Explainable Artificial Intelligence for Mechanism-Based Discovery of Multi-Component Herbal Therapies

Ananya Parekh
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
ananya.work@unr.edu

Abstract

Multi-component herbal therapies present a distinctive challenge for computational discovery because their therapeutic effects are rarely reducible to single molecular targets. Instead, they emerge from complex interactions among phytochemicals, biological pathways, disease states, and patient-specific contexts. Explainable artificial intelligence offers a promising route toward mechanism-based discovery in this domain, yet its application requires careful integration of network pharmacology, systems biology, causal reasoning, and regulatory knowledge. This paper provides a systems-level analysis of explainable artificial intelligence for multi-component herbal therapy discovery. It examines the structural trade-offs between predictive performance and interpretability, the role of network-based representations, the epistemic demands of mechanism-based explanation, and the deployment constraints imposed by clinical and regulatory environments. The discussion emphasizes that explanation in this setting cannot be reduced to feature attribution alone. It must connect model outputs to modular biological mechanisms that can be interrogated, replicated, and evaluated within a governance framework. The paper further considers robustness, fairness, and reproducibility as properties of the broader socio-technical infrastructure rather than as after-the-fact checks. By framing explainable artificial intelligence as part of an integrated discovery system, the analysis clarifies how computational models can support, rather than displace, the interpretive practices of pharmacologists, clinicians, and herbal medicine researchers. The paper concludes with forward-looking observations on the institutional and technical conditions needed for credible translation of multi-component herbal therapies into clinical settings.

Keywords

explainable artificial intelligence, herbal medicine, mechanism-based discovery, network pharmacology, systems biology, causal inference, interpretability, governance.

1. Introduction

The growing interest in artificial intelligence for drug discovery has been driven by rapid advances in deep learning, large-scale molecular data, and computational infrastructure. Deep learning models now achieve impressive predictive performance across many biomedical tasks, including molecular property prediction, drug-target interaction inference, and disease classification [1]. However, predictive accuracy alone is insufficient when the goal is to understand why a biological intervention works, how multiple components interact, and under what conditions a therapeutic effect might generalize. This limitation is particularly acute in the study of multi-component herbal therapies, where the object of inquiry is not a single molecule but a heterogeneous mixture of phytochemicals acting across multiple scales. In

such contexts, black-box predictive models may identify candidate combinations without offering actionable mechanistic insight [2]. Post hoc explanation methods such as Shapley value estimation and local surrogate models can provide feature attributions, but these attributions often remain disconnected from the causal structure of biological systems [3][4].

The conceptual foundation for studying such complex pharmacological mixtures increasingly rests on network pharmacology and network medicine. Network pharmacology treats drug action as a perturbation of molecular interaction networks rather than as a one-to-one binding event [5]. Network medicine extends this perspective to human disease, arguing that pathophenotypes arise from interconnected molecular modules whose disruption or restoration can be understood through network topology [6]. Herbal medicines are particularly well suited to network-based analysis because their constituent compounds frequently modulate multiple targets in a coordinated manner. A systems-level view of herbal formulations therefore aligns naturally with the explanatory ambitions of mechanism-based discovery, but it also introduces considerable modeling complexity [7].

Computational approaches in this space rely on heterogeneous representations, including molecular descriptors, protein-protein interaction networks, pathway annotations, and text-mined knowledge extracted from the biomedical literature. Feature selection methods have long been used to reduce this complexity, but they can obscure the relational structure that is essential for understanding multi-target interventions [8]. More recent techniques use unsupervised language representations to extract latent chemical and biological knowledge from scientific text [9]. Graph neural networks have further enabled learning from molecular graphs and biological networks, creating new opportunities for drug discovery [10]. Network-based machine learning has also been applied in precision oncology, illustrating how graph-theoretic algorithms can inform therapeutic reasoning in complex disease contexts [11].

Mechanism-based discovery, however, demands more than predictive modeling over graph-structured inputs. It requires explicit causal language because mechanisms are fundamentally about how interventions propagate through biological systems. Causal inference provides a formal vocabulary for distinguishing association from causation, for reasoning about confounders, and for predicting the effects of interventions that have not yet been observed [12]. Causal representation learning has recently sought to integrate causal assumptions into deep learning architectures, moving toward models that learn stable, manipulable representations rather than purely statistical patterns [13]. Within medicine, these ideas intersect with a growing literature on the scientific foundations of interpretability and the epistemic risks of relying on explanations that are not themselves validated [14]. Broad surveys of explainable artificial intelligence document a heterogeneous landscape of methods, but they also reveal persistent confusion about what counts as an explanation in high-stakes domains [15]. In medical applications, explainability is not merely a technical convenience; it is entangled with clinical trust, professional accountability, and institutional decision-making [16].

This paper advances a systems-level argument for the role of explainable artificial intelligence in mechanism-based discovery of multi-component herbal therapies. It does not propose a single algorithm or model architecture. Instead, it analyzes the structural conditions under which computational explanations can become biologically meaningful, clinically useful, and institutionally credible. The subsequent sections examine conceptual foundations, representational choices, causal reasoning, system architecture, deployment governance, and robustness. Throughout, the focus is on trade-offs rather than singular solutions.

2. Conceptual Framing: Herbal Therapies as Complex Systems

Herbal therapies challenge the reductionist assumptions that have historically guided pharmaceutical discovery. Conventional drug development often proceeds from a clear molecular target, a well-characterized ligand, and a relatively stable pathway model. Multi-component herbal formulations violate each of these assumptions. A single herbal formula may contain dozens or hundreds of phytochemicals, each with different pharmacokinetic properties, target affinities, and tissue distributions. When such a formula is administered, the observed therapeutic effect is not simply the sum of independent molecular actions. It is an emergent property of a multiscale interaction system involving absorption, metabolism, gene regulation, protein signaling, cellular phenotypes, tissue physiology, and host-microbe interactions.

The systems perspective thus becomes an analytical necessity rather than a rhetorical preference. Network pharmacology reframes therapeutic action as a perturbation that propagates across a biological network [5]. The same compound may be promiscuous in beneficial ways, modulating several disease-associated proteins simultaneously. Conversely, a combination may produce synergy through complementary actions on different nodes within a pathway module. Network medicine emphasizes that diseases themselves are rarely localized to a single gene or protein. They emerge from disruptions in modular network neighborhoods, and therapeutic interventions can be understood as attempts to restore network function [6]. Herbal medicines, with their inherent multi-target character, represent an especially instructive case for this perspective [7].

However, the systems framing also raises difficult epistemic questions. A network model can identify a set of interacting nodes, but it does not automatically explain the mechanism by which a clinical effect emerges. Mechanism-based discovery requires identifying the causal chain or network of events that connects an intervention to an outcome. In herbal therapy research, this chain may be plastic, context-dependent, and only partially observable. The same formula may act differently depending on the disease stage, genetic background, diet, microbiome, and co-administered drugs. A robust explanatory framework must therefore tolerate uncertainty while still offering actionable insight.

The concept of synergy is central to multi-component herbal therapy. Synergy implies that the whole is greater than the sum of its parts, but this assertion can be made only against a formal baseline model of additivity or independence. Without such a baseline, claims of synergy are often rhetorical. Computational models must specify what counts as an expected additive effect and what constitutes a meaningful departure from that expectation. This is not merely a statistical problem; it is a modeling problem that involves assumptions about molecular independence, pathway crosstalk, and biological redundancy. The interpretation of departures from additivity must be tied to mechanistic hypotheses that can be tested in follow-up experiments.

Herbal therapies also operate within a distinctive evidentiary tradition. Many herbal formulations have long histories of empirical use, but this historical evidence is heterogeneous in quality, documentation, and standardization. Translating such evidence into computational representations requires careful attention to provenance, nomenclature, and chemical composition. Different geographic regions may use different plant species under similar names, and the same species may exhibit chemical variability due to cultivation, harvest time, and processing methods. These issues are not peripheral. They determine

whether a computational model is trained on coherent biological objects or on noisy labels that undermine mechanism-based inference.

3. Explainable AI in Biomedical Discovery: Architectures and Epistemic Requirements

Explainability in artificial intelligence is not a single property but a family of epistemic objectives. A model may be considered explainable if it is inherently transparent, if it can be approximated by a simpler surrogate, if its decisions can be attributed to input features, or if it supports counterfactual reasoning about alternative inputs. The appropriate form of explanation depends on the audience, the decision context, and the downstream action. In biomedical discovery, explanations are often intermediate products that guide experimental design. They are valuable not because they satisfy a human observer at the moment of prediction, but because they generate hypotheses that can be tested in independent systems.

Post hoc explanation methods such as local interpretable model-agnostic explanations and Shapley value attribution provide a convenient layer of interpretability over complex models [3][4]. These methods can reveal which features influence a prediction, but they do not necessarily reveal the biological mechanism through which those features produce an outcome. In herbal therapy discovery, a feature attribution might highlight a particular compound or target as important to a prediction. Yet this attribution remains silent about whether the compound acts upstream or downstream of a disease module, whether its effect depends on another compound, or whether the apparent importance is an artifact of correlated features.

A deeper challenge is that many popular explanation methods assume a relatively simple mapping between input features and model output. Biological systems violate this assumption through nonlinear interactions, feedback loops, and redundancy. Herbal interventions often depend on the simultaneous presence of multiple compounds, which means that individual feature attributions can be misleading. A compound may appear unimportant when evaluated alone but become important in combination with others. This implies that explanation methods for multi-component therapies must be sensitive to interactions and modular structures rather than isolated feature importances [2][8].

Recent work on graph neural networks has expanded the representational capacity of biomedical models by allowing them to learn directly from molecular graphs and biological networks [10]. Such models can capture relational inductive biases that are absent from fixed feature vectors. For herbal therapies, graph-based representations are attractive because they align naturally with the multi-target structure of phytochemical action. However, graph neural networks are often as opaque as other deep learning architectures. Their explanations require specialized methods that identify relevant subgraphs, node neighborhoods, and message-passing pathways. These explanations can be more mechanistically suggestive than feature attributions, but they still require validation against established biological knowledge.

The epistemic requirements of biomedical explanation go beyond technical interpretability. Doshi-Velez and Kim argued for a rigorous science of interpretability in which explanation methods are evaluated according to their ability to support meaningful tasks, rather than according to subjective notions of transparency [14]. In the context of herbal therapy discovery, the relevant tasks include identifying plausible mechanisms, prioritizing candidates for experimental validation, detecting confounded predictions, and communicating uncertainty to domain experts. A broad survey of explainable artificial intelligence methods reveals that many approaches are assessed on proxy metrics that do not capture these

downstream epistemic functions [15]. Moreover, medical applications introduce institutional and ethical considerations that shape what counts as an acceptable explanation [16].

An important architectural distinction concerns intrinsically interpretable models versus post hoc explanation of opaque models. Intrinsically interpretable models, such as sparse linear models, decision trees, or attention-based architectures with constrained structure, may sacrifice predictive performance but offer direct access to their reasoning. Opaque models may provide superior generalization in large data regimes but require additional explanation machinery. In many biomedical settings, the trade-off is not absolute. The right choice depends on whether the model is intended primarily for candidate generation, for mechanistic inference, or for clinical decision support. Mechanism-based discovery often favors architectures that can be more easily mapped onto biological modularity, even if this entails some loss of predictive accuracy.

4. Representing Multi-Component Herbal Interventions: Data, Ontologies, and Network Pharmacology

The quality of any computational explanation depends on the underlying representation. Multi-component herbal therapies require representations that span chemical structures, molecular targets, pathway annotations, disease phenotypes, and clinical contexts. This heterogeneity poses substantial integration problems. Phytochemical databases must be linked to protein interaction networks through compound-target associations that are often incomplete or uncertain. Pathway databases impose their own ontological choices about what counts as a pathway, and these choices may not align with the modular organization of disease processes. Herbal medicine additionally requires careful handling of plant taxonomy, extraction methods, and formulation composition, which are not always represented in standard pharmacological databases.

Network pharmacology provides a natural organizing framework for these heterogeneous data sources. By representing compounds, targets, pathways, and diseases as nodes in a multilayer network, researchers can apply graph-theoretic methods to identify candidate mechanisms and combinations [5][6][7]. However, the construction of such networks requires many inferential steps. Compound-target interactions may be predicted computationally, extracted from the literature, or measured experimentally. Each source has distinct error profiles and biases. Text-mined knowledge can introduce false positives and fail to capture negative results. Experimental binding data may be reliable for a narrow set of targets but unrepresentative of the full biological context. These uncertainties propagate through the network and influence downstream mechanistic claims.

Feature selection and dimensionality reduction are often necessary when working with high-dimensional molecular and network data [8]. Yet these techniques can disrupt the relational structure that is essential for understanding combinatorial effects. For example, selecting the most predictive individual features may discard exactly those variables whose joint activity defines synergy. Alternatively, embedding methods can preserve relational information while making it less directly interpretable. Unsupervised representation learning from scientific corpora has shown that latent vectors can capture chemically and biologically meaningful regularities [9]. Such representations can support retrieval of related compounds and targets, but they do not by themselves yield mechanism-level explanations. They must be connected to explicit network structures and causal models.

Graph neural networks have been proposed as a way to learn from molecular graphs and biological networks simultaneously [10]. These models can propagate information across molecular and network neighborhoods, enabling the integration of chemical structure and systems-level context. In principle, a graph model trained on multi-component interventions could learn to identify subgraphs associated with synergy or antagonism. In practice, the interpretation of such models requires careful attention to the learned embeddings and attention weights. Network-based machine learning has demonstrated value in precision oncology, where patient-specific molecular networks guide therapeutic decision-making [11]. Similar approaches can be adapted to herbal medicine, although the chemical and botanical complexity introduces additional representational challenges.

A key issue is the alignment between computational modules and biological modules. Mechanism-based discovery assumes that the system can be decomposed into relatively stable functional units, such as protein complexes, signaling cascades, or metabolic modules. Network community detection methods can identify such units, but the resulting partitions may be sensitive to network density, edge confidence, and algorithm parameters. Community boundaries are not natural kinds; they are model-dependent constructs that must be evaluated against biological evidence. Recent work on network community-based models illustrates both the promise and the difficulty of identifying synergistic herb combinations through modular decomposition [17]. Although such models can generate plausible candidates, their explanatory value depends on whether the detected communities correspond to causal functional units rather than to artifacts of data coverage.

Transcriptomic profiling adds another layer of mechanistic evidence. The Connectivity Map established a framework for connecting gene-expression signatures across small molecules, genes, and disease states [18]. Similar logic can be used to compare herbal formulations with known perturbagens, thereby generating hypotheses about shared mechanisms. However, transcriptomic signatures are context-dependent. They capture cellular states at a particular time point and under particular experimental conditions. Extrapolating from such signatures to whole-organism therapeutic effects requires integrative reasoning across pharmacokinetics, tissue distribution, and physiological feedback.

5. Mechanism-Based Inference Under Uncertainty: Causal Reasoning and Modular Explanations

Mechanism-based discovery is a causal enterprise. To claim that a herbal therapy works through a particular mechanism is to claim that specific components, targets, and pathways stand in a causal relation to the observed outcome. Causal inference supplies the formal tools needed to reason about such relations, including graphical models, interventions, and counterfactuals [12]. These tools allow researchers to distinguish between variables that are merely associated with an outcome and variables that are causally responsible for it. In observational biomedical data, this distinction is essential because confounding is pervasive. Disease severity, comorbidity, demographic factors, and treatment history can all distort naive associations.

Causal representation learning seeks to integrate causal structure into the representations learned by deep models [13]. The goal is to construct latent variables that correspond to stable causal factors rather than to statistical artifacts. For herbal therapy discovery, this would mean representing compounds and targets in terms of their capacity to produce specific interventions in biological systems. Such representations could support more reliable generalization across experimental contexts and patient populations. However, causal

representation learning remains an open research area, and its application to complex mixtures is far from mature.

Modularity is central to both causal reasoning and mechanism-based explanation. A mechanism can often be decomposed into a sequence or network of modular processes, each of which can be independently manipulated. In herbal medicine, modules might include absorption of a phytochemical, binding to a target, activation of a signaling pathway, and downstream gene expression changes. A credible explanation identifies not only the overall input-output relation but also the intermediate modules that mediate it. This is why purely predictive models are insufficient: they may reproduce the input-output mapping without representing the mediating processes.

The identification of modular mechanisms from observational and interventional data is challenging because biological modules are not directly observed. They must be inferred from networks whose edges and nodes are themselves uncertain. Community detection in biological networks can suggest candidate modules, but it cannot establish that these modules are causally coherent [17]. Additional evidence is needed, such as experimental perturbations, genetic knockdowns, or pharmacological interventions that isolate specific modules. Network-based pharmacological approaches have begun to integrate such evidence, but the integration remains partial [20]. The causal coherence of a module must be treated as a hypothesis to be tested, not a conclusion to be assumed.

Uncertainty quantification is another essential component of mechanism-based inference. A model that predicts a synergistic combination without reporting uncertainty offers little guidance for experimental follow-up. Uncertainty can arise from data sparsity, measurement error, model misspecification, and biological variability. In herbal therapy research, uncertainty is compounded by chemical variability in plant materials and by incomplete annotation of phytochemical targets. Explanation methods should therefore communicate not only which mechanism is most likely but also how sensitive that conclusion is to plausible perturbations of the model and data. This requirement aligns with broader calls for rigorous interpretability in high-stakes domains [14][16].

Counterfactual reasoning is particularly valuable for mechanism-based discovery because it aligns with experimental thinking. Researchers often ask what would happen if a particular component were removed, if a target were inhibited, or if a pathway were constitutively active. Counterfactual explanations can formalize such questions and identify the minimal changes that would alter the predicted outcome. In multi-component herbal therapies, counterfactual analysis can support formulation design by revealing which components are necessary, which are redundant, and which are synergistic. However, counterfactual predictions are only as reliable as the causal model that supports them. Without validated causal structure, counterfactual statements can be misleading.

6. System Architecture for Explainable Discovery Platforms

A credible explainable discovery platform for multi-component herbal therapies requires more than a set of models. It requires an integrated system architecture that coordinates data ingestion, knowledge representation, predictive modeling, explanation generation, and experimental feedback. The architecture must be designed to support iterative refinement of mechanistic hypotheses. It should not treat explanation as a terminal output but as an intermediate artifact that enters into dialogue with experimental researchers and clinical experts.

The data layer of such a platform must accommodate chemical structures, botanical sources, formulation metadata, target interactions, pathway annotations, transcriptomic profiles, and clinical observations. These data types have different formats, error profiles, and update cycles. The architecture should therefore include robust provenance tracking and versioning. Without provenance, it is impossible to determine whether a mechanistic claim depends on a particular database version, a specific experimental platform, or a text-mining pipeline. Reproducibility depends on the ability to reconstruct the exact data environment in which a model was trained and evaluated.

The modeling layer should include a portfolio of approaches rather than a single monolithic model. Intrinsically interpretable models can provide baseline mechanistic hypotheses, while graph neural networks can capture relational structure and support candidate generation [10]. Feature selection and embedding methods can manage dimensionality while preserving relevant biological signal [8][9]. Causal models can formalize the difference between associative and interventional claims [12][13]. The outputs of these models should be integrated within a common representation that allows cross-model comparison. A discovery platform that cannot compare explanations across different model families will struggle to distinguish robust mechanistic signals from model-specific artifacts.

The explanation layer should produce not only feature attributions but also network-level explanations, counterfactual scenarios, and uncertainty estimates. Network-level explanations are especially important for herbal therapies because therapeutic effects may depend on coordinated activity across multiple targets and pathways. Such explanations might take the form of subgraphs that connect a formulation to a disease outcome through intermediate pathway nodes. Counterfactual scenarios can identify which components or targets are most causally important. Uncertainty estimates can inform prioritization by flagging predictions that rest on weak evidence. These explanation types are complementary, and their relative value depends on the stage of discovery.

The human-in-the-loop dimension is crucial. Explanation is a communicative act between a computational system and a human researcher. The platform must therefore support interactive exploration, allowing researchers to query why a particular combination was predicted, how the prediction changes when inputs are perturbed, and what evidence supports each step in the proposed mechanism. This requirement has architectural consequences. The system must maintain traceable links between raw data, processed representations, model outputs, and explanatory claims. It must also support auditability, so that a third party can evaluate whether an explanation is grounded in evidence or is merely a plausible narrative generated after the fact.

Transcriptomic profiling and chemogenomic data can enhance the mechanistic grounding of such platforms [18]. By comparing herbal formulation signatures to reference perturbagens, researchers can generate hypotheses about shared modes of action. However, transcriptomic comparisons must be interpreted carefully because signatures are context-dependent. The system should therefore encode metadata about cell lines, time points, and experimental conditions. Network-based pharmacological frameworks can help integrate these diverse evidence types into a coherent picture [20]. Yet any integration must remain transparent about its assumptions and limitations.

7. Deployment and Translational Governance

The deployment of explainable artificial intelligence in herbal therapy discovery occurs within a broader governance environment that includes regulatory agencies, clinical institutions, ethics review boards, and public health systems. In many jurisdictions, medical artificial intelligence is subject to emerging requirements for transparency, accountability, and human oversight. The European Union's regulatory discussions around algorithmic decision-making have highlighted the tension between automated inference and individual rights to explanation [21]. Although these regulations do not directly resolve scientific questions about mechanism discovery, they shape the institutional expectations that a discovery platform must satisfy if its outputs are to influence clinical practice.

Machine learning in medicine has achieved notable successes in image analysis, risk prediction, and molecular classification [22]. However, translating these successes into clinical decision support for multi-component herbal therapies faces additional obstacles. Herbal formulations are often regulated as dietary supplements or traditional medicines rather than as conventional pharmaceuticals. Their chemical composition may vary across manufacturers and batches. Clinical evidence may be heterogeneous, and standardization may be incomplete. In this setting, an explainable discovery system must be especially attentive to the limits of its training data and the generalizability of its predictions. A model trained on one formulation cannot be assumed to apply to a nominally similar formulation without chemical verification.

Explainability in clinical deployment is not solely about satisfying regulatory requirements. It is also about supporting professional judgment. Clinicians need to understand enough about a model's reasoning to decide when to trust it, when to override it, and when to seek additional evidence. This requires explanations that are not merely technically faithful but also clinically meaningful. A list of important targets is less useful than a mechanistic narrative that situates those targets within disease pathophysiology and treatment context. The interdisciplinary literature on explainability in healthcare emphasizes that explanations must be designed with attention to the cognitive and institutional roles of their users [19].

Governance also includes questions of data governance and intellectual property. Herbal knowledge is often embedded in traditional medical systems and community practices. Computational discovery platforms that mine this knowledge must consider issues of prior informed consent, benefit sharing, and cultural authority. Failure to address these issues can undermine the legitimacy of the entire enterprise, even if the underlying models are technically sound. A systems-level perspective therefore treats governance not as a legal afterthought but as a design constraint that shapes data collection, model development, and dissemination practices.

The translational pathway for computational herbal discovery likely requires staged validation. Early-stage candidate generation can be relatively permissive, producing mechanistic hypotheses for experimental follow-up. Mid-stage validation should involve independent experimental systems, such as cell-based assays, organoids, or animal models, to test the predicted mechanisms. Late-stage translation requires clinical evidence that is commensurate with the claims being made. Explainable models can support each stage by making their assumptions explicit and by generating testable predictions. However, they cannot substitute for experimental and clinical validation.

8. Robustness, Fairness, and Reproducibility in Herbal AI Systems

Robustness in herbal artificial intelligence systems is not merely a matter of adversarial attacks or distributional shift. It also involves chemical and biological variability. A model trained on a particular herbal formulation may fail when applied to a different preparation of the same plant because the chemical composition has changed. Robustness therefore requires explicit modeling of composition variability, extraction methods, and batch effects. It also requires stress-testing models across plausible data perturbations rather than assuming that performance on a fixed benchmark will transfer to new settings.

Fairness is an underappreciated dimension of herbal therapy discovery. Training data may overrepresent certain geographic regions, plant species, disease populations, or clinical settings. If these biases are not addressed, the resulting models may systematically overlook effective therapies from underrepresented traditions or may generate predictions that are poorly calibrated for diverse patient groups. Fairness in this context is not simply about demographic parity in algorithmic decisions. It is about epistemic equity in the generation and use of biomedical knowledge. Herbal medicine traditions from different parts of the world should not be rendered invisible by data availability asymmetries.

Reproducibility is a persistent challenge in artificial intelligence research, and it is amplified in interdisciplinary domains where data are heterogeneous and pipelines are complex. A mechanism-based discovery claim should be reproducible in the sense that independent researchers, using the same data and code, can obtain comparable results. Yet reproducibility alone is insufficient. The biological implication of a computational result may depend on database versions, annotation choices, and preprocessing decisions. Reproducible workflows must therefore capture not only code but also the data environment and the interpretive assumptions that inform model construction.

The interpretability of a model can itself become a source of false confidence. A visually compelling explanation, such as a highlighted subgraph or a ranked list of targets, may appear more definitive than it actually is. This risk is especially pronounced when explanations are generated post hoc without validation. The literature on explainable artificial intelligence has repeatedly cautioned that explanation methods can produce plausible but misleading rationales [2][15]. In medical contexts, the consequences of such misleading explanations can be significant because they may shape subsequent experiments, funding decisions, and clinical judgments [16]. Mechanism-based discovery therefore requires a culture of explanation validation, in which explanatory claims are treated as hypotheses subject to independent test.

The institutional infrastructure for herbal artificial intelligence must include mechanisms for ongoing monitoring and model updating. Biological knowledge is not static, and neither are the datasets that encode it. A discovery platform that is not updated will gradually diverge from current evidence. At the same time, frequent updates can undermine reproducibility by making it difficult to identify which model version generated a particular claim. Versioning, documentation, and audit trails are therefore essential. These practices support accountability and enable post hoc analysis of model failures and successes.

Cross-domain comparison offers useful lessons. In precision oncology, network-based machine learning has shown that graph-theoretic approaches can generate clinically relevant insights when they are embedded within an ecosystem of experimental validation and clinical expertise [11]. Similar ecosystems are needed for herbal medicine, but they must be adapted to the distinctive evidentiary and cultural features of this domain. The goal is not to force

herbal medicine into the template of conventional drug discovery, but to build computational infrastructures that respect its complexity while enabling rigorous mechanistic inquiry.

9. Conclusion

Explainable artificial intelligence has the potential to transform the discovery of multi-component herbal therapies by connecting predictive models to mechanistic hypotheses. However, this potential cannot be realized by treating explainability as a superficial layer added to opaque models. Mechanism-based discovery requires a deeper integration of network pharmacology, causal reasoning, modular decomposition, and experimental validation. It demands representations that preserve relational structure, explanations that capture interactions and uncertainty, and architectures that support iterative human-in-the-loop inquiry.

This paper has argued that the central challenge is not solely technical. It is also epistemic and institutional. The structural trade-offs between predictive accuracy and interpretability must be evaluated in relation to the goals of discovery. Feature attributions may be useful for some tasks, but they are insufficient for understanding synergy, redundancy, and causal mediation in multi-component interventions. Network-level explanations, counterfactual analyses, and causal models offer more appropriate resources for mechanism-based reasoning, but they bring their own assumptions and uncertainties.

The deployment of such systems must be governed by principles of transparency, reproducibility, fairness, and accountability. Herbal therapies occupy a distinctive position within medicine, combining empirical traditions with modern molecular research. Computational discovery platforms must navigate this complexity without erasing the epistemic contributions of traditional knowledge systems. They must also remain humble about the limits of their training data and the generalizability of their predictions.

Future work should develop integrated platforms that couple graph-based learning, causal modeling, and explanation generation with experimental feedback loops. These platforms should be evaluated not only by their predictive performance but also by their ability to produce mechanistic hypotheses that survive independent validation. In this way, explainable artificial intelligence can become a genuine partner in the discovery of multi-component herbal therapies, supporting the kind of rigorous, transparent, and cumulative inquiry that the field requires.

References

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
2. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1-42.
3. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765-4774).
4. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144).
5. Hopkins, A. L. (2008). Network pharmacology: The next paradigm in drug discovery. Nature Chemical Biology, 4(11), 682-690.

6. Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. Nature Reviews Genetics, 12(1), 56-68.

7. Li, S., & Zhang, B. (2013). Traditional Chinese medicine network pharmacology: Theory, methodology and application. Chinese Journal of Natural Medicines, 11(2), 110-120.

8. Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. In Data Classification: Algorithms and Applications (pp. 37-64). CRC Press.

9. Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature, 571(7763), 95-98.

10. Sun, M., Zhao, S., Gilvary, C., Elemento, O., Zhou, J., & Wang, F. (2020). Graph convolutional networks for computational drug development and discovery. Briefings in Bioinformatics, 21(3), 919-935.

11. Zhang, W., Chien, J., Yong, J., & Kuang, R. (2017). Network-based machine learning and graph theory algorithms for precision oncology. npj Precision Oncology, 1(1), 25.

12. Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.

13. Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612-634.

14. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

15. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey of explainable artificial intelligence (XAI). IEEE Access, 6, 52138-52160.

16. Ratti, E., & Graves, M. (2022). Explainable machine learning practices: Opening another black box for reliable medical AI. AI and Ethics, 2(4), 801-814.

17. 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).

18. Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., ... & Golub, T. R. (2006). The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. Science, 313(5795), 1929-1935.

19. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 310.

20. Boezio, B., Audouze, K., Ducrot, P., & Taboureau, O. (2017). Network-based approaches in pharmacology. Molecular Informatics, 36(10), 1700048.

21. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38(3), 50-57.

22. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.