AI-Powered Multi-Omics Integration for Exploring Disease-Associated Molecular Networks and Therapeutic Intervention Pathways

Cailin McEnaney
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
cailin.mcenaney@uab.edu

Kyra Phillipsen
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
kyrawork@ku.edu

Abstract

Disease-associated molecular networks are shaped by interactions across genomic, transcriptomic, proteomic, epigenomic, and metabolomic layers, yet most integrative analyses still reduce multi-omics evidence to static gene lists or single-modality predictive scores. This limitation constrains both mechanistic interpretation and the identification of therapeutic intervention pathways, because a disease module, a target, and a combination of targets can be statistically associated with a phenotype without forming a coherent, modifiable network. The manuscript develops an analytical framework for AI-powered multi-omics integration that links latent-variable fusion to graph-based network inference and pathway-level intervention scoring. The proposed framework represents each sample through modality-specific likelihoods with missingness indicators, learns a shared latent state, and couples this state to a modality-weighted message-passing network built from prior interactions and data-derived edges. Disease association is expressed as a posterior over graphs and modules, while therapeutic hypotheses are ranked through pathway perturbation scores and a constrained combination objective that penalizes network dispersion, toxicity, and cost. A Bayesian uncertainty layer distinguishes confident network predictions from those that are primarily prior-driven, and sensitivity analysis is specified for graph structure, modality weighting, and missing-data mechanisms. The contribution is methodological: it provides a coherent way to connect multi-omics evidence, network topology, and intervention reasoning while preserving the distinction between predictive association and causal mechanism. The framework offers a technically defensible basis for prioritizing disease modules and combinatorial therapeutic hypotheses for subsequent experimental validation.

Keywords

multi-omics integration; disease molecular networks; graph neural networks; pathway intervention; Bayesian uncertainty; network medicine

1. Introduction

Multi-omics studies have changed the scale at which disease mechanisms can be examined, but they have not resolved the central inferential problem of how to integrate heterogeneous measurements into a coherent molecular network [1]. Genomic variation,

transcript abundance, protein state, epigenetic marks, and metabolite concentrations are not independent readouts of a single underlying process. They reflect distinct regulatory layers, measurement scales, and error structures, and their relationships to disease can differ by tissue, cell type, developmental stage, and environment [2]. A model that simply concatenates omics matrices may improve prediction in a narrow setting while obscuring the network relationships that make a therapeutic hypothesis biologically interpretable.

Network medicine provides a useful organizing principle because it treats disease as a perturbation of interacting molecular systems rather than as an isolated gene or pathway defect [3]. Disease proteins and modules can be proximal within molecular interaction networks, and disease-disease relationships can be analyzed through shared or adjacent network regions [4]. These observations motivate integrative models that infer disease-associated networks instead of classifying samples from aggregate features alone. Machine learning has been used for biological data integration and network inference, but the choice of model class matters: latent-factor models capture shared variation across modalities, while graph-based models propagate information across relational structure [5, 6]. The two are complementary, because latent variables can denoise modality-specific measurements and graph propagation can impose biological connectivity.

Unsupervised integration methods such as multi-omics factor analysis provide a principled way to estimate shared latent factors across data types [7], and similarity network fusion shows how multiple data types can be combined through patient or feature graphs [8]. Network-community formulations for synergistic combinations from herbal medicines and complex systems further motivate module-level reasoning about multi-target interventions [9]. What remains less developed is an explicit bridge from multimodal latent representation to an uncertain disease network and then to a constrained intervention path. This manuscript addresses that gap by specifying a unified analytical framework in which multi-omics fusion, graph inference, disease-module posterior estimation, and pathway intervention scoring share the same notation and assumptions.

The central contribution is a methodological framework, not a claim of empirical validation. The framework defines inputs, latent variables, graph construction, objective terms, network posteriors, intervention scores, combination constraints, and uncertainty logic. It is designed to answer a precise question: given partial and heterogeneous multi-omics measurements and prior molecular interactions, how can an AI model represent a disease-associated network and rank modifiable pathway interventions while making its assumptions and uncertainty explicit? The subsequent sections develop the data model, the graph-learning architecture, the disease-network posterior, and the therapeutic intervention objective. The discussion then examines identifiability, confounding, failure modes, and the boundary between predictive network association and causal therapeutic effect.

2. Methodological Framework

2.1 Data representation and latent-variable fusion

Let X denote a collection of n samples measured across M omics modalities. For sample i and modality m, let $x_i^{(m)}$ be the observed feature vector, with p_m features and a missingness indicator r_{im} that records whether modality m is available for sample i. The model assumes a shared latent state z_i that summarizes the biological and technical variation common to the available modalities. Modality-specific likelihoods accommodate different scales and distributions, such as count models for sequencing data, continuous models for protein

abundance, and categorical or threshold models for variant-derived features. The generative relationship is:

$$p(X|Z,\Theta) = \prod_i \prod_m p(x_i^{(m)}|z_i,\Theta_m)^{\wedge}(r_{im})$$

Here Z contains all z_i, Θ contains modality-specific parameters Θ_m, and r_{im} weights only the observed modalities. The factorization assumes conditional independence of modalities given the shared latent state, which is a modeling convenience rather than a biological claim. Residual dependence can be addressed by adding covariate terms for tissue, batch, age, sex, or treatment status, and by allowing modality-specific nuisance factors when sample size permits. The latent state is not interpreted as a causal variable; it is an intermediate representation that supports de-noising, missing-modality handling, and integration of heterogeneous evidence.

2.2 Graph construction and message passing

The second component couples latent states to a molecular graph G=(V,E). Nodes V represent genes, proteins, metabolites, or pathway modules, and edges E represent prior interactions, regulatory relationships, pathway membership, or data-derived co-variation. Edge priors can incorporate circuit-discovery and pathway-enrichment evidence [10, 11], as well as pathway-based interpretation of genetic association [12]. The graph is not treated as complete or infallible. Instead, edge weights are modeled as uncertain quantities that can be updated by observed multi-omics evidence.

A modality-weighted message-passing layer propagates latent information across neighbors. For node i and layer l, a graph attention or graph convolution operation can be written as:

$$h_i^{(l+1)} = \sigma\left(\Sigma_m \alpha_m \Sigma_j \in N(i) A_{ij} W_m h_j^{(l)} + W_0 h_i^{(l)}\right)$$

In this expression, $h_i^{(l)}$ is the node representation at layer l, N(i) is the neighbor set of node i, A_{ij} is the adjacency weight, W_m is a modality-specific transformation, W_0 is a self-connection term, α_m is a nonnegative modality weight, and σ is a nonlinear activation. The modality weights allow the model to emphasize layers with stronger evidence while retaining contributions from weaker or sparser data. Graph attention mechanisms provide a flexible way to learn neighbor importance [13], whereas graph convolutional layers offer a simpler and more stable propagation rule when edges are dense [14]. The choice between them is a bias-variance trade-off: attention increases capacity but also increases the risk of overfitting small or noisy graphs.

Training minimizes a composite objective that balances reconstruction, regularization, and graph fidelity:

$$L(\Theta,\alpha) = -\Sigma_i log\ p(x_i|z_i,\Theta) + \lambda_1 \|W\|_2^2 + \lambda_2 \Sigma_m \alpha_m log \alpha_m + \lambda_3 R(G)$$

The first term is the negative log-likelihood of observed data under the latent model. The second term penalizes large weights, the third discourages degenerate modality weighting by favoring a non-collapsed distribution over α_m, and R(G) represents a graph regularizer such as sparsity, smoothness, or agreement with prior edges. The coefficients λ_1, λ_2, and λ_3 control the relative strength of these preferences. This objective does not establish that learned edges are causal; it specifies the conditions under which a network representation is stable, sparse, and consistent with available evidence.

2.3 Disease association and network posterior

Disease association is represented as a posterior distribution over graphs and modules rather than as a single point estimate. Let D denote phenotype or clinical evidence and let X denote the multi-omics measurements. The posterior can be written as:

$$p(G|X,D) \propto p(D|G)p(X|G,\Theta)p(G)$$

Here p(G) is a prior over graph structures, p(X|G,Θ) is the multi-omics likelihood under the graph-dependent model, and p(D|G) links network structure to disease evidence. The phenotype likelihood may use genetic burden, differential expression, clinical outcomes, or other endpoints, but its form must be specified before inference. The posterior supports module-level summaries such as the marginal probability that a node or pathway belongs to a disease-associated region. Such summaries are useful because individual edge probabilities can be unstable even when module-level evidence is consistent.

The graph posterior also exposes a key limitation: many graphs can explain the same observed data. Identifiability depends on sample size, edge priors, modality coverage, and the strength of phenotype evidence. A posterior that is dominated by p(G) should not be interpreted as data-driven discovery. For this reason, the model should report posterior contraction, sensitivity to prior edge sets, and the fraction of module evidence attributable to observed rather than prior structure. These diagnostics are part of the analytical framework, not optional post hoc checks.

3. Therapeutic Intervention Pathways

3.1 Pathway perturbation and target prioritization

Therapeutic intervention is modeled as a perturbation of the inferred molecular network. For a candidate target set g, the model estimates a perturbation effect $\Delta z_v(g)$ on node or latent state v. A pathway-level score aggregates these effects over pathway nodes V_p:

$$S_p(g) = \Sigma_\{v \in V_p\} w_v |\Delta z_v(g)| / \sqrt{|V_p|}$$

Here w_v is a node weight reflecting network centrality, disease association, or evidence quality, and $|V_p|$ is the number of nodes in pathway p. The square-root normalization prevents large pathways from dominating solely through size. A positive score indicates that the candidate perturbation is predicted to alter pathway activity; it does not by itself indicate therapeutic benefit. Benefit requires a directional criterion, such as reversal of a disease-associated expression signature, restoration of a disrupted module, or movement of a latent disease state toward a reference state. The model should therefore separate perturbation magnitude from beneficial direction.

Target prioritization also requires tractability and safety constraints. A target with strong network effect but poor accessibility, high on-target toxicity risk, or broad systemic expression may be a poor intervention candidate. These constraints can be represented as penalty terms or as admissibility filters. The purpose is not to replace pharmacological judgment but to ensure that network scores are not interpreted outside the biological and clinical conditions under which an intervention would be feasible.

3.2 Combination interventions and network-community selection

Combination therapy is a natural extension of network-based reasoning because single-target interventions may fail when a disease module is buffered by redundant pathways. Graph-based models of polypharmacy have shown that relational structure can be used to model combined effects and adverse interactions [15]. Transcriptomic connectivity resources

provide complementary evidence by linking compounds to gene-expression signatures [16], and large-scale perturbation resources such as L1000 support signature-based hypotheses at broader scale [17]. Pharmacogenomic interaction maps further show that drug response depends on molecular context and cannot be reduced to a single target property [18]. These resources are used here as evidence sources for intervention scoring, not as proof that a predicted combination will be effective.

For a candidate target set g drawn from a tractable target pool T, the combination objective can be formulated as:

$$max_\{g \subseteq T, |g| \leq k\} J(g) = \Sigma_p \beta_p S_p{}^{+}(g) - \lambda_d \Sigma_\{u, v \in g\} d_G(u, v) - \lambda_c C(g)$$

In this expression, $S_p^+(g)$ is the predicted beneficial pathway effect after directional filtering, β_p weights pathway importance, d_G(u,v) is a network distance or dispersion penalty, C(g) is a cost or toxicity proxy, and λ_d and λ_c control the trade-off. The cardinality constraint $|g| \leq k$ limits combinatorial complexity. The dispersion penalty encodes the idea that targets acting within a coherent module may produce more interpretable combination effects than targets scattered across unrelated pathways. This is a hypothesis-generating criterion, not evidence of synergy. Synergy would require experimental dose-response data or a validated predictive model trained on combination outcomes.

Network-community structure is useful at this stage because it groups nodes with dense internal connectivity. A combination that spans multiple communities may engage compensatory pathways, whereas a combination concentrated in one community may suffer from redundancy. The optimization should therefore be evaluated under alternative community partitions and edge thresholds. If the selected combination changes substantially under small graph perturbations, the result is not robust. Sensitivity of this kind is more informative than a single ranked list of targets.

3.3 Uncertainty, confounding, and causal boundaries

Predictive uncertainty must be separated into aleatoric variation, which reflects irreducible measurement or biological noise, and epistemic uncertainty, which reflects limited data or model ambiguity. For a candidate intervention g and sample-level covariates x, epistemic uncertainty can be summarized as:

$$U(x, g) = \mathrm{Var}_{p(\theta|D)}[p(y = 1|x, g, \theta)]$$

Here θ denotes model parameters, D denotes observed evidence, and p(y=1|x,g,θ) is the predicted probability of a beneficial or disease-relevant response. A large U(x,g) indicates that the model's prediction depends strongly on parameter uncertainty. This quantity should be reported alongside the point prediction, because a high-scoring intervention with broad posterior uncertainty is a weak basis for prioritization.

Causal interpretation requires stronger assumptions than predictive performance. The model can estimate associations between network perturbations and phenotype, but it cannot identify a causal effect without assumptions about confounding, exchangeability, and intervention stability. Batch effects, tissue composition, treatment history, and reverse causation can all generate network associations that are not modifiable. The framework therefore distinguishes three evidence levels: network association, predicted perturbation response, and experimentally or clinically supported causal effect. Only the third level justifies a therapeutic claim. The first two are useful for prioritization if their uncertainty and assumptions are explicit.

4. Analytical Assessment and Technical Discussion

The main strength of the proposed framework is that it makes the path from multi-omics data to therapeutic hypothesis explicit. Latent-variable fusion handles missing modalities and heterogeneous scales; graph message passing incorporates prior molecular interactions; the graph posterior represents uncertainty over disease-associated modules; and the intervention objective converts network evidence into constrained target-set scores. This integration is more informative than a pipeline that first selects differentially expressed genes, then overlays them on a pathway database, because it allows evidence from multiple modalities and network topology to influence module inference jointly.

The framework also has identifiable failure modes. If edge priors are biased toward well-studied proteins, graph propagation can amplify popularity effects and neglect understudied disease genes. If latent factors absorb batch or tissue composition, downstream module scores may reflect technical structure rather than disease biology. If phenotype labels are noisy or confounded, the posterior $p(D|G)$ can favor large modules that correlate with sample collection rather than mechanism. If the graph is dense and the sample size is small, message passing can overfit and produce unstable edge weights. These risks are not solved by model capacity; they require design controls such as batch-aware likelihoods, stratified validation, edge-dropout sensitivity, and negative-control pathways.

Model evaluation should therefore be structured around the intended claim. For predictive claims, held-out samples and external cohorts are needed, with calibration and decision-curve analysis rather than discrimination alone. For network claims, edge-level and module-level recovery should be assessed against independent interaction data, and posterior contraction should be reported. For intervention claims, perturbation experiments or clinically annotated response data are required; a ranked pathway list is not sufficient. The framework should be compared with credible baselines, including single-modality models, concatenation-based classifiers, latent-factor integration without graph propagation, and graph propagation without multi-omics fusion. Ablation of modality weights and edge priors is essential because these components can create apparent interpretability without improving out-of-sample validity.

Single-cell and variant-level evidence can strengthen the framework when available. Deep learning models for inferring gene relationships from single-cell expression data can provide context-specific regulatory edges [19], and sequence-based models can predict variant effects on expression and disease risk [20]. These evidence sources should enter as modality-specific likelihoods or edge priors with explicit uncertainty, not as unquestioned ground truth. The same principle applies to pathway databases and compound signatures: they are curated evidence sources with coverage and ascertainment biases. The model's output is therefore best interpreted as a calibrated synthesis of available evidence, not as a complete map of disease biology.

5. Implementation and Scientific Boundary Conditions

Implementation requires attention to computational scaling and reproducibility. Graph propagation over genome-scale networks can be memory-intensive, and latent-variable inference with missing modalities can be sensitive to initialization. Mini-batch training, sparse adjacency representations, and modular likelihoods are practical choices, but they introduce approximations that should be documented. Preprocessing must be fixed before evaluation to avoid information leakage, and feature selection must be performed within training folds.

Hyperparameters such as λ_1, λ_2, λ_3, k, β_p, λ_d, and λ_c should be selected through nested validation when labels are available, or through stability analysis when they are not.

The framework is intended for hypothesis prioritization in diseases with sufficient molecular annotation and accessible perturbation evidence. It is less suitable when the relevant tissue is unavailable, when the phenotype is heterogeneous, or when the network prior is dominated by interactions from unrelated cell types. Under these conditions, the model may produce coherent but biologically displaced modules. The appropriate response is not to increase model complexity but to narrow the claim, report uncertainty, and require independent validation before therapeutic interpretation.

6. Conclusion

This manuscript has developed an analytical framework for AI-powered multi-omics integration that connects latent-variable fusion, modality-weighted graph inference, disease-network posterior estimation, pathway perturbation scoring, and constrained combination selection. The framework treats disease-associated networks as uncertain objects rather than fixed maps, and it separates predictive association from causal therapeutic effect. Its technical value lies in the explicit linkage between data likelihoods, graph structure, intervention scores, and uncertainty, which allows each assumption to be examined and tested. The framework does not assert that any particular network or combination is effective; it specifies how such hypotheses can be generated, compared, and bounded. Future empirical work should evaluate calibration, external validity, and perturbation recovery against independent cohorts and experimental systems. Used with appropriate restraint, the approach provides a disciplined route from multi-omics evidence to testable molecular network hypotheses and therapeutic intervention pathways.

References

1. Hasin, Y., Seldin, M., & Lusis, A. J. (2017). Multi-omics approaches to disease. Genome Biology, 18, 83. https://doi.org/10.1186/s13059-017-1215-1

2. Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. Nature Reviews Genetics, 19(5), 299–310. https://doi.org/10.1038/nrg.2018.8

3. Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. Nature Reviews Genetics, 12(1), 56–68. https://doi.org/10.1038/nrg2918

4. Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., & Barabási, A.-L. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science, 347(6224), 1257601. https://doi.org/10.1126/science.1257601

5. Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion, 50, 71–91. https://doi.org/10.1016/j.inffus.2018.09.012

6. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-generation machine learning for biological networks. Cell, 173(7), 1581–1592. https://doi.org/10.1016/j.cell.2018.05.015

7. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology, 14(6), e8124. https://doi.org/10.15252/msb.20178124

8. Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nature Methods, 11(3), 333–337. https://doi.org/10.1038/nmeth.2810

9. HerbSyner_Finder: a network community-based model for identifying synergistic combinations from herbal medicines and complex systems. Targetome. 2026;2(2):0-0. doi: 10.48130/targetome-0026-001

10. Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics, 18(Suppl. 1), S233–S240. https://doi.org/10.1093/bioinformatics/18.suppl_1.S233

11. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

12. Wang, K., Li, M., & Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. Nature Reviews Genetics, 11(12), 843–854. https://doi.org/10.1038/nrg2884

13. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations. https://arxiv.org/abs/1710.10903

14. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations. https://arxiv.org/abs/1609.02907

15. Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457–i466. https://doi.org/10.1093/bioinformatics/bty294

16. Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., & Golub, T. R. (2006). The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. Science, 313(5795), 1929–1935. https://doi.org/10.1126/science.1132939

17. Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I. C., Lam, D., ... Golub, T. R. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell, 171(6), 1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049

18. Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., van

Dyk, E., Chang, H., de Silva, H., Heyn, H., Deng, X., Egan, R. K., Liu, Q., ... McDermott, U. (2016). A landscape of pharmacogenomic interactions in cancer. Cell, 166(3), 740–754. https://doi.org/10.1016/j.cell.2016.06.017

19. Yuan, Y., & Bar-Joseph, Z. (2019). Deep learning for inferring gene relationships from single-cell expression data. Proceedings of the National Academy of Sciences, 116(52), 27151–27158. https://doi.org/10.1073/pnas.1915536116

20. Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., & Troyanskaya, O. G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature Genetics, 50(8), 1171–1179. https://doi.org/10.1038/s41588-018-0160-6