Multi-Omics Integration for Elucidating Herb–Gut Microbiota Interactions in Traditional Chinese Medicine

Grent Heward
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
grantmail@uc.edu

Pavel Alvaorez
Department of Computer Science, University of Houston, Houston, TX, USA.
pavel.alvarez@uh.edu

Hector L. Watkins
Department of Computer Science, George Mason University, Fairfax, VA, USA.
hectorlwatkins634@gmu.edu

Davide M. Carpenter
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
carpenter1992@oregonstate.edu

Abstract

Traditional Chinese Medicine (TCM) exerts therapeutic effects through multi-component herbal formulations whose activities are mediated by host metabolic and immune networks and by biotransformations carried out by the gut microbiota. Multi-omics integration offers a system-level framework for disentangling these interactions, but the field still lacks robust architectures that connect genomic, transcriptomic, proteomic, metabolomic, and metagenomic measurements with clinically interpretable outcomes. This paper develops a systematic analysis of multi-omics integration for herb-gut microbiota research. It reviews the conceptual foundations of TCM systems pharmacology, the analytical challenges posed by data heterogeneity and scale, and the structural trade-offs in current computational platforms. The discussion emphasizes governance, FAIR data management, reproducibility, fairness across populations and experimental modalities, and sustainable deployment. It further examines how network-based modeling, community detection, and integrative pipelines can link chemical complexity to microbial function. Although multi-omics data generation has accelerated, integrative frameworks often treat TCM and microbiome research as separate domains, leading to fragmented evidence and limited clinical translation. By framing herb-gut microbiota interactions as a socio-technical infrastructure problem rather than a purely molecular puzzle, the paper provides a forward-looking agenda for academic, clinical, and regulatory stakeholders.

Keywords

multi-omics integration; gut microbiota; traditional Chinese medicine; systems pharmacology; network pharmacology; data governance; computational infrastructure.

1. Introduction

The gut microbiome is now recognized as a dynamic metabolic and immunological interface that shapes host responses to diet, environment, and pharmaceuticals [1]. In parallel, multi-omics measurement technologies have enabled simultaneous profiling of host genomes, transcriptomes, proteomes, metabolomes, and microbial communities, generating unprecedented opportunities to understand disease mechanisms at a systems level [2]. These advances are particularly relevant for Traditional Chinese Medicine, which is characterized by chemically complex herbal formulations, context-dependent efficacy, and multi-target modes of action. Network pharmacology has long argued that TCM interventions should be studied as perturbations of biological networks rather than as single ligand-receptor events [3]. However, the gut microbiota introduces another layer of complexity because many herbal constituents are poorly absorbed and undergo microbial transformation before they exert systemic effects [4].

The relationship between TCM herbs and the gut microbiota is bidirectional. Herbal compounds can alter the composition and function of microbial communities, while microbial enzymes can convert glycosides, polyphenols, and alkaloids into active or inactive metabolites. This mutual interaction creates a complex adaptive system in which molecular activity depends on host genotype, diet, microbial ecology, and temporal dynamics. Population-scale metagenomic studies have demonstrated that gut microbial composition varies widely across individuals, geographic regions, and dietary patterns, and that this variation can influence drug metabolism and disease susceptibility [5]. Integrative omics approaches therefore become essential because no single data type can resolve the causal chains that link herbal chemical complexity to microbial transformation and host response [6].

Despite the conceptual appeal of multi-omics integration, the translation of TCM herb-gut microbiota research into reliable mechanistic insight and clinical decision support remains difficult. Data are generated across different laboratories, platforms, sequencing depths, mass spectrometry instruments, and experimental designs. Biological samples from human cohorts, animal models, and in vitro fermentations are often combined without consistent metadata standards. The resulting analytical pipelines may produce statistically significant but biologically fragile associations. Moreover, the sociotechnical infrastructure needed to share, govern, and sustain these data resources is still underdeveloped. This paper therefore adopts a systems perspective that moves beyond individual molecular findings to examine the architectural, infrastructural, and policy conditions under which multi-omics integration can become a dependable knowledge-generating system for TCM and gut microbiome science.

2. Conceptual Foundations and Systems-Level Challenges

Multi-omics integration in herb-gut microbiota research is best understood as a problem of aligning heterogeneous data modalities across multiple levels of biological organization. The host genome provides a stable background that influences enzyme expression, immune function, and disease susceptibility. The transcriptome and proteome reflect dynamic regulatory states. The metabolome captures downstream chemical phenotypes, while metagenomic and metatranscriptomic data describe the taxonomic and functional potential of microbial communities. Statistical methods such as multi-omics factor analysis have been developed to identify latent factors that explain shared variation across these layers, offering a principled route to reduce dimensionality while preserving biological interpretability [7]. However, such methods assume a meaningful common structure across data types, an assumption that may be violated when microbial taxa and host-derived metabolites are governed by fundamentally different timescales and regulatory logics.

A further conceptual challenge is the distinction between correlation and causality in multi-omics studies. Integrative analyses frequently identify associations between microbial taxa, herbal metabolites, and host phenotypes, but these associations do not establish whether microbial transformation is necessary for therapeutic efficacy. Multi-omics integration frameworks therefore need to be embedded within experimental designs that include longitudinal sampling, germ-free models, controlled dietary interventions, and perturbation experiments [8]. Without such designs, integrative models may simply reflect the simultaneous response of host and microbiome to shared environmental factors. This issue is especially acute in TCM research, where herbal preparations are multi-component and treatment regimens are often individualized, making standard randomization and exposure quantification difficult.

The analytical pipeline for microbial community profiling illustrates the broader systems challenge. Metagenomic and amplicon sequencing data require extensive quality control, taxonomic assignment, functional annotation, and normalization before they can be combined with host-derived omics data [9]. Metabolomics data pose additional challenges related to chemical identification, quantification, and the separation of host-derived metabolites from microbially transformed products [10]. The metabolome is particularly important because it sits at the interface between herbal constituents, microbial enzymes, and host receptors. Emerging metabolomics methods can systematically map drug and diet-related metabolites, yet the annotation rate for untargeted metabolomics remains a limiting factor [11]. These technical limitations propagate through integrative models and can distort downstream biological conclusions if not explicitly addressed.

From a systems perspective, multi-omics integration is not a single computational step but a multi-stage socio-technical process. It begins with study design and sample collection, proceeds through molecular profiling and data cleaning, continues with statistical integration and network modeling, and ends with biological interpretation and clinical translation. Each stage involves trade-offs between throughput, accuracy, cost, and interpretability. For example, increasing the number of omics layers may improve coverage but can also inflate the multiple-testing burden and reduce statistical power. Systems-level analysis therefore requires careful attention to the structural properties of the entire pipeline rather than isolated optimization of any single analytical component.

3. Multi-Omics Architecture and Network-Based Integration

The architectural design of a multi-omics integration platform for herb-gut microbiota research must accommodate heterogeneous data types, diverse experimental models, and evolving biological knowledge. One productive direction is the use of network-based representations that treat herbal compounds, microbial species, genes, proteins, and metabolites as nodes in a multi-layer biological network. Community detection and network propagation can then identify modules that link herbal interventions to microbial transformations and host responses. A network community-based model for identifying synergistic combinations from herbal medicines and complex systems has been proposed to capture these higher-order interactions [12]. Within such frameworks, synergy is not evaluated solely through pairwise chemical assays but through the coordinated perturbation of network modules that span host and microbial domains.

Host-gut microbiota metabolic interactions provide a natural substrate for network-based integration because many microbial transformations can be represented as metabolic reactions with defined substrates and products [13]. For instance, the conversion of dietary

polysaccharides into short-chain fatty acids is a well-characterized microbial pathway that influences host energy metabolism and immune regulation. Herbal polysaccharides may undergo similar microbial degradation, generating metabolites that act through distinct host receptors. By integrating metagenomic functional profiles with metabolomic measurements, researchers can begin to associate specific microbial gene families with the production of bioactive herbal metabolites. This systems-level mapping is more informative than taxonomic profiling alone because it emphasizes the functional capacity of the microbiome rather than its species composition.

Dietary patterns exert strong selective pressures on gut microbial communities, and this dietary context is critical for interpreting herb-microbiota interactions. Long-term dietary studies have shown that gut microbial enterotypes are associated with diet and may influence metabolic responses to food and drugs [14]. TCM research must therefore account for dietary variation when modeling herb-microbiota effects, especially because dietary advice is often integrated into TCM treatment. A systems architecture that incorporates dietary metadata, habitual intake records, and microbial community states can improve the robustness of multi-omics models. Without such contextual data, associations between herbal compounds and microbial taxa may be confounded by dietary differences across study populations.

The integration of multi-omics data across host and microbial compartments also raises the problem of data alignment. Host transcriptomic and proteomic measurements are mapped to human reference genomes, whereas microbial sequencing data are mapped to large and rapidly evolving reference catalogs. Metabolomic features, by contrast, are not genome-anchored in the same way and require separate chemical databases for identification. A coherent architecture must therefore include mapping layers that translate between genome-anchored features, chemical features, and clinical phenotypes. This translation requires shared identifiers, semantic ontologies, and standardized exchange formats. Network models can serve as a common formal language because nodes can represent entities from different data domains and edges can represent statistical, metabolic, or regulatory relationships.

The computational trade-offs involved in network-based multi-omics integration are substantial. Large-scale network inference can become computationally expensive when the number of microbial genes, metabolites, and host genes reaches millions. Community detection algorithms may require heuristic approximations that can introduce instability. In addition, many network methods rely on prior knowledge databases that are incomplete or biased toward well-studied organisms and pathways. These limitations do not invalidate network approaches, but they emphasize the need for algorithmic transparency and uncertainty quantification. A systems-level architecture should therefore combine network inference with rigorous validation on independent datasets and experimental perturbations.

4. Governance, Infrastructure, and Reproducibility

The value of multi-omics integration in TCM herb-gut microbiota research depends on the availability of well-governed, interoperable data resources. The FAIR principles for scientific data management provide a foundational framework for making data findable, accessible, interoperable, and reusable [15]. Applying these principles to TCM and microbiome research requires more than depositing raw data in public repositories. It requires consistent metadata schema that capture herbal material identity, processing methods, dosage, collection time, host phenotype, diet, and microbiome sampling protocol. Without this contextual information, multi-omics datasets cannot be meaningfully integrated across institutions or over time.

Computational infrastructure for microbiome data science has advanced substantially through platforms that support reproducible sequence processing, feature table construction, and statistical analysis [16]. These platforms enable researchers to share analysis workflows and versioned data objects, thereby reducing the risk of ad hoc preprocessing decisions. Metabolomics has similarly benefited from transparent web-based platforms that support compound identification, statistical analysis, and pathway mapping [17]. However, the integration of these domain-specific tools remains incomplete. A full multi-omics pipeline for TCM herb-gut microbiota research must connect microbiome processing, metabolomics, host transcriptomics, and clinical metadata within a unified computational environment. This integration imposes significant demands on data models, software interoperability, and computing capacity.

The challenge of big data in biology extends beyond storage volume to include heterogeneity, quality variation, and the complexity of analysis workflows [18]. For TCM research, these challenges are amplified by the need to represent herbal nomenclature, traditional preparation methods, and multi-herb formulations in machine-readable forms. Different names for the same plant species, regional processing variations, and historical text references create semantic ambiguity. The development of standardized terminologies and taxonomies is therefore a governance priority. Such standardization should be open, community-driven, and linked to existing biomedical ontologies so that multi-omics datasets can be integrated across Chinese, international, and cross-domain resources.

Reproducibility is a systemic property rather than an individual virtue. It requires version control for data, code, and reference databases, as well as documented parameter choices and random seeds. In multi-omics integration, small changes in normalization or feature selection can alter downstream biological conclusions. Research infrastructures should therefore support the capture of complete provenance records that link each analytical result to the exact data and code that produced it. This level of traceability is especially important for regulatory acceptance of TCM-derived findings, because evidence may eventually be used to support clinical claims about herbal interventions.

Governance structures must also address data sharing across international borders, particularly because TCM research involves genetic data, microbial genomic data, and potentially identifiable human information. Different jurisdictions have different requirements for informed consent, data protection, and secondary use. A federated architecture in which raw data remain at contributing institutions while analytical queries travel across nodes can reduce privacy risks and support compliance. Federated multi-omics integration, however, introduces additional complexity in model training, harmonization, and validation. The design of federated systems must balance privacy protection with the need for sufficient data access to generate reproducible biological knowledge.

5. Robustness, Fairness, and Sustainability

Robustness in multi-omics integration refers to the ability of an analytical system to produce stable and meaningful results across different datasets, platforms, and populations. Network-based predictions of drug combinations have demonstrated that model performance can be highly sensitive to network completeness and to the choice of validation strategy [19]. In the context of TCM herb-gut microbiota interactions, robustness must be assessed not only on held-out data but also across different microbial contexts and host genetic backgrounds. A model that predicts therapeutic synergy in one cohort may fail in another cohort with different

dietary habits or microbial community states. Cross-cohort validation should therefore be treated as a first-class requirement in the development of integrative models.

Fairness is closely related to robustness but has distinct ethical and scientific dimensions. Multi-omics models trained primarily on data from one population may encode biological assumptions that do not generalize to other populations. Gut microbiome composition is shaped by ancestry, geography, diet, and social determinants of health. If training datasets are biased toward urban or high-resource cohorts, then integrative models may produce less reliable predictions for rural or underrepresented populations that use TCM as a primary healthcare resource. Fairness therefore requires deliberate sampling strategies, transparent reporting of dataset demographics, and the development of transfer learning methods that can adapt models to new populations without requiring complete retraining.

The gut microbiota is also a potential new territory for drug targeting, but its therapeutic manipulation raises questions about ecological stability and unintended consequences [20]. Herbal compounds that selectively promote certain microbial taxa may produce beneficial effects in one host context while causing dysbiosis in another. Multi-omics models should therefore include ecological robustness metrics that assess how predicted interventions might shift microbial community structure over time. Sustainability in this context includes both biological sustainability and infrastructural sustainability. Biologically, interventions should be evaluated for their potential to maintain or restore microbial diversity rather than simply enrich single taxa. Infrastructurally, multi-omics research resources require long-term funding, maintenance, and community governance to remain useful beyond the lifetime of individual grants.

The sustainability of multi-omics research infrastructures depends on standards adoption, open data practices, and incentives for data contributors. Many public repositories suffer from incomplete metadata and uneven data quality, which reduces their long-term value. Research communities should develop lightweight but enforceable metadata standards that capture the minimum information needed for herb-gut microbiota studies. Incentive mechanisms, including data citation and co-authorship policies, can encourage researchers to invest in high-quality data deposition. Without such incentives, the multi-omics ecosystem will remain fragmented and dominated by small, siloed datasets that cannot support robust systems-level inference.

The environmental dimension of sustainability also deserves attention. Large-scale multi-omics data generation and computation consume substantial energy and laboratory resources. Thoughtful experimental design can reduce waste by prioritizing longitudinal sampling, multiplexed sequencing, and targeted metabolomics where appropriate. Systems-level research should therefore consider the full life cycle of data generation, from sample collection through storage and computation, rather than treating data as an unlimited resource. This perspective aligns with broader movements in biomedical research toward responsible data infrastructure and low-carbon scientific computing.

6. Deployment, Translation, and Policy Implications

The clinical translation of multi-omics findings in TCM herb-gut microbiota research requires a deployment architecture that supports decision-making under uncertainty. Unlike single-molecule drug development, TCM interventions are often multi-component and context-dependent, making it difficult to define a single biomarker or target. Multi-omics integration can provide a probabilistic systems view of how an herbal formula is expected to behave in a

given patient, but this view must be integrated with clinical judgment and patient preferences. Decision support tools should therefore present uncertainty transparently and avoid deterministic claims about efficacy or safety.

Deployment across different healthcare systems also raises regulatory questions. Regulatory frameworks for botanical products, dietary supplements, and microbiome-directed interventions vary widely across jurisdictions. Some jurisdictions recognize multi-herb formulations as approved medicines, while others regulate them as dietary supplements with lower evidentiary requirements. Multi-omics data may eventually provide mechanistic evidence that supports regulatory claims, but only if the data are generated under rigorous quality standards and analyzed with validated methods. Regulatory agencies increasingly expect data provenance, reproducibility, and defined analytical pipelines. Multi-omics research communities should engage with regulators early to align standards and avoid producing evidence that is scientifically interesting but not decision-relevant.

The governance of artificial intelligence and machine learning in biomedical research also applies to multi-omics models for TCM. Algorithms that predict herb-microbiota interactions may be used to recommend herbal formulations, adjust dosages, or identify patients who are likely to respond. These models can produce harm if they are deployed without appropriate validation across diverse populations. Policy frameworks should require documentation of model limitations, training data composition, and performance metrics stratified by relevant demographic and clinical variables. This is not merely a technical requirement but a governance mechanism to ensure accountability in complex systems.

Cross-domain comparisons can inform the development of multi-omics deployment strategies. In environmental health, the integration of exposure data, genomic profiles, and metabolomic markers has enabled more nuanced risk assessment by linking external exposures to internal biological responses. In pharmacogenomics, genetic variants are used to predict drug metabolism and guide dosing. TCM herb-gut microbiota research sits at the intersection of these fields because it must account for both host genetic variation and microbial metabolic capacity. A generalizable deployment architecture should therefore support modular integration of different omics layers, allowing clinical sites to adopt only those components that are feasible in their local setting while preserving compatibility with larger research networks.

The long-term policy goal should be the creation of a learning health system for TCM that links multi-omics research, clinical observation, and real-world evidence. In such a system, anonymized patient data, herbal prescription records, microbiome profiles, and clinical outcomes are continuously collected and analyzed to improve treatment strategies. This vision requires careful attention to privacy, consent, and data security. It also requires the development of common data models that can represent traditional diagnostic categories and herbal interventions without losing their semantic richness. Multi-omics integration is one component of this larger socio-technical infrastructure, but its success depends on the surrounding governance and policy environment.

7. Conclusion

Multi-omics integration provides a powerful but demanding framework for elucidating herb-gut microbiota interactions in Traditional Chinese Medicine. The scientific promise lies in the ability to connect the chemical complexity of herbal formulations to microbial transformation and host molecular responses through layered biological data. However, the realization of this

promise depends on more than advanced statistical methods. It requires coherent data architectures, robust governance structures, interoperable standards, reproducible computational workflows, and sustained attention to fairness and population diversity. Network-based models offer a useful formal vocabulary for representing interactions across host and microbial domains, but their value depends on validation in independent cohorts and experimental systems.

The systems perspective developed in this paper emphasizes that multi-omics integration is a socio-technical process embedded in institutional, regulatory, and cultural contexts. Data generation, analysis, and interpretation are shaped by decisions about which populations are studied, which data types are prioritized, and how findings are translated into clinical practice. A narrow focus on algorithmic performance may overlook these structural determinants of success. By treating herb-gut microbiota research as an infrastructure problem, the field can build shared resources that support cumulative knowledge production and responsible translation.

Future work should prioritize the development of federated multi-omics platforms that respect privacy while enabling cross-cohort learning, the establishment of community-driven metadata standards for TCM and microbiome studies, and the integration of ecological and evolutionary perspectives into model development. It should also address the training of interdisciplinary researchers who can move between molecular biology, computational science, clinical practice, and regulatory science. Within this broader agenda, multi-omics integration can become a reliable foundation for understanding the complex therapeutic logic of Traditional Chinese Medicine in a modern biomedical context.

References

1. Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. Nature Medicine, 24(4), 392–400.

2. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome Biology, 18(1), 83.

3. Li, S., Zhang, B., & Zhang, N. (2011). Network target for screening synergistic drug combinations with applications to traditional Chinese medicine. BMC Systems Biology, 5(Suppl 1), S10.

4. Zhao, L., Nicholson, J. K., Lu, A., Wang, Z., Tang, H., Holmes, E., Shen, J., Zhang, X., Lindon, J. C., & Li, S. (2012). Targeting the human genome–microbiome axis for drug discovery: inspirations from global systems biology and traditional Chinese medicine. Journal of Proteome Research, 11(4), 2440–2450.

5. Zhernakova, A., Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A. V., Falony, G., Vieira-Silva, S., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science, 352(6285), 565–569.

6. Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. Nature Reviews Genetics, 19(5), 299–310.

7. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for

unsupervised integration of multi-omics data sets. Molecular Systems Biology, 14(6), e8124.

8. Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. Bioinformatics and Biology Insights, 14, 1177932219899051.

9. Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., & Huttenhower, C. (2015). Sequencing and beyond: integrating molecular omics for microbial community profiling. Nature Reviews Microbiology, 13(6), 360–372.

10. Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., & Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. Trends in Biotechnology, 22(5), 245–252.

11. Wishart, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. Nature Reviews Drug Discovery, 15(7), 473–484.

12. Wang, Y., Yao, J., Sui, Y., Jiang, H., Ma, B., Lai, S., ... & Tan, N. (2026). HerbSyner_Finder: a network community-based model for identifying synergistic combinations from herbal medicines and complex systems. Targetome, 2(2).

13. Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., & Pettersson, S. (2012). Host-gut microbiota metabolic interactions. Science, 336(6086), 1262–1267.

14. Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. Science, 334(6052), 105–108.

15. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018.

16. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology, 37(8), 852–857.

17. Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D. S., & Xia, J. (2018). MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Research, 46(W1), W486–W494.

18. Marx, V. (2013). Biology: The big challenges of big data. Nature, 498(7453), 255–260.

19. Cheng, F., Kovács, I. A., & Barabási, A. L. (2019). Network-based prediction of drug combinations. Nature Communications, 10(1), 1197.

20. Jia, W., Li, H., Zhao, L., & Nicholson, J. K. (2008). Gut microbiota: a potential new territory for drug targeting. Nature Reviews Drug Discovery, 7(2), 123–129.

21. Wang, Y., & Ling, C. (2025). Comparing SAS® and R Approaches in Reshaping data. In PharmaSUG 2025 Conference Proceedings.