Knowledge Graph Construction for Systematic Discovery of Bioactive Compounds and Therapeutic Targets in Natural Products

Akash A. Melhetria
Department of Computer Science, George Mason University, Fairfax, VA, USA.
malhotra554@gmu.edu

Jeramy J. Beiley
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
jeremywork@unh.edu

Abstract

The discovery of bioactive compounds and therapeutic targets from natural products remains a central challenge in biomedical research because the relevant evidence is scattered across heterogeneous chemical, genomic, pharmacological, and clinical data sources. This paper examines the construction and use of knowledge graphs as an integrative systems infrastructure for systematic discovery in natural product research. Rather than proposing a single algorithm, the paper develops a system-level account of how knowledge graph architectures can reconcile chemical structure, species provenance, bioactivity, protein interactions, pathway membership, disease associations, and phenotypic outcomes within one queryable semantic fabric. We analyze the structural trade-offs among schema-centric, property-graph, and linked-data designs, and we discuss entity resolution, relation extraction, provenance modeling, and quality assurance as governance challenges rather than purely technical tasks. The paper further addresses scalability, updating, deployment, fairness, and regulatory implications when such graphs support target prioritization and compound repurposing. Throughout, we emphasize that the value of a knowledge graph lies not only in inference accuracy but also in its capacity to make evidence transparent, contestable, and maintainable across research communities. We argue that sustainable knowledge graph infrastructures for natural products require a balance between automated extraction and expert curation, between open interoperability and local control, and between exploratory discovery and rigorous validation. The discussion offers forward-looking perspectives on how these systems can support more reproducible, equitable, and policy-aware natural product drug discovery.

Keywords

knowledge graphs; natural products; bioactive compounds; therapeutic target discovery; data integration; semantic interoperability; fairness; infrastructure.

1. Introduction

Natural products have long supplied molecules with therapeutic value, from antibiotics to anticancer agents, and they continue to offer chemical scaffolds that are often underrepresented in synthetic compound libraries. However, the evidence required to connect a natural compound to a bioactive mechanism and a therapeutic target is scattered across chemical databases, genomic resources, pharmacological assays, ethnobotanical records, and

clinical literature. This fragmentation creates a systems-level problem: individual data sources may be reliable in isolation, yet their combined interpretation demands reconciliation of identifiers, semantics, experimental contexts, and confidence levels. The Semantic Web vision of machine-interpretable linked data [1] offers a conceptual foundation for addressing this fragmentation, because it shifts the integration burden from ad hoc pipelines to shared graph-based vocabularies and resolvable identifiers.

A knowledge graph represents entities and their typed relationships in a graph structure, enabling both query-time traversal and learning-time feature construction. In the natural products domain, entities may include compounds, plant or microbial sources, proteins, genes, pathways, diseases, assays, publications, and institutions. Relations may encode structural similarity, bioactivity measurements, protein binding, pathway membership, disease association, geographic origin, and provenance. The linked-data principles articulated for open knowledge integration [2] emphasize the use of resolvable identifiers and standardized vocabularies, which are especially important when knowledge graphs span multiple scientific communities with different naming practices. Without such principles, a natural product knowledge graph may become another silo rather than an integrative substrate.

This paper provides a system-level analysis of knowledge graph construction for discovering bioactive compounds and therapeutic targets in natural products. We do not focus on a single extraction algorithm or a particular predictive model. Instead, we examine the architectural choices, data governance arrangements, inference methods, and deployment conditions that shape whether such a graph can support robust discovery. The central argument is that knowledge graph construction is simultaneously an engineering, ontological, and institutional problem. Technical performance measures such as link prediction accuracy are necessary but insufficient; the system must also preserve provenance, expose uncertainty, accommodate disagreement, and remain maintainable as source databases evolve.

The remainder of the paper is organized as follows. Section 2 reviews related work and background systems. Section 3 presents a conceptual architecture for natural product knowledge graphs. Section 4 discusses data acquisition and normalization. Section 5 addresses entity and relation extraction. Section 6 examines graph representation and semantic integration. Section 7 analyzes systematic discovery of bioactive compounds and therapeutic targets. Section 8 considers governance, robustness, fairness, and quality assurance. Section 9 discusses deployment, sustainability, and policy implications. Section 10 concludes.

2. Background and Related Work

Several large-scale resources have shaped the data landscape for biomedical knowledge graphs. DrugBank [3] provides structured information on approved and investigational drugs, including chemical structures, targets, transporters, and pharmacokinetic properties. KEGG [4] contributes pathway, disease, and drug records with strong cross-organism coverage, making it widely used for interpreting genomic and metabolomic experiments. These resources are comparatively drug-oriented, however, and their coverage of natural products can be uneven. Integrating them into a natural product knowledge graph requires mapping their identifiers to compound structures and biological entities without losing the context of their original curation.

Natural product research has benefited from dedicated databases and chemistry resources. Protein interaction networks such as STRING [5] provide a broad map of functional

associations that can contextualize compound-target relationships. NPASS [6] was developed to provide natural product activity and species source data, supporting the exploration of traditionally used organisms and their constituent molecules. ChEMBL [7] aggregates bioactivity data from medicinal chemistry literature, including many natural product-derived compounds, and thereby offers dose-response measurements and target annotations that are useful for supervised relation extraction. ChEBI [8] supplies a controlled vocabulary for chemical entities of biological interest, enabling hierarchical classification and the alignment of synonyms across resources. Together these sources offer complementary coverage, but they differ in identifier systems, curation standards, provenance granularity, and update cycles.

Early knowledge graph applications in biomedicine often emphasized link prediction and drug repurposing. Those efforts showed that graph-based representations can reveal indirect paths between compounds, genes, and diseases that are not explicit in any single database. However, natural products introduce additional complexity because their source organisms, extraction methods, biosynthetic contexts, and traditional usage patterns are part of the evidence. A graph that omits these contextual dimensions may perform well on benchmark tasks while remaining difficult to interrogate by natural product chemists or pharmacognosy researchers. Consequently, the construction of a natural product knowledge graph must be treated as a socio-technical design problem rather than a simple pipeline.

3. Conceptual Architecture for Natural Product Knowledge Graphs

A knowledge graph for natural product discovery must balance expressiveness, query performance, and maintainability. One architectural option is a schema-centric design, in which an ontology defines the allowed entity types and relation types before ingestion. This approach supports strong semantic validation and reduces the chance of inconsistent edges, but it can be brittle when new source data introduce relations that were not anticipated. A property-graph design, by contrast, permits flexible node and edge properties while retaining a smaller core schema. This flexibility accelerates ingestion from multiple natural product sources, but it may shift the burden of semantic consistency to application code and downstream consumers. A linked-data architecture emphasizes resolvable identifiers and shared vocabularies, which facilitates federation across institutional boundaries but can be more difficult to optimize for high-throughput analytical queries.

The entity model typically includes compounds, natural product extracts, source organisms, collection events, proteins, genes, pathways, diseases, assays, publications, and institutions. The relation model must distinguish chemical similarity from bioactivity, and direct binding from pathway modulation, because these distinctions carry different epistemic weight. A particularly important design decision is whether the graph treats a natural product extract as a first-class entity separate from its constituent compounds. If extracts are collapsed into compound sets, the graph may lose information about polypharmacology and traditional formulation effects. If extracts are retained as separate nodes, the graph can represent complex mixtures but must also manage ambiguous composition and batch variation.

Provenance is also an architectural concern. Every edge in a natural product knowledge graph should ideally carry evidence about its source, extraction method, experimental assay, publication, and confidence. This requirement influences the storage layer because provenance properties can be large and highly repetitive. Some architectures attach provenance directly to edges, while others reify edges as nodes so that multiple evidence records can point to the same claim. The reification approach is more expressive and supports contested or conflicting assertions, but it increases graph size and query complexity. The

choice between these models determines whether the system can cleanly represent disagreement, uncertainty, and negative results, all of which are common in natural product research.

Another structural trade-off concerns centralization. A single centralized graph simplifies consistency enforcement and versioning, but it may not reflect the distributed governance of natural product data. A federated architecture allows different communities to maintain their own source graphs and exposes them through shared interfaces, yet it complicates global reasoning and can introduce latency and partial availability. In practice, many sustainable systems adopt a hybrid model in which a core graph is centrally maintained for high-value entities, while external linked resources are accessed on demand for specialized queries. Such a design supports institutional participation while retaining a coherent analytical core for machine learning and systematic discovery.

4. Data Acquisition and Normalization

Data acquisition for natural product knowledge graphs begins with an inventory of source databases, publication repositories, and experimental data collections. The sources vary widely in their licensing, update frequency, download formats, and degree of structured annotation. Some resources provide bulk downloads and stable identifiers, while others require web services or text extraction from supplementary materials. A robust acquisition layer musthandle versioned snapshots, incremental updates, and license-compliant redistribution. Because source databases change frequently, the acquisition layer should treat each import as an event with a timestamp and a source version. This event log supports reproducibility and allows the graph to be rolled back when a source release introduces incompatible identifiers or erroneous annotations. Normalization begins with identifier resolution across chemical databases. A compound may appear under multiple synonyms, CAS numbers, InChI strings, and vendor identifiers. Canonicalization through structural representations available in PubChem [9] reduces redundancy but must not erase legitimate distinctions between stereoisomers or salt forms, which can have different biological activities. Taxonomic names require reconciliation against authoritative botanical and microbial nomenclature because the same species may be cited under outdated synonyms or spelling variants. Protein and gene identifiers demand genome-build awareness, since target identity can shift when reference assemblies change. Assay-based bioactivity records require normalization of dose, time, cell line, and assay technology, because raw potency values alone do not capture the conditions under which a compound-target relation holds.

Disease mentions require mapping to structured vocabularies such as the Disease Ontology [11] and phenotype ontologies [12] to align clinical descriptions with experimental models. This mapping is difficult when source literature uses symptom-level language rather than formal diagnostic categories. For example, a plant extract may be described as anti-inflammatory without specifying a molecular mediator. A knowledge graph can preserve both the original textual assertion and its normalized interpretation, allowing users to trace the transformation from source claim to structured edge. This traceability is essential because normalization often involves interpretative choices. Different curators may map the same phrase to different disease classes, and the graph should record these alternatives rather than selecting a single default. Such an approach increases the size of the graph but substantially improves its epistemic transparency.

Normalization also intersects with chemical structure representation. Natural product databases may store compounds as two-dimensional structures, three-dimensional

conformations, or stereochemically undefined depictions. The choice of representation influences downstream similarity measures and binding predictions. A knowledge graph that stores only two-dimensional fingerprints may miss stereochemical differences that are critical for target binding, while a graph that stores full three-dimensional structures may become computationally expensive. A practical middle ground is to store multiple representations linked to the same compound node, allowing users to choose the appropriate level of detail for their analysis. This design illustrates a recurrent theme in knowledge graph construction: the need to preserve multiple views of the same entity while maintaining a stable identity.

5. Entity and Relation Extraction

Once normalized identifiers exist, the next challenge is extracting entities and relations from unstructured text. Natural product literature includes not only abstracts but also full-text articles, patents, and traditional medicine monographs. Transformer-based language models [13] have improved named entity recognition for chemicals and genes, but they require domain adaptation because natural product nomenclature is highly irregular and multilingual. Joint entity and relation extraction can reduce error propagation compared to pipeline approaches, but it also demands larger annotated corpora. Since manually annotated data in pharmacognosy are scarce, distant supervision from existing databases [3,6,7] can generate weak labels, at the cost of introducing systematic biases toward well-studied compounds and targets.

Relation extraction must distinguish direct molecular interactions from indirect functional associations. A statement that a compound inhibits an enzyme carries different meaning from a statement that a compound modulates a pathway through unknown intermediates. The graph schema should therefore include edge types with explicit causal semantics, and extraction models should be trained to retain negation, speculation, and experimental conditions. Textual hedging such as may inhibit or was not active against is often discarded in binary relation extraction, but omitting these modalities can create false confidence in downstream target discovery. Multi-task learning approaches [14] can treat entity recognition and relation extraction jointly, but they still require careful treatment of contextual modifiers and cross-sentence dependencies.

Entity and relation extraction from structured sources also requires careful semantic mapping. Bioactivity databases use different units and activity thresholds; for example, a compound may be recorded as active if its IC50 falls below a particular cutoff, but that cutoff may vary by assay type. The knowledge graph should retain the raw measurement and the threshold, rather than only a binarized active or inactive label. This design choice allows later users to apply their own decision boundaries and to weight evidence according to experimental quality. Provenance-aware extraction is especially important for natural products because the same compound can be isolated from multiple organisms and tested under divergent protocols, leading to apparently conflicting conclusions.

Beyond pairwise relation extraction, community-level patterns can reveal synergistic combinations among herbal components. A network community-based model for identifying synergistic combinations from herbal medicines and complex systems has been proposed [10], and it suggests that graph community structure may capture multi-component interactions that isolated binary edges cannot. Integrating such community signals into a knowledge graph requires representing both individual compounds and multi-component herbal formulas as relationally distinct classes, with edges linking formulas to their constituents and to observed combination effects. This representation supports queries about which combinations appear in

traditional formulations and whether their observed synergy can be explained by shared targets or complementary pathway modulation.

Extraction quality is also shaped by the choice of training data. Supervised models trained on general biomedical text may perform poorly on natural product articles because of genre differences, chemical naming conventions, and the frequent use of Latin binomials. Fine-tuning on domain-specific corpora is necessary, but the creation of such corpora is labor intensive. A sustainable extraction pipeline may use active learning, where human curators review the most uncertain model predictions, thereby improving performance where it matters most. This feedback loop should be captured in the graph as provenance, so that later users can determine whether an edge was generated by a model, a curator, or a combination of both.

6. Graph Representation and Semantic Integration

Graph representation and semantic integration determine how well a knowledge graph can serve both human interrogation and machine learning. Once entities and relations are extracted, the graph must be embedded into a representation that preserves multi-relational structure. Translational embedding methods [16] and graph neural approaches [15,17] have been widely used for link prediction, but their behavior in sparse and heterogeneous natural product graphs is different from their behavior in dense social networks or homogeneous protein interaction graphs. Natural product graphs often contain long-tailed distributions of compounds, where a small number of well-studied molecules have many edges and a large number of newly isolated compounds have few or no edges. This imbalance creates a structural bias that standard embedding models may amplify, because embeddings for rare entities are poorly constrained and often lie near the origin or in diffuse regions of the latent space.

Semantic integration across ontologies is equally important. A graph may combine chemical classes from ChEBI [8], disease concepts from the Disease Ontology [11], and anatomical concepts from Uberon [12] or related phenotype resources. These ontologies have different levels of granularity and different design philosophies. Some are developed for clinical coding, while others are designed for model organism biology. Mapping between them is not purely syntactic; it requires judgment about whether two classes are truly equivalent or merely overlapping. For example, a disease class in a human phenotype ontology may have no exact match in a traditional medicine syndrome vocabulary. Forcing a match can distort the evidence, while leaving the terms unaligned can fragment the graph. The infrastructure should therefore support both equivalence mappings and weaker similarity links, with explicit provenance for each alignment decision.

Schematic decisions also affect inference. A property graph can encode relation weights, time stamps, and confidence scores, but these values may not be interoperable across sources. If one source assigns confidence based on manual curation and another uses automated text mining, a simple numeric threshold may not be meaningful. One solution is to store confidence as a typed provenance node rather than a scalar edge weight. This allows the graph to represent that a claim was made by a specific method, at a specific time, with a specific level of support. Downstream learners can then aggregate evidence according to their own epistemologies. This approach is more difficult to implement than a flat confidence score, but it supports transparency and contestability, which are critical when the graph informs target prioritization or compound repurposing decisions.

Versioning is a related concern in semantic integration. When an ontology releases a new version, classes may be merged, split, or retired. A knowledge graph that imports ontology terms must decide whether to update all edges immediately or to retain version-specific references. Immediate updates can break traceability, while retaining old versions can create inconsistent semantics across the graph. A version-aware graph can associate each edge with the ontology version used at ingestion time, but this increases storage and query complexity. The choice depends on whether the graph is primarily used for historical reproducibility or for real-time querying. In many cases, a hybrid approach is appropriate: maintain a current view for fast queries and a versioned archive for audit and reproducibility.

7. Systematic Discovery of Bioactive Compounds and Therapeutic Targets

Systematic discovery in a natural product knowledge graph typically proceeds through link prediction, path search, and network proximity analysis. Relational machine learning [18] can infer missing compound-target or compound-disease edges, but such predictions must be interpreted with caution. In the natural products domain, the absence of an edge may reflect a lack of experimental testing rather than a negative result. A knowledge graph that records negative assay outcomes and experimental contexts can reduce this ambiguity. Without negative evidence, link prediction models may systematically overpredict associations for compounds that have simply been studied more often. This is a fairness concern because traditional medicines and understudied taxa may appear less promising, not because they lack bioactivity but because they lack curated data.

Polypharmacology is another use case where knowledge graphs offer distinct advantages. Natural compounds often interact with multiple proteins, and their therapeutic effects may arise from the simultaneous modulation of several targets. A graph can reveal these multi-target patterns through shared neighbors and motifs, but it must distinguish between direct binding, enzymatic inhibition, allosteric modulation, and downstream expression changes. If these relation types are collapsed into a generic interacts with edge, the graph may produce plausible-looking yet mechanistically misleading predictions. The infrastructure should therefore preserve relation semantics and allow users to query by mechanism class. This requirement interacts with the extraction challenge described earlier, because relation typing in literature is difficult and often requires expert curation.

Target discovery can also be framed as a path-finding problem between a compound and a disease. Shortest paths may pass through well-connected hub proteins, but hubs are often biased by experimental popularity. Alternative measures such as degree-weighted proximity or network propagation can reduce this bias, yet they introduce their own assumptions about how biological signals flow. A systems perspective suggests that different proximity measures may be appropriate for different disease areas, and the knowledge graph should make these choices visible. When a graph is used to rank candidate targets for a natural product, the report should include not only the ranked list but also the path classes, evidence types, and provenance that contributed to each score.

For example, a graph integrating NPASS [6] and ChEMBL [7] might identify a plant-derived alkaloid with weak activity against a kinase and strong activity against a G protein-coupled receptor. A path-based analysis could connect both targets to a disease module, suggesting a polypharmacological explanation for the compound's observed phenotypic effect. This case illustrates how graph structure can generate hypotheses that single-target screens would miss. However, the hypothesis remains provisional until the underlying assay conditions, species

source, and chemical identity are verified. The graph should therefore present such findings as ranked hypotheses with evidence trails, rather than as confirmed facts.

8. Governance, Robustness, Fairness, and Quality Assurance

Governance is a central but often neglected component of natural product knowledge graph construction. Because the graph aggregates data from many communities, it needs clear rules about who can edit, validate, or deprecate assertions. A purely automated pipeline can become stale and unreliable if source databases change their formats or identifiers. A purely manual curation model can become a bottleneck and may reflect the biases of a small group of curators. Hybrid governance models, in which automated updates are reviewed by domain experts and community contributors, offer a more sustainable path, but they require workflows for conflict resolution and versioning. These workflows should be designed so that contested claims remain visible rather than being silently overwritten. For example, if one assay reports a compound active and another reports it inactive, the graph should preserve both records with their experimental conditions, allowing users to resolve the conflict in context.

Robustness also depends on the management of incomplete and biased data. Biomedical knowledge graphs are known to be biased toward well-studied organisms, genes, and diseases [20]. This bias arises from funding priorities, publication practices, and database curation policies. In natural product research, the bias is compounded by geographic and linguistic factors, because muchtraditional medicinal knowledge is documented in languages other than English and may not be indexed in major biomedical databases. This underrepresentation can result in a knowledge graph that privileges compounds from regions with strong digital research infrastructures while overlooking potentially valuable compounds from biodiversity-rich but data-poor areas. Fairness in natural product knowledge graphs therefore requires deliberate efforts to incorporate multilingual literature, ethnobotanical records, and community-curated data sources [19]. It also requires recognizing that a graph is not a neutral mirror of biological reality but a curated artifact shaped by historical and institutional priorities. Addressing these biases involves not only algorithmic adjustments but also partnership models with local researchers and knowledge holders, ensuring that data sovereignty and attribution are respected.

Quality assurance in knowledge graph construction extends beyond entity resolution and relation extraction to include consistency checking, contradiction detection, and confidence assessment. Automated consistency rules can flag edges that violate domain constraints, such as a compound having two different canonical structures or a protein being assigned to incompatible species. However, not all contradictions are errors; some reflect genuine biological variability or context-dependent effects. The graph should therefore distinguish between hard constraints and soft expectations, allowing exceptions to be recorded with explanations. Confidence assessment typically combines source reputation, experimental quality, and agreement across independent studies. A high-confidence edge may be derived from multiple orthogonal assays, while a low-confidence edge may originate from a single text-mining mention. This stratification enables downstream users to filter evidence according to their risk tolerance and to prioritize compounds and targets for experimental validation. Governance frameworks should specify how confidence levels are updated when new evidence arrives, ensuring that the graph does not become a static accumulation of claims but a living resource that reflects the current state of knowledge.

9. Deployment, Sustainability, and Policy Implications

Deploying a natural product knowledge graph at scale requires careful attention to computational infrastructure, interface design, and long-term maintenance. The graph may contain tens of millions of nodes and hundreds of millions of edges, especially when provenance records and versioned assertions are retained. Serving interactive queries over such a graph demands efficient indexing and caching strategies, while supporting graph learning workloads requires batch access patterns and the ability to export subgraphs for external model training. A deployment architecture that separates a transactional write layer from an analytical read layer can balance these demands. The write layer handles curation, normalization, and provenance capture, while the read layer provides low-latency access for exploratory browsing, hypothesis generation, and graph embedding. This separation also improves robustness because maintenance operations on the write layer do not interrupt discovery-oriented reads. However, it introduces a synchronization delay between the curated current view and the analytical snapshot, a trade-off that must be documented for users who require real-time evidence.

Interoperability with external tools is another deployment consideration. Researchers may wish to query the graph through standard interfaces such as SPARQL endpoints, or they may prefer a property graph query language. Supporting both paradigms is possible but increases maintenance burden and can create semantic drift if the same query returns different results across interfaces. One mitigation is to expose a canonical application programming interface that abstracts over the storage back end, allowing the graph to evolve internally while preserving stable query semantics for external consumers. The FAIR data principles [21] provide a useful guide here: the graph should be findable through persistent identifiers, accessible through open protocols, interoperable through community vocabularies, and reusable with clear licensing and provenance. Adhering to these principles does not guarantee scientific utility, but it significantly lowers the barriers to reuse and reduces the risk of the graph becoming an isolated artifact with a short lifespan.

Sustainability of a natural product knowledge graph is a socio-economic challenge as much as a technical one. Community databases such as ChEMBL and DrugBank rely on a mixture of institutional funding, grant support, and commercial licensing arrangements. A knowledge graph that aggregates multiple sources must navigate the licensing terms of each contributor, especially when redistributing derived edges or embeddings. If the graph is intended for open science, it should adopt an open license for its original contributions while clearly attributing third-party data. Sustainability models may include tiered access, where the core graph is open but value-added services such as hosted analytical environments or custom export pipelines are offered to commercial users. Alternatively, a consortium of academic and industry partners can share maintenance costs in exchange for early access to updates and governance voice. Regardless of the model, the governance structure must be transparent about how decisions are made, who controls the roadmap, and how conflicts between commercial and scientific interests are resolved.

Policy implications arise when knowledge graphs inform drug discovery decisions that may eventually affect patient safety and public health. Regulators increasingly expect computational predictions to be accompanied by interpretable evidence trails, especially when a graph is used to prioritize compounds for in vivo testing or to justify target selection in preclinical studies. A knowledge graph can support this expectation by preserving the provenance of each edge and by offering explanation interfaces that reconstruct the paths and evidence behind a prediction. In some jurisdictions, machine-learning outputs used in drug

development may be subject to audit requirements under software as a medical device or clinical decision support regulations [22]. Even when such regulations do not directly apply, research funders and journal editors are raising expectations for reproducibility and data availability. A well-governed knowledge graph can serve as a reproducibility asset, but only if its version history, source mappings, and curation decisions are publicly documented.

Ethical considerations also touch the sourcing of natural product data. Many bioactive compounds are derived from plants and microorganisms collected in regions with rich traditional knowledge. The Nagoya Protocol on access and benefit-sharing establishes that benefits arising from the utilization of genetic resources should be shared fairly with the countries and communities providing those resources. A knowledge graph that links a compound to its source organism and geographic origin can support benefit-sharing obligations, but it can also inadvertently facilitate biopiracy if provenance data are used without appropriate consent frameworks. Governance policies should therefore require that source organisms are recorded with geospatial and legal metadata where available, and that access permissions are tracked. Community partnership models, in which local researchers co-curate and retain control over their data, can help align the knowledge graph with principles of data sovereignty and equitable research collaboration [23].

Explainability is another policy-relevant requirement. When a knowledge graph ranks candidate therapeutic targets, the ranking may be based on graph neural embeddings that are inherently opaque. To meet scientific and regulatory expectations, the system should offer post hoc explanations that map high-level embedding similarities to specific graph paths, evidence types, and source records. Techniques from explainable artificial intelligence [24] can identify which nodes and edges contributed most to a prediction, but these techniques must be adapted to heterogeneous multi-relational graphs with provenance-rich edges. An explanation that cites a particular bioactivity assay and its experimental conditions is more useful to a natural product chemist than an abstract feature attribution. Therefore, the deployment layer should include explanation services that translate model outputs into human-readable evidence summaries, allowing domain experts to assess whether a prediction is biologically plausible or merely an artifact of data bias.

Long-term sustainability also requires continuous evaluation against external benchmarks and real-world validation outcomes. A knowledge graph that is never tested against new experimental findings becomes a historical archive rather than a discovery tool. Integrating feedback from wet-lab validation studies allows the graph to learn which types of evidence are most predictive and which extraction methods introduce the least noise. This feedback loop should be designed to avoid circularity: if the graph is used to prioritize compounds for testing, and then the test results are added to the graph, the evaluation must account for selection bias. Holdout experiments, in which some relations are withheld and later checked against new publications, provide a more rigorous assessment. Deployment teams should publish such evaluation results regularly, including negative findings, to maintain trust in the infrastructure.

10. Conclusion

Knowledge graph construction for natural product discovery is a multidisciplinary infrastructure problem that extends well beyond the assembly of data triples. It requires careful architectural choices that balance semantic richness with maintainability, flexible ingestion with provenance preservation, and automated extraction with human curation. The structural trade-offs identified in this paper, including schema design, normalization, relation

typing, versioning, and confidence stratification, collectively determine whether the graph can support robust and transparent hypothesis generation. A purely technical focus on link prediction accuracy may obscure deeper issues of bias, fairness, and governance that are especially acute in natural product research, where traditional knowledge and biodiversity-rich regions are often underrepresented in digital resources.

The systematic discovery of bioactive compounds and therapeutic targets benefits from a graph that can represent complex, multi-relational evidence and support path-based reasoning. However, the value of such reasoning depends on the quality of the underlying provenance and on the ability of users to interrogate the evidence behind each edge. By treating knowledge graphs as socio-technical systems rather than static databases, researchers can build infrastructures that are not only powerful for inference but also accountable to the communities that provide data and the patients who may ultimately benefit from new therapies. Future work should focus on developing standardized evaluation protocols, integrating multilingual and community-held knowledge, and designing governance models that sustain open, equitable, and scientifically rigorous natural product knowledge graphs over the long term.

References

1. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. Scientific American, 284(5), 34-43.

2. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data: The story so far. International Journal on Semantic Web and Information Systems, 5(3), 1-22.

3. Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., ... Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Research, 46(D1), D1074-D1082.

4. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research, 45(D1), D353-D361.

5. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering, C. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research, 47(D1), D607-D613.

6. Zeng, X., Zhang, P., He, W., Qin, C., Chen, S., Tao, L., Wang, Y., Tan, Y., Gao, D., Wang, B., Chen, Z., Chen, W., Jiang, Y. Y., & Chen, Y. Z. (2018). NPASS: Natural product activity and species source database for natural product research, discovery and tool development. Nucleic Acids Research, 46(D1), D1217-D1222.

7. Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrian-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magarinos, M. P., Overington, J. P., Papadatos, G., Smit, I., & Leach, A. R. (2017). The ChEMBL database in 2017. Nucleic Acids Research, 45(D1), D945-D954.

8. Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., & Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research, 44(D1), D1214-D1219.

9. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2021). PubChem in 2021: New data content and improved web interfaces. Nucleic Acids Research, 49(D1), D1388-D1395.

10. 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).

11. Schriml, L. M., Munro, J. B., Schor, M., Olley, D., McCracken, C., Felix, V., Baron, J. A., Jackson, R., Bello, S. M., Bearer, C., Lichenstein, R., Bisordi, K., Dialo, N. C., Giglio, M., & Greene, C. (2020). The Human Disease Ontology 2020 update: Classification, content and workflow expansion. Nucleic Acids Research, 48(D1), D1007-D1013.

12. Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., ... Robinson, P. N. (2021). The Human Phenotype Ontology in 2021. Nucleic Acids Research, 49(D1), D1207-D1217.

13. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

14. Miwa, M., & Bansal, M. (2016). End-to-end relation extraction using LSTMs on sequences and tree structures. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1105-1116.

15. Schlichtkrull, M., Kipf, T. N., Bloem, P., Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. In The Semantic Web: 15th International Conference, ESWC 2018, 593-607.

16. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Advances in Neural Information Processing Systems 26, 2787-2795.

17. Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems 30, 1024-1034.

18. Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1), 11-33.

19. Chandak, P., Huang, K., & Zitnik, M. (2022). Building a knowledge graph to enable precision medicine. Scientific Data, 9, Article 114.

20. Rodriguez-Esteban, R. (2020). Biomedical knowledge graphs: Bias, availability, and applications. Frontiers in Research Metrics and Analytics, 5, Article 593317.

21. 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., Bouwman, J.,

Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, Article 160018.

22. Ling, C., & Wang, Y. (2025). TLFQC: A High-compatible R Shiny based Platform for Automated and Codeless TLFs Generation and Validation. In *PharmaSUG 2025 conference proceedings*.

23. Carroll, S. R., Garba, I., Figueroa-Rodriguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for indigenous data governance. Data Science Journal, 19(1), 43.

24. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Muller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

25. Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., & Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. Briefings in Bioinformatics, 20(4), 1085-1093.