Computers and Artificial Intelligence; Vol.3, No.4; 2026
E-ISSN: 3006-3728 P-ISSN: 3006-371X
DOI: https://doi.org/10.70267/cai.v3n4.6081
Published by: Zeus Press

Drug–Drug Interaction Prediction: Aligning Evidence, Models, Validation, and Clinical Use

Jingxiang Yang*

School of Computer Science and Technology, Tongji University, Shanghai, 200092, P. R. China

**Corresponding author: Jingxiang Yang.*

Abstract

Drug–drug interactions (DDIs) are a major source of preventable medication-related harm, but the evidence used to anticipate them spans molecular structure, pharmacology, biomedical text, knowledge graphs, and longitudinal clinical records. Computational studies often treat distinct targets as a single benchmark, even though binary interaction screening, event prediction, relation extraction, and patient-specific risk require different evidence and validation. This critical review introduces a claim-centered framework that aligns the clinical question, evidence source, prediction unit, model inductive bias, validation design, and intended use. Representative methods are compared by what they encode, the assumptions they impose, and their characteristic failure modes rather than by incomparable leaderboard scores. The analysis covers similarity and factorization methods, molecular and relational graph learning, pair-conditioned substructure models, contrastive and pretrained learning, multimodal fusion, and language-based systems. A validation-to-translation ladder then connects leakage-aware splitting, calibration, uncertainty, external testing, mechanistic corroboration, and prospective workflow evaluation. The synthesis indicates that architecture-level gains are conditional on label construction and distribution shift. It also shows that useful explanations require evidence alignment and actionability in addition to visual plausibility. This framework clarifies which claims current DDI benchmarks can support and what additional evidence is needed before predictions can guide prescribing or monitoring.

Keywords

drug–drug interaction; molecular representation; graph learning; benchmark validity; uncertainty; clinical translation

1. Introduction

Polypharmacy is common among older adults, people with multiple chronic conditions, and patients receiving complex anticancer or anti-infective regimens. Co-administration can alter absorption, distribution, metabolism, or excretion. It can also produce additive toxicity, antagonistic effects, formulation incompatibility, or patient-specific contraindications. These mechanisms differ in timing, dose dependence, reversibility, and clinical severity. A database association is therefore not equivalent to a clinically actionable risk.

Experimental interaction studies remain indispensable, but they cannot enumerate every pair, dose, formulation, genotype, disease state, and treatment sequence. Computational methods can prioritize pairs for

testing and combine evidence that is otherwise reviewed separately. Early systems relied on chemical similarity, manually designed pharmacological features, and network-based link prediction [1-3]. Representation learning reduced dependence on fixed descriptors, while graph neural networks (GNNs) encoded molecular topology and polypharmacy relations [4,5]. Subsequent methods introduced pair-conditioned substructures, heterogeneous graph Transformers, contrastive objectives, multimodal fusion, and language-based representations [6-10].

Existing reviews have organized the field around datasets, deep architectures, graph-based methods, and clinical applications [11-15]. These perspectives establish the breadth of computational DDI research. However, they do not fully resolve how a reported score should be interpreted when task definitions, evidence sources, candidate-pair universes, and deployment settings differ. The same architecture can estimate a curated database link, a typed pharmacological event, a relation stated in text, or an outcome in a patient record. These outputs do not support the same scientific or clinical claim.

The interpretive problem begins with the label. DrugBank, TWOSIDES, DDInter, and clinical sources encode different observation processes and different mixtures of positive, uncertain, and missing evidence [16-18]. Random pair splits may also place related compounds, repeated interaction templates, or records from the same source in both training and test sets. Performance can then reflect interpolation within an evidence-generating process rather than generalization to a new drug, scaffold, hospital, or time period. Model architecture, dataset construction, split design, and intended use must therefore be evaluated together.

The central contribution is a claim-centered framework for six layers of DDI research (Figure 1). It aligns the clinical question, evidence source, prediction unit, model inductive bias, validation design, and intended use. Three linked outputs make this framework operational. First, a task–evidence map separates interaction screening, event or mechanism prediction, text extraction, polypharmacy risk, and patient-specific decision support. Second, a mechanism-level taxonomy compares representative models by their inputs, computations, assumptions, and failure modes. Third, a validation-to-translation ladder links leakage-aware evaluation and calibration to mechanistic corroboration, external testing, and prospective workflow assessment. Together, these outputs clarify which conclusions are supported by current benchmarks and which require additional evidence.

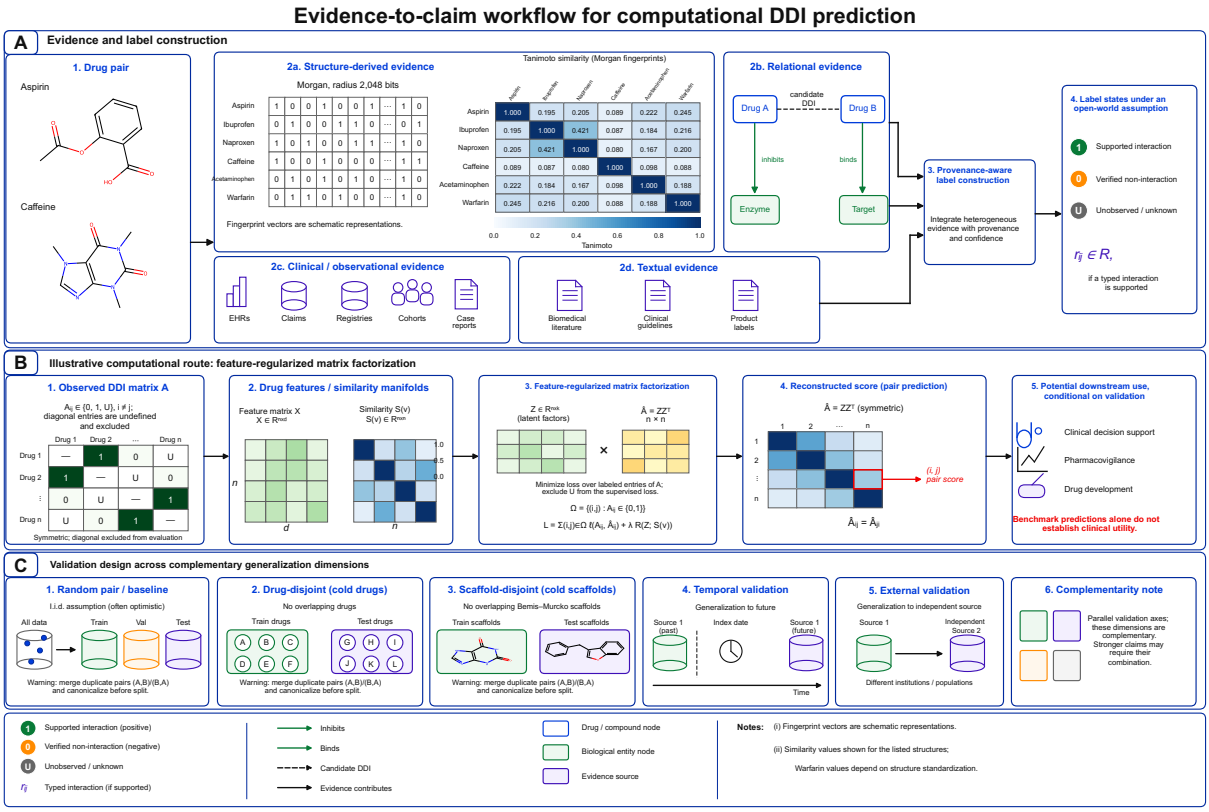
2. Review Scope and Methodology

This critical review focuses on computational methods for pairwise and polypharmacy DDI prediction, interaction extraction, mechanism attribution, and patient-specific risk estimation. It also draws on work in molecular representation, graph learning, uncertainty, calibration, and clinical prediction when those topics change how a DDI claim should be interpreted. The synthesis is organized by the clinical question, the evidence available to answer it, the inductive bias imposed by the representation and model, and the validation required for the intended use.

Representative studies were prioritized when their target, input evidence, prediction unit, and evaluation setting could be identified. Seminal models are included to establish methodological transitions, while later methods illustrate distinct mechanisms, assumptions, or validation challenges. The literature uses “interaction” for chemically, pharmacologically, textually, and clinically different outcomes. We therefore retain each study’s task definition and avoid comparisons when label semantics, candidate universes, or split protocols are incompatible.

The analysis is structured but not meta-analytic. Dataset versions, label semantics, negative sampling, candidate-pair universes, and splitting protocols are too heterogeneous for a pooled performance estimate to be meaningful. Each method is instead assessed through four questions: what it represents, what evidence it consumes, which distribution shift it has faced, and which failure would matter in practice.

Figure 1: Evidence-to-claim workflow for computational DDI prediction



Note. A, Evidence and label construction begins with a defined drug pair and combines structure-derived, relational, clinical or observational, and textual evidence under an open-world label model. B, An illustrative feature-regularized matrix-factorization route maps an observed DDI matrix and drug similarity information to a symmetric pair score; unobserved entries are excluded from the supervised loss. C, Random-pair, drug-disjoint, scaffold-disjoint, temporal, and external validation probe complementary forms of generalization. Molecular structures are rendered from RDKit, whereas fingerprints, matrices, and validation icons are schematic.

3. Clinical Questions and Computational Targets

3.1 Mechanisms and Consequences

Clinical DDI resources commonly distinguish pharmacokinetic, pharmacodynamic, pharmaceutical, and contraindication-related interactions [15,19,20]. Pharmacokinetic interactions change exposure through absorption, distribution, metabolism, or excretion. They often involve enzyme or transporter inhibition and induction, and their magnitude depends on dose, timing, organ function, and patient genotype. Pharmacodynamic interactions arise when drugs act on the same physiological pathway or produce convergent toxicity, such as additive bleeding or QT prolongation. Pharmaceutical interactions occur before administration, for example through incompatibility in a solution or infusion line. Contraindications encode a clinical decision boundary that may depend on disease, pregnancy, age, or a co-medication rather than on a molecular contact alone.

The mechanism determines the appropriate evidence. A structure-based model can identify shared motifs or physicochemical complementarity, but it cannot by itself infer an enzyme's induction time course. A knowledge graph can connect a drug to a target, pathway, or adverse event, but the edge may represent a hypothesis, a literature report, or a regulated label. An electronic health record (EHR) association can capture a clinically expressed signal, yet it is vulnerable to confounding by indication, prescribing selection, and documentation bias. A robust review therefore treats mechanisms as complementary views, not interchangeable labels.

3.2 Prediction Tasks

Let d_i and d_j denote two drugs, c the available context (e.g., dose, indication, genotype, laboratory values, or care setting), and t the observation time. The most general target is a conditional risk.

$$p(y \mid d_i, d_j, c, t), \tag{1}$$

where y may be a binary indicator, an interaction type, a severity category, a mechanism, a text relation, or a patient-level outcome. The following targets should be reported separately:

- **Pairwise screening:** estimate whether any interaction has been reported or is plausible for (d_i, d_j). This is useful for triage but is not a severity or causality judgment.
- **Event or mechanism prediction:** estimate a named consequence (for example, increased concentration, enzyme inhibition, or a specific adverse event). This is naturally multi-label because one pair can have several mechanisms and outcomes.
- **Interaction extraction:** identify drug entities and relations in biomedical text. The unit of prediction is a sentence, document, or evidence span rather than an unordered drug pair [21-23].
- **Polypharmacy risk:** estimate the incremental risk of a set of drugs, potentially conditioned on sequence, dose, and patient covariates. Pairwise scores are not additive and should not be interpreted as a complete regimen risk.
- **Patient-specific decision support:** estimate a calibrated probability or expected utility for a defined patient and action, such as reviewing an order or selecting monitoring intensity. This target requires outcome time windows, comparator definitions, and prospective workflow evaluation.

Table 1 summarizes the output and evidence requirements. A central principle is that a model should not be rewarded for predicting a less specific target than the intended clinical use.

Table 1: DDI tasks, outputs, and minimum evidence requirements.

Task	Typical output	Evidence that can support it	Evaluation priority
Pairwise screening	$\hat{y}_{ij} \in \{0,1\}$ or a risk score	curated labels, product information, literature, knowledge graphs	scaffold/temporal split, area under the precision–recall curve, calibration
Event/mechanism prediction	multi-label event vector or ranked mechanisms	pharmacology, exposure, adverse-event and mechanistic relations	per-label macro metrics, rare-event recall, evidence audit
Textual DDI extraction	entity spans, relation type, evidence span	annotated sentences/documents	entity/relation F1, span faithfulness, document-level generalization
Polypharmacy risk	regimen-level risk or incremental effect	longitudinal EHR, claims, dose and timing	patient-level split, confounding analysis, decision utility
Clinical decision support	probability plus recommended action/monitoring	validated labels and workflow constraints	calibration, alert burden, prospective or silent deployment

4. Evidence Resources and the Label-generation Problem

4.1 What a DDI Label Means

A positive label can originate from a regulated product document, controlled exposure study, case report, or relation extracted from text. These sources do not provide equivalent evidence. Negative labels are even more heterogeneous: a pair may have been tested with no detected interaction, or it may simply be absent from a database. Treating all unobserved pairs as negatives converts an open-world problem into a closed-world classification problem and can bias both training and evaluation [24].

For a dataset $D = \{(d_i, d_j, y_{ij})\}$, it is useful to distinguish three label states:

$$y_{ij} \in \mathcal{Y} = \{1, 0, \mathrm{U}\}, \tag{2}$$

where 1 denotes supported interaction evidence, 0 denotes evidence of no interaction under a defined protocol, and U denotes an unknown or unobserved pair. Positive–unlabeled or propensity-aware objectives are more defensible than indiscriminate negative sampling when the third state dominates. If a benchmark nevertheless uses sampled negatives, the sampling frame and its relation to the deployment universe must be reported.

4.2 Resource Families

Table 2 groups common evidence sources by what they measure. DrugBank and DDInter provide curated drug facts and interaction records. TWOSIDES captures adverse-event signals from spontaneous reports, while textual corpora provide relation mentions and evidence spans. EHR and claims data capture outcomes under real prescribing behavior. Molecular resources and knowledge graphs provide structural or mechanistic context [16-18]. These resources can be fused, but fusion does not remove their biases. A graph dominated by one source can encode source identity, while a post-treatment adverse event can leak outcome information into the input.

Table 2: Evidence resources and recurrent threats to validity.

Resource family	Observed unit	What it contributes	Caveat that must be disclosed
Curated interaction databases	drug pair and event/mechanism	normalized names, labels, and regulatory or expert evidence	coverage is selective; version changes alter the label universe
Spontaneous-report signals	pair and adverse-event signal	population-scale safety hypotheses	reporting, indication, and co-medication confounding; not direct causality
Biomedical text	mention, relation, evidence span	mechanisms, qualifiers, and temporal language	duplicate documents and annotation disagreement can leak across splits
Molecular resources	structure, assay, target, transporter, enzyme	mechanistic inductive bias and cold-start features	a structural proxy is not a clinical effect; salt and stereochemistry need normalization
Knowledge graphs	typed entities and relations	links among drugs, targets, pathways, phenotypes, and diseases	relation provenance and confidence are often mixed
EHR/claims data	patient, exposure, outcome, time	clinical relevance and context dependence	confounding by indication, missing adherence, coding drift, and site shift

Table 3. Representative benchmark resources and the question each can answer.

Resource	Prediction unit	Useful signal	Interpretation limit
DrugBank	curated drug pair and interaction description	normalized entities, mechanisms, and label text	curation and release history define coverage; absence is not a negative
DDInter	drug pair with typed interaction evidence	broader normalized interaction types and evidence links	identifier mapping and evidence aggregation can change the label set
TWOSIDES and related signal resources	drug pair–adverse event	population-scale safety hypotheses and rare-event signals	spontaneous-reporting and indication confounding; signal is not proof of causality
DDI corpus and related text datasets	sentence/document relation	interaction mentions, relation types, and evidence spans	annotation scope and document duplication constrain generalization
Molecular and biomedical graphs	drug, target, enzyme, transporter, pathway	mechanistic context and cold-start features	predicted or weak edges may be mistaken for validated evidence
EHR/claims cohorts	patient–exposure–outcome–time	clinically expressed risk under real prescribing	confounding, missing adherence, coding drift, and site shift

4.3 Dataset Versioning and Provenance

Every experiment should record the database release, identifier mapping, salt/brand normalization, deduplication rule, and label aggregation policy. When multiple records describe the same pair, the aggregation rule should preserve event type and evidence strength instead of silently collapsing them to a single positive. The same discipline applies to text: documents, sentences, and near-duplicate abstracts must be grouped before splitting. Versioned data cards should report the candidate-pair universe, the proportion of unknown pairs, the number of unique scaffolds, and whether labels were available before the nominal prediction time.

Figure 2 summarizes the resulting task–evidence relationship. Its purpose is not to rank data sources. Instead, it shows that a source can provide primary evidence for one target, supporting context for another, and insufficient evidence for a third.

5. Representing Drugs, Interactions, and Context

Representation determines what a model can learn before any optimization takes place. The useful question is not whether a representation is "advanced," but whether its invariances match the target and its missing information is acknowledged.

5.1 Sequences and Engineered Descriptors

Simplified molecular-input line-entry system (SMILES) strings, International Chemical Identifier (InChI) strings, fingerprints, physicochemical descriptors, and target or protein-sequence annotations offer compact and reproducible inputs. Convolutional or recurrent encoders can learn local sequence patterns, while fingerprints and similarity kernels provide strong low-variance baselines [3,25]. These representations are efficient and often effective for known-drug interpolation. They become brittle when stereochemistry, salts, formulation, dose, or time-dependent enzyme regulation is clinically important.

5.2 Molecular Graphs and 3-D Information

In a molecular graph $G = (V, E)$, nodes encode atoms and edges encode bonds or distance relations. A generic message-passing layer first aggregates messages from the neighbors of node v and then updates its representation:

$$\boldsymbol{m}_v^{(\ell)} \& = \underset{u \in \mathcal{N}(v)}{\mathrm{AGG}} \psi_\ell \left(\boldsymbol{h}_v^{(\ell)}, \boldsymbol{h}_u^{(\ell)}, \boldsymbol{e}_{uv} \right), \tag{3}$$

$$\boldsymbol{h}_v^{(\ell+1)} \& = \phi_\ell \left(\boldsymbol{h}_v^{(\ell)}, \boldsymbol{m}_v^{(\ell)} \right). \tag{4}$$

Here, $\mathcal{N}(v)$ is the neighborhood of node v, $\boldsymbol{m}_v^{(\ell)}$ is the aggregated message at layer ℓ, and AGG is a permutation-invariant operator such as a sum, mean, attention-weighted sum, or learned set function. Graph convolutional networks (GCNs) are stable and economical, while graph attention networks (GATs) allow differentiated neighbor weights. Substructure or motif models focus capacity on chemically meaningful regions [4,6,26]. Three-dimensional coordinates can encode geometry and stereochemical constraints, but they introduce conformer uncertainty and require careful handling of missing or multiple conformations.

5.3 Relational and Heterogeneous Context

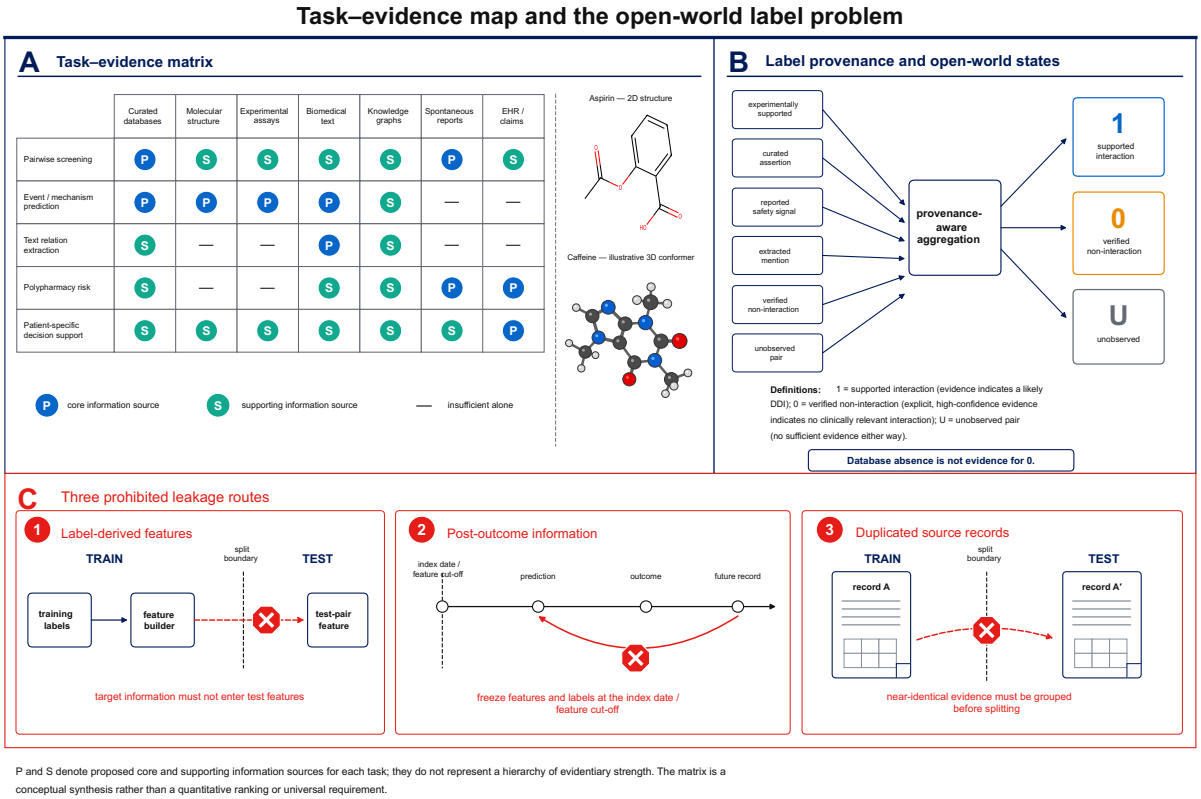
DDI prediction often needs a pair of drug graphs plus a relational context graph containing targets, enzymes, transporters, pathways, phenotypes, and diseases. Heterogeneous graph models can propagate typed evidence and use relation-specific parameters; knowledge-subgraph methods can also expose a candidate explanation path [7,27,28]. The main risk is provenance collapse: a model may treat a weak, predicted, or text-mined edge as equivalent to a validated relation. Edge confidence, source, and temporal availability should therefore be encoded or used for sensitivity analysis.

5.4 Text, Language, and Multimodal Representations

Text carries qualifiers that structures do not: "may increase," "no clinically significant change," dose ranges, timing, and patient subgroups. Relation extraction systems can identify interaction mentions and their evidence

spans [21-23]. Transformer encoders and language models can represent long-range context, while multimodal models combine text, structure, phenotypes, and graphs [9,29]. Their apparent flexibility increases the burden of provenance and factuality checks: a generated explanation is not evidence unless it is traceable to an input document or a validated mechanistic relation.

Figure 2: Task-evidence map and the open-world label problem



Note. A, The task-evidence matrix distinguishes primary information sources (P), supporting information sources (S), and sources that are insufficient on their own. B, Provenance-aware aggregation separates supported interactions (1), verified non-interactions (0), and unobserved pairs (U); database absence is not evidence for state 0. C, Three prohibited leakage routes are shown: label-derived test features, post-outcome information, and duplicate source records crossing the split boundary.

6. Computational Methods for DDI Prediction

Computational DDI methods can be separated along two axes. The first is the unit on which representation learning is performed: a descriptor vector, a molecular string, an atom–bond graph, a DDI network, a heterogeneous biomedical graph, or a text document. The second is the source of supervision: observed pair labels, typed interaction events, self-supervised views, or patient outcomes. Architectures with the same name may therefore solve different problems. For example, a GCN over atoms estimates a structure-dependent pair representation, whereas a GCN over the DrugBank interaction network primarily smooths information among drugs with known neighbors. The following classification preserves this distinction and relates each model to the evidence it actually consumes.

6.1 Similarity Inference and Feature-engineered Classifiers

Similarity-based methods begin with the neighborhood principle: if d_i resembles d_k and d_k interacts with d_j, the pair (d_i, d_j) receives supporting evidence. With modality-specific similarities S_m and known interaction matrix Y, a generic score can be written as

$$\hat{y}_{ij} = \sum_{m=1}^{M} \alpha_m \frac{\sum_{k \neq i} S_m(i,k) Y_{kj}}{\sum_{k \neq i} S_m(i,k) + \epsilon}, \qquad \alpha_m \geq 0, \quad \sum_m \alpha_m = 1. \tag{5}$$

The equation makes the inductive bias explicit: an unknown pair is estimated from the interaction profiles of similar drugs. INDI converts chemical, ligand, side-effect, therapeutic, and related similarities into pair-level features and uses logistic regression to infer both interactions and management recommendations [1]. Vilar et al. use molecular fingerprints and interaction-profile similarity for large-scale screening, making the contributing neighbors inspectable [3]. Cheng and Zhao combine phenotypic, therapeutic, chemical, and genomic properties with conventional classifiers, while Fokoue et al. formulate the problem as large-scale similarity-based link prediction [2,25]. Sridhar et al. go beyond independent pair classification by using a probabilistic collective model, so related DDI variables can influence one another [30].

These models remain important because they are inexpensive, relatively transparent, and difficult to beat when the test set contains drugs close to the training set. Their apparent simplicity also exposes the principal risk: similarity may reflect database coverage rather than pharmacological equivalence. Structural analogs can differ in stereochemistry, metabolism, dose-response, or transporter affinity, and interaction-profile similarity is unreliable for a new drug with few recorded neighbors. Consequently, every neural model should be compared with a tuned similarity baseline under both random-pair and drug-disjoint evaluation.

6.2 Matrix Factorization, Tensor Factorization, and Matrix Completion

Matrix-factorization methods regard the DDI matrix as a partially observed low-rank object. A common objective is

$$\min_{U,V} \left\| W \odot (Y - UV^{\top}) \right\|_F^2 + \lambda(\|U\|_F^2 + \|V\|_F^2) + \sum_m \gamma_m \operatorname{tr}(U^{\top} L_m U), \tag{6}$$

where W marks observed or sampled entries, U and V are latent drug factors, and L_m is a graph Laplacian derived from one drug-similarity modality. For symmetric binary DDI matrices, the reconstruction is often constrained toward $UU^{\top}$; typed events require a relation-specific matrix or a drug–drug–event tensor. The graph term preserves neighbourhoods defined by structures, targets, enzymes, pathways, indications, or adverse effects.

The variants differ mainly in how they constrain the latent space. MRMF introduces feature-derived manifold regularization into matrix completion, testing separate manifolds constructed from substructures, targets, enzymes, transporters, pathways, indications, and side effects [31]. BRSNMF uses a semi-nonnegative factorization with a balance constraint to model signed enhancive and degressive relations, and adds drug-binding protein information for cold-start inference [32]. ISCMF first integrates several similarity networks, including Gaussian interaction-profile similarity, and then constrains the factors by the fused similarity geometry [33]. TMFUF uses triple matrix factorization to connect side-effect information with signed interaction changes and explicitly targets new-drug prediction [34]. GRPMF casts the symmetric DDI matrix probabilistically and learns a graph regularizer informed by expert chemical similarity [35].

Factorization is most defensible when the scientific target is completion of a fixed interaction matrix. It becomes less suitable when the output depends on atom-level motifs, exposure, or patient context. Latent dimensions are not automatically biological mechanisms, and a random entry-wise split leaks drug-specific factors into both training and test sets. Cold-start claims require withholding all interactions of the target drug and restricting side information to evidence available at the prediction date.

6.3 Feed-forward, Convolutional, and Recurrent Molecular Models

Early deep DDI models replaced the final linear classifier while retaining engineered molecular inputs. DeepDDI converts each compound into a structural similarity profile against a reference drug set, reduces the profile, concatenates an ordered drug pair, and applies a multilabel deep neural network (DNN). The model predicts 86 DrugBank-derived interaction types expressed through sentence templates [5]. This design was influential because it moved beyond binary screening, but its input remains a similarity profile. Performance can depend on the coverage and redundancy of the reference panel, and drug order must be treated consistently for directional event descriptions.

DDIMDL expands event prediction to four DrugBank feature families—chemical substructures, targets, enzymes, and pathways. Separate subnetworks learn modality-specific pair representations, which are joined by a DNN to predict 65 events [29]. CNN–DDI adds a feature-selection stage and applies convolutions to the

resulting multimodal representation for event classification [36]. These models demonstrate that a dense vector can support fine-grained event prediction, but the input features are often derived from the same curated resource that supplies the labels. A valid split must therefore remove label-derived attributes and duplicated interaction descriptions from the feature-generation pipeline.

Sequence encoders instead operate on SMILES tokens. One-dimensional CNNs detect recurring local token patterns, whereas bidirectional RNNs aggregate dependencies in both directions; Transformers allow direct long-range attention. SMILES are compact and widely available, but they are not a unique representation of a molecule. Randomized SMILES can be useful augmentation, while canonicalization, stereochemical tokens, disconnected salts, and maximum sequence length should be reported. A sequence model should also be compared with a molecular-graph encoder under scaffold-disjoint evaluation, because high random-split accuracy does not establish that token motifs generalize to new chemistry.

6.4 DDI Extraction from Biomedical Text

Text extraction is related to, but not interchangeable with, molecular DDI prediction. Its input is a sentence or document containing drug mentions, and its output is a relation label and ideally an evidence span. The DDI corpus and SemEval-2013 Task 9 define four positive classes—mechanism, effect, advice, and an underspecified interaction class—plus a non-interaction class [21,22]. Thus, a text model can recover what a document states; it does not independently establish that the stated interaction is causal or clinically important.

The CNN system of Liu et al. represents each candidate pair with word and relative-position embeddings, applies convolution and max pooling, and classifies all within-sentence drug pairs without extensive syntactic feature engineering [37]. Character–word RNNs add character-level morphology to word-level recurrent context, which helps with variable biomedical names [38]. Later attention-based BiLSTM systems focus on pair-relevant tokens or shortest dependency paths, while biomedical Transformer encoders provide contextual representations pretrained on domain text [23,39]. The critical evaluation unit is the source document: sentence-level randomization can place nearly identical label language or repeated DrugBank descriptions in both partitions. End-to-end systems should separately report entity-recognition errors, relation-classification errors, cross-sentence failures, negation, speculation, and evidence-span faithfulness.

6.5 Molecular GNNs and Pairwise Interaction Networks

For a molecular graph $G = (V, E)$, a message-passing layer updates atom v by

$$\boldsymbol{h}_v^{(\ell+1)} = \phi_\ell \left(\boldsymbol{h}_v^{(\ell)}, \sum_{u \in \mathcal{N}(v)} a_{uv}^{(\ell)} \, \psi_\ell(\boldsymbol{h}_u^{(\ell)}, \boldsymbol{e}_{uv}) \right), \tag{7}$$

where $a_{uv} = 1/c_{uv}$ yields a GCN-like normalized aggregation and learned a_{uv} yields attention or gating. A readout produces one embedding per drug, after which a symmetric or relation-conditioned pair decoder predicts the DDI. This architecture learns atom environments directly, but standard message passing is local and can oversmooth after many layers.

DDI-network GNNs operate at a different scale. Decagon constructs a multimodal graph containing drugs, proteins, drug–target edges, protein–protein edges, and 964 polypharmacy side-effect relation types. A multirelational GCN encoder propagates across this graph and a tensor-factorization decoder scores drug–drug–side-effect triples [4]. The design shares statistical strength across rare side effects, but it is transductive for the drug nodes present in the graph and inherits reporting bias from TWOSIDES. SmileGNN combines a SMILES-derived structural representation with a drug's topological representation in a knowledge graph [40]. ACDGNN constructs a heterogeneous biological network and uses cross-domain transformations and attention to align information from drugs and other biomedical entities; it explicitly evaluates both transductive and inductive settings [41]. AutoDDI searches GNN architecture choices for the DDI graph rather than fixing the number and type of aggregation layers manually [42].

These models must be described by their graph, not only by the term "GNN." Molecular-graph encoders can represent a previously unseen drug if its structure is available. DDI-network encoders may fail for a zero-degree drug. Heterogeneous graph models can alleviate that failure using targets or genes, but only if those relations are genuinely available at deployment. Random edge splits are especially vulnerable to hub and two-hop leakage; drug-disjoint, relation-disjoint, scaffold-disjoint, and time-sliced tests answer different questions.

6.6 Substructure and Pair-conditioned Molecular Reasoning

Whole-molecule pooling can dilute the small functional groups that determine a pair-specific interaction. CASTER first mines a vocabulary of recurring chemical substructures, learns an autoencoded representation from labeled and unlabeled compounds, and uses dictionary-learning coefficients to expose which substructures support a binary prediction [43]. SSI–DDI removes the fixed dictionary: graph-attention layers at different receptive fields generate candidate substructures for each drug, a co-attention matrix weights every cross-drug substructure pair, and a relation-specific bilinear decoder aggregates their contributions [6]. Its score has the form

$$s_r(d_i, d_j) = \sum_{p,q} \alpha_{pq}^{(r)} \, (\boldsymbol{g}_{i,p})^{\top} M_r \boldsymbol{g}_{j,q}, \tag{8}$$

where $\boldsymbol{g}_{i,p}$ and $\boldsymbol{g}_{j,q}$ represent receptive-field substructures and $\alpha_{pq}^{(r)}$ is pair relevance.

R²-DDI makes the predicted relation part of representation learning: drug features and the candidate relation refine one another, while multi-branch consistency training regularizes the prediction [44]. CGIB takes a complementary information-theoretic approach, selecting a minimal subgraph of one molecule that remains sufficient for the target conditional on the paired molecule [45]. Customized subgraph selection and motif-oriented topology refinement further replace a fixed receptive field with target-pair-dependent selection [26,46]. These methods provide a more plausible unit of explanation than an undifferentiated drug embedding, but attention or selection weights are still model evidence, not confirmation of a metabolic or pharmacodynamic mechanism. Fidelity requires counterfactual removal, stability across seeds and conformers, and agreement with independently sourced target, enzyme, or transporter evidence.

6.7 Knowledge-graph Embedding and Pair-specific Subgraphs

Knowledge-graph methods represent facts as typed triples (h, r, t) and score candidate drug–relation–drug triples. Translational, bilinear, and neural decoders are efficient, but their performance depends on negative sampling and on whether inverse or logically equivalent edges cross the split. Dai et al. use a Wasserstein adversarial autoencoder to generate hard negative representations for DDI knowledge-graph embedding, addressing the weakness of uniformly sampled corruptions [27]. Such negatives are computational hypotheses; they are not verified non-interactions.

Pair-specific subgraph methods attempt to retain the paths that connect two target drugs. KnowDDI merges a DDI graph with an external biomedical knowledge graph, extracts a drug-flow subgraph, and iteratively learns edge strengths to retain a smaller explanatory knowledge subgraph [28]. TIGER processes the molecular graph and heterogeneous biomedical graph in separate relation-aware Transformer channels and exchanges features at node and graph levels [7]. The two designs address different limitations: KnowDDI emphasizes local relational paths and sparsity, whereas TIGER preserves both molecular structure and high-order biomedical context. For either approach, provenance is essential. A path is uninformative if it includes a DDI assertion derived from the test label, a post-outcome adverse event, or a predicted edge whose uncertainty is hidden.

6.8 Contrastive, Self-supervised, and Pretrained Learning

Contrastive learning uses agreement between views as an auxiliary signal. For positive views (z_i, z_i^+) and a comparison set $\mathcal{B}$, a common objective is

$$\mathcal{L}_{\text{con}} = -\sum_i \log \frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\sum_{k \in \mathcal{B}} \exp(\text{sim}(z_i, z_k)/\tau)}, \tag{9}$$

with temperature τ. MIRACLE treats the DDI network as a graph whose nodes are themselves molecular graphs: a GCN encodes inter-drug network context, a bond-aware message-passing network encodes intramolecular structure, and contrastive learning aligns the two views [8]. CGIB uses conditional compression rather than simple view agreement, while more recent visual-molecule pretraining constructs self-supervised structural views before DDI fine-tuning [10,45].

The central design decision is what should remain invariant. Randomly deleting a bond, stereocentre, functional group, or rare relation can erase the causal signal; treating two same-drug views as positive can also

enforce source identity more strongly than interaction mechanism. Studies should specify every augmentation, measure how often it changes chemically meaningful motifs, and include non-contrastive pretrained and randomly initialized controls. Pretraining and test data must also be time-audited when the pretrained graph or corpus may already contain the target interaction.

6.9 Multimodal Fusion, Attention, and Siamese Comparison

Multimodal methods combine evidence that is incomplete in different ways. A general late-fusion model encodes modality m as $z_i^{(m)}$ and computes

$$z_i = \sum_m \beta_i^{(m)} z_i^{(m)}, \qquad \beta_i^{(m)} = \mathrm{softmax}_m(q_i^\top k_i^{(m)}), \tag{10}$$

where the weights may depend on the drug or pair. Early concatenation is simpler but assumes all modalities are present and commensurate; late fusion permits modality-specific encoders; cross-attention models interactions among modalities but increases capacity and leakage risk.

AttentionDDI applies a shared Siamese Transformer encoder separately to each drug's chemical, target, pathway, gene-expression, and related similarity vectors. Feature attention pools the modalities, and the classifier combines the two drug embeddings with their distance [47]. DDIMDL instead trains one subnetwork per feature family before joint event classification [29]. MDF–SA–DDI forms four pair representations through a Siamese network, a CNN, and two autoencoders, then uses Transformer self-attention to fuse the latent vectors for multi-type event prediction [9]. These architectures can exploit complementary evidence, but an attention weight is not a calibrated measure of biological importance. Missing-modality masks, modality-dropout tests, single-modality baselines, and source-disjoint evaluation are necessary to determine whether fusion adds information rather than redundancy.

6.10 Explanations, Rare Events, and Language Interfaces

DDI explanations occur at several levels. CASTER exposes dictionary coefficients; SSI–DDI and CGIB identify pair-conditioned molecular regions; KnowDDI selects relational paths; and ExDDI generates natural-language rationales for predicted relations [6,28,43,45,48]. These outputs answer different questions. A fragment or graph path may be faithful to a model without being clinically causal, while a fluent sentence may be plausible without being entailed by any retrieved evidence. Explanation evaluation should therefore separate predictive fidelity, chemical or relational plausibility, evidence entailment, stability, and usefulness to a pharmacist.

Long-tailed relation frequencies add a second challenge. RareDDIE formulates rare-event prediction as metric-based meta-learning. Chemical substructure encoding and task-guided biological-neighborhood aggregation form dual-granular drug features, while a variational pair representation maps support and query pairs into a shared event metric space [49]. Its language-augmented variant aligns event descriptions with this space for zero-shot prediction. Such methods should report per-class precision–recall curves, macro-averaged metrics, and uncertainty for rare classes. Micro averages can conceal complete failure on clinically important events. Distribution-aware benchmarks such as DDIBen also show that model rankings can change with drug and relation distributions [50]. Language models are therefore best used as retrieval, normalization, or explanation interfaces around a validated predictor. Citations and abstention are necessary when evidence is incomplete.

Table 4: Representative DDI models and the methodological distinction contributed by each. Reported scores are not pooled because datasets, negative sampling, and split protocols differ.

Model	Task	Input evidence	Core computation	Primary audit question
INDI (2012)	binary DDI and recommendation	seven drug-similarity families	pair features with logistic regression	Are recommendations independently supported?
Vilar (2014)	binary screening	fingerprints and interaction profiles	nearest-neighbor similarity inference	Do close analogs cross the split?
MRMF (2018)	matrix completion	DDI matrix and feature manifolds	low-rank factors with Laplacian regularization	Is the test drug represented during training?

Model	Task	Input evidence	Core computation	Primary audit question
TMFUF (2018)	signed DDI	side effects and interaction matrices	triple matrix factorization	Are enhancive and degressive labels reliable?
DeepDDI (2018)	multi-type event	SMILES-derived similarity profiles	dimension reduction and multilabel DNN	Is performance tied to the reference drug panel?
Decagon (2018)	polypharmacy side effect	drug–protein–side-effect graph	multirelational GCN plus tensor decoder	Are reporting-signal edges treated as causal facts?
BRSNMF (2019)	signed DDI/cold start	DDI network and binding proteins	balanced semi-nonnegative matrix factorization	Is cold start fully drug-disjoint?
DDIMDL (2020)	65 event types	substructures, targets, enzymes, pathways	modality-specific DNNs plus joint fusion	Were label-derived features removed?
CASTER (2020)	binary DDI	mined chemical substructures	autoencoding and dictionary learning	Are selected fragments stable and faithful?
ISCMF (2020)	binary DDI	fused similarity networks	similarity-constrained factorization	Does similarity fusion include interaction profiles?
SSI–DDI (2021)	typed DDI	paired molecular graphs	multi-scale GAT readouts, co-attention, bilinear decoder	Do selected substructure pairs survive perturbation?
MIRACLE (2021)	binary DDI	molecular graphs and DDI network	dual encoders with graph contrastive alignment	Are graph augmentations chemically valid?
AttentionDDI (2021)	binary DDI	multimodal similarity vectors	Siamese self-attention encoder	Are attention weights stable under missing modalities?
MDF–SA–DDI (2022)	event type	substructure, target, enzyme features	four pair encoders plus Transformer fusion	Does higher capacity exploit duplicated sources?
SmileGNN (2022)	binary DDI	SMILES features and knowledge-graph topology	structural/topological GNN fusion	Can a zero-degree drug be represented?
R^2-DDI (2023)	typed DDI	molecular graphs and relation labels	relation-aware feature refinement and consistency	Are gains retained for rare/unseen relations?
ACDGNN (2023)	binary DDI	biological heterogeneous network	cross-domain transformation and attention	Which entity types are available at deployment?
CGIB (2023)	molecular relation	paired molecular graphs	conditional information bottleneck	Does the selected core subgraph remain sufficient?
KnowDDI (2024)	typed DDI and rationale	DDI graph and external KG	drug-flow extraction and learned knowledge subgraph	Do explanatory paths contain target leakage?
TIGER (2024)	typed DDI	molecular and heterogeneous graphs	dual-channel relation-aware graph Transformer	What is gained by each channel under ablation?
AutoDDI (2024)	binary DDI	DDI graph	automated GNN architecture search	Was architecture selection nested within training data?
RareDDIE (2025)	few-/zero-shot event type	molecular, biological-neighborhood, and event-text features	metric meta-learning with variational pair representation	Are event-disjoint macro precision–recall and uncertainty reported?
ExDDI (2025)	textual explanation	predicted DDI and language evidence	generated natural-language rationale	Is each claim entailed by cited evidence?

Table 5: Method families, inductive biases, and appropriate claims.

Family	Primary bias	Strength	Failure mode	Defensible use
Similarity / factorization	sharing among related drugs or latent factors	strong low-variance baseline under sparse labels	profile leakage, popularity bias, weak mechanism specificity	completion and candidate prioritization
Molecular sequence / graph	compositional tokens or local atom topology	supports unseen compounds from structure	scaffold shift, stereochemical loss, oversmoothing	structure-conditioned screening
DDI / heterogeneous graph	relational propagation among biomedical entities	integrates interaction and biological context	hub leakage, missing entities, circular edges	typed link prediction with provenance audit
Substructure / pair reasoning	partner-dependent motifs and cross-drug interactions	localized chemical hypotheses	unstable or non-causal explanations	mechanistic hypothesis generation
Contrastive / pretrained	invariance across constructed views	label efficiency and multimodal alignment	invalid augmentations and pretraining contamination	low-label learning with augmentation audits
Text / language	contextual token and document evidence	captures qualifiers, negation, dose, and advice	document leakage and unsupported generation	evidence extraction and grounded summarization
Patient-context model	conditional risk under dose, time, and covariates	closest to a clinical decision	confounding, missingness, site shift, alert burden	prospective or silent clinical evaluation

Figures 3 and 4 convert the detailed comparisons above into a two-part mechanism-level map. Figure 3 summarizes established computational paradigms, including similarity-profile, interaction-matrix, pair-conditioned chemical, and knowledge-augmented graph methods. Figure 4 continues the same representation-to-claim logic for long-tail learning, text-guided explanation, and patient-contextualized temporal risk modeling. The families are not chronological and do not imply that later or more complex models are universally preferable.

7. Benchmark Validity: From a Score to a Defensible Claim

7.1 The Split Defines the Question

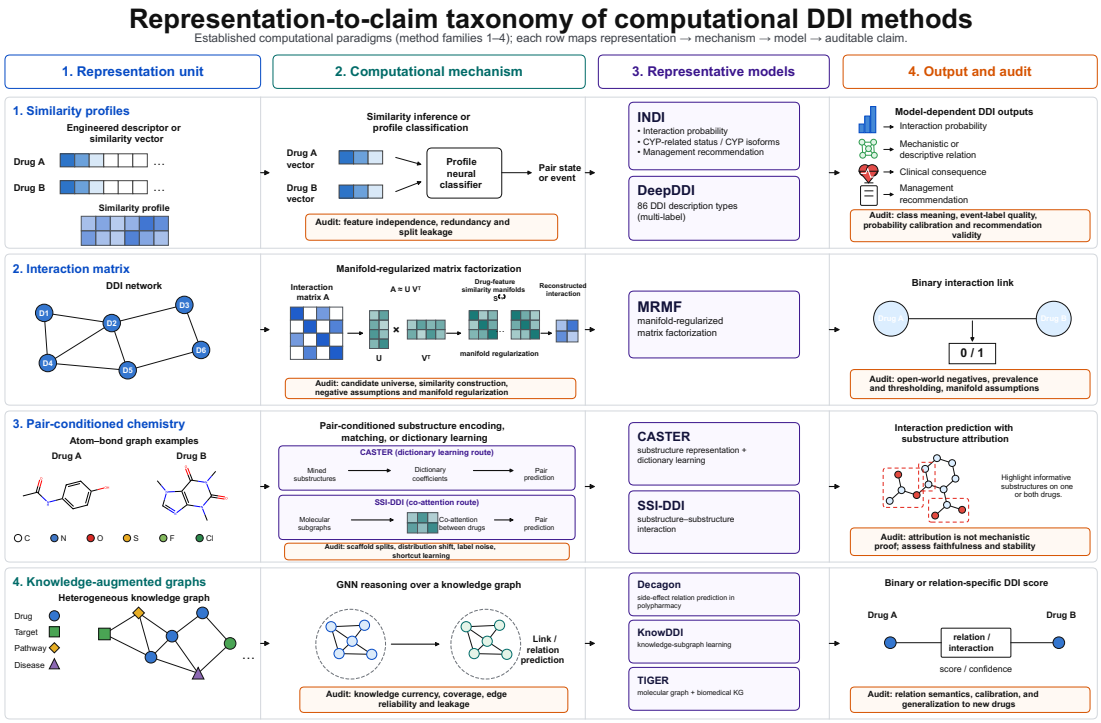
The test split is part of the scientific claim. A random pair split asks whether a model can interpolate among drugs and evidence patterns that resemble training examples. A drug-disjoint split tests generalization to an unseen drug, while a scaffold-disjoint split tests unfamiliar chemistry. A temporal split tests future evidence using only information available at an earlier date. External validation across institutions or data sources probes dependence on local documentation and prescribing habits. These settings should be reported side by side when a method is intended for discovery and clinical use.

Distribution-shift benchmarks show that rankings can change substantially when the candidate universe or relation distribution changes [50]. Cold-start evaluation is particularly important for new molecular entities, combination products, and rare interaction types. Few-shot or out-of-graph methods can help, but they do not eliminate the need to define what information is available at prediction time [51].

7.2 Leakage Controls

At minimum, a DDI study should: (i) deduplicate salts, brands, stereoisomers, and near-identical structures; (ii) group documents and source records before splitting; (iii) construct features using only data available before the prediction timestamp; (iv) prevent target-derived edges or adverse-event labels from entering the input graph; and (v) publish the exact candidate-pair universe and negative-sampling procedure. If a graph is built once before splitting, its edges can connect a test drug to training labels and silently invalidate the estimate. Inductive pipelines that rebuild the graph per fold are safer.

Figure 3: Representation-to-claim taxonomy of established computational DDI methods (families 1-4), linking representations, mechanisms, representative models, output claims, and audit requirements



Note. RDKit-generated structures and atom-bond graphs are illustrative; red attribution boxes are schematic, not mechanistic evidence.

7.3 Metrics, Uncertainty, and Calibration

The area under the receiver-operating-characteristic curve (AUROC) is insensitive to prevalence and can look impressive for rare outcomes. Precision–recall curves, recall at a fixed review budget, precision among the top *k* alerts, and macro-averaged scores across event types are more informative for triage. Probability outputs should be calibrated, for example by reporting reliability diagrams, expected calibration error, and calibration slope/intercept [52]. Confidence intervals should account for pair or patient clustering, and repeated splits or bootstrap estimates should be used when datasets are small. A model that abstains on out-of-distribution pairs may be safer than one that produces a confident score for every pair.

7.4 A validation Ladder

Table 6 gives a minimum reporting ladder. A manuscript need not perform every experiment, but it should state which claim each experiment supports and which claim remains untested. Statistical significance is not a substitute for clinical relevance: a small gain in AUROC may be immaterial if calibration or alert burden worsens.

Table 6: Validation ladder for DDI prediction studies.

Level	Question	Required disclosure
S0: internal random	Can the method fit the observed benchmark?	pair universe, sampling, repeated seeds, uncertainty
S1: entity/scaffold disjoint	Can it extrapolate to unfamiliar drugs or chemistry?	disjointness definition, scaffold algorithm, cold-start coverage
S2: temporal	Can it predict evidence that was not available at training time?	index date, feature cut-off, database versions
S3: external	Does it transfer across source, institution, or geography?	site/source shift, label harmonization, missingness
S4: prospective or silent deployment	Does it improve a defined clinical or review workflow?	alert burden, override rate, calibration, safety monitoring

Figure 5 extends the validation levels into an evidence ladder. Advancement between levels requires a distinct test rather than a new interpretation of the same retrospective score. The ladder therefore separates internal interpolation, chemical or entity extrapolation, temporal and external transport, prospective workflow feasibility, and comparative impact evaluation.

8. Interpretability, Mechanistic Evidence, and Uncertainty

An explanation has at least four distinct properties. It should be *plausible* to a domain expert, *faithful* to the computation that produced the score, *stable* under small irrelevant perturbations, and *useful* for deciding what to verify or monitor. Attention maps, highlighted atoms, or generated prose may satisfy the first property while failing the others. Local interpretable model-agnostic explanations (LIME) and GNNExplainer are useful probes, but they do not guarantee causal or pharmacological validity [53,54].

We recommend reporting a four-part explanation audit:

1) **Faithfulness:** remove or mask the reported atoms, relations, or text spans and measure the change in the prediction.
2) **Stability:** repeat the explanation across seeds, equivalent molecular encodings, and small perturbations that should not change the mechanism.
3) **Evidence alignment:** compare the explanation with an independent source, such as an enzyme, transporter, pathway, or cited evidence span, while preserving disagreements rather than forcing consensus.
4) **Actionability:** specify what a pharmacist, clinician, or experimenter should do differently because of the explanation.

Uncertainty should be decomposed into at least aleatoric uncertainty (ambiguous or variable biology), epistemic uncertainty (limited training support), and label uncertainty (conflicting or incomplete evidence). Ensembles, conformal prediction, Bayesian approximations, or calibrated abstention can quantify some of these components, but uncertainty intervals are meaningful only when the calibration population resembles the deployment population. A confidence score attached to a memorized drug pair is not evidence of confidence for a new scaffold.

9. Clinical Translation

9.1 From Interaction Probability to a Clinical Action

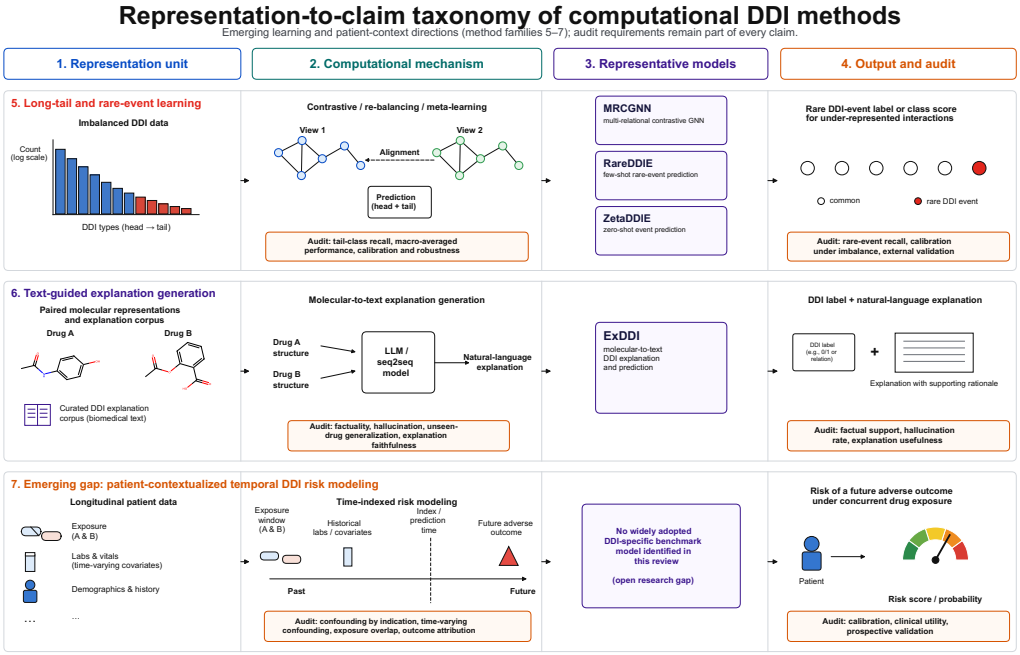
Clinical decision support requires an action threshold, not just a ranking. The threshold should reflect the relative cost of a missed serious interaction, an unnecessary alert, and an avoidable treatment substitution. For a patient p, a useful output is a calibrated risk together with the evidence type, expected onset, severity, and recommended verification or monitoring action. A pairwise score should be presented as one component of this record, not as an autonomous contraindication.

The pharmacokinetic and pharmacodynamic mechanisms described above imply different monitoring plans. An exposure-mediated interaction may require dose adjustment or a laboratory measurement. Additive toxicity may require an electrocardiogram or symptom surveillance, while formulation incompatibility may be prevented by changing the administration route. Systems that omit mechanism and timing force clinicians to reconstruct the reasoning and increase alert fatigue.

9.2 EHR Prediction and Causal Caution

Clinical data can reveal interactions not represented in curated resources, but observational associations are affected by treatment indication, disease severity, co-prescribing, adherence, and ascertainment. A model may therefore predict that two drugs are associated with an adverse event without showing that their combination caused it. Target-trial thinking, negative-control analyses, time-varying exposure definitions, and site-level external validation are appropriate safeguards. Subgroup performance should be reported for age, renal/hepatic function, sex, ethnicity when available, and care setting; aggregate performance can hide clinically important failures.

Figure 4: Representation-to-claim taxonomy of emerging DDI directions (families 5-7): rare-event learning, text-guided explanation, and patient-contextualized temporal risk modeling. The patient-context row denotes an open research gap rather than an established DDI benchmark



9.3 Prospective Evaluation and Governance

The translation ladder should proceed from retrospective validation, to silent deployment, to a monitored decision-support trial. Outcomes should include calibration, time-to-review, alert acceptance and override, medication changes, and patient safety indicators. Governance requires versioned models, provenance for each evidence edge, an abstention policy, an audit trail, and a mechanism for correcting or retracting an interaction record. Privacy-preserving training can be valuable across institutions, but it does not remove the need to harmonize labels and workflows.

10. Findings and Critical Synthesis

Five findings recur across evidence sources and method families. Each is conditional on the task definition and evaluation setting.

1) **Representation gains are task-dependent.** Molecular graphs and pair-conditioned substructures are informative when the target depends on local chemistry. Text and relational evidence become necessary when qualifiers, mechanisms, or clinical context define the output [6,7,28]. No representation dominates across all DDI tasks.

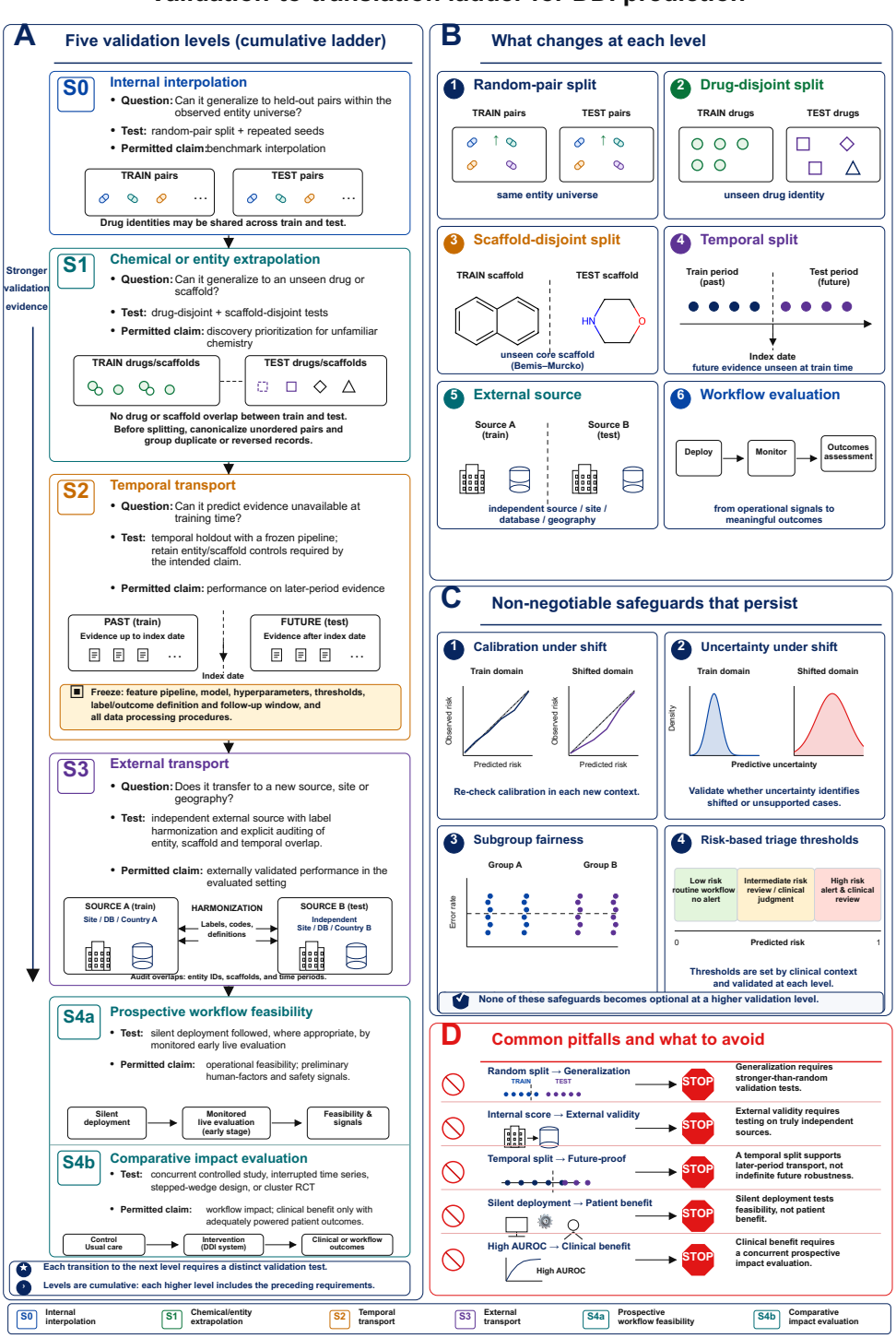
2) **Benchmark construction can dominate model ranking.** Candidate-pair selection, negative sampling, graph construction, and split design can alter rankings independently of architecture [50]. Claims of superiority therefore require leakage audits and distribution-shift tests.

3) **Rare-event performance requires dedicated evaluation.** Macro-averaged precision–recall, recall under a review budget, calibration, and uncertainty reveal failures hidden by micro-averaged discrimination [49]. This distinction matters when clinical cost is concentrated in a small number of severe events.

4) **Interpretability is an evidence problem.** A useful explanation links a prediction to a verifiable substructure, relation, or evidence span and identifies what remains uncertain [28,45,48]. Visual saliency or fluent prose alone does not establish faithfulness or pharmacological validity.

5) **Clinical value requires workflow evidence.** Retrospective discrimination is necessary but does not show that a system reduces harmful exposure, review time, or alert burden. These claims require calibration under deployment shift and prospective or silent evaluation [15,55].

Figure 5: Validation-to-translation ladder for DDI prediction.



Note. A, S0 evaluates interpolation within an observed benchmark; S1 tests unseen drugs or scaffolds; S2 freezes features and labels at an index date; S3 evaluates an independent source, site, or geography; and S4 separates prospective workflow feasibility from comparative impact evaluation. B, Six compact modules show the distribution or evidence change introduced by random-pair, drug-disjoint, scaffold-disjoint, temporal, external-source, and workflow evaluation. C, Calibration, uncertainty assessment, subgroup fairness, and risk-based triage thresholds remain necessary at every level. D, Stop symbols mark common inferential jumps that cannot be justified by reinterpreting the same retrospective score. Curves and icons are conceptual rather than empirical results.

11. Open Problems and Research Agenda

11.1 Open-world and Evolving Labels

Future benchmarks should represent unknown pairs explicitly, preserve evidence strength, and support positive–unlabeled or selective prediction objectives. Versioned temporal splits are needed to measure how quickly a model adapts when labels and prescribing practices change.

11.2 Mechanism-aware Multimodal Learning

The next generation of models should align structure, target biology, exposure, text, and patient context at the level of a named mechanism and time window. A useful design is a modular system in which each evidence source can be ablated, provenance-traced, and assigned a calibrated reliability. This is preferable to an opaque fusion layer that cannot reveal which source drove a high-risk alert.

11.3 Data-efficient and Distribution-aware Learning

Self-supervised pretraining, active learning, conformal prediction, and domain adaptation are promising for rare interactions and new scaffolds. Their evaluation should report the amount and type of novelty, not only the number of held-out pairs. Synthetic augmentation should preserve stereochemistry, pharmacophore constraints, and clinically meaningful dose or timing information.

11.4 Causal and patient-specific Reasoning

Pairwise association is an intermediate representation. Patient-specific systems need explicit exposure windows, comparator treatment strategies, time-varying covariates, and causal sensitivity analyses. Hybrid models that combine mechanistic constraints with observational learning may offer a principled way to separate plausible biological effects from prescribing artifacts.

11.5 Reproducibility and Reporting

A reproducible DDI paper should release code, preprocessing, identifier maps, split files, model checkpoints, and a data card or executable description of label construction. It should report all major baselines, failed ablations, confidence intervals, calibration, and the exact information available at prediction time. Transparent reporting principles from clinical prediction research provide a useful minimum standard [55].

12. Conclusion

DDI prediction is best understood as a claim-alignment problem. Clinical questions, evidence sources, prediction units, model assumptions, validation designs, and intended uses must be specified together. The literature supports the value of molecular, relational, textual, and patient-level representations, but their advantages remain conditional on label construction and distribution shift.

The framework developed here provides a practical way to interpret those conditions. Task–evidence mapping identifies which inputs can support a target. Mechanism-level taxonomy exposes what a model assumes and where it can fail. The validation-to-translation ladder defines the additional evidence required for claims about new drugs, future data, external settings, or workflow benefit. This review does not establish clinical effectiveness or justify pooled rankings across incompatible benchmarks. It instead defines when a computational score is an auditable hypothesis and when stronger mechanistic or prospective evidence is still required.

References

[1] Assaf Gottlieb, Gideon Y. Stein, Yoram Oron, Eytan Ruppin, and Roded Sharan. INDI: A computational framework for inferring drug interactions and their associated recommendations. Molecular Systems Biology, 8: 592, 2012. https://doi.org/10.1038/msb.2012.26.

[2] Feixiong Cheng and Zhongming Zhao. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association*, 21 (e2): e278–e286, 2014. https://doi.org/10.1136/amiajnl-2013-002512.

[3] Santiago Vilar, Eugenio Uriarte, Lourdes Santana, Tal Lorberbaum, George Hripcsak, Carol Friedman, and Nicholas P. Tatonetti. Similarity-based modeling in large-scale prediction of drug–drug interactions. *Nature Protocols*, 9: 2147–2163, 2014. https://doi.org/10.1038/nprot.2014.151.

[4] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34 (13): i457–i466, 2018. https://doi.org/10.1093/bioinformatics/bty294.

[5] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences*, 115 (18): E4304–E4311, 2018. https://doi.org/10.1073/pnas.1803294115.

[6] Arnold K. Nyamabo, Hui Yu, and Jian-Yu Shi. SSI–DDI: Substructure–substructure interactions for drug–drug interaction prediction. *Briefings in Bioinformatics*, 22 (6): bbab133, 2021. https://doi.org/10.1093/bib/bbab133.

[7] Xiaorui Su, Pengwei Hu, Zhu-Hong You, Philip S. Yu, and Lun Hu. Dual-channel learning framework for drug–drug interaction prediction via relation-aware heterogeneous graph transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (1): 249–256, 2024. https://doi.org/10.1609/aaai.v38i1.27777.

[8] Yingheng Wang, Yaosen Min, Xin Chen, and Ji Wu. Multi-view graph contrastive representation learning for drug–drug interaction prediction. In *Proceedings of The Web Conference 2021*, pages 2921–2933, 2021. https://doi.org/10.1145/3442381.3449786.

[9] Shenggeng Lin, Yanjing Wang, Lingfeng Zhang, Yanyi Chu, Yatong Liu, Yitian Fang, Mingming Jiang, Qiankun Wang, Bowen Zhao, Yi Xiong, and Dong-Qing Wei. MDF-SA-DDI: Predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. *Briefings in Bioinformatics*, 23 (1): bbab421, 2022. https://doi.org/10.1093/bib/bbab421.

[10] Tengfei Ma, Kun Chen, Yongsheng Zang, Yujie Chen, Xuanbai Ren, Bosheng Song, Hongxin Xiang, Yiping Liu, and Xiangxiang Zeng. Self-supervised blending structural context of visual molecules for robust drug interaction prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=7DyKKOU6HR.

[11] Xuan Lin et al. Comprehensive evaluation of deep and graph learning on drug–drug interactions prediction. *Briefings in Bioinformatics*, 24 (4): bbad235, 2023. https://doi.org/10.1093/bib/bbad235.

[12] Xinyue Li, Zhankun Xiong, Wen Zhang, and Shichao Liu. Deep learning for drug–drug interaction prediction: A comprehensive review. *Quantitative Biology*, 12 (1): 30–52, 2024. https://doi.org/10.1002/qub2.32.

[13] Ning-Ning Wang, Bei Zhu, Xin-Liang Li, Shao Liu, Jian-Yu Shi, and Dong-Sheng Cao. Comprehensive review of drug–drug interaction prediction based on machine learning: Current status, challenges, and opportunities. *Journal of Chemical Information and Modeling*, 64 (1): 96–109, 2024. https://doi.org/10.1021/acs.jcim.3c01304.

[14] Zhen Gao et al. Drug–drug interaction prediction: Databases, web servers and computational models. *Briefings in Bioinformatics*, 25 (1): bbad445, 2024. https://doi.org/10.1093/bib/bbad445.

[15] Yuqing Lu, Jing Chen, Nini Fan, Wenchao Song, Haiyang Sheng, Yinfeng Yang, and Jinghui Wang. Machine learning models for drug–drug interaction prediction from computational discovery to clinical application. *npj Digital Medicine*, 9: 198, 2026. https://doi.org/10.1038/s41746-026-02400-3.

[16] Craig Knox, Mike Wilson, Christen M. Klinger, et al. DrugBank 6.0: The DrugBank knowledgebase for 2024. *Nucleic Acids Research*, 52 (D1): D1265–D1275, 2024. https://doi.org/10.1093/nar/gkad976.

[17] Yao Tian et al. DDInter 2.0: An enhanced drug interaction resource with expanded data coverage, new interaction types, and improved user interface. Nucleic Acids Research, 53 (D1): D1356–D1362, 2025. https://doi.org/10.1093/nar/gkae726.

[18] Santiago Vilar, Carol Friedman, and George Hripcsak. Detection of drug–drug interactions through data mining studies using clinical sources, scientific literature and social media. *Briefings in Bioinformatics*, 19 (5): 863–877, 2018. https://doi.org/10.1093/bib/bbx010.

[19] Caterina Palleria, Antonello Di Paolo, Chiara Giofrè, Chiara Caglioti, Giacomo Leuzzi, Antonio Siniscalchi, Giovambattista De Sarro, and Luca Gallelli. Pharmacokinetic drug–drug interaction and their implication in clinical management. Journal of Research in Medical Sciences, 18 (7): 601–610, 2013.

[20] Thomayant Prueksaritanont et al. Drug–drug interaction studies: Regulatory guidance and an industry perspective. *The AAPS Journal*, 15: 629–645, 2013. https://doi.org/10.1208/s12248-013-9470-x.

[21] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, 46 (5): 914–920, 2013. https://doi.org/10.1016/j.jbi.2013.07.011.

[22] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. SemEval-2013 task 9: Extraction of drug–drug interactions from biomedical texts. In *Second Joint Conference on Lexical and Computational Semantics*, pages 341–350, 2013. URL https://aclanthology.org/S13-2056/.

[23] Tianlin Zhang, Jiaxu Leng, and Ying Liu. Deep learning for drug–drug interaction extraction from the literature: A review. *Briefings in Bioinformatics*, 21 (5): 1609–1627, 2020. https://doi.org/10.1093/bib/bbz087.

[24] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109: 719–760, 2020. https://doi.org/10.1007/s10994-020-05877-5.

[25] Achille Fokoue, Mohammad Sadoghi, Oktie Hassanzadeh, and Ping Zhang. Predicting drug–drug interactions through large-scale similarity-based link prediction. In The Semantic Web: Latest Advances and New Domains—13th International Conference, ESWC 2016, Lecture Notes in Computer Science, volume 9678, pages 774–789. Springer, 2016. https://doi.org/10.1007/978-3-319-34129-3_47.

[26] Ran Zhang, Xuezhi Wang, Guannan Liu, Pengyang Wang, Yuanchun Zhou, and Pengfei Wang. Motif-oriented representation learning with topology refinement for drug–drug interaction prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (1): 1102–1110, 2025. https://doi.org/10.1609/aaai.v39i1.32097.

[27] Yuanfei Dai, Chenhao Guo, Wenzhong Guo, and Carsten Eickhoff. Drug–drug interaction prediction with Wasserstein adversarial autoencoder-based knowledge graph embeddings. Briefings in Bioinformatics, 22 (4): bbaa256, 2021. https://doi.org/10.1093/bib/bbaa256.

[28] Yaqing Wang, Zaifei Yang, and Quanming Yao. Accurate and interpretable drug–drug interaction prediction enabled by knowledge subgraph learning. *Communications Medicine*, 4: 59, 2024. https://doi.org/10.1038/s43856-024-00486-y.

[29] Yifan Deng, Xinran Xu, Yang Qiu, Jingbo Xia, Wen Zhang, and Shichao Liu. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics*, 36 (15): 4316–4322, 2020. https://doi.org/10.1093/bioinformatics/btaa501.

[30] Dhanya Sridhar, Shobeir Fakhraei, and Lise Getoor. A probabilistic approach for collective similarity-based drug–drug interaction prediction. *Bioinformatics*, 32 (20): 3175–3182, 2016. https://doi.org/10.1093/bioinformatics/btw342.

[31] Wen Zhang, Yanlin Chen, Dingfang Li, and Xiang Yue. Manifold regularized matrix factorization for drug–drug interaction prediction. *Journal of Biomedical Informatics*, 88: 90–97, 2018. https://doi.org/10.1016/j.jbi.2018.11.005.

[32] Jian-Yu Shi, Kui-Tao Mao, Hui Yu, and Siu-Ming Yiu. Detecting drug communities and predicting comprehensive drug–drug interactions via balance regularized semi-nonnegative matrix factorization. *Journal of Cheminformatics*, 11 (1): 28, 2019. https://doi.org/10.1186/s13321-019-0352-9.

[33] Narjes Rohani, Changiz Eslahchi, and Ali Katanforoush. ISCMF: Integrated similarity-constrained matrix factorization for drug–drug interaction prediction. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 9 (1): 11, 2020. https://doi.org/10.1007/s13721-019-0215-3.

[34] Jian-Yu Shi, Hua Huang, Jia-Xin Li, Peng Lei, Yan-Ning Zhang, Kai Dong, and Siu-Ming Yiu. TMFUF: A triple matrix factorization-based unified framework for predicting comprehensive drug–drug interactions of new drugs. *BMC Bioinformatics*, 19 (Suppl 14): 411, 2018. https://doi.org/10.1186/s12859-018-2379-8.

[35] Stuti Jain, Emilie Chouzenoux, Kriti Kumar, and Angshul Majumdar. Graph regularized probabilistic matrix factorization for drug–drug interactions prediction. *IEEE Journal of Biomedical and Health Informatics*, 27 (5): 2565–2574, 2023. https://doi.org/10.1109/JBHI.2023.3246225.

[36] Chengcheng Zhang, Yao Lu, and Tianyi Zang. CNN–DDI: A learning-based method for predicting drug–drug interactions using convolution neural networks. *BMC Bioinformatics*, 23: 88, 2022. https://doi.org/10.1186/s12859-022-04612-2.

[37] Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. Drug–drug interaction extraction via convolutional neural networks. *Computational and Mathematical Methods in Medicine*, 2016: 6918381, 2016. https://doi.org/10.1155/2016/6918381.

[38] Ramakanth Kavuluru, Anthony Rios, and Tung Tran. Extracting drug–drug interactions with word and character-level recurrent neural networks. In *2017 IEEE International Conference on Healthcare Informatics*, pages 5–12, 2017. https://doi.org/10.1109/ICHI.2017.15.

[39] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36 (4): 1234–1240, 2020. https://doi.org/10.1093/bioinformatics/btz682.

[40] Xueting Han, Ruixia Xie, Xutao Li, and Junyi Li. SmileGNN: Drug–drug interaction prediction based on the SMILES and graph neural network. Life, 12 (2): 319, 2022. https://doi.org/10.3390/life12020319.

[41] Hui Yu, Kangkang Li, Wenmin Dong, Shuanghong Song, Chen Gao, and Jian-Yu Shi. Attention-based cross domain graph neural network for prediction of drug–drug interactions. *Briefings in Bioinformatics*, 24 (4): bbad155, 2023. https://doi.org/10.1093/bib/bbad155.

[42] Jianliang Gao, Zhenpeng Wu, Raeed Al-Sabri, Babatounde Moctard Oloulade, and Jiamin Chen. AutoDDI: Drug–drug interaction prediction with automated graph neural network. *IEEE Journal of Biomedical and Health Informatics*, 28 (3): 1773–1784, 2024. https://doi.org/10.1109/JBHI.2024.3349570.

[43] Kexin Huang, Cao Xiao, Trong Hoang, Lucas Glass, and Jimeng Sun. CASTER: Predicting drug interactions with chemical substructure representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (1): 702–709, 2020. https://doi.org/10.1609/aaai.v34i01.5412.

[44] Jiacheng Lin, Lijun Wu, Jinhua Zhu, Xiaobo Liang, Yingce Xia, Shufang Xie, Tao Qin, and Tie-Yan Liu. R^2-DDI: Relation-aware feature refinement for drug–drug interaction prediction. Briefings in Bioinformatics, 24 (1): bbac576, 2023. https://doi.org/10.1093/bib/bbac576.

[45] Namkyeong Lee, Dongmin Hyun, Gyoung S. Na, Sungwon Kim, Junseok Lee, and Chanyoung Park. Conditional graph information bottleneck for molecular relational learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18852–18871, 2023. URL https://proceedings.mlr.press/v202/lee23e.html.

[46] Haotong Du, Quanming Yao, Juzheng Zhang, Yang Liu, and Zhen Wang. Customized subgraph selection and encoding for drug–drug interaction prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. https://doi.org/10.52202/079017-3478.

[47] Kyriakos Schwarz, Ahmed Allam, Nicolas Andres Perez Gonzalez, and Michael Krauthammer. AttentionDDI: Siamese attention-based deep learning method for drug–drug interaction predictions. *BMC Bioinformatics*, 22: 412, 2021. https://doi.org/10.1186/s12859-021-04325-y.

[48] Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, and Yulan He. ExDDI: Explaining drug–drug interaction predictions with natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (24): 25228–25236, 2025. https://doi.org/10.1609/aaai.v39i24.34709.

[49] Zhonghao Ren, Xiangxiang Zeng, Yizhen Lao, Zhu-Hong You, Yifan Shang, Quan Zou, and Chen Lin. Predicting rare drug–drug interaction events with dual-granular structure-adaptive and pair variational representation. *Nature Communications*, 16: 3997, 2025. https://doi.org/10.1038/s41467-025-59431-9.

[50] Zhenqian Shen, Mingyang Zhou, Yongqi Zhang, and Quanming Yao. Benchmarking drug–drug interaction prediction methods: A perspective of distribution changes. *Bioinformatics*, 41 (11): btaf569, 2025. https://doi.org/10.1093/bioinformatics/btaf569.

[51] Jinheon Baek, Dong Bok Lee, and Sung Ju Hwang. Learning to extrapolate knowledge: Transductive few-shot out-of-graph link prediction. In *Advances in Neural Information Processing Systems*, volume 33, pages 546–560, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0663a4ddceacb40b095eda264a85f15c-Abstract.html.

[52] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *Proceedings of Machine Learning Research*, 70: 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

[53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144, 2016. https://doi.org/10.1145/2939672.2939778.

[54] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[55] Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel G. M. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Annals of Internal Medicine*, 162 (1): 55–63, 2015. https://doi.org/10.7326/M14-0697.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights