Dual-Pathway Fusion of EHRs and Knowledge Graphs for Predicting Unseen Drug-Drug Interactions

Franklin Lee FRANKLIN.LEE@STONYBROOK.EDU
Jericho Senior High School, USA

Tengfei Ma TENGFEI.MA@STONYBROOKMEDICINE.EDU
Stony Brook University, USA

Abstract

Drug–drug interactions (DDIs) remain a major source of preventable harm, and many clinically important mechanisms are still unknown. Existing models either rely on pharmacologic knowledge graphs (KGs), which fail on unseen drugs, or on electronic health records (EHRs), which are noisy, temporal, and site-dependent. We introduce, to our knowledge, the first system that conditions KG relation scoring on patient-level EHR context and distills that reasoning into an EHR-only model for zero-shot inference. A fusion "Teacher" learns mechanism-specific relations for drug pairs represented in both sources, while a distilled "Student" generalizes to new or rarely used drugs without KG access at inference. Both operate under a shared ontology (set) of pharmacologic mechanisms (drug relations) to produce interpretable, auditable alerts rather than opaque risk scores. Trained on a multi-institution EHR corpus paired with a curated DrugBank DDI graph, and evaluated using a a clinically aligned, decision-focused protocol with leakage-safe negatives that avoid artificially easy pairs, the system maintains precision across multi-institutuion test data, produces mechanism-specific, clinically consistent predictions, reduces false alerts (higher precision) at comparable overall detection performance (F1), and misses fewer true interactions compared to prior methods. Case studies further show zero-shot identification of clinically recognized CYP-mediated and pharmacodynamic mechanisms for drugs absent from the KG, supporting real-world use in clinical decision support and pharmacovigilance.

Keywords: Drug–drug interactions, Knowledge graphs, EHRs, Multimodal fusion, Zero-shot generalization, Knowledge distillation, Clinically aligned evaluation

Data and Code Availability We use a DrugBank knowledge graph with R=86 relation types over 1,710 drugs. Per-drug EHR embeddings are computed from ∼200M de-identified TriNetX (learn more about TriNetX at trinetx.com) events via PyHealth. TriNetX provides harmonized, multi-institution EHR data (standard code sets such as ICD-10, RxNorm, LOINC; routinely refreshed) across 220+ healthcare organizations in 30 countries. Experiments restrict to drugs in both sources. For a candidate pair (h, t), we form pairwise EHR features $x{=}x_{\text{EHR}}(h, t)$ shared by all EHR baselines and the student (Yang et al., 2023; Wishart et al., 2018; Palchuk et al., 2023; Wang et al., 2024b). The code is available at https://tinyurl.com/mu3tuc6w. Additional information about TriNetX is in Appendix A.

Institutional Review Board (IRB) Our de-identified data does not require Institutional Review Board approval.

1. Introduction

Drug-drug interactions (DDIs) can cause serious adverse reactions and pose a significant challenge to healthcare. In recent years, many unknown DDIs are emerging and predicting potentially new drug interactions has attracted increasing attention (Wang et al., 2024a,b). At the bedside, pharmacists need precision-first, mechanism-specific alerts (e.g., "CYP inhibition") that reflect the current patient context (Shang et al., 2019; Yang et al., 2021) and remain effective for new or rarely used drugs lacking prior evidence (Huang et al., 2024; Wang et al., 2023).

Previous computational approaches often rely on existing DDI knowledge graphs (KGs) and transform DDI prediction into a link-prediction problem (Zitnik et al., 2018). KG link prediction interpolates among seen drugs; however, such models struggle in zero-shot settings when a target drug lacks KG edges (Huang et al., 2024; Wang et al., 2023). Text-only

© F. Lee & T. Ma.

zero-shot methods infer interactions from curated drug descriptions (e.g., TextDDI) but depend on the availability and quality of those texts (performance can degrade when descriptions are missing or noisy) limiting robustness in practice (Zhu et al., 2023). Researchers have used electronic health record (EHR) data, a rich source for recording drug side effects, to screen DDIs; however, signals mined from EHRs are noisy, temporal, and cohort-dependent, and often transfer poorly across sites and timeframes (Choi et al., 2016; Wu et al., 2021). More fundamentally, a static–dynamic mismatch separates population-level, timeless KG relations from patient-specific, time-varying EHR signals; reasoning from either source in isolation rarely aligns with pharmacist review (Baltrušaitis et al., 2019; Arevalo et al., 2017). The result is a practical gap: systems either ignore patient context (good interpolation, poor extrapolation) or overfit cohort noise (weak mechanism specificity), producing alerts that are imprecise or inapplicable bedside.

We introduce, to our knowledge, the first DDI framework that injects EHR context into knowledge-graph relation scoring and, within the same system, distills KG structure into an EHR-only model for zero-shot use, supporting interpolation on seen drugs and zero-shot extrapolation on unseen drugs in one system. For drugs in the graph, a fusion model combines KG and EHR features, modulating mechanism scores (Arevalo et al., 2017; Perez et al., 2018; Srivastava et al., 2015); for new or rarely used drugs, a distilled EHR-only path reproduces the teacher's mechanism behavior without KG (Hinton et al., 2015; Lopez-Paz et al., 2015), where EHR drug embeddings are derived from patient-level RETAIN and aggregated per drug before pairwise feature construction (Appendix B & C) (Choi et al., 2016). Our system can be applied to tasks such as order verification, admission and discharge reconciliation, and formulary screening. It is designed as a single clinician-facing interface that provides clear, mechanism-specific alerts pharmacists can verify, helping to reduce unnecessary warnings and avoid generic risk scores.

2. Methods

2.1. Clinical objective and framework

In this work, our goal is to predict unsafe DDIs for unseen or newly ordered drugs that are absent from KGs but present in EHRs. The primary outcome is exact mechanism matching to DrugBank's 86-way ontology to ensure DDI alerts are actionable. We also evaluate high-precision detection (alert vs. no alert) and report recall-derived metrics like F1.

KG-only methods interpolate among seen drugs but falter without edges, while EHR-only methods capture bedside context but are noisy and cohort-dependent. We propose a KG–EHR teacher–student framework: a Fusion Module mixes KG and aggregated per-drug EHR embeddings derived from RETAIN (Choi et al., 2016) (Appendix C) on overlapped drugs, and a distilled EHR-only student, an MLP on the same pairwise EHR features, handles drugs without KG access while preserving mechanism behavior. Thus the EHR-based student operates on drugs absent from KGs while benefiting from KG knowledge via distillation. We evaluate under two regimes (see Experimental Splits): edge hold-out, where the drug set is fixed and a fraction of edges from KGs is withheld for test while KG embeddings for all drugs remain available; and node hold-out, where a disjoint set of drugs and their incident edges are withheld during training so any test pair involving a held-out drug is evaluated zero-shot to the KG (no usable KG embedding, but may have an EHR embedding). A high-precision detection gate governs whether to alert; on alerts, the system returns mechanism-level flags for pharmacist review.

2.2. Fusion teacher (embedding-level & clinically motivated)

KG and EHR entity representations are fused by a dimension-wise gate before the KG scoring model. Let $e_h, e_t \in \mathbb{R}^d$ be KG entity embeddings and $v_h, v_t \in \mathbb{R}^d$ the per-drug EHR embeddings. For $i \in \{h, t\}$,

$$\hat{e}_i = P_k e_i, \quad \hat{v}_i = P_e v_i, \quad g_i = \sigma\big(W_2\, \rho(W_1[\hat{e}_i \,\|\, \hat{v}_i])\big), \tag{1}$$

$$\tilde{e}_i = g_i \odot \hat{e}_i + (1 - g_i) \odot \hat{v}_i. \tag{2}$$

The KG scoring model then scores each relation r from the fused entities,

$$s_r(h,t) = \phi(\tilde{e}_h,\, e_r,\, \tilde{e}_t), \qquad r = 1, \ldots, R\ (R{=}86) \tag{3}$$

where e_r is the (unchanged) relation embedding and ϕ is the AutoSF or R-GNN scoring model. We train with single-label cross-entropy over the R logits $s = \{s_r\}_{r=1}^{R}$.

2.3. Knowledge-distilled student (zero-shot)

The student only uses EHR feature embeddings and predicts over the same R relations. The KG head exists only in the teacher. We use two heads in code: a DistMult-style head when entity embeddings are

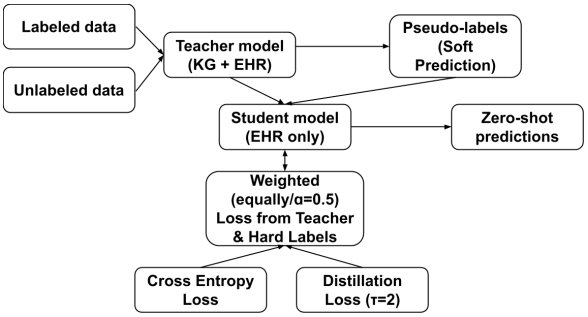

Figure 1: Training pipeline for the distillation model

available, and an EHR head for pure EHR features:

$$z_s = W_2\,\rho(W_1 x + b_1) + b_2 \in \mathbb{R}^R. \tag{4}$$

Distillation uses teacher logits z_t converted to soft targets via a teacher-only temperature; in all experiments τ=1 so $q = \sigma(z_t)$. Training minimizes a mixture of KD and supervised losses using BCE-with-logits on the student logits with no temperature:

$$\mathcal{L}_{\text{KD}} = \frac{1}{R}\sum_{r=1}^{R}\Big[-q_r \log\sigma([z_s]_r) - (1-q_r)\log\big(1-\sigma([z_s]_r)\big)\Big],$$
$$\mathcal{L}_{\text{sup}} = \frac{1}{R}\sum_{r=1}^{R}\Big[-y_r \log\sigma([z_s]_r) - (1-y_r)\log\big(1-\sigma([z_s]_r)\big)\Big],$$
$$\mathcal{L} = \alpha\,\mathcal{L}_{\text{KD}} + (1-\alpha)\,\mathcal{L}_{\text{sup}}, \quad \alpha{=}0.5,\ \tau{=}1. \tag{5}$$

If a sample lacks a hard label, the loss reduces to $\alpha\,\mathcal{L}_{\text{KD}}$ (Hinton et al., 2015; Lopez-Paz et al., 2015). A detailed rationale for using distillation in this setting is provided in Appendix D. Further implementation details (optimizer, batch size, training epochs) are provided in Appendix G.

3. Experiments

3.1. Experimental splits

Transductive (edge hold-out) We hold out 10% edges for test and use 80% for training and then evaluate exact-match relation ID on the held-out edges. This setting acts as a secondary check of the fusion Teacher's effectiveness.

Zero-shot (node hold-out) To assess generalization to unseen drugs, we use a node-level split: 80% of drug nodes for training and 10% for test; test drugs and all incident edges are removed so they are unseen at train time. Before distilling the student, we verified no leakage by remapping indices to the overlap space, building teacher-train and KD-on-train edge arrays, deduplicating and range-checking, and asserting that both endpoints of every KD edge lie in the training-node set; both checks returned OK.

We minimize per-mechanism BCE to match the teacher's unnormalized confidences across 86 mechanisms; we found KL on softmaxed logits suppressed rare-mechanism recall.

3.2. Detection, metrics, & Baselines

Two-head inference We use detection→ classification for both regimes: a binary detector decides alert vs no-alert; on alerts, the student predicts the mechanism class.

Negatives Detection uses leakage-safe random negatives (k=2 per positive on train, k=10 per positive on test), excluding any pair positive in any split. For detection we fix per-fold negative pools (k=10), use a single set of shared hyperparameters, and enforce leakage-safe sampling (no negatives touching test drugs).

Reported metrics We report exact-mechanism precision (95% Wilson CIs) and F1 (same splits/negs).

Baselines We compare to KG-only (AutoSF), EHR-only, SMILES-only, and a late-fusion variant. Main metrics are precision and F1 on seen/unseen splits (Table 1).

AutoSF or R-GNN (KG only) We used the AutoSF (Zhang et al., 2019) or R-GNN (Schlichtkrull et al., 2018) for KG embedding and link prediction as our ablation model and baseline.

EHR MLP This is the model using MLP on concatenated per-drug EHR vectors for (h, t) without distillation. We use the same splits as (Yang et al., 2023).

SMILES MLP Using chemical structures has also been a common method for predicting drug-drug interactions and it is also generalizable to unseen drugs, so we take SMILES as input and use ChemBERTa to obtain per-drug embeddings for DDI prediction. (Chithrananda et al., 2020).

4. Results

Overview We report exact–mechanism precision with 95% Wilson CIs and binary F1 with constant splits and negatives. Results are presented for two evaluation regimes: edge hold-out (seen drugs, unseen edges) and node hold-out (unseen drugs). Additional formulas for calculations in this section are provided in Appendix H.

Table 1: Edge hold-out (seen drugs, unseen edges): precision with 95% CIs.

Pipeline	+ Precision	Margin of Error	F1
Fusion Module	**0.9008**	**0.0053**	**0.3198**
AutoSF KG	0.8133	0.0069	0.1682
EHR MLP	0.5712	0.0087	0.1718
SMILES MLP	0.4541	0.0088	0.1667

Edge hold-out (seen drugs) The two-head Fusion (detector→mechanism) attains **0.9008** precision (95% CI **0.895–0.906**), an **11%** point gain over AutoSF KG at **0.8133** (95% CI **0.806–0.820**) and far above unimodal EHR (**0.5712**) and SMILES (**0.4541**). F1 rises to **0.3198**, nearly doubling AutoSF (**0.1682**) and EHR (**0.1718**). A detection head gates edges before the mechanism classifier fires, Fusion filters away the bulk of false positives that inflate graph-only and unimodal baselines. The non-overlapping CIs establish a clear, statistically robust advantage for seen drugs.

Node hold-out We highlight three contrasts that matter in practice. *Precision:* the **Fusion Student** reaches 0.713 ± 0.035 versus **SMILES-only** 0.426 ± 0.019 $(+0.287)$ and **EHR-only** 0.587 ± 0.024 $(+0.126)$, cutting the false-alert fraction at the same alert volume by roughly half vs. SMILES $(1 - 0.713$ vs. $1 - 0.426)$ and by $\sim 30\%$ vs. EHR; the $\pm$ bands do not overlap. *F1:* the Student attains 0.355 vs. **EHR-only** 0.280 $(+0.075)$ and **SMILES-only** 0.293 $(+0.062)$, so higher precision at the shared threshold is accompanied by stronger detection quality. *Stability/cold-start:* the **KG-only R-GCN** drops to 0.177 ± 0.011 precision with highly variable F1 (0.153 ± 0.094) under $k = 10$ node hold-out. Threshold-free corroboration: ROC-AUC is 0.622 ± 0.001 (**EHR-only** 0.511 ± 0.001, **SMILES-only** 0.541 ± 0.009) and AP (stepwise) is 0.217 ± 0.001 (**EHR-only** 0.170 ± 0.000, **SMILES-only** 0.179 ± 0.003), both well above the prevalence baseline ~ 0.051, meaning earlier retrieval of true interactions at a fixed review budget.

Case Studies Mechanism-level case studies for Dapagliflozin and Tamoxifen evaluate pairs whose partner drugs were absent from DrugBank KG at train time. EHR-only Distilled Student flagged these interactions and assigned mechanisms consistent with clinical evidence (e.g., CYP2D6-mediated reduction of endoxifen; pharmacodynamic volume-depletion with diuretics), providing external signal beyond DrugBank and complementing the aggregate metrics.

Table 3: External signal beyond DrugBank: mechanism-level case studies.

Pair	Mechanism
Dapagliflozin x Furosamide	PD volume depletion or hypertension risk
Dapagliflozin x Hydrochlorothiazide	PD volume depletion or hypotension risk
Tamoxifen x Diphenhydramine	CYP2D6 inhibition → Lower endoxifen
Tamoxifen x Tarbinafine	Potent CYP2D6 inhibitor → lower endoxifen

(continued)

Model Call	Citation
Prediction = 3 (Correct)	(U.S. Food and Drug Administration, 2023)
Prediction = 48 (Correct)	(U.S. Food and Drug Administration, 2024)
Prediction = 46 (Correct)	(Hamelin et al., 2000)
Prediction = 46 (Correct)	(Dürrbeck and Nenoff, 2016)

Despite EHRs specialized in kidney disease, the distilled Student still identifies non-renal pharmacokinetic pairs (tamoxifen), suggesting the learned pattern transfers beyond the training domain.

Cross-setting ablations Experiments and baselines act as an ablation of information sources. KG topology interpolates effectively among seen drugs (KG on edge hold-out) but fails to extrapolate to unseen nodes (node hold-out collapse). EHR features generalize to unseen drugs but are weaker than Fusion on seen drugs. Fusion integrates both, cutting false alerts while raising mechanism precision in the known-drug setting, and the KD-trained Student transfers this calibration into an EHR-only model that boosts zero-shot precision and F1 without KG access at inference. Together, these findings demonstrate that only structured integration yields consistently reliable performance across both regimes.

Robustness All key gains are supported by 95% and bootstrap Wilson CIs with no interval overlap (Fusion vs KG; Student vs EHR). Across 3 random seeds for the zero-shot "unseen" setting (and 1 for the transductive "seen" setting), improvements hold across both mechanism precision and detection F1, underscoring that the observed effects reflect true signal rather than sampling noise. *Margins of Error:* For precision, 95% margins of error from Table 2 are: Fusion 0.0345, EHR-only MLP 0.0238, SMILES-only MLP 0.0192, Raw R-GCN (KG-only) 0.0111. *Detection protocol:* For F1, we fix per-fold negative pools (k=10), use a shared threshold, and enforce leakage-safe sampling (no negatives

Table 2: **est k=10 negatives per positive; same splits/negatives across models.**

Model	ROC-AUC	AP (stepwise)	AP (baseline)	F1 (detection)	Precision
Distillation (teacher: Fusion head; student distilled from R-GNN)	0.6223 ± 0.0009	0.2172 ± 0.0011	0.0506	0.3545 ± 0.0014	0.7126 ± 0.0345
EHR-only MLP	0.5114 ± 0.0008	0.1697 ± 0.0003	0.0506	0.2798 ± 0.0006	0.5869 ± 0.0238
SMILES-only MLP	0.5408 ± 0.0094	0.1794 ± 0.0028	0.0506	0.2934 ± 0.0042	0.4255 ± 0.0192
Raw R-GCN (KG-only)	0.5176 ± 0.0240	0.1724 ± 0.0078	0.0506	0.1526 ± 0.0936	0.1772 ± 0.0111

that touch test drugs). *Summary by regime:* At θ^*, F1 is: Fusion 0.3545 ± 0.0014, EHR-only MLP 0.2798 ± 0.0006, SMILES-only MLP 0.2934 ± 0.0042, Raw R-GCN (KG-only) 0.1526 ± 0.0936; the ordering Fusion > SMILES > EHR > Raw R-GCN holds in both regimes (edge hold-out, node hold-out).

Seeds Experiments use S=3 seeds for the unseen and S=1 for seen; precision and F1 are Wilson intervals and bootstrap CIs from pooled predictions.

5. Discussion and Outlook

Clinical & algorithmic takeaways Fusion A naïve EHR+SMILES concatenation (Appendix F) baseline underperformed unimodal models, underscoring that simple feature-level merging cannot reconcile misaligned statistical structure and pharmacological semantics (Appendix F). Thus, we take the strong graph model for known drugs and add patient checks: EHR signals (e.g., co-medications, labs, QTc) amplify mechanisms they support and dampen those they contradict. This preserves relational structure while filtering patient-implausible links, resulting in higher precision, narrower 95% confidence intervals, and lower false-positive rate.

EHR-only Student For new or rarely used drugs, the graph offers no help. We train an EHR-only model to mimic the teacher (distillation, τ=1), learning its mechanism pattern and applying it with routine EHR features—no KG at runtime. The Student generalizes to unseen drugs with higher precision and better F1 than a plain EHR model: decisions are mechanism-aligned, and stable even when DrugBank KG embeddings are missing or stale.

Future bedside implications & applications The system emits mechanism-specific, patient-conditioned flags at workload-aligned precision; Fusion covers seen formulary drugs and the Student provides zero-shot coverage for new or rare drugs (deployment details and service metrics in Appendix I).

6. Conclusion

We target precise, auditable DDI flags for pharmacist review. Prior work has largely treated KGs and EHRs in isolation; here we couple KG relation scoring with patient context and distill the structure into an EHR-only model for zero-shot use, producing the same 86-mechanism ontology. Trained and evaluated on a large EHR corpus paired with a curated DDI KG under leakage-safe edge and node hold-outs with $k = 2$ negatives for training and $k = 10$ negatives for testing, the system is clinically impactful, statistically sound, and deployable.

Precision and F1 improve while false positives decrease relative to AutoSF, EHR-MLP, and SMILES-MLP. Fusion raises precision and reduces bedside false alerts for seen drugs, while the student surpasses EHR-only and SMILES on unseen drugs without KG access at inference. Outputs are mechanism-aware, with robust estimates (non-overlapping CIs, tight margins, stable F1), yielding more precise alerts and lower clinical burden through patient-conditioned interpolation and distilled extrapolation.

Future work focuses on a silent-mode hospital deployment to assess precision and actionability on unseen drugs. Looking ahead, we aim to calibrate uncertainty and add abstention mechanisms to improve reliability, refresh KG structure and enable lightweight model editing over time, and conduct external and temporal validation. We will also continue to modernize the models used, scaling our pipeline with SAIL and GrAIL, and optimize hyperparameters like distillation temperature (τ) through more ablation studies. These extensions will enhance generalizability and strengthen readiness for real-world deployment. To align the system with bedside use, we will operate at a clinically chosen threshold (targeting TPR $\geq$ 0.90 and maximizing precision on a validation cohort) while routing lower-confidence candidates to non-interruptive queues; surface interpretable evidence—the predicted mechanism and dimension-wise fusion gate weights (EHR vs. KG) with links to sources; support adoption via versioned models and a CPOE feedback loop ("useful" / "not useful" / "missed"); and harden safety rails with abstain/OOD flags for unfamiliar drugs and a "silent-mode" logging phase before any interruptive alerts.

References

John Arevalo, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017. URL https://arxiv.org/abs/1702.01992.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020. URL https://arxiv.org/abs/2010.09885.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua A. Kulas, Andy Schuetz, and Walter F. Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Neural Information Processing Systems*, 2016. URL https://api.semanticscholar.org/CorpusID:948039.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Abhinav Vaid, Jure Leskovec, Girish N. Nadkarni, Benjamin S. Glicksberg, Nils Gehlenborg, and Marinka Zitnik. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30(12):3601–3613, 2024. doi: 10.1038/s41591-024-03233-x. TxGNN.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015. URL https://arxiv.org/abs/1511.03643.

Bradley Malin. Summary of the assessment of trinetx data privacy principles (expert determination). Technical report, TriNetX, LLC, 2020. URL https://trinetx.com/wp-content/uploads/2022/04/TriNetX-Empirical-Summary-by-Brad-Malin-2020-branded.pdf. External Expert Determination assessment under HIPAA Privacy Rule §164.514(b)(1).

Matvey B. Palchuk, Jack W. London, David Perez-Rey, Zuzanna J. Drebert, Jessamine P. Winer-Jones, Courtney N. Thompson, John Esposito, and Brecht Claerhout. A global federated real-world data and analytics platform for research. *JAMIA Open*, 6(2):ooad035, 2023. doi: 10.1093/jamiaopen/ooad035.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web*, pages 593–607, Cham, 2018. Springer International Publishing. ISBN 978-3-319-93417-4.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1126–1133, 2019. doi: 10.1609/aaai.v33i01.33011126.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015. URL https://arxiv.org/abs/1505.00387.

Yaqing Wang, Zaifei Yang, and Quanming Yao. Accurate and interpretable drug-drug interac-

tion prediction enabled by knowledge subgraph learning. *Communications Medicine*, 4(1):59, Mar 2024b. ISSN 2730-664X. doi: 10.1038/s43856-024-00486-y. URL https://doi.org/10.1038/s43856-024-00486-y.

Yuxuan Wang, Ying Xia, Junchi Yan, Ye Yuan, Hong-Bin Shen, and Xiaoyong Pan. Zerobind: a protein-specific zero-shot predictor with subgraph matching for drug–target interactions. *Nature Communications*, 14(1):7861, 2023. doi: 10.1038/s41467-023-43597-1.

Ziyan Wang, Zhankun Xiong, Feng Huang, Xuan Liu, and Wen Zhang. Zeroddi: A zero-shot drug-drug interaction event prediction method with semantic enhanced learning and dual-modal uniform alignment, 2024a. URL https://arxiv.org/abs/2407.00891.

David S. Wishart, Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018. doi: 10.1093/nar/gkx1037.

Po-Yen Wu, Yu-Chih Chen, Peter Chow, Jing Hao, Bradley Malin, Joshua C. Denny, and Wei-Qi Wei. Ddiwas: High-throughput electronic health record-based screening of drug–drug interactions. *Journal of the American Medical Informatics Association*, 28(10):2147–2157, 2021. doi: 10.1093/jamia/ocab035.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pages 3735–3741. International Joint Conferences on Artificial Intelligence Organization, 2021. doi: 10.24963/ijcai.2021/514. URL https://www.ijcai.org/proceedings/2021/514.

Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin P. Danek, and Jimeng Sun. Pyhealth: A deep learning toolkit for healthcare applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, pages 5788–5789. Association for Computing Machinery, 2023. doi: 10.1145/3580305.3599178.

Yongqi Zhang, Quanming Yao, Wenyuan Dai, and Lei Chen. Autosf: Searching scoring functions for knowledge graph embedding, 2019. URL https://arxiv.org/abs/1904.11682.

Fangqi Zhu, Yongqi Zhang, Lei Chen, Bing Qin, and Ruifeng Xu. Learning to describe for predicting zero-shot drug-drug interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 14855–14870. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.918. URL http://dx.doi.org/10.18653/v1/2023.emnlp-main.918.

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

Appendix A. Information about TriNetX (Palchuk et al., 2023)

What it is. TriNetX is a federated network of de-identified electronic health records (EHRs) used for cohort discovery and real-world evidence studies, described in *JAMIA Open* as a global commercial network linking healthcare organizations (HCOs) with industry and academia.

Scale (as of 2025). TriNetX reports 300M+ patients across 200+ HCOs and 8,800+ sites worldwide (press release, Sept. 29, 2025). An earlier update (Jan. 2025) cited 283M patients, and the peer-reviewed history documents growth from **55** HCOs / 7 countries (2017) to 220+ HCOs / 30 countries (2022)—counts vary as sites onboard.

Data domains & standards. Contributed EHRs are harmonized to common terminologies (e.g., ICD-10-CM, ICD-10-PCS/HCPCS/CPT, RxNorm, LOINC), with cross-region mappings described in TriNetX terminology materials.

Privacy & governance. Data shared to researchers are de-identified under HIPAA Expert Determination. Protections include a minimum cell size of 10 for aggregates and obfuscations for person-level extracts (e.g., month–year for death, restricted geography), per an external assessment by Bradley Malin, PhD Malin (2020).

U.S. footprint & refresh. Public materials note ˜117M U.S. patients, with datasets refreshed about every 2–4 weeks (offerings vary by product/network).

How we used it. We queried the de-identified Research Network to construct cohorts and features using standard codes; no identified data left TriNetX. Many institutions treat such analyses as IRB-exempt, but local policy governs.

Appendix B. EHR Feature Groups Derived from TriNetX

We derive per-drug vectors from `PyHealth` RETAIN (exported `model.embeddings["drugs"].weight`) and augment them with structured EHR features computed pairwise for candidate drug pairs (h, t). Table 4 lists the source tables and the pairwise features we construct. These features are concatenated to form $\mathbf{x}_{\text{EHR}}(h, t)$ for the EHR-only baseline and as inputs to the fusion gate.

We list the EHR sources used to derive per-drug EHR embeddings/features. This table is for transparency only; no pairwise feature definitions here.

Table 4: EHR feature groups and source tables (codes).

Group	Source tables (codes)
Demographics	`patient.csv`
Encounters	`encounter.csv`
Diagnoses	`diagnosis.csv` (ICD-10)
Meds (orders/dispense)	`medication_drug.csv` (RxNorm/NDC)
Meds (ingredients)	`medication_ingredient.csv`
Procedures	`procedure.csv` (CPT)
Labs	`lab_result.csv` (LOINC)
Vitals	`vital_signs.csv`

Appendix C. Deriving EHR Drug Embeddings from `PyHealth` RETAIN

We train RETAIN with `PyHealth` using `feature_keys=["drugs","procedures"]` and a multi-label objective. During training, RETAIN learns a global token embedding table for the `"drugs"` stream:

$$W \in \mathbb{R}^{V \times D}, \qquad D = 64 \text{ (our setting)}.$$

After training, we export this layer exactly as in code:

$$\texttt{emb\_layer} := \texttt{model.embeddings["drugs"]}$$

$$W := \texttt{emb\_layer.weight}.$$

Let $\text{id}(i)$ be the PyHealth vocabulary index for drug i (built from the training split). The *EHR embedding* we use for drug i is the corresponding row of W:

$$\mathbf{v}_i \;=\; W_{\text{id}(i)} \in \mathbb{R}^D.$$

Downstream pairwise EHR features for a candidate pair (h, t) are constructed from these exported vectors:

$$\mathbf{x}_{\text{EHR}}(h, t) \;=\; \big[\, \mathbf{v}_h \,\|\, \mathbf{v}_t \,\big],$$

or optionally

$$\mathbf{x}_{\text{EHR}}(h, t) \;=\; \big[\, \mathbf{v}_h \,\|\, \mathbf{v}_t \,\|\, |\mathbf{v}_h - \mathbf{v}_t| \,\|\, \mathbf{v}_h \odot \mathbf{v}_t \,\big],$$

after ℓ_2 normalization of $\mathbf{v}_i$. These $\mathbf{x}_{\text{EHR}}$ features are used both in the EHR-only MLP baseline and as inputs to the fusion gate.

Appendix D. Rationale for Fusion & Distillation

Rationale for ChemBERTa Why ChemBERTa (zinc-base-v1). It is a stable, off-the-shelf SMILES encoder with good coverage and straightforward integration for our ablations.

Fusion objective: Why this gate The gate is parameter-efficient, easy to optimize under multi-relation supervision, and interpretable: it turns up KG or EHR signal dimension-by-dimension in ways clinicians can inspect (mechanism evidence ↑, contradictory evidence ↓), while preserving KG relational geometry. Empirically, naive concatenation + MLP underperforms both unimodal models, motivating an explicit gate rather than "just stacking features."

Mathematical form already in the paper.

$$\hat{e}_i = P_k e_i \tag{6}$$

$$\hat{v}_i = P_e v_i \tag{7}$$

$$g_i = \sigma\big(W_2\, \rho\big(W_1[\hat{e}_i; \hat{v}_i]\big)\big) \tag{8}$$

$$s_r(h,t) \;=\; \phi\big(\tilde{e}_h,\, e_r,\, \tilde{e}_t\big), \qquad r = 1, \ldots, 86 \tag{9}$$

Backbone robustness (not tied to a single scorer). Under train/test node splits of 80% and 10%, a PyG R-GCN backbone (KG-only) yields higher precision than AutoSF, while preserving the ordering Fusion > KG-only and the Student's zero-shot advantages:

R-GCN per-seed precision: 0.7352, 0.7297, 0.6729
→ mean 0.7126 ± 0.0345

AutoSF: 0.6983, 0.6707, 0.6443 → mean 0.6711±0.0270.

Multi-seed results for baselines will be in the camera-ready version. **Distillation objective** DrugBank mechanisms are predominantly single-label, but teacher confidences over R=86 are not a normalized distribution and can place mass on multiple plausible mechanisms. BCE treats mechanisms independently and avoids the probability-mass competition imposed by $\mathrm{KL}(\mathrm{softmax}_\tau(z_t) \,\|\, \mathrm{softmax}(z_s))$, which we observed to suppress rare-mechanism recall. BCE was also more numerically stable with many negatives.

Appendix E. Method Positioning vs. Baselines

Table 5: **Model capabilities and inputs.** Our Fusion teacher combines KG + EHR with a gating module; the Student is an EHR-only model distilled from the teacher. All models use the same 86-mechanism labels and splits/negatives.

Method	Uses KG	Uses EHR	Uses SMILES
Ours: Fusion (teacher)	✓	✓	×
Ours: Student (distilled)	×	✓	×
KG-only: AutoSF	✓	×	×
KG-only: R-GCN	✓	×	×
EHR-only: MLP (RETAIN W)	×	✓	×
SMILES-only: MLP	×	×	✓

Method	Fusion gate	KD (teacher→student)	86 mech.	Zero-shot drugs
Ours: Fusion (teacher)	✓	×	✓	✓
Ours: Student (distilled)	×	✓	✓	✓
KG-only: AutoSF	×	×	✓	×
KG-only: R-GCN	×	×	✓	×
EHR-only: MLP (RETAIN W)	×	×	✓	✓
SMILES-only: MLP	×	×	✓	×

Notes. "Uses EHR" denotes structured/EHR-derived per-drug vectors (exported from `PyHealth` RETAIN token embedding table W) and pairwise features; "Zero-shot drugs" indicates the model can score pairs where both drugs are unseen in the KG (enabled via EHR inputs for our teacher/student).

Appendix F. Additional Baseline: EHR+SMILES

Motivation. We additionally tested a naïve concatenation baseline combining per-drug EHR embeddings with ChemBERTa-derived SMILES embeddings, followed by an MLP classifier. This baseline was included to assess whether molecular structure embeddings could enhance patient-derived features without KG support.

Results. The EHR+SMILES baseline underperformed both unimodal models. In edge hold-out (seen drugs, unseen edges), it achieved a precision of 0.2818, below both EHR and SMILES-only. In node hold-out (unseen drugs), it yielded 0.3027 with $F1$=0, indicating that the model admitted no alerts. These results highlight the difficulty of integrating heterogeneous embeddings whose statistical structure and pharmacological semantics are misaligned, reinforcing that simple feature-level merging does not substitute for KG-grounded relational alignment. Statistics, appended to existing data, are shown in Tables 6 and 7.

Table 6: Edge hold-out (seen drugs, unseen edges): precision with 95% CIs.

Pipeline	+ Precision	Margin of Error	F1
Fusion Module	**0.9008**	**0.0053**	**0.3198**
AutoSF KG	0.8133	0.0069	0.1682
EHR MLP	0.5712	0.0087	0.1718
SMILES MLP	0.4541	0.0088	0.1667
EHR+SMILES MLP	0.2818	0.0056	—

Appendix G. Implementation & Training

G.1. KG-only Hyperparameters

```
python train.py --task_dir=KG_Data/DrugBank \
  --optim=Adam --lamb=0.000282 --lr=0.3775 \
  --n_dim=64 --n_epoch=250 --n_batch=2048 \
  --epoch_per_test=250 --test_batch_size=50 \
  --thres=0.0 --parrel=1 --decay_rate=0.99456
```

G.2. Fusion Module Optimization (R-GNN)

- Learning rate: **0.001**
- Training epochs: **10**

G.3. EHR MLP (scikit-learn)

```
from sklearn.neural_network \
import MLPClassifier

clf = MLPClassifier(
    alpha=0.05,
    hidden_layer_sizes=(32, 16, 8, 4, 2),
    max_iter=30,
    random_state=1
)
```

Hyperparameters

1. alpha: learning rate
2. hidden_layer_sizes: self-explanatory
3. max_iter: number of epochs
4. random_state: random seed (changed for multi-seed experiments)

G.4. SMILES MLP (scikit-learn)

```
from sklearn.neural_network \
import MLPClassifier

clf = MLPClassifier(
    alpha=0.10,
    hidden_layer_sizes=(32, 16, 8, 4, 2),
    max_iter=15,
    random_state=1
)
```

G.5. EHR+SMILES MLP (scikit-learn)

```
from sklearn.neural_network \
import MLPClassifier

clf = MLPClassifier(
    alpha=0.05,
    hidden_layer_sizes=(32, 16, 8, 4, 2),
    max_iter=50,
    random_state=1
)
```

G.6. SMILES Encoder (HuggingFace Transformers)

```
from transformers \
import AutoTokenizer, AutoModel

MODEL = "seyonec/ChemBERTa-zinc-base-v1"
tok = AutoTokenizer.from_pretrained(MODEL)
enc = AutoModel.from_pretrained(MODEL)
```

G.7. KD Student Configuration

```
KD_ALPHA   = 0.5      #
TAU        = 1        # temperature
LR         = 0.1
WD         = 1e-4
BATCH_SIZE = 1024
EPOCHS     = 30
n_rel      = 86
n_dim      = 64
DEVICE     = torch.device("cuda" \
if torch.cuda.is_available() else "cpu")
```

Reproducibility note. All hyperparameters and commands above reflect the exact values used in the experiments reported in the main text.

G.8. Training Time

1. AutoSF (KG-only): 250 epochs on NVIDIA T4, 5 minutes total.
2. Relational GNN (R-GCN): 30 epochs on Colab High-RAM CPU, 3 minutes total.
3. Both runs use 112,008 KG triples (head–relation–tail) as training rows. Implementation commands/hyperparameters are already listed in Appendix G.

Appendix H. Results: Additional Details

Metrics and confidence intervals. We report (i) exact-mechanism precision

Table 7: **k=10 test on the same splits/negatives for all models—ROC-AUC, AP (stepwise), AP (baseline), F1, and precision.**

Model	ROC-AUC	AP (stepwise)	AP (baseline)	F1 (detection)	Precision
Distillation (teacher: Fusion head; student distilled from R-GNN)	0.6223 ± 0.0009	0.2172 ± 0.0011	0.0506	0.3545 ± 0.0014	0.7126 ± 0.0345
EHR-only MLP	0.5114 ± 0.0008	0.1697 ± 0.0003	0.0506	0.2798 ± 0.0006	0.5869 ± 0.0238
SMILES-only MLP	0.5408 ± 0.0094	0.1794 ± 0.0028	0.0506	0.2934 ± 0.0042	0.4255 ± 0.0192
EHR+SMILES MLP	—	—	—	0.0000	0.3027±0.0116
Raw R-GCN (KG-only)	0.5176 ± 0.0240	0.1724 ± 0.0078	0.0506	0.1526 ± 0.0936	0.1772 ± 0.0111

$\text{Prec} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\arg\max_{r\in\{1,\dots,R\}} \hat{p}_i(r) = y_i\}$
$\text{Prec}_{\text{bin}} = \frac{\text{TP}}{\text{TP+FP}}$, $\text{Rec} = \frac{\text{TP}}{\text{TP+FN}}$, $\text{F1} = \frac{2\,\text{Prec}_{\text{bin}}\cdot\text{Rec}}{\text{Prec}_{\text{bin}}+\text{Rec}}$

Wilson 95% CI for a proportion $\hat{p}$ over n samples:

$$\text{CI}_{\text{low,high}} = \frac{\hat{p} + \frac{z^2}{2n} \mp z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1+\frac{z^2}{n}}, \quad z = 1.96.$$

For the edge-vs-none task we construct a test set with a ratio k negatives per positive (k=2 in the main results). A sampled pair (h, t) is *excluded* if it is positive in any split or direction; self-loops and inverse duplicates are removed to avoid leakage. All models are evaluated on the same sets.

Relative false-positive reduction. At fixed true positives across methods, $\text{FP} = \text{TP}\left(\frac{1}{p} - 1\right)$ for precision p. The relative FP reduction of Fusion over a baseline is

$$1 - \frac{\frac{1}{p_{\text{fusion}}} - 1}{\frac{1}{p_{\text{base}}} - 1}.$$

Seeds and uncertainty. Experiments were run with S=3 random seeds in the zero-shot "unseen" setting and S=1 in the transductive "seen" setting. For single-seed results, we report Wilson 95% CIs for precision from pooled predictions, and F1 is reported as a point estimate. For multi-seed results, we incorporate both Wilson CIs and bootstrap CIs for each metric.

Tables referenced in the body. Table 1 (Seen/edge hold-out): Fusion vs AutoSF, precision with 95% CIs.

Table 2 (Unseen/node hold-out): Student vs EHR-MLP, SMILES-MLP, AutoSF; precision with 95% CIs and F1 (± across seeds).

Appendix I. Deployment details

For pharmacist order verification, alerts must be *actionable* and *rare*: the system will emit a top mechanism with a short, patient-conditioned rationale (e.g., co-medications, labs) and an audit trail; validation-fixed thresholds keep operating characteristics stable for governance (Shang et al., 2019; Yang et al., 2021; Choi et al., 2016). Use cases include formulary screening, coverage for investigational or rare drugs (Student), antimicrobial stewardship and oncology checks (Fusion), and admission/discharge reconciliation when KG embeddings may be stale (Huang et al., 2024; Wang et al., 2023; Zitnik et al., 2018). Integration is EHR-native (Student uses tabular features only) and supports service metrics such as false alerts/100 orders, time-to-clear, mitigation acceptance, plus model metrics for QA and safety review (Pedregosa et al., 2011; Yang et al., 2023).