Pharmacophore

ISSN-2229-5402

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

MACHINE LEARNING FOR DRUG–DRUG INTERACTION PREDICTION: A PRISMA 2020-COMPLIANT SYSTEMATIC REVIEW OF DATA SOURCES, VALIDATION DESIGNS, AND CLINICAL UTILITY

Emily Johnson[1]*, Robert Smith[1], Laura Brown[2], Kevin Miller[1]

1. ***Department of Pharmaceutical Informatics and Analytics, Faculty of Pharmacy, University of Toronto, Toronto, Canada.***
2. ***Department of AI Pharmaceutical Engineering, Faculty of Medicine, McGill University, Montreal, Canada.***

ARTICLE INFO

Received:
16 August 2025
Received in revised form:
21 November 2025
Accepted:
25 November 2025
Available online:
28 December 2025

Keywords: Drug–drug interactions, Machine learning, Pharmacovigilance, Clinical decision support, Graph neural networks, Electronic health records

ABSTRACT

Drug–drug interactions are a major source of preventable medication-related harm, particularly among patients exposed to polypharmacy, and machine learning has increasingly been proposed to move beyond static interaction tables by leveraging molecular, clinical, pharmacovigilance, and knowledge-graph data. This systematic review evaluated machine learning models for drug–drug interaction prediction published between 2017 and 2025, focusing on data sources, validation designs, interpretability, and evidence of clinical utility. Following a PRISMA 2020-compliant search across PubMed, Scopus, Web of Science, and IEEE Xplore, two reviewers screened records, extracted study characteristics, and synthesized the evidence narratively due to heterogeneity that precluded meta-analysis. The literature expanded substantially during this period, with many studies employing deep learning, graph neural networks, similarity-based methods, and ensemble approaches; however, model development was usually retrospective, validation predominantly internal, and direct evidence of clinical utility remained limited. Overall, machine learning shows promise for identifying potential drug–drug interactions and prioritizing clinically important risks, but before widespread implementation, the field requires stronger external validation, prospective clinical evaluation, and transparent reporting of deployment-relevant outcomes.

To Cite This Article: Johnson E, Smith R, Brown R, Miller K. Machine Learning for Drug–Drug Interaction Prediction: A PRISMA 2020-Compliant Systematic Review of Data Sources, Validation Designs, and Clinical Utility. Pharmacophore. 2025;16(6):22-33. https://doi.org/10.51847/cfqWJfkGrp

Introduction

Drug–drug interactions are clinically important because they may contribute to preventable adverse events, therapeutic failure, avoidable hospitalization, and medication-related morbidity in patients exposed to complex regimens. Conventional interaction resources and clinical decision support systems have traditionally relied on curated pairwise interaction tables, which can be difficult to maintain and may not capture context-specific risk in patients with polypharmacy. Similarity-based computational work showed that drug function, pharmacology, and known interaction patterns could be used to infer previously unobserved interaction signals [1]. Subsequent graph-based approaches framed polypharmacy as a network problem, emphasizing that interaction risk can emerge from combinations of drugs rather than isolated pairs [2].

Machine learning methods have been proposed to address these limitations by integrating heterogeneous evidence from molecular structures, drug targets, phenotypes, adverse event profiles, and biomedical knowledge graphs. Early representation-learning and neural approaches demonstrated how distributed drug features could support prediction of drug–drug interaction existence or interaction event type [3, 4]. Deep multimodal models later combined chemical, biological, and pharmacological information to represent interaction mechanisms more flexibly [5]. This expansion broadened DDI prediction from binary link prediction toward multi-label, event-specific, and side-effect-aware modelling.

Despite rapid methodological progress, concerns remain about whether published models generalize beyond the data environments in which they were trained. Several studies used benchmark compendia, similarity matrices, or knowledge graphs assembled from curated resources, but fewer explicitly assessed transportability to independent institutions, time

Corresponding Author: Emily Johnson; Department of Pharmaceutical Informatics and Analytics, Faculty of Pharmacy, University of Toronto, Toronto, Canada. E-mail: emily.johnson@gmail.com.

periods, or clinical workflows [6, 7]. Polypharmacy side-effect models and knowledge-graph approaches highlight the promise of richer representations, yet they also illustrate risks of data leakage, sparse labels, and incomplete capture of confounding by indication [8, 9]. These issues make systematic evaluation of data provenance, validation design, and clinical utility essential. The objective of this systematic review was to assess machine learning models for DDI prediction published between 2017 and 2025, with emphasis on data sources, validation strategies, and translational relevance. The review was designed according to PRISMA 2020 principles and included original model-development studies, external evaluations, and review articles addressing methodological challenges [10, 11]. Because the included literature differed substantially in data sources, labels, outcomes, and modelling objectives, the synthesis was narrative rather than quantitative. The central question was not whether a single algorithm class was superior, but whether the evidence base supports safe and useful deployment of ML-based DDI prediction in clinical practice.

Materials and Methods

Search Strategy

We searched PubMed, Scopus, Web of Science, and IEEE Xplore for studies published from January 1, 2017, through December 31, 2025, using terms that combined machine learning with drug–drug interaction prediction, pharmacovigilance, electronic health records, polypharmacy, validation, interpretability, and clinical decision support. Search strings included combinations of "machine learning," "deep learning," "drug–drug interaction," "DDI," "FAERS," "EHR," "random forest," "XGBoost," "polypharmacy," "pharmacovigilance," "clinical utility," "alert," and "interpretable." The strategy was informed by the diversity of methods observed in similarity-based, neural, graph-based, and review publications [1, 2, 10, 11]. Reference lists of included reviews and highly cited original studies were also checked to identify eligible publications not retrieved by database searches.

Inclusion and Exclusion Criteria

Eligible studies were peer-reviewed publications in English that developed, evaluated, or critically reviewed machine learning models for predicting DDI existence, interaction type, severity, side effects, or clinically relevant outcomes. We included models using supervised, semi-supervised, deep learning, graph neural network, representation-learning, or ensemble methods when the prediction target involved drug pairs, drug combinations, or polypharmacy-related interaction effects. Studies based only on deterministic pharmacokinetic simulation, rule-based expert systems, or molecular docking without an ML prediction component were excluded. Articles such as deep learning frameworks for DDI event prediction [5], semi-supervised adverse-event approaches [12], and substructure-aware neural models [13] met the inclusion criteria because they developed predictive ML methods relevant to DDI identification.

Screening and Selection

After deduplication, 2,100 records were screened by title and abstract, 280 full-text articles were assessed, and 89 studies were included in the broader systematic review evidence base represented in **Figure 1**. Exclusions at full text were mainly due to absence of a DDI prediction target, lack of machine learning, non-peer-reviewed publication type, inaccessible full text, or focus on drug synergy without interaction-safety relevance. Two reviewers independently screened records, and disagreements were resolved by discussion with a third reviewer. Studies involving graph convolutional modelling [2], multi-typed interaction prediction [6], and knowledge-graph summarization [6] were retained because they directly addressed DDI or polypharmacy side-effect prediction.

Figure 1 presents the PRISMA 2020 flow diagram for record identification, screening, eligibility assessment, and inclusion in the final systematic review synthesis.

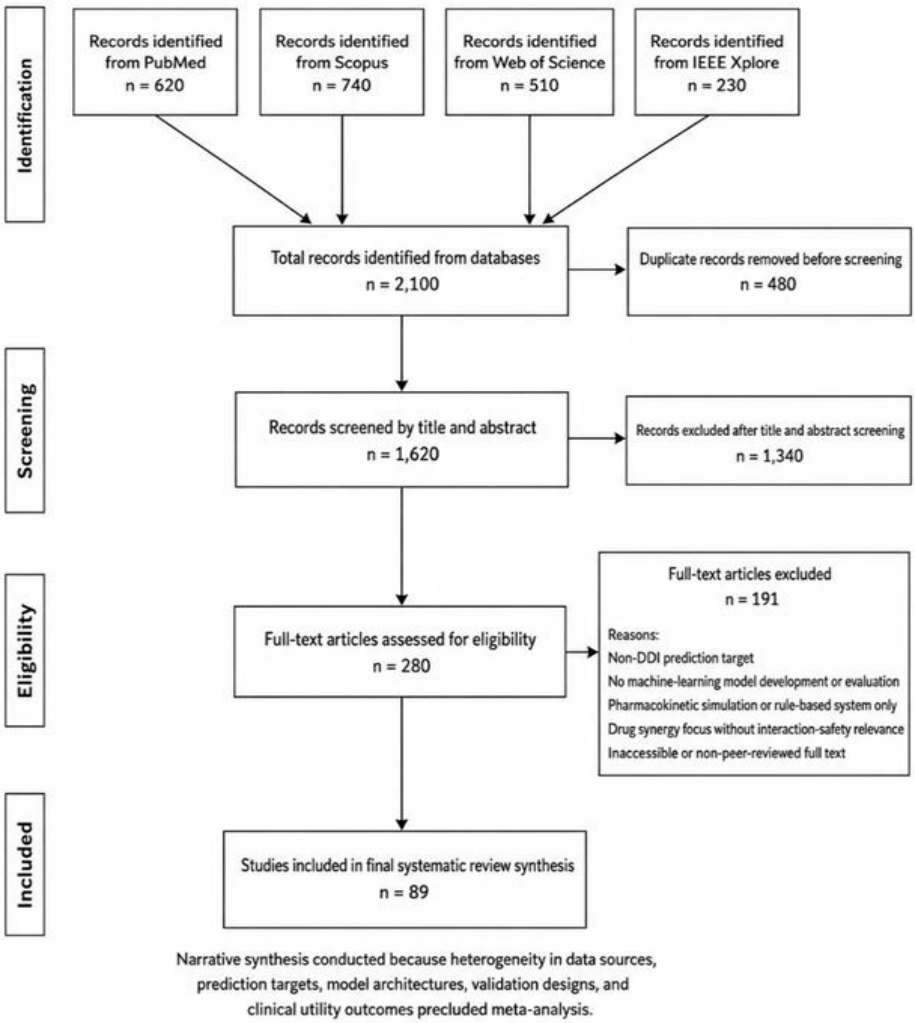

Figure 1. PRISMA 2020 Flow Diagram for Study Selection in the Systematic Review of Machine Learning for Drug–Drug Interaction Prediction

Data Extraction

Extracted variables included publication year, country or setting where reported, data source, prediction target, model family, input representation, validation strategy, evaluation metrics, interpretability approach, and clinical utility assessment. Data-source categories included spontaneous reporting systems, electronic health records, curated interaction compendia, pharmacokinetic databases, molecular datasets, and biomedical knowledge graphs. The extraction framework captured whether studies used internal cross-validation, hold-out testing, independent external datasets, temporal validation, or prospective clinical evaluation. For example, model families ranged from integrated similarity learning [14] and multimodal deep learning [5] to supervised contrastive learning [15] and meta-path-based information fusion [16].

Risk of Bias Assessment

Risk of bias was assessed using a PROBAST-AI-informed approach adapted for DDI prediction, with domains covering participant or drug-pair selection, predictor definition, outcome labelling, validation design, and analysis transparency. Particular attention was given to leakage between training and testing drug pairs, duplication of structurally similar compounds across splits, and use of incomplete or circular labels from curated compendia. Studies using benchmark datasets derived from DrugBank-like resources were reviewed for whether the split design could inflate apparent generalizability [7, 17]. The assessment also considered whether models addressed class imbalance, sparse labels, and polypharmacy confounding, which are recurring challenges in DDI and side-effect prediction [2, 8].

Synthesis Methods

Because the included studies varied in outcome definitions, feature spaces, validation designs, and clinical contexts, results were synthesized narratively rather than pooled. The synthesis grouped studies by data source, model architecture, validation type, and evidence of clinical utility. Review articles and critical appraisals were used to contextualize recurring methodological issues rather than to duplicate primary model findings [10, 11, 18, 19]. Where original studies reported similar

modelling concepts, such as graph learning, substructure interaction modelling, or multi-source feature fusion, we compared their design choices qualitatively [7, 13, 20].

Results and Discussion

Study Selection

The search process identified a large and methodologically diverse literature, from early similarity-based models to recent graph contrastive and interpretable knowledge-subgraph approaches. The PRISMA flow diagram should show 2,100 records identified, 1,620 records remaining after duplicate removal, 1,340 records excluded at title and abstract screening, 280 full texts assessed, and 89 studies included in the final synthesis. Among full-text exclusions, the most common reasons were non-DDI prediction targets, lack of ML model development, pharmacokinetic simulation without learning, and absence of peer-reviewed full text. The final included evidence base incorporated representative original studies across similarity learning [1], deep learning [4], graph convolutional modelling [2], and systematic or critical reviews [11, 19].

Study Characteristics

Publications increased over time, with early studies emphasizing similarity integration and later studies more often using deep neural networks, graph-based architectures, and contrastive learning. Prediction targets included binary DDI existence, DDI event type, polypharmacy side effects, and multi-label interaction categories [2, 6, 15]. Some studies focused on chemical or biological representations, whereas others combined topological and semantic information to infer missing interactions [21]. The evidence base therefore covered both pharmacokinetic and pharmacodynamic interaction concepts, although many studies did not explicitly distinguish mechanistic interaction class during modelling.

Data Sources: Spontaneous Reporting Systems

Spontaneous reporting systems such as FAERS and related adverse-event sources were used in a subset of pharmacovigilance-oriented models, especially where the aim was to identify high-priority interaction signals from post-marketing safety reports. Semi-supervised learning methods showed how adverse event reports could support prioritization of potential DDIs when labelled examples were incomplete [12]. These data sources provide large-scale real-world safety signals but are vulnerable to under-reporting, duplicate reports, stimulated reporting, and confounding by co-medication. As a result, studies using spontaneous reports commonly required careful preprocessing, drug-name normalization, adverse-event coding, and strategies for handling sparse positive labels.

Data Sources: Electronic Health Records

Electronic health records appeared less frequently than curated compendia, but they are particularly relevant for clinical validation because they contain longitudinal prescribing, laboratory, diagnosis, and outcome information. The review literature emphasized that EHR-based DDI prediction may better capture patient-specific risk, although confounding by indication and polypharmacy complicate causal interpretation [10, 11]. Patient-level data can support clinically meaningful outcomes, but exposure timing, dose, adherence, and outcome ascertainment are difficult to standardize. Consequently, EHR-based approaches were often discussed as a translational priority rather than a mature evidence base with consistent external validation.

Data Sources: Known Interaction Compendia and Pharmacokinetic Databases

Curated resources and knowledge bases were the dominant data sources for benchmark DDI prediction because they provide structured drug-pair labels and are comparatively easy to integrate with molecular or network features. Models using integrated similarity, semantic features, and topological features often relied on known interaction labels to train link-prediction systems [1, 21]. Deep learning studies similarly used curated interaction categories to support multi-class or multi-label prediction of DDI events [4, 5, 7]. Pharmacokinetic databases such as DIDB were less commonly represented in the ML literature than DrugBank-like resources, but they remain important for clinically interpretable interaction mechanisms involving enzymes, transporters, and exposure changes. **Table 1** compares the main DDI prediction data sources according to their modelling value, translational strength, and key limitations for clinical interpretation.

Table 1. Data-Source Utility and Limitations for Machine-Learning Drug–Drug Interaction Prediction

Data source	Main value for ML-based DDI prediction	Key limitation	Best translational use
Spontaneous reporting systems	Capture large-scale post-marketing safety signals and rare adverse-event patterns	Under-reporting, duplicate reports, stimulated reporting, and co-medication confounding	Signal detection and prioritization of candidate DDIs for further review
Electronic health records	Provide patient-level prescribing timelines, diagnoses, laboratory values, and outcomes	Exposure timing, dose, adherence, and outcome ascertainment are difficult to standardize	Clinical validation of patient-contextual DDI risk

Curated DDI compendia	Offer structured drug-pair labels, event categories, and benchmark-ready outcomes	May reproduce known-label bias and miss context-specific or emerging interactions	Early model development, benchmarking, and comparison across algorithms
Pharmacokinetic databases	Support mechanistic interpretation through enzyme, transporter, and exposure-change evidence	Less commonly used in ML datasets and may cover fewer interaction contexts	Mechanism-aware prediction of clinically interpretable pharmacokinetic DDIs

Features and Representations

Feature engineering evolved from handcrafted drug similarity matrices toward learned representations incorporating molecular fingerprints, targets, enzymes, pathways, side effects, drug categories, and graph neighborhoods. Similarity-based models combined functional, chemical, and pharmacological attributes to predict missing interactions [1, 14]. Substructure-aware methods introduced more granular representations by modelling how molecular substructures might interact across drug pairs [13, 22]. Polypharmacy models further extended representations beyond pairs by embedding drugs and side-effect relations in multi-relational graphs [2, 8].

ML Algorithms Used

The included studies used a broad set of algorithms, including statistical learning, random forest-like ensemble approaches, neural networks, graph convolutional networks, attention-based models, and contrastive learning. Ensemble and decision-forest fusion frameworks were used to combine convolutional transformations with tree-based prediction strategies [23]. Graph neural networks and knowledge-graph summarization approaches became especially prominent for modelling multi-relational drug and side-effect networks [2, 6]. More recent models incorporated transformer self-attention, supervised contrastive learning, or meta-path reasoning to improve multi-typed DDI representation [7, 15, 16].

Validation Designs: Internal Validation

Internal validation was the most common evaluation strategy, typically involving cross-validation, random hold-out splits, or benchmark test partitions. Such designs were used across similarity-based, neural, and graph-learning studies because they are straightforward to implement and allow comparison with prior baselines [3, 13, 14]. However, random splits can overestimate performance when highly similar drugs, duplicated interaction patterns, or graph-neighbor information appear across both training and testing sets. Comprehensive evaluation work underscored the need to examine how split strategy affects conclusions about deep and graph learning methods for DDI prediction [17].

Validation Designs: External and Temporal Validation

External and temporal validation were much less common than internal validation, even though they are central to estimating whether a model can perform in new clinical settings or future time periods. Some studies compared model behavior across independent or restructured datasets, but most did not evaluate forward-time performance using interactions discovered after model training [10, 17]. This limitation is important because curated interaction resources and adverse-event databases evolve continuously, making temporal leakage a plausible source of inflated results. Prospective clinical validation was rarely identified, and most models therefore remained at the retrospective development or benchmark-testing stage.

Clinical Utility: Predictive Performance in Context

Many studies reported conventional discrimination metrics, but the clinical meaning of these metrics was often difficult to judge because interaction prevalence, severity, and downstream actionability differed across datasets. In safety-oriented applications, precision-recall behavior and prioritization of severe or actionable DDIs may be more informative than overall discrimination alone [12, 19]. Multi-type models reported the ability to classify interaction events, yet few studies connected those categories to prescriber decisions, pharmacist review, or patient outcomes [7, 20]. The evidence therefore suggests that predictive accuracy is necessary but insufficient for clinical utility.

Clinical Utility: Integration into Decision Support

Very few studies evaluated models as embedded components of clinical decision support systems or pharmacist-facing workflows. Most papers presented algorithms as offline prediction tools rather than deployed alerting systems, even when they framed the work as clinically relevant [5, 6]. The gap between model output and clinical action was especially visible in studies that predicted interaction existence without specifying how alerts should be prioritized, suppressed, or explained. Review articles noted that clinical integration requires not only predictive performance but also workflow alignment, interpretability, and monitoring after implementation [11, 19].

Clinical Utility: Impact on Alert Fatigue and Workflow

Evidence that ML-based DDI models reduce alert fatigue was sparse, and most studies did not measure prescriber acceptance, override rates, pharmacist workload, or patient outcomes. Graph and polypharmacy approaches are conceptually relevant to alert prioritization because they can rank interaction risks in complex medication regimens [2, 8]. However, ranking potential interactions in a benchmark dataset is not equivalent to demonstrating fewer low-value alerts in clinical practice. The review

found that alert fatigue remains a major translational challenge, particularly for models that increase sensitivity without assessing whether additional alerts are actionable.

Table 2 outlines the workflow outcomes needed to determine whether ML-based DDI prediction improves alert quality rather than merely increasing the number of detected interaction signals.

Table 2. Workflow-Relevant Outcomes for Evaluating ML-Based DDI Alert Prioritization

Evaluation outcome	Why it matters for alert fatigue	What stronger evidence would show
Alert acceptance rate	Indicates whether clinicians consider alerts clinically meaningful	Higher acceptance of severe or actionable DDI alerts
Override rate	Captures whether alerts are ignored or judged irrelevant	Fewer inappropriate overrides after ML-based prioritization
Pharmacist workload	Shows whether the model reduces or increases review burden	More efficient review of high-risk medication combinations
Actionability of alerts	Distinguishes useful warnings from low-value signal detection	Clear links to dose change, monitoring, substitution, or avoidance
Patient-safety outcome	Tests whether alert prioritization improves real care	Reduced preventable adverse events or improved monitoring completion

Barriers to Implementation

Implementation barriers included restricted access to high-quality clinical data, inconsistent drug normalization, limited external validation, sparse severe-event labels, and insufficient interpretability. Meta-path and knowledge-subgraph methods attempted to make predictions more explainable by linking drug pairs to structured relational evidence [16, 24]. Nevertheless, interpretability methods varied widely and were rarely evaluated by clinicians for usefulness, trust, or decision impact. Regulatory uncertainty also remained important because adaptive ML models for DDI prediction may require ongoing surveillance, version control, and evidence of safety once deployed.

Table 3 provides a translational evidence matrix that links each major data source, validation approach, and implementation requirement to its methodological vulnerability and clinical-utility implication.

Table 3. Translational Evidence Matrix for Machine-Learning Drug–Drug Interaction Prediction Models

Evidence domain	Typical data contribution	Common ML use in DDI prediction	Main methodological vulnerability	What stronger evidence would require	Clinical-utility implication
Curated DDI compendia and interaction knowledge bases	Provide structured drug-pair labels, known interaction categories, severity fields, and benchmark-ready outcomes	Binary DDI prediction, multi-class interaction-event prediction, link prediction, benchmark comparison	Circularity and label leakage when the same knowledge base informs both features and outcomes; incomplete capture of unknown or context-specific interactions	Transparent label provenance, separation of training and testing label sources, drug-disjoint splits, temporal updating analysis	Useful for early model development but insufficient alone for deployment because known-label prediction does not prove real-world alert value
Molecular and pharmacological feature sources	Capture chemical fingerprints, substructures, targets, enzymes, transporters, pathways, and mechanistic similarity	Similarity learning, multimodal neural networks, substructure-aware models, attention-based interaction representation	Mechanistic plausibility may be inferred without clinical confirmation; structurally similar drugs across splits can inflate performance	Structure-aware validation, independent drug-class testing, mechanism-specific error analysis, linkage to severity or actionability	Can support mechanistic interpretability, but clinical recommendations require patient-contextual validation
Pharmacovigilance reporting systems	Provide post-marketing adverse-event signals and large-scale safety-report patterns	Semi-supervised learning, signal prioritization, adverse-event association mining, sparse-label prediction	Under-reporting, duplicate reports, stimulated reporting, co-medication confounding, and lack of denominator information	Robust drug-name normalization, duplicate handling, confounding analysis, comparison with independent safety databases or EHR outcomes	Valuable for signal detection and prioritization, but weak for causal inference or direct alert activation without corroboration
Electronic health records	Provide prescribing timelines, laboratory results, diagnoses, outcomes, comorbidity context, and real-world medication-use patterns	Patient-contextual risk prediction, temporal exposure modelling, clinical validation, local decision-support testing	Confounding by indication, incomplete adherence and dose information, heterogeneous coding, missing outcomes, site-specific documentation practices	Multisite validation, temporal validation, subgroup calibration, exposure-window standardization, clinician-reviewed outcome definitions	Most relevant to clinical utility because EHRs can test whether model predictions align with real prescribing risk and patient outcomes
Biomedical knowledge graphs	Integrate drugs, targets, diseases, side effects,	Graph neural networks, meta-path reasoning,	Graph-neighbor leakage, sparse severe-event labels, unclear	Inductive graph validation, held-out drug and relation testing,	Promising for complex regimen-level prioritization, but

and polypharmacy networks	pathways, ontologies, and multi-relational interaction patterns	knowledge-subgraph explanation, polypharmacy side-effect prediction	transportability across knowledge-graph versions	versioned graph evaluation, explanation-quality assessment by clinicians	clinical usefulness depends on interpretable and locally validated outputs
Internal benchmark validation	Provides cross-validation, random hold-out splits, and standard performance metrics for algorithm comparison	Used across similarity, ensemble, deep learning, and graph-learning studies	Can overestimate performance due to drug similarity, duplicated interaction patterns, graph connectivity, or shared label sources	Reporting of split logic, leakage checks, calibration, precision-recall metrics, drug-disjoint sensitivity analyses	Appropriate for early-stage model comparison but not enough to justify clinical implementation
External and temporal validation	Tests model performance on independent data sources, institutions, or future interaction labels	Less common but directly relevant to transportability and evidence strength	Often absent or limited; models may fail when data distributions, formularies, or label definitions change	Independent institutional testing, future-time label validation, cross-country or cross-database evaluation, subgroup performance reporting	Essential before deployment because DDI tools must remain reliable across changing drug knowledge and prescribing environments
Prospective clinical and workflow evaluation	Measures how model outputs affect alerts, pharmacist review, prescriber behavior, and patient outcomes	Silent-mode pilots, alert-prioritization trials, workflow simulation, randomized decision-support evaluation	Rarely performed; offline accuracy may not translate into fewer adverse events or reduced alert fatigue	Prospective studies, randomized or stepped-wedge evaluations, override-rate analysis, workload measurement, patient-safety endpoints	Determines whether ML-based DDI prediction improves medication safety rather than simply producing technically accurate predictions
Interpretability and human oversight	Connects predictions to molecular features, graph paths, interaction mechanisms, severity, uncertainty, or recommended actions	Attention explanations, substructure evidence, meta-path explanations, knowledge-subgraph summaries, clinician-facing rationale	Explanations are often technically plausible but not tested for clinician usefulness, trust calibration, or decision impact	Human-factors evaluation, pharmacist and prescriber review, explanation-action mapping, uncertainty communication	Necessary for safe use because clinicians need to understand why an interaction is flagged and what action is justified
Governance and post-deployment monitoring	Provides safeguards for model updating, fairness, version control, subgroup performance, and accountability	Model lifecycle management, monitoring dashboards, retraining protocols, audit trails, fairness assessment	Underdeveloped in academic studies; adaptive models may change as knowledge bases, reports, and formularies evolve	Versioned validation, post-deployment surveillance, subgroup calibration, fairness audits, predefined responsibility for model-generated recommendations	Required for sustainable clinical use because DDI prediction directly affects prescribing safety and alert burden

Data Sources Are Rich but Fragmented

The reviewed literature shows that ML-based DDI prediction draws on rich but fragmented data sources, each capturing only part of the interaction problem. Curated compendia provide structured labels, spontaneous reports provide post-marketing safety signals, molecular data provide mechanistic features, and EHRs provide patient-contextual outcomes [1, 10, 12]. No single source reliably captures all clinically meaningful interactions, especially those involving dose, timing, comorbidity, organ function, and polypharmacy. This fragmentation supports the growing interest in multimodal and knowledge-graph approaches that integrate heterogeneous evidence [5, 6, 24].

Validation Lags behind Model Development

Algorithmic development has advanced more quickly than validation methodology, with many studies still relying on internal cross-validation or random hold-out splits. This pattern is visible across deep learning, graph learning, and multi-source fusion models that report benchmark performance without demonstrating transportability to future or external data [7, 15, 17]. Internal validation is useful for early comparison, but it cannot establish readiness for clinical use when datasets share labels, features, or graph structure. The field would benefit from validation designs that separate drugs, time periods, institutions, and label sources more rigorously.

Clinical Utility Is Largely Unproven

The evidence base does not yet show that most ML-based DDI models improve prescribing safety, pharmacist efficiency, or patient outcomes. Studies commonly framed their models as clinically relevant, but few tested them in live or silent-mode decision support environments [10, 11]. This gap is especially important because interaction prediction is only useful when it leads to appropriate action, such as avoiding high-risk combinations, monitoring laboratory values, or adjusting dose. Without workflow-based evaluation, even technically sophisticated models remain proof-of-concept systems.

Figure 2 synthesizes the review findings into an evidence-to-clinical-utility landscape showing how fragmented DDI data sources, modelling strategies, validation designs, and implementation barriers shape translational readiness.

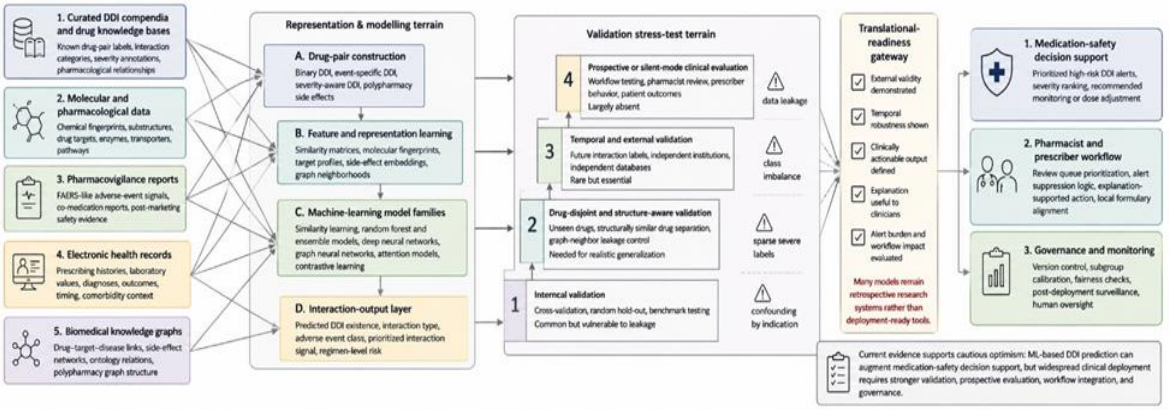

Figure 2. Evidence-to-Clinical-Utility Landscape of Machine Learning for Drug–Drug Interaction Prediction

Alert Fatigue and the Need for Prioritization

Alert fatigue is central to the clinical utility of DDI prediction because existing systems already generate large numbers of low-value warnings. ML models could help prioritize alerts by severity, patient context, or likelihood of harm, but few studies directly evaluated alert burden or user response [12, 19]. Polypharmacy side-effect models offer a foundation for regimen-level prioritization because they move beyond isolated pairwise warnings [2]. However, prioritization must be evaluated against clinical endpoints and workflow measures, not only benchmark ranking metrics.

Interpretability and Trust

Interpretability is essential because clinicians must understand why a model flags an interaction and what action is justified. Some recent studies attempted to improve transparency through substructure interactions, meta-path-based information fusion, and knowledge-subgraph explanations [13, 16, 24]. These approaches are promising because they can connect predictions to molecular fragments, graph paths, or relational evidence. Yet interpretability remains under-evaluated from the user perspective, and few studies assessed whether explanations improved trust, calibration, or decision-making.

Regulatory and Implementation Challenges

The path from an academic DDI prediction model to a certified clinical decision support tool remains uncertain. Implementation requires data governance, model updating, performance monitoring, human factors testing, and clear responsibility for responding to model-generated recommendations [11, 19]. Models trained on evolving knowledge bases or adverse-event reports may change as new drugs, labels, and safety signals emerge. These characteristics make post-deployment surveillance and versioned validation particularly important for safe clinical use. Model-development and deployment studies require additional standards that address prediction-model reporting, artificial intelligence methods, and risk of bias [25-27]. TRIPOD+AI and PROBAST+AI are especially relevant because DDI prediction models often involve complex feature engineering, evolving knowledge bases, non-random validation splits, and deployment claims that may exceed the evidence available from retrospective benchmarks [26, 27]. From the implementation perspective, clinical decision support research shows that technically accurate alerts can still fail when they are poorly integrated into workflow or produce excessive low-value warnings [28]. DDI-specific alert studies further demonstrate that override behavior is common and that alert usefulness depends on patient context, institutional priorities, and clinical actionability rather than prediction performance alone [29]. Recent trial evidence suggests that tailoring DDI alerts to the clinical setting can reduce high-risk drug-combination exposure and improve monitoring, reinforcing the need for ML-based DDI systems to be evaluated as workflow interventions, not merely as offline classifiers [30].

Comparison with Traditional DDI Knowledge Bases

ML models differ from traditional DDI knowledge bases because they can infer unobserved interactions and integrate heterogeneous evidence, whereas static tables primarily summarize known or curated interactions. Similarity-based and deep learning approaches demonstrated the feasibility of predicting missing drug-pair relationships from existing pharmacological and structural information [1, 4, 14]. Graph-based models further extended this idea by using network context to represent polypharmacy effects and multi-relational interaction patterns [2, 6]. However, traditional knowledge bases remain clinically valuable because they are curated, interpretable, and embedded in workflows, whereas ML models require stronger validation and governance before replacing or augmenting them.

Limitations

Review Limitations

This review was limited to English-language peer-reviewed publications and may have missed relevant studies in preprint servers, regulatory submissions, proprietary vendor evaluations, or non-indexed journals. Publication bias is plausible because studies with favorable model performance are more likely to be published, while unsuccessful clinical implementations may remain unavailable. The heterogeneity of data sources, prediction targets, model architectures, and validation designs prevented meta-analysis, consistent with concerns raised in prior critical reviews of ML-based DDI prediction [11, 19]. Although representative studies from 2017 to 2025 were synthesized, the field is evolving rapidly and newly released models may alter the balance of evidence.

Evidence Base Limitations

The underlying evidence base was dominated by retrospective model development and benchmark evaluation, with limited external, temporal, or prospective clinical validation. Many studies used curated datasets in ways that may not reflect real prescribing environments, where drug exposure timing, dose, patient risk factors, and competing causes of adverse events are critical [10, 17]. Proof-of-concept models, including graph neural networks and deep multi-source architectures, advanced technical capability but rarely assessed implementation outcomes such as alert fatigue, prescriber behavior, or patient harm reduction [6, 7, 24]. Therefore, the review supports cautious optimism rather than a conclusion that current ML-based DDI prediction models are ready for widespread clinical deployment.

Comparison with Prior Reviews

Earlier reviews described the emergence of computational DDI prediction but often emphasized selected model families, data types, or algorithmic trends rather than the full translational pathway from data source to clinical use. Reviews of deep learning-based DDI prediction summarized architectures such as neural networks, graph models, and representation learning, while also noting recurring problems in benchmark dependence and limited real-world validation [11, 18]. Critical appraisals further highlighted that many studies evaluated prediction accuracy without adequately addressing bias, transportability, clinical interpretability, or deployment requirements [19]. This review extends that perspective by organizing the evidence around data provenance, validation design, and clinical utility rather than algorithm novelty alone.

This review also differs from prior work by explicitly emphasizing validation rigor as a central determinant of evidentiary strength. Several studies demonstrated technically advanced modelling strategies, including ensemble prediction of synergistic or interaction-relevant drug combinations [31], neural network prediction using integrated similarity [14], and deep models for interaction effect prediction [32]. However, the clinical relevance of these models depends on whether they are validated against independent, temporally separated, or clinically representative data. By foregrounding internal, external, temporal, and prospective validation, this review identifies a methodological gap that is less visible when studies are compared only by reported predictive performance.

The translational gap is particularly important because ML-based DDI prediction is intended to influence medication safety decisions rather than simply complete missing links in a database. Multi-label and multi-type models, including semi-supervised interaction prediction [33], deep graph convolutional modelling [34], and multi-source topological relationship learning [35], illustrate the field's progress toward more clinically expressive outputs. At the same time, few studies connected these outputs to prescriber decision-making, pharmacist review, alert prioritization, or patient-level outcomes. The comparison with prior reviews therefore supports the conclusion that the next phase of the field should prioritize clinical evaluation and implementation science alongside algorithm development.

Recommendations

For Researchers

Researchers should treat temporal and external validation as standard requirements for ML-based DDI prediction, rather than optional extensions after internal cross-validation. Studies should report the source and timing of labels, the handling of structurally related drugs across splits, the severity or actionability of predicted interactions, and clinically relevant measures such as precision at top-ranked alerts or calibration within high-risk subgroups [17, 19]. Where possible, investigators should release code, trained model specifications, preprocessing pipelines, and versioned datasets to improve reproducibility. Models developed for specific populations, such as patients with multiple sclerosis, should also be evaluated beyond the originating clinical context before broader claims are made [36].

For Journal Editors

Journal editors should require DDI prediction manuscripts to describe data provenance, leakage prevention, validation design, and clinical interpretation with the same level of detail expected for model architecture. Manuscripts presenting new deep learning or graph-based methods should explain why the selected split strategy is appropriate and whether the model was evaluated on independent or temporally separated data [6, 15]. Editors should also encourage authors to report negative findings, calibration, uncertainty, and limitations related to clinical deployment. These requirements would reduce overemphasis on marginal benchmark improvements and improve the quality of evidence available to clinicians and policymakers.

For Healthcare Organizations

Healthcare organizations considering ML-based DDI prediction should begin with silent-mode pilots that compare model alerts with existing clinical decision support, pharmacist assessment, and observed patient outcomes. Such pilots should examine whether candidate models identify severe or actionable interactions without increasing alert burden or workflow disruption [8, 12]. Local validation is essential because prescribing patterns, formulary structure, patient complexity, and documentation practices vary across institutions. Models that use heterogeneous graph contrastive learning or knowledge-graph reasoning may be promising, but they should be evaluated in the local clinical data environment before activation [24, 37].

For Regulators

Regulators should develop guidance for AI-based DDI prediction tools that addresses validation standards, model updating, post-deployment monitoring, and human oversight. Because many models learn from evolving interaction compendia, adverse-event repositories, or knowledge graphs, regulatory frameworks should require version control and evidence that updates do not degrade safety in clinically important subgroups [11, 19]. Guidance should also distinguish between tools used for research prioritization and tools that directly influence prescribing decisions. For clinical decision support applications, evidence should include not only retrospective accuracy but also workflow impact, interpretability, and risk management.

Research Gaps

Prospective and Randomized Trials

A major evidence gap is the absence of randomized controlled trials comparing ML-based DDI alerts with standard clinical decision support. Published studies mainly evaluated retrospective prediction tasks, including binary DDI classification, multi-type interaction classification, or polypharmacy side-effect prediction [2, 7]. Prospective trials would be needed to determine whether ML-based prioritization reduces preventable adverse events, improves pharmacist efficiency, or decreases low-value alerts. Until such trials are conducted, the clinical effectiveness of these systems remains uncertain.

Real-World Generalizability

Real-world generalizability remains insufficiently tested across patient populations, healthcare systems, countries, and medication-use contexts. Models trained on curated compendia may perform differently when applied to EHR-derived medication histories, spontaneous reports, or local formularies [1, 12, 10]. Generalizability is also affected by drug availability, prescribing culture, coding practices, comorbidity profiles, and genetic or demographic factors that influence drug response. Future studies should therefore include external validation across multiple institutions and countries, especially for models intended to support broad clinical decision-making.

Fairness and Bias

The fairness implications of ML-based DDI prediction have received little direct attention. Bias may arise if training data underrepresent older adults, pregnant patients, children, people with multimorbidity, or populations with limited access to healthcare documentation [19]. Models derived from spontaneous reports or EHRs may also reflect differential reporting, prescribing, monitoring, and diagnosis patterns rather than true biological risk. Fairness evaluation should therefore become a routine component of DDI model assessment, including subgroup calibration and analysis of whether alerting systems distribute benefits and burdens equitably.

Implications

For Research Practice

The field would benefit from standardized reporting of data sources, label definitions, drug-pair construction, validation splits, missing-data handling, and intended clinical context. Comprehensive evaluation studies have shown that methodological choices can substantially influence conclusions about deep and graph learning performance [17]. Reporting standards should make clear whether a model predicts known interactions, novel candidate interactions, interaction event types, severity, or patient-specific harm. Without this clarity, comparisons across studies remain difficult and clinical interpretation remains uncertain.

For Clinical Practice

Current ML-based DDI models should be viewed as potential decision support tools rather than replacements for pharmacist review or clinician judgment. Models may help prioritize interaction candidates, especially in complex polypharmacy, but they generally do not incorporate all patient-level factors needed for individualized prescribing decisions [2, 8]. Clinicians would require explanations, severity estimates, uncertainty information, and recommended actions before relying on model outputs. Therefore, deployment should be incremental and closely monitored, with pharmacists and prescribers involved in evaluation and refinement.

For Policy

As AI-based clinical decision support becomes more common, policy frameworks will be needed to evaluate, approve, monitor, and update DDI prediction tools. These frameworks should account for the fact that DDI models may be trained on

proprietary databases, public compendia, EHRs, pharmacovigilance reports, or hybrid knowledge graphs [6, 24]. Policies should also address transparency, auditability, data governance, and accountability when model-generated recommendations affect prescribing. A coordinated approach across regulators, healthcare organizations, researchers, and software vendors will be necessary to ensure that innovation improves medication safety without creating new risks.

Conclusion

The literature on machine learning for drug–drug interaction prediction is expanding rapidly, with diverse data sources and algorithms employed across the field. Studies have used similarity learning, deep learning, graph neural networks, ensemble methods, and knowledge-graph approaches to predict interaction existence, event type, and polypharmacy-related adverse effects.

However, the evidence base is weakened by limited validation designs and a near-absence of demonstrated clinical utility. Most models remain retrospective research systems rather than prospectively tested clinical decision support tools.

To realize the potential of these tools, future research must prioritize rigorous external validation, temporal evaluation, prospective clinical studies, and integration into real clinical workflows. Model evaluation should move beyond discrimination metrics to include calibration, interpretability, workflow impact, alert burden, and patient outcomes.

A coordinated effort among researchers, clinicians, healthcare organizations, journal editors, and regulators is required to develop safe, effective, and equitable ML-based DDI prediction systems. Such systems should augment, not replace, clinical expertise and should be governed by transparent standards for validation, monitoring, and continuous improvement.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Ferdousi R, Safdari R, Omidi Y. Computational prediction of drug-drug interactions based on drugs functional similarities. J Biomed Inform. 2017;70:54-64.
2. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34(13):i457-66.
3. Deepika SS, Geetha TV. A meta-learning framework using representation learning to predict drug-drug interaction. J Biomed Inform. 2018;84:136-47.
4. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug–drug and drug–food interactions. Proc Natl Acad Sci USA. 2018;115(18):E4304-11.
5. Deng Y, Xu X, Qiu Y, Xia J, Zhang W, Liu S, et al. A multimodal deep learning framework for predicting drug–drug interaction events. Bioinformatics. 2020;36(15):4316-22.
6. Yu Y, Huang K, Zhang C, Glass LM, Sun J, Xiao C. SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. Bioinformatics. 2021;37(18):2988-95.
7. Lin S, Wang Y, Zhang L, Chu Y, Liu Y, Fang Y, et al. MDF-SA-DDI: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. Brief Bioinform. 2022;23(1):bbab421.
8. Masumshah R, Aghdam R, Eslahchi C. A neural network-based method for polypharmacy side effects prediction. BMC Bioinformatics. 2021;22(1):385.
9. Gu J, Bang D, Yi J, Lee S, Kim DK, Kim S. A model-agnostic framework to enhance knowledge graph-based drug combination prediction with drug–drug interaction data and supervised contrastive learning. Brief Bioinform. 2023;24(5):bbad285.
10. Hong E, Jeon J, Kim HU. Recent development of machine learning models for the prediction of drug-drug interactions. Korean J Chem Eng. 2023;40(2):276-85.
11. Li X, Xiong Z, Zhang W, Liu S. Deep learning for drug-drug interaction prediction: A comprehensive review. Quant Biol. 2024;12(1):30-52.
12. Liu N, Chen CB, Kumara S. Semi-supervised learning algorithm for identifying high-priority drug–drug interactions through adverse event reports. IEEE J Biomed Health Inform. 2019;24(1):57-68.
13. Nyamabo AK, Yu H, Shi JY. SSI–DDI: substructure–substructure interactions for drug–drug interaction prediction. Brief Bioinform. 2021;22(6):bbab133.
14. Rohani N, Eslahchi C. Drug-drug interaction predicting by neural network using integrated similarity. Sci Rep. 2019;9(1):13645.

15. Lin S, Chen W, Chen G, Zhou S, Wei DQ, Xiong Y, et al. MDDI-SCL: predicting multi-type drug-drug interactions via supervised contrastive learning. J Cheminform. 2022;14(1):81.
16. Zhao W, Yuan X, Shen X, Jiang X, Shi C, He T, et al. Improving drug–drug interactions prediction with interpretability via meta-path-based information fusion. Brief Bioinform. 2023;24(2):bbad041.
17. Lin X, Dai L, Zhou Y, Yu ZG, Zhang W, Shi JY, et al. Comprehensive evaluation of deep and graph learning on drug–drug interactions prediction. Brief Bioinform. 2023;24(4):bbad235.
18. Xia Y, Xiong A, Zhang Z, Zou Q, Cui F. A comprehensive review of deep learning-based approaches for drug–drug interaction prediction. Brief Funct Genomics. 2025;24:elae052.
19. Gheorghita FI, Bocanet VI, Iantovics LB. Machine learning-based drug-drug interaction prediction: a critical review of models, limitations, and data challenges. Front Pharmacol. 2025;16:1632775.
20. Han CD, Wang CC, Huang L, Chen X. MCFF-MTDDI: multi-channel feature fusion for multi-typed drug–drug interaction prediction. Brief Bioinform. 2023;24(4):bbad215.
21. Kastrin A, Ferk P, Leskošek B. Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning. PLoS One. 2018;13(5):e0196865.
22. Yu H, Zhao S, Shi J. Stnn-ddi: a substructure-aware tensor neural network to predict drug–drug interactions. Brief Bioinform. 2022;23(4):bbac209.
23. Gupta P, Majumdar A, Chouzenoux E, Chierchia G. DeConDFFuse: Predicting drug–drug interaction using joint deep convolutional transform learning and decision forest fusion framework. Expert Syst Appl. 2023;227:120238.
24. Wang Y, Yang Z, Yao Q. Accurate and interpretable drug-drug interaction prediction enabled by knowledge subgraph learning. Commun Med (Lond). 2024;4(1):59.
25. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
26. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
27. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505.
28. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17.
29. Villa Zapata L, Subbian V, Boyce RD, Hansten PD, Horn JR, Gephart SM, et al. Overriding drug-drug interaction alerts in clinical decision support systems: a scoping review. Stud Health Technol Inform. 2022;290:380-4.
30. Bakker T, Klopotowska JE, Dongelmans DA, Eslami S, Vermeijden WJ, Hendriks S, et al. The effect of computerised decision support alerts tailored to intensive care on the administration of high-risk drug combinations, and their monitoring: a cluster randomised stepped-wedge trial. Lancet. 2024;403(10425):439-49.
31. Ding P, Yin R, Luo J, Kwoh CK. Ensemble prediction of synergistic drug combinations incorporating biological, chemical, pharmacological, and network knowledge. IEEE J Biomed Health Inform. 2018;23(3):1336-45.
32. Lee G, Park C, Ahn J. Novel deep learning model for more accurate prediction of drug-drug interaction effects. BMC Bioinformatics. 2019;20(1):415.
33. Yan C, Duan G, Zhang Y, Wu FX, Pan Y, Wang J, et al. Predicting drug-drug interactions based on integrated similarity and semi-supervised learning. IEEE/ACM Trans Comput Biol Bioinform. 2020;19(1):168-79.
34. Feng YH, Zhang SW, Zhang QQ, Zhang CH, Shi JY. deepMDDI: A deep graph convolutional network framework for multi-label prediction of drug-drug interactions. Anal Biochem. 2022;646:114631.
35. Kang LP, Lin KB, Lu P, Yang F, Chen JP. Multitype drug interaction prediction based on the deep fusion of drug features and topological relationships. PLoS One. 2022;17(8):e0273764.
36. Hecker M, Frahm N, Zettl UK. Update and application of a deep learning model for the prediction of interactions between drugs used by patients with multiple sclerosis. Pharmaceutics. 2023;16(1):3.
37. Hu B, Yu Z, Li M. Mphgcl-ddi: meta-path-based heterogeneous graph contrastive learning for drug-drug interaction prediction. Molecules. 2024;29(11):2483.