial boards.

**Q: The Rapid Statistical Review did not reveal significant gaps in my work. Since this is not of use to me, will you provide me a refund?**
A: The Rapid Statistical Review will be carried out to meet the full scope of the service. We will only make suggestions for rework when it is warranted and is needed to improve the statistical robustness of your study. We will not provide a refund in such cases, since the service scope has been met. If your manuscript is returned after peer review with comments that point out gaps in statistical methods or analysis that could have been identified during this service, we will offer you a full refund.

**Q: Is there post service support?**
A: This is a one-round service. However, if you have any queries about any of the deliverables, you can get in touch with us at any time.