Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

GRAPH LEARNING FROM CHEMICAL BONDS TO THERAPEUTIC SYSTEMS: A SCOPING REVIEW OF PHARMACEUTICAL REPRESENTATIONS AND REASONING TASKS

Kerstin Richter[1]*, Mark Thompson[2], Dmitry Volkov[3], Cecilia Torres[4]

1. ***Department of Graph Learning and Chemical Representations, Faculty of Pharmacy, Technical University of Dresden, Dresden, Germany.***
2. ***Department of Pharmaceutical Reasoning and Therapeutic Systems, Faculty of Pharmacy, University of British Columbia, Vancouver, Canada.***
3. ***Department of Scoping Review of Graph Learning in Pharma, Faculty of Pharmaceutical Sciences, Saint Petersburg State University, Saint Petersburg, Russia.***
4. ***Department of Molecular-to-System Graph Reasoning, Faculty of Pharmacy, University of Chile, Santiago, Chile.***

ARTICLE INFO

Received:
02 May 2025
Received in revised form:
04 August 2025
Accepted:
08 August 2025
Available online:
28 August 2025

Keywords: Graph neural networks, Molecular representations, Biomedical knowledge graphs, Drug discovery, Computational pharmacology, Explainable artificial intelligence

ABSTRACT

Graph learning has become an important computational strategy for representing chemical structures, molecular interactions, biological networks, and heterogeneous therapeutic knowledge. However, the pharmaceutical meaning of a graph model depends not only on its architecture or predictive performance but also on what its nodes and relations encode, which reasoning task is attempted, how evaluation is designed, and whether evidence extends beyond retrospective benchmarks. This scoping review maps graph representations and learning tasks across molecular, biomolecular, network, disease, and knowledge-system scales. A transparent eligibility and evidence-charting framework was used to distinguish representation choices, prediction levels, reasoning claims, validation settings, uncertainty practices, and translational boundaries. The synthesis indicates that molecular property prediction and relation prediction constitute the most methodologically developed areas, whereas pathway reasoning, cross-scale transfer, evidence-linked explanation, and decision-context evaluation remain less mature. Increasing graph breadth can expand the scope of computable relationships, but it also introduces semantic heterogeneity, provenance problems, missing relations, confounding, and greater validation requirements. The review therefore proposes a multiaxial interpretation organized by graph scale, pharmaceutical task, and evidence maturity. This framework is intended as a review-coding and reasoning device rather than a validated readiness scale. It separates benchmark capability from generalization, experimental confirmation, pharmaceutical usefulness, and routine implementation. The principal implication is that graph learning should be evaluated as a representation-dependent and claim-specific scientific instrument. Stronger therapeutic relevance will require realistic data partitions, explicit applicability domains, external and temporal validation, calibrated uncertainty, experimentally testable hypotheses, and evaluation within defined pharmaceutical decision contexts.

To Cite This Article: Richter K, Thompson M, Volkov D, Torres C. Graph Learning from Chemical Bonds to Therapeutic Systems: A Scoping Review of Pharmaceutical Representations and Reasoning Tasks. Pharmacophore. 2025;16(4):72-80. https://doi.org/10.51847/OuepYWKrTh

Introduction

Pharmaceutical research is inherently relational. Atoms are connected by bonds, ligands interact with macromolecular environments, drugs perturb targets and pathways, diseases emerge from interacting biological processes, and therapeutic knowledge is distributed across heterogeneous experimental and clinical records. Graph learning provides a computational language for these structures by representing entities as nodes, relations as edges, and local or global context as learnable patterns. Its pharmaceutical applications consequently extend beyond molecular property prediction to protein–ligand modeling, drug–target inference, adverse-interaction prediction, network-based repurposing, and biomedical knowledge

Corresponding Author: Kerstin Richter; Department of Graph Learning and Chemical Representations, Faculty of Pharmacy, Technical University of Dresden, Dresden, Germany. E-mail: kerstin.richter@tu-dresden.de.

integration [1]. This breadth makes graph learning scientifically attractive, but it also complicates comparison because models operating at different graph scales do not answer equivalent questions or carry equivalent evidentiary burdens.

The construction of a graph is not a neutral preprocessing step. A molecular graph based on two-dimensional connectivity exposes different information from a three-dimensional geometric graph, a protein–ligand interaction graph, a multirelational drug–target network, or a disease-centered knowledge graph. Decisions concerning node identity, edge definition, directionality, relation typing, spatial encoding, feature selection, and pooling determine which regularities can be learned and which phenomena remain absent from the model [2]. Representation choice therefore functions as a scientific assumption about the entities and relations considered relevant to a task. A model may perform well within that assumption while remaining insensitive to conformational variability, assay conditions, biological context, competing mechanisms, manufacturability, exposure, safety, or other determinants of pharmaceutical value.

The unresolved problem is not whether graph models can generate predictions, but how their reported capabilities should be interpreted across representational and pharmaceutical scales. Graph learning now spans molecular structures, biomolecular relations, and multi-omic or therapeutic networks; nevertheless, the defensibility of a claim remains conditional on the chosen representation, data quality, validation design, and intended decision context [1-3]. Similar predictive metrics may therefore support markedly different conclusions. A retrospective molecular benchmark can demonstrate task-specific discrimination or error reduction, but it does not by itself establish chemical-space generalization, mechanistic explanation, experimental reproducibility, clinical utility, or routine development readiness. Conversely, broader knowledge graphs may connect more evidence types while simultaneously increasing vulnerability to semantic inconsistency, incompleteness, provenance loss, and confounded associations.

This scoping review evaluates how pharmaceutical graph representations are constructed, what prediction and reasoning tasks they support, how evidence is assessed, and where interpretation exceeds validation. Experimental follow-up has shown that graph-based molecular prediction can contribute to testable discovery hypotheses, although an influential prospective example cannot establish routine readiness across targets, modalities, or development settings [4]. More generally, useful computational drug design requires fit-for-purpose data, multi-objective reasoning, interpretable hypothesis generation, and integration with experimental and pharmaceutical workflows rather than isolated optimization of model metrics [5]. The review's original contribution is a proposed multiaxial synthesis connecting graph scale, reasoning task, and evidence maturity. This synthesis is not presented as an empirically validated classification or regulatory framework; it is used to clarify what has been demonstrated, what remains plausible, and what validation is required before graph-derived outputs can influence pharmaceutical decisions.

Scoping-Review Questions and Eligibility Framework

The review was designed using PRISMA-ScR-compatible principles of transparent objective setting, eligibility specification, source selection, data charting, and synthesis reporting [6]. Its purpose was to map the breadth and structure of pharmaceutical graph-learning evidence rather than to calculate a pooled effect. The unit of evidence was a peer-reviewed journal article that directly addressed graph construction, graph-based learning, evaluation, explanation, transfer, or therapeutic knowledge reasoning in drug discovery, development, computational pharmacology, or closely related pharmaceutical contexts. Eligibility required sufficient methodological or conceptual detail to identify the graph representation, the attempted task, the form of validation, and the limits of the resulting claim.

A scoping design was selected because the literature is heterogeneous in representation scale, endpoint, dataset, architecture, validation setting, and intended use. Such heterogeneity makes a single quantitative effect estimate conceptually inappropriate and potentially misleading. Scoping reviews are particularly suitable when the objective is to map terminology, evidence types, conceptual boundaries, and knowledge gaps rather than determine the average efficacy of a uniform intervention [7]. Accordingly, the present review does not rank graph-learning methods through a pooled performance hierarchy. It instead examines how differences in graph construction and pharmaceutical task alter what can reasonably be inferred from reported evidence.

Eligibility, searching, study selection, charting, and synthesis were treated as iterative but auditable components. Refinement was permitted when terminology or representation types encountered during review revealed relevant concepts not captured adequately by an initial label; however, these refinements were required to remain consistent with the prespecified pharmaceutical scope and exclusion criteria. This approach combines transparent reporting, an appropriate scoping rationale, and documented iterative conduct without implying effect synthesis [6-8]. Articles were excluded when graphs were incidental to the study, when the work did not permit interpretation of the represented entities and relations, or when claims could not be mapped to a pharmaceutical task or validation setting. Resource descriptions were eligible only when they contributed substantive evidence about representation, reasoning, or evaluation rather than merely announcing data availability.

The review questions were organized around seven connected domains: operationalization of the eligibility framework; effects of search and selection decisions on the representation taxonomy; progression from atom–bond graphs to biomedical knowledge systems; tasks spanning property, interaction, pathway, and disease inference; practices for explanation, transfer, uncertainty, and evaluation; evidence gaps arising at different graph scales; and priorities for cross-scale therapeutic research. This breadth remains bounded by graph scale, pharmaceutical task, evidence type, and intended claim, preserving the defining function of a scoping question while preventing unrestricted discussion of artificial intelligence in drug discovery [9]. Charting distinguished information explicitly reported by an article from interpretations proposed in this review, and absent

methodological details were treated as unreported rather than inferred [10]. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop scoping-review questions and eligibility framework within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Scoping-review questions and eligibility framework

Evidence domain	Eligible evidence or method	Reported capability or claim	Strength of support	Reproducibility or bias concern	Unresolved gap
Review design and reporting	Scoping-review guidance, transparent eligibility criteria, documented selection, structured evidence charting	The literature can be mapped systematically across heterogeneous concepts, methods, and evidence types	Strong methodological support for transparent mapping; not evidence of model effectiveness	Incomplete reporting of selection decisions or post hoc modification of eligibility boundaries	Consistent reporting of graph-specific inclusion, exclusion, and charting decisions
Representation construction	Molecular graphs, geometric graphs, complexes, interaction networks, pathway graphs, disease networks, and knowledge graphs	Graphs can encode relational structures not represented directly by conventional tabular inputs	Strong evidence that representation affects accessible information; no universal representation is established	Hidden preprocessing choices, inconsistent node or edge definitions, and undocumented feature engineering	Comparative studies isolating the effect of graph construction from architecture and dataset choice
Pharmaceutical task	Property prediction, interaction inference, target identification, pathway prioritization, disease association, repurposing, explanation, and transfer	Graph learning can support prediction or prioritization within a defined task	Variable and task-dependent; strongest for retrospective prediction and ranking	Endpoint leakage, proxy labels, class imbalance, and task definitions disconnected from decisions	Validation against endpoints that reflect actual pharmaceutical questions and failure costs
Validation and generalization	Internal testing, realistic data partitions, external validation, temporal validation, cross-domain evaluation, and experimental confirmation	Reported performance may indicate task-specific capability or generalization under tested conditions	Conditional on data partition, comparator, domain coverage, and endpoint quality	Random splits, redundant chemical series, closed-world assumptions, and underreported negative findings	Prospective, temporally separated, and decision-context evaluations
Explanation and reasoning	Feature attribution, subgraph identification, relation paths, counterfactual analysis, and knowledge-supported interpretation	Models may identify influential structures, relations, or concepts	Useful for inspection and hypothesis generation; causal or mechanistic support is generally limited	Instability, low faithfulness, selective presentation, and conflation of attribution with mechanism	Explanation validation against experimental evidence and domain-relevant reasoning criteria
Pharmaceutical translation	Experimentally testable hypotheses, workflow integration, uncertainty communication, and decision-impact assessment	Graph outputs may support prioritization within a defined development context	Sparse and context-specific beyond retrospective evaluation	Publication bias, unmodeled developability constraints, and mismatch between model metrics and decision value	Evidence that model use improves a defined scientific or development decision without unacceptable new risks

Search Strategy, Study Selection, and Representation Taxonomy

The search strategy combined graph-method terminology with pharmaceutical entities, applications, and evidentiary concepts. Core searches covered graph neural networks, graph learning, molecular graphs, geometric deep learning, knowledge graphs, and network medicine together with drug discovery, therapeutics, molecular properties, drug–target relations, pathways, diseases, representation, reasoning, explainability, transfer, uncertainty, and evaluation. Searches were structured for reproducible application across biomedical, multidisciplinary, engineering, and publisher databases, followed by DOI and journal-status verification. During evidence charting, graph-learning studies were classified by graph construction, message-passing or aggregation mechanism, pooling level, learning objective, and prediction level [11]. This architectural classification was complemented by a geometric perspective because graph and geometric models can encode permutation, locality, symmetry, and relational inductive biases that conventional Euclidean representations do not express directly [12]. These properties were charted as representational capabilities, not as proof of pharmaceutical validity.

Study selection prioritized direct claim fit rather than the presence of graph terminology alone. Included articles had to permit identification of the entities represented, the relations encoded, the task performed, and the validation supporting the stated conclusion. Comparative molecular studies were interpreted with particular attention to whether learned graphs were assessed against fixed descriptors, whether data partitions challenged structural generalization, and whether conclusions remained stable across heterogeneous endpoints [13]. Applicability-domain reasoning was also treated as central to selection because predictive performance outside the represented chemical or biological domain cannot be assumed from aggregate benchmark results. Reproducible modeling therefore requires explicit descriptions of data provenance, preprocessing, endpoint construction, domain coverage, model comparison, and conditions under which predictions should not be trusted [14]. Articles reporting

narrow benchmarks without sufficient information to evaluate these conditions were not used to support broad pharmaceutical claims.

The resulting representation taxonomy separates five analytically distinct levels: atom–bond molecular graphs; three-dimensional atomistic or molecular graphs; protein–ligand and other molecular-complex graphs; multirelational interaction, pathway, or disease networks; and heterogeneous biomedical knowledge graphs. Within each level, charting recorded node and edge semantics, spatial or experimental context, task, dataset or application setting, validation design, explanation or uncertainty method, prospective confirmation, and permissible claim. The taxonomy also distinguishes two-dimensional connectivity, three-dimensional geometry, periodic or spatial context, and higher-order relations rather than treating every graph as an interchangeable instance of the same data type [15]. This organization is an original synthesis proposed for the review. It does not imply that graph scales form a linear progression of model quality, that larger graphs necessarily reason more effectively, or that representational breadth establishes mechanism, experimental validity, or therapeutic usefulness.

Graph Representations from Atoms and Bonds to Biomedical Knowledge

At the smallest scale considered in this review, graphs represent atoms as nodes and chemical bonds or spatial relationships as edges. Two-dimensional molecular graphs preserve connectivity and local chemical environments, whereas atomistic geometric models can additionally encode continuous interatomic distances and symmetry-related constraints. SchNet illustrates how continuous-filter convolutions can construct representations from atomic positions while respecting properties relevant to three-dimensional molecular environments [16]. Such representations are well aligned with geometry-sensitive physicochemical tasks, but geometric fidelity does not by itself capture assay conditions, biological adaptation, formulation constraints, exposure, toxicity, or therapeutic context. Atomistic detail should therefore be interpreted as increased physical expressiveness rather than automatic pharmaceutical completeness.

Complex-level graphs extend molecular representation by distinguishing covalent neighborhoods from non-covalent or spatially proximal environments. PotentialNet uses staged graph convolutions to model intramolecular structure and spatial interactions, permitting application to ligand-only and protein–ligand systems [17]. This separation is scientifically important because a molecular graph does not directly represent the binding environment in which a compound is evaluated. Nevertheless, complex graphs remain conditional on the availability, quality, and relevance of structural information. A static complex may omit conformational ensembles, water-mediated interactions, protonation uncertainty, target flexibility, competing binding modes, kinetic behavior, and assay-specific effects. Complex representation can therefore support interaction-aware prediction without establishing binding mechanism or in vivo activity.

Beyond individual molecules and complexes, multirelational graphs represent several entity and relation types simultaneously. In polypharmacy modeling, relation-specific graph convolution has been used to encode drug pairs and distinct side-effect relations rather than treating all interactions as a single homogeneous edge type [18]. This formulation expands the learning problem from predicting a property of one molecule to predicting a typed relation between entities. It can reveal structured patterns across observed combinations, but the inferred relation remains dependent on the completeness and construction of the underlying network. Unobserved interactions are not necessarily true negatives, and associations may reflect indication, prescribing, surveillance, or population structure rather than a direct pharmacological mechanism.

Heterogeneous biomedical networks further connect drugs, genes, diseases, pathways, phenotypes, and other evidence-bearing entities. Path-based integration across such networks can prioritize repurposing hypotheses by evaluating relational patterns that would be inaccessible within a molecular graph alone [19]. Disease-centered knowledge graphs can extend this organization across molecular, biological-process, anatomical, phenotypic, and therapeutic contexts while retaining typed relations and source-specific knowledge [20]. The review therefore treats graph scale as an expansion in the scope of representable questions, not as a hierarchy in which larger graphs are inherently superior. As graph breadth increases, semantic alignment, provenance, missingness, temporal consistency, and confounding become more consequential, and the validation burden shifts from molecular prediction toward evidence-supported therapeutic reasoning.

Learning Tasks across Property Prediction, Interactions, Pathways, and Diseases

Molecular property prediction is the most standardized graph-learning task in the reviewed evidence. It includes classification and regression for physicochemical, biochemical, pharmacokinetic, and toxicity-related endpoints. MoleculeNet demonstrates why performance must be interpreted through task-specific datasets, metrics, and data partitions rather than a single aggregate concept of molecular-learning accuracy [21]. The same architecture can behave differently across endpoints because label quality, chemical diversity, class balance, assay design, and error costs differ. Benchmark performance therefore establishes capability only within the represented task and evaluation design. Pharmaceutical usefulness requires additional evidence that the endpoint is decision-relevant and that the model generalizes to the chemical series or development setting in which it would be applied.

Interaction prediction shifts the unit of inference from a molecular property to a relation between entities. Heterogeneous-network integration can combine drug, target, disease, and side-effect information and learn low-dimensional representations for drug–target interaction prediction and computational repositioning [22]. Such models may identify relational regularities that are difficult to recover from isolated molecular descriptors. Their outputs, however, are rankings or estimated links within a constructed network. The strength of an inferred association depends on how positives, negatives, and unknown relations are

defined, whether evidence sources are independent, and whether the network contains information that would be unavailable at the intended prediction time.

Knowledge-graph embedding similarly supports candidate target discovery by representing entities and relation types in a shared latent space. Candidate drug–target links can then be ranked according to learned relational compatibility [23]. This capability is valuable for prioritization because it can integrate distributed evidence and expose relationships that warrant investigation. It does not alone establish that a target is causally implicated in disease, that modulation will produce the desired phenotype, or that a candidate compound can engage the target safely and effectively. Link prediction should therefore be described as evidence-guided hypothesis ranking. Mechanistic interpretation requires orthogonal biological evidence, while target validation requires experimental interrogation in a relevant context.

At disease scale, network medicine connects therapeutic hypotheses to modules, pathways, and disease-associated regions of the interactome. Disease-specific network analyses can generate repurposing priorities and can gain credibility when computational rankings are followed by orthogonal experimental assessment [24]. Such follow-up represents a stronger evidentiary stage than retrospective ranking alone, but it remains specific to the disease context, experimental system, compounds, and endpoints examined. The review consequently distinguishes property prediction, interaction inference, target prioritization, pathway interpretation, and disease-level repurposing as related but non-equivalent tasks. Each requires a claim-matched validation chain, and success at one level cannot be presumed to propagate automatically to therapeutic exposure, efficacy, safety, or clinical value.

Reasoning, Explainability, Transfer, and Evaluation Practices

The term “reasoning” is used inconsistently across pharmaceutical graph learning. In some studies it refers to message propagation, relation composition, path ranking, or latent-space inference; in others it implies a scientifically interpretable account of why a prediction was made. Explainable artificial-intelligence methods can identify attributed atoms, substructures, examples, relations, or concepts that influence model output [25]. These outputs may support inspection, error analysis, and hypothesis generation, but they should not automatically be interpreted as mechanisms. Attribution can be unstable, sensitive to the explanation method, or faithful only to the model rather than to the biological system. Mechanistic claims require correspondence with independently established chemical or biological evidence.

Transfer learning addresses a different limitation: labeled pharmaceutical data are often scarce or unevenly distributed across tasks. Contrastive pretraining on molecular graphs can learn representations from unlabeled structures and improve downstream performance after task-specific fine-tuning [26]. This supports the plausibility of reusable graph representations, particularly where supervised examples are limited. Transfer, however, remains conditional on alignment between the pretraining corpus, augmentations, molecular domain, downstream endpoint, and evaluation split. Improved benchmark performance does not demonstrate that a representation transfers across assay technologies, therapeutic modalities, chemical projects, biological contexts, or decision settings. Claims of generality require external and domain-shift evaluation rather than repeated success on closely related benchmarks.

Uncertainty estimation is necessary when graph predictions are used to prioritize experiments or exclude candidates, but uncertainty is not synonymous with model confidence. Comparative analysis of scalable uncertainty methods for molecular property prediction shows that methods can differ materially in calibration and error-detection behavior [27]. A numerically confident prediction may still be unreliable outside the represented domain, near sparse regions of chemical space, or under changes in assay and label construction. Useful uncertainty practices should therefore evaluate calibration, coverage, selective prediction, and failure under shift. They should also specify how uncertainty changes a decision, such as triggering abstention, requesting additional evidence, or modifying experimental priority.

Evaluation practices determine whether an apparent model advance reflects architecture, data, task definition, or comparison design. Large-scale comparisons across drug–target prediction tasks demonstrate the importance of applying consistent datasets, metrics, baselines, and validation protocols when comparing methods [28]. Even technically fair comparisons remain scientifically incomplete when the benchmark endpoint is only weakly connected to pharmaceutical decisions. Evaluation should therefore proceed in layers: controlled retrospective comparison, external or temporal generalization, local failure analysis, uncertainty assessment, prospective confirmation, and decision-context impact. This layered approach is proposed as a review interpretation rather than a universal validation standard, but it clarifies why statistical improvement and pharmaceutical relevance should be reported separately.

Evidence Gaps by Graph Scale and Pharmaceutical Task

A major molecular-scale gap concerns the relationship between benchmark separation and genuine chemical generalization. Ligand-based classification benchmarks may contain structural redundancy that allows models to retrieve familiar chemical patterns rather than infer transferable structure–activity relationships [29]. Random partitions can therefore make a method appear robust while leaving performance on new scaffolds, evolving projects, or temporally later compounds uncertain. This problem is not limited to graph architectures, but relational models are not exempt from it. Stronger evaluation requires partitions that reflect the intended use, together with explicit analyses of chemical similarity, domain coverage, and performance deterioration outside well-represented neighborhoods.

At interaction, pathway, and therapeutic-system scales, the central gap is the distance between a computational proxy and the downstream decision it is assumed to support. Improvements in prediction, ranking, or generation do not establish effects on

compound quality, experimental productivity, development time, attrition, efficacy, or safety [30]. Higher-scale graphs can connect more entities, yet the additional edges may combine evidence with different meanings, reliability, temporal status, and susceptibility to confounding. A model can therefore produce coherent relational outputs without resolving whether the represented links are causal, actionable, experimentally reproducible, or relevant to a specific development program.

Chemical and biological data also differ from the large, relatively standardized corpora commonly associated with image or language learning. Measurements are sparse, selectively generated, condition-dependent, noisy, and shaped by changing scientific priorities, while missing labels frequently mean "not tested" rather than "negative." Benchmark redundancy, proxy endpoints, and domain-specific data limitations can jointly inflate apparent pharmaceutical readiness [29-31]. The proposed scale atlas therefore maps increasing representational scope alongside increasing requirements for provenance, semantic alignment, calibration, experimental confirmation, and decision-impact evidence. **Figure 1** presents the pharmaceutical graph-learning scale atlas, showing how the article's key components and boundaries are connected within evidence gaps by graph scale and pharmaceutical task.

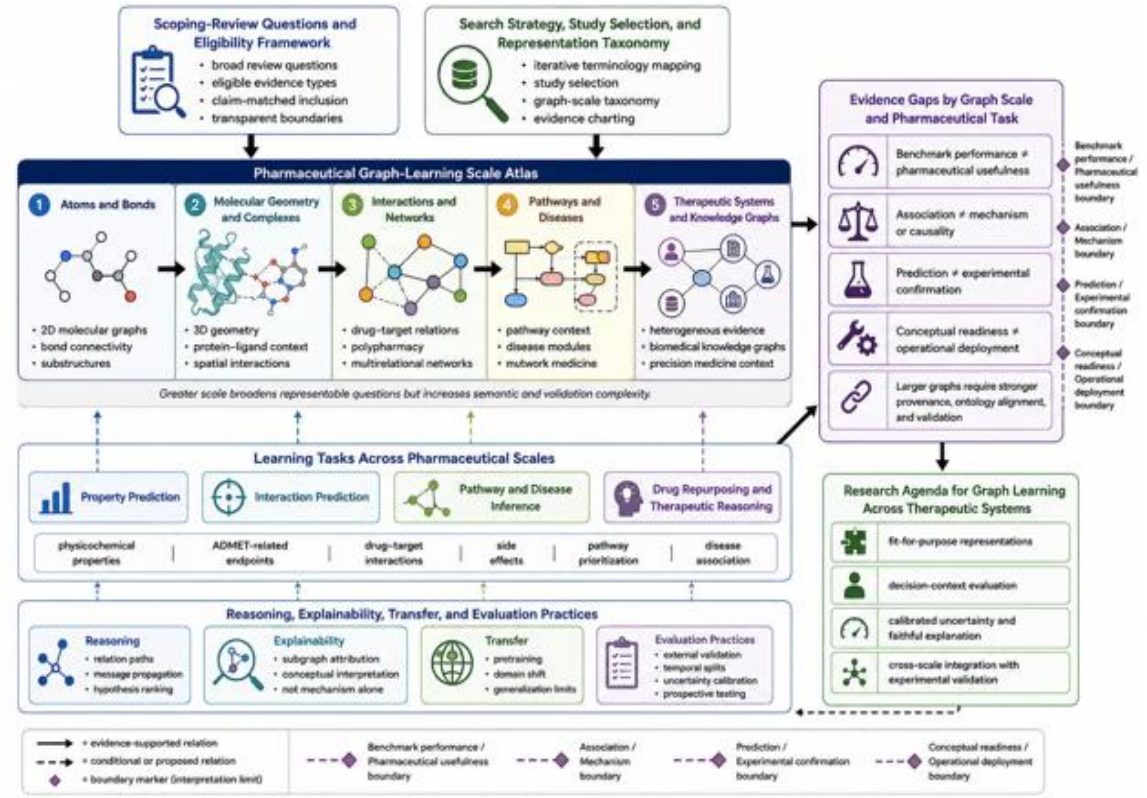

Figure 1. Pharmaceutical Graph-Learning Scale Atlas

The figure is an original conceptual synthesis that organizes the article's central contribution across scoping-review questions and eligibility framework, search strategy, study selection, and representation taxonomy, search strategy, study selection, representation taxonomy, graph representations from atoms and bonds to biomedical knowledge. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

The remaining molecular-scale gap is local rather than global. High aggregate performance can coexist with substantial errors around activity cliffs, where small structural changes are associated with large changes in activity [32]. These chemically sensitive neighborhoods are especially important during lead optimization because they challenge smooth representation assumptions and can alter prioritization among closely related compounds. At broader graph scales, analogous local failures may occur around rare relation types, underrepresented mechanisms, newly emerging evidence, or poorly aligned ontologies. The common requirement is evaluation at the resolution of the intended decision. Overall metrics should be supplemented by local error analysis, domain-specific calibration, explicit unknown-relation handling, and prospective testing of cases that are difficult rather than merely representative.

Research Agenda for Graph Learning Across Therapeutic Systems

The first priority is to connect molecular prediction and generation to explicit pharmaceutical constraints. Graph-based generation or property optimization should not be treated as an endpoint when a candidate must also satisfy synthesizability, selectivity, exposure, safety, formulation, stability, and development requirements. Reviews of deep learning in molecule generation and property prediction emphasize the need to condition design on relevant objectives and to retain experimental

evaluation within the discovery loop [33]. Future studies should therefore define the intended decision before model construction, identify which constraints are represented directly, and state which remain outside the graph. Multi-objective outputs should be framed as prioritization hypotheses rather than complete candidate assessments.

The second priority is a task-aware evaluation ecosystem spanning molecular, interaction, pathway, disease, and development questions. Therapeutic-science benchmark infrastructure can improve comparability by aligning datasets, tasks, splits, evaluation criteria, and documentation with distinct scientific objectives [34]. Shared resources should not, however, institutionalize narrow proxies or reward repeated optimization against familiar test sets. Benchmark design should include provenance, temporal versioning, chemically realistic partitions, relation uncertainty, negative-evidence policies, and explicit context of use. Where possible, evaluation should trace whether a graph-derived output remains useful as it moves from retrospective prediction to external validation, experimental testing, and a defined pharmaceutical decision.

The third priority is to integrate domain knowledge without treating knowledge inclusion as proof of explanation. Chemical knowledge can be incorporated into graph pretraining and functional prompts to shape representations and improve selected downstream tasks [35]. Such methods may increase data efficiency or align learned features with recognizable chemical concepts, but their explanations and transfer gains still require independent validation. Research should test whether knowledge-enhanced models remain faithful under perturbation, whether highlighted concepts are stable across retraining, and whether experts can convert outputs into experimentally discriminating hypotheses. Knowledge sources should also retain provenance and uncertainty so that inherited errors or outdated relations remain visible.

The fourth priority is integration across geometric modeling, knowledge graphs, uncertainty, explainability, scalable learning, and real discovery constraints. Contemporary synthesis of graph neural networks in AI-aided drug discovery indicates that future progress depends on combining these capabilities rather than optimizing isolated architectures [36]. The review proposes that integration be assessed through linked evidence stages: representation adequacy, task-specific benchmark capability, external or temporal generalization, experimental confirmation, decision-context utility, and routine readiness. These stages are an interpretive framework, not a validated scale or regulatory classification. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop research agenda for graph learning across therapeutic systems within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Graph Learning from Chemical Bonds to Therapeutic Systems: A Scoping Review of Pharmaceutical Representations and Reasoning Tasks

Approach or application	Data and task context	Evaluation practice	Evidence maturity	Translational limitation	Review interpretation
Molecular property prediction	Molecular graphs linked to physicochemical, biochemical, pharmacokinetic, or toxicity endpoints	Task-specific metrics, chemically realistic partitions, external or temporal testing, and local error analysis	Retrospective benchmarking is established; broader generalization remains variable	Dataset redundancy, endpoint noise, activity cliffs, and incomplete applicability-domain assessment	Suitable for bounded prediction and prioritization when the endpoint and domain are explicit; not sufficient alone for candidate value
Three-dimensional molecular and protein–ligand modeling	Atomistic geometry, conformers, complexes, and spatial interaction neighborhoods	Geometry-aware comparisons, conformer sensitivity analysis, target-transfer evaluation, and orthogonal experimental testing	Strong methodological development with predominantly retrospective validation	Structural availability, conformational uncertainty, assay context, and limited connection to exposure or developability	May improve physical and interaction context for suitable tasks; does not establish biological efficacy or binding mechanism
Drug–target and multirelational interaction prediction	Heterogeneous drug, protein, disease, side-effect, and pathway networks	Relation-specific validation, positive–unlabeled analysis, temporally separated edges, and independent biological assays	Retrospective link prediction is common; prospective confirmation is limited	Missing relations, uncertain negatives, confounding, and relation imbalance	Useful for ranking candidate interactions; inferred edges require biological and pharmacological confirmation
Disease networks and biomedical knowledge graphs	Multi-source evidence connecting drugs, genes, pathways, phenotypes, anatomy, and diseases	Provenance audits, ontology alignment, source-level ablation, temporal reconstruction, and experimental hypothesis testing	Evidence supports integration and prioritization; therapeutic decision evidence remains sparse	Semantic inconsistency, source dependence, temporal leakage, and causal ambiguity	Strong organizational and hypothesis-generation value; graph connectivity does not establish mechanism or clinical utility
Explainability, transfer, and uncertainty	Model attribution, contrastive pretraining, knowledge-enhanced learning, and predictive uncertainty	Faithfulness and stability tests, external transfer, calibration analysis, abstention evaluation, and expert-grounded review	Mainly method demonstrations and retrospective comparisons	Explanations may be non-causal, transfer may fail under domain shift, and confidence may be miscalibrated	These practices can improve inspection and risk management only when independently evaluated for the intended task
Cross-scale therapeutic reasoning	Linked molecular, interaction, pathway, disease, and development decisions	Sequential computational, experimental, and decision-impact validation with predefined context of use	Conceptual integration and fragmented task-specific demonstrations	No unified validation standard and limited evidence that predictions improve downstream decisions	A priority for research rather than an established operational capability

Pharmaceutical workflow implementation	Candidate prioritization, experiment selection, portfolio decisions, and evidence review	Prospective silent studies, controlled workflow comparisons, documented human–model interaction, and decision-relevant endpoints	Limited decision-context evidence	Workflow dependence, unmodeled constraints, automation bias, maintenance requirements, and unclear accountability	Implementation claims should be limited to defined contexts and should not be inferred from benchmark superiority

Conclusion

Graph learning provides a coherent computational language for pharmaceutical relationships ranging from chemical bonds and spatial molecular environments to drug–target networks, disease modules, and heterogeneous biomedical knowledge. Its scientific value, however, is inseparable from the representation selected, the task defined, the evidence encoded, and the validation used to support a claim. This scoping review proposes a multiaxial synthesis linking graph scale, reasoning task, and evidence maturity. The synthesis is intended to organize a heterogeneous literature and expose validation requirements; it is not an empirically validated readiness scale. The available evidence most clearly supports bounded retrospective prediction, relation ranking, and evidence organization, with more limited support for prospective transfer, mechanistic interpretation, decision impact, and routine therapeutic use. Progress therefore depends less on treating increasingly broad graphs as intrinsically intelligent and more on constructing auditable evidence chains. Representation adequacy, realistic generalization, calibrated uncertainty, faithful explanation, experimental confirmation, and defined pharmaceutical context must be evaluated separately. Graph learning may become more consequential when its outputs are framed as testable, uncertainty-aware contributions to scientific decisions rather than substitutes for chemical, biological, translational, or clinical validation.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, et al. Utilizing graph machine learning within drug discovery and development. Brief Bioinform. 2021;22(6). doi:10.1093/bib/bbab159
2. David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: A review and practical guide. J Cheminform. 2020;12(1):56. doi:10.1186/s13321-020-00460-5
3. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5
4. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688-702. doi:10.1016/j.cell.2020.01.021
5. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3
6. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467-73. doi:10.7326/M18-0850
7. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 2018;18(1):143. doi:10.1186/s12874-018-0611-x
8. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. 2020;18(10):2119-26. doi:10.11124/JBIES-20-00167
9. Munn Z, Pollock D, Khalil H, Alexander L, McInerney P, Godfrey CM, et al. What are scoping reviews? Providing a formal definition of scoping reviews as a type of evidence synthesis. JBI Evid Synth. 2022;20(4):950-2. doi:10.11124/JBIES-21-00483
10. Pollock D, Peters MDJ, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth. 2023;21(3):520-32. doi:10.11124/JBIES-22-00123
11. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst. 2021;32(1):4-24. doi:10.1109/TNNLS.2020.2978386
12. Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P. Geometric deep learning: Going beyond Euclidean data. IEEE Signal Process Mag. 2017;34(4):18-42. doi:10.1109/MSP.2017.2693418
13. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88. doi:10.1021/acs.jcim.9b00237

14. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov VV, et al. QSAR without borders. Chem Soc Rev. 2020;49(11):3525-64. doi:10.1039/D0CS00098A
15. Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, et al. Graph neural networks for materials science and chemistry. Commun Mater. 2022;3(1):93. doi:10.1038/s43246-022-00315-6
16. Schutt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Muller KR. SchNet: A deep learning architecture for molecules and materials. J Chem Phys. 2018;148(24):241722. doi:10.1063/1.5019779
17. Feinberg EN, Sur D, Wu Z, Husic BE, Mai H, Li Y, et al. PotentialNet for molecular property prediction. ACS Cent Sci. 2018;4(11):1520-30. doi:10.1021/acscentsci.8b00507
18. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34(13). doi:10.1093/bioinformatics/bty294
19. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife. 2017;6. doi:10.7554/eLife.26726
20. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023;10(1):67. doi:10.1038/s41597-023-01960-3
21. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A
22. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun. 2017;8(1):573. doi:10.1038/s41467-017-00680-8
23. Mohamed SK, Novacek V, Nounu A. Discovering protein drug targets using knowledge graph embeddings. Bioinformatics. 2020;36(2):603-10. doi:10.1093/bioinformatics/btz600
24. Morselli Gysi D, do Valle I, Zitnik M, Ameli A, Gan X, Varol O, et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. Proc Natl Acad Sci U S A. 2021;118(19). doi:10.1073/pnas.2025581118
25. Jimenez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573-84. doi:10.1038/s42256-020-00236-4
26. Wang Y, Wang J, Cao Z, Barati Farimani A. Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell. 2022;4(3):279-87. doi:10.1038/s42256-022-00447-x
27. Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model. 2020;60(6):2697-717. doi:10.1021/acs.jcim.9b00975
28. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441-51. doi:10.1039/C8SC00148K
29. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403
30. Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009
31. Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data used for AI in drug discovery. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
32. van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model. 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073
33. Walters WP, Barzilay R. Applications of deep learning in molecule generation and molecular property prediction. Acc Chem Res. 2021;54(2):263-70. doi:10.1021/acs.accounts.0c00699
34. Huang K, Fu T, Gao W, Zhao Y, Roohani Y, Leskovec J, et al. Artificial intelligence foundation for therapeutic science. Nat Chem Biol. 2022;18(10):1033-6. doi:10.1038/s41589-022-01131-2
35. Fang Y, Zhang Q, Zhang N, Chen Z, Zhuang X, Shao X, et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. Nat Mach Intell. 2023;5(5):542-53. doi:10.1038/s42256-023-00654-0
36. Zhang O, Lin H, Zhang X, Wang X, Wu Z, Ye Q, et al. Graph neural networks in modern AI-aided drug discovery. Chem Rev. 2025;125(20):10001-103. doi:10.1021/acs.chemrev.5c00461