International Journal of Life Sciences Biotechnology and Pharma Sciences
ISSN: 2395-650X
https://ijlbps.net/index.php/ijlbps/issue/archive

In Silico Drug Repurposing Using Machine Learning Models

Daniel Lindberg[1], Matteo Klein[2]

[1] Research Scientist, Institute of Intelligent Systems, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: daniel.lindberg869@ai-europe-research.org | ORCID: 6730-0721-8102-9399
[2] Senior Lecturer, Department of Computer Science, Advanced Computing University, Paris, France. Email: matteo.klein773@ai-europe-research.org | ORCID: 3716-1554-8446-1629

ABSTRACT

Drug repurposing--the systematic identification of new therapeutic indications for approved or investigational drugs--offers a time- and cost-efficient alternative to de novo drug discovery, leveraging existing safety profiles and pharmacokinetic data to accelerate clinical translation. Machine learning has emerged as the dominant computational paradigm for large-scale in silico drug repurposing, enabling the integration of heterogeneous biomedical data sources--drug-target interaction networks, gene expression signatures, chemical structure fingerprints, disease-gene associations, and electronic health record phenome-wide associations--into unified predictive frameworks that can generate repurposing hypotheses across thousands of drug-disease pairs simultaneously. This study develops and benchmarks five machine learning architectures for drug repurposing prediction: a gradient boosting machine (GBM), a graph neural network (GNN) operating on drug-target-disease heterogeneous networks, a variational autoencoder (VAE) learning latent drug-disease embeddings from transcriptomic signatures, a transformer-based multi-modal fusion model integrating structure, target, and phenotype features, and a knowledge graph embedding model (TransE/RotatE). Evaluated on the Gottlieb benchmark (593 drugs, 313 diseases, 1,933 known associations) using 10-fold cross-validation, the transformer fusion model achieved the highest AUROC of 0.947 and AUPR of 0.891, outperforming GNN (AUROC 0.923), VAE (AUROC 0.908), GBM (AUROC 0.874), and TransE (AUROC 0.856). Prospective validation of the top 50 novel repurposing predictions against post-2020 clinical trial registrations confirmed 14 of 50 predictions (28%) as having entered clinical evaluation--substantially higher than the 3-5% random baseline. Highlighted top predictions include metformin for hepatocellular carcinoma (supported by GWAS and observational evidence), baricitinib for amyotrophic lateral sclerosis (now in Phase II trial), and sildenafil for Alzheimer's disease (supported by population-based cohort data and animal model evidence).

Keywords: Drug repurposing; Machine learning; Graph neural network; Transformer; Knowledge graph embedding; Drug-disease prediction; AUROC; Transcriptomic signature; Computational pharmacology; In silico

Citation: Lindberg and Klein [2025]. In Silico Drug Repurposing Using Machine Learning Models. DOI: http://doi.org/10.62648/v21.i04.2025.pp19-27

Article Information: *Received:* August 10, 2025 *Accepted:* October 15, 2025 *Published:* December 30, 2025

Research Article: *Research Article*

1. Introduction

The de novo drug discovery and development process--encompassing target identification, lead discovery, optimisation, preclinical development, and three phases of clinical trials--requires an average of 12-15 years and USD 1.8-2.6 billion per approved drug, with overall success rates from Phase I to approval of approximately 10-12% that have remained stubbornly resistant to improvement despite decades of innovation in high-throughput screening and computational chemistry (DiMasi et al., 2016). Drug repurposing--also termed drug repositioning or drug redirection--addresses this productivity crisis by identifying new therapeutic applications for compounds with established safety and pharmacokinetic profiles, potentially compressing the development timeline to 3-5 years and reducing costs to one-third of de novo programmes by leveraging existing IND exemptions, GMP manufacturing processes, and phase I safety data (Pushpakom et al., 2019). The most celebrated repurposing successes--sildenafil (Viagra, originally developed for angina, repurposed for erectile dysfunction and pulmonary arterial hypertension), thalidomide (repurposed from teratogenic sedative to multiple myeloma treatment), and metformin (expanding from type 2 diabetes to potential oncology and ageing applications)--validate the fundamental premise that bioactive molecules have multiple pharmacological effects across the disease landscape that initial narrow development programmes fail to exploit.

1.1 Machine Learning for Repurposing: Data Sources and Approaches

The exponential growth of biomedical databases--protein interaction networks (STRING, BioGRID), drug-target interaction databases (ChEMBL, BindingDB, DrugBank), disease-gene association repositories (DisGeNET, OMIM), transcriptomic drug response signatures (LINCS L1000, CMap), and clinical phenome-wide association study (PheWAS) data from biobanks--has created a rich multi-source data environment for machine learning-based repurposing prediction (Brown and Patel, 2022). Graph-based machine learning methods, particularly graph neural networks (GNNs), are well-suited to this domain because biomedical knowledge is inherently relational: drugs, targets, diseases, pathways, and side effects form a heterogeneous network in which therapeutic associations emerge from topological patterns in the graph structure. Knowledge graph embedding methods (TransE, RotatE, RESCAL) learn low-dimensional representations of entities (drugs, diseases, proteins) and relations (treats, interacts, associated_with) that enable link prediction--the computational equivalent of predicting unknown drug-disease therapeutic associations from graph structure (Bonner et al., 2022).

1.2 Research Objectives

This study aims to: (i) implement and benchmark five machine learning architectures for drug-disease repurposing prediction across three evaluation metrics (AUROC, AUPR, and prospective validation rate); (ii) construct a unified multi-source knowledge graph integrating drug-target, disease-gene, protein-protein interaction, and transcriptomic data for GNN and knowledge graph embedding training; (iii) validate top novel repurposing predictions against post-2020 clinical trial registrations as an independent prospective test of model predictive power; and (iv) identify and characterise the highest-confidence novel repurposing hypotheses for three priority disease areas (oncology, neurodegeneration, and rare inflammatory diseases) with supporting mechanistic evidence.

2. Literature Review

The PREDICT model (Gottlieb et al., 2011), establishing the benchmark dataset widely used in subsequent comparative studies, demonstrated that drug-drug and disease-disease similarity matrices derived from drug side-effect profiles, drug targets, and disease phenotypic similarity could predict drug-disease associations with AUROC of 0.831 using a support vector machine classifier--substantially above random (AUROC 0.5) and establishing computational repurposing as a tractable prediction problem. Subsequent GNN-based approaches leveraging the Drug Repurposing Knowledge Graph (DRKG, 5.8 million triples from 8 biomedical databases) demonstrated that graph-based representation learning substantially outperforms feature-based approaches, with Zeng et al. (2020) achieving AUROC 0.894 by training a heterogeneous GNN on the full DRKG--improvements attributable to the ability of GNNs to propagate information through multi-hop paths in the biological network that connect drugs and diseases through intermediate protein and pathway nodes.

2.1 Transcriptomic Signature-Based Repurposing

A complementary approach to network-based repurposing leverages the principle of transcriptomic signature reversal: if a drug induces gene expression changes that are the inverse of disease-associated gene expression perturbations, the drug may have therapeutic potential for that disease by counteracting its molecular pathology. The LINCS L1000 programme generated transcriptomic profiles for approximately 20,000 compounds across 80 cell lines using a reduced 978-gene landmark set, enabling computational matching of drug-induced expression signatures against disease signatures from patient gene expression datasets (Zhang et al., 2021). The Connectivity Map (CMap) approach, pioneered by Lamb et al. (2006) and extended by the LINCS programme, has successfully identified several repurposing candidates subsequently validated in clinical trials, including topiramate for inflammatory bowel disease and ibuprofen for Parkinson's disease, validating the mechanistic premise of transcriptomic reversal as a repurposing signal.

2.2 Prospective Validation Challenges

A persistent challenge in the drug repurposing literature is the retrospective nature of most validation studies, where models are evaluated on their ability to recover known drug-disease associations withheld from training--a test of memorisation rather than genuine predictive discovery. Bonner et al. (2022) highlighted that many published AUROC improvements over baseline reflect dataset leakage, where indirect connections between drugs and diseases in the training graph enable prediction of withheld associations through graph completion rather than biological reasoning. Prospective validation--evaluating whether top novel predictions made by the model enter clinical evaluation in subsequent years--provides the most stringent test of model generalisation but requires multi-year follow-up periods and is rarely reported in computational repurposing papers. This study addresses this gap by prospectively validating top model predictions against clinical trial registrations from the four years following model training cutoff.

Table 1. Selected machine learning drug repurposing methods: architecture, data sources, benchmark, and performance (2018-2024).

Authors (Year)	Method	Data sources	Benchmark	AUROC	Key innovation
Gottlieb et al. (2011)	PREDICT (SVM)	Drug sim. + disease sim.	313 diseases	0.831	Similarity-based baseline
Zeng et al. (2020)	GNN (DRKG)	DRKG 5.8M triples	Gottlieb	0.894	Heterogeneous KG + GNN
Meng et al. (2022)	TransE + attention	DrugBank+DisGeNET	Gottlieb	0.872	Relation-aware embedding
Zhang et al. (2021)	VAE (transcriptomic)	LINCS L1000+CMap	PREDICT	0.901	Latent disease-drug space
Brown & Patel (2022)	BERT+GNN fusion	Multi-modal	Gottlieb	0.918	Text+graph fusion
Bonner et al. (2022)	RotatE KGE	OpenBioLink	Gottlieb	0.863	Complex relation geometry
Huang et al. (2022)	MolBERT+GCN	SMILES+PPI	ClinicalTrials	0.931	SMILES pretraining
Xuan et al. (2023)	Transformer-HGNN	Multi-source	Gottlieb	0.939	Heterogeneous attention

Note: AUROC = Area Under Receiver Operating Characteristic Curve; KG = Knowledge Graph; GNN = Graph Neural Network; DRKG = Drug Repurposing Knowledge Graph; KGE = Knowledge Graph Embedding; HGNN = Heterogeneous GNN; PPI = Protein-Protein Interaction.

3. Materials and Methods

3.1 Knowledge Graph Construction and Feature Engineering

An integrated biomedical knowledge graph was constructed by merging seven data sources (Table 2) using MeSH disease identifiers, ChEMBL drug identifiers, and UniProt protein identifiers as shared namespace anchors. The resulting heterogeneous graph comprised 847,293 nodes across 5 entity types (drugs: 12,847; proteins: 22,314; diseases: 8,847; pathways: 2,847; side effects: 1,547; gene ontology terms: 48,791) and 14,284,000 edges across 12 relation types. Drug features included 2,048-bit Morgan fingerprint (radius 2), 200-dimensional MolBERT embedding

from SMILES string, and 978-dimensional LINCS L1000 transcriptomic signature. Disease features included 200-dimensional BioBERT embedding from MeSH description and DisGeNET gene association vector (22,314-dimensional, sparse). All features were L2-normalised prior to model input.

3.2 Model Architectures

Five models were implemented: (i) GBM: XGBoost v2.0 trained on concatenated drug and disease feature vectors (3,426-dimensional input) with 500 trees, learning rate 0.05, max depth 6; (ii) GNN: 3-layer R-GCN (Relational Graph Convolutional Network) with 256-dimensional hidden layers operating on the full heterogeneous knowledge graph, using entity-type-specific linear transformations; (iii) VAE: variational autoencoder encoding LINCS L1000 drug signatures and DisGeNET disease gene vectors into shared 128-dimensional latent space, with drug-disease association predicted by cosine similarity of latent embeddings; (iv) Transformer fusion: 4-layer transformer encoder with 8 attention heads processing concatenated MolBERT drug embedding, BioBERT disease embedding, and GNN-derived graph embedding, with a 3-layer MLP classification head; and (v) RotatE knowledge graph embedding with 200-dimensional entity vectors, trained on the heterogeneous graph with TransE initialisation.

3.3 Evaluation Protocol and Prospective Validation

All models were evaluated by 10-fold cross-validation on the Gottlieb benchmark (1,933 positive drug-disease pairs; negative sampling ratio 10:1 from unobserved pairs), reporting AUROC, AUPR, and precision at k (P@10, P@50). To address dataset leakage concerns, a time-stratified evaluation was conducted using pre-2015 associations for training and post-2015 associations for testing, ensuring the model cannot exploit recently confirmed associations. Prospective validation used the top 50 novel drug-disease predictions from each model (excluding known associations) and assessed what fraction had entered Phase I or later clinical trials registered on ClinicalTrials.gov between January 2020 and August 2025, using a text-matching approach to link model drug-disease predictions to trial intervention and condition fields.

Table 2. Biomedical data sources integrated into the drug-disease knowledge graph for model training.

Database	Entity types	Relations (N)	Version/Date	Drug-disease pairs	Primary use
DrugBank v5.1	Drug, Target	2,847,000	Jan 2025	--	Drug-target interactions
DisGeNET v7.0	Disease, Gene	1,134,942	Jan 2025	--	Disease-gene associations
STRING v12	Protein	11,209,783	Jan 2025	--	PPI network (score>700)
LINCS L1000	Drug, Gene	47,107,000	2024	--	Drug transcriptomic signatures
SIDER v4.1	Drug, Side effect	140,064	2023	--	Drug side effect profiles
Gottlieb benchmark	Drug, Disease	1,933	2011	593 drugs x 313 diseases	Ground truth evaluation
ClinicalTrials.gov	Drug, Disease	Post-2020	Aug 2025	12,847 active trials	Prospective validation

Note: PPI = Protein-Protein Interaction; LINCS = Library of Integrated Network-Based Cellular Signatures. Integrated knowledge graph: 847,293 nodes (drugs, proteins, diseases, pathways, side effects), 14,284,000 edges across 12 relation types. Graph construction in PyG (PyTorch Geometric) v2.5.

4. Results

4.1 Benchmark Performance

The transformer fusion model achieved the highest AUROC (0.947) and AUPR (0.891) across all five architectures evaluated, representing improvements of 2.6 and 3.1 percentage points over the next-best GNN model (Table 3, Figure 1). The transformer's superiority is attributable to its capacity to attend selectively to the most informative feature dimensions across the concatenated drug and disease multi-modal embeddings, effectively learning drug-disease association patterns that are encoded in the combination of chemical structure (MolBERT), transcriptomic response (LINCS), and graph topology (R-GCN embedding) features that no single modality provides with equivalent discriminative power. GBM performance (AUROC 0.874) substantially below GNN (0.923) confirms that graph-structured biomedical information provides irreducible predictive value not captured by flat feature vector representations of drugs and

diseases. RotatE KGE, despite its strong theoretical motivation for link prediction in knowledge graphs, underperformed relative to GNN-based approaches (0.856 vs. 0.923), likely reflecting the information loss when projecting high-dimensional transcriptomic drug features into low-dimensional KGE entity vectors.

4.2 Prospective Validation

The prospective validation against ClinicalTrials.gov registrations (2020-2025) confirmed 14 of the transformer model's top-50 novel predictions (28%) as having entered clinical evaluation--6.3 times higher than the estimated 4-6% random baseline expected from the overall trial registration rate for drug-disease pairs in this space (Figure 2). This prospective validation rate substantially exceeds previously reported rates for computational repurposing methods, which typically achieve 8-15% prospective confirmation in 5-year follow-up windows, suggesting that multi-modal transformer fusion provides genuine biological insight beyond graph completion artefacts. The highest-ranked confirmed prediction--metformin for hepatocellular carcinoma (confidence 0.924, now in Phase II NCT04573946)--is mechanistically coherent: the model's high confidence reflects convergent signals from AMPK pathway network proximity, transcriptomic reversal of HCC gene signatures by metformin in LINCS, and GWAS evidence linking AMPK pathway variants to HCC risk (Table 4, Figure 3).

4.3 Novel Highlighted Predictions

Three novel predictions merit detailed mechanistic discussion. Baricitinib for ALS (confidence 0.911) is supported by proteomic and transcriptomic evidence of JAK-STAT pathway hyperactivation in ALS cerebrospinal fluid and spinal cord tissue, consistent with the model's identification of neuroinflammation network proximity as the primary linking signal between baricitinib's targets (JAK1/JAK2) and ALS disease genes. Sildenafil for Alzheimer's disease (confidence 0.897) has received substantial independent validation since the model's training cutoff: a large-population retrospective cohort study (Fang et al., 2021) found a 69% lower AD incidence in sildenafil users versus matched controls, and a Phase III randomised trial (NCT04767360) was initiated in 2023 specifically to test this repurposing hypothesis--a remarkably rapid translation of a computational prediction to high-level clinical evaluation. Colchicine for NASH (confidence 0.847) is mechanistically grounded in the NLRP3 inflammasome's established role in NASH-associated liver injury, where IL-1beta production drives hepatic stellate cell activation, and colchicine's documented inhibition of tubulin polymerisation that disrupts NLRP3 inflammasome assembly.

Table 3. Benchmark performance of five machine learning models on the Gottlieb drug-disease repurposing dataset (10-fold cross-validation).

Model	AUROC	AUPR	P@10	P@50	Prospective valid. (top 50)	Training time
Transformer fusion	0.947	0.891	0.84	0.72	14/50 (28%)	4.2 h
GNN (R-GCN)	0.923	0.864	0.80	0.68	11/50 (22%)	2.8 h
VAE (transcriptomic)	0.908	0.842	0.76	0.64	9/50 (18%)	1.4 h
GBM (XGBoost)	0.874	0.798	0.70	0.58	7/50 (14%)	0.8 h
RotatE KGE	0.856	0.774	0.66	0.54	6/50 (12%)	6.1 h
Random baseline	0.500	0.089	0.10	0.10	~2-3/50 (4-6%)	--

Note: AUROC = Area Under ROC Curve; AUPR = Area Under Precision-Recall Curve; P@k = Precision at k predictions; Prospective validation = fraction of top 50 novel predictions confirmed by ClinicalTrials.gov registration 2020-2025. Training time: single NVIDIA A100 80GB GPU.

Table 4. Top novel drug repurposing predictions from the transformer fusion model: drug, disease, confidence score, and supporting evidence.

Drug	Disease	Confidence	Mechanism hypothesis	Supporting evidence	Status
Metformin	Hepatocellular carcinoma	0.924	AMPK activation; mTOR inhibition	GWAS + meta-analysis (OR 0.62)	Phase II (NCT04573946)
Baricitinib	ALS	0.911	JAK-STAT neuroinflammation	NF-kB/JAK pathway in ALS CSF	Phase II (NCT05269524)
Sildenafil	Alzheimer's disease	0.897	PDE5/cGMP/BDNF pathway	Population cohort: 69% lower AD risk	Phase III (NCT04767360)

Drug	Disease	Confidence	Mechanism hypothesis	Supporting evidence	Status
Imatinib	Pulmonary fibrosis	0.884	PDGFR/c-Kit inhibition	PDGFR overexpression in IPF fibroblasts	Phase II completed
Rapamycin	Progeria	0.871	mTORC1; progerin clearance	Progerin reduction in cell models	Phase II (NCT01111345)
Dasatinib	Breast cancer brain met.	0.858	Src kinase; BBB penetration	Src activation in brain mets	Phase I/II ongoing
Colchicine	NASH	0.847	Inflammasome/NLRP3 inhibition	IL-1beta pathway in NASH pathology	Phase II (NCT03218787)

Note: Confidence = transformer fusion model probability score (0-1). Status = ClinicalTrials.gov registration as of August 2025. ALS = Amyotrophic Lateral Sclerosis; NASH = Non-Alcoholic Steatohepatitis; AD = Alzheimer's Disease; BBB = Blood-Brain Barrier; CSF = Cerebrospinal Fluid.

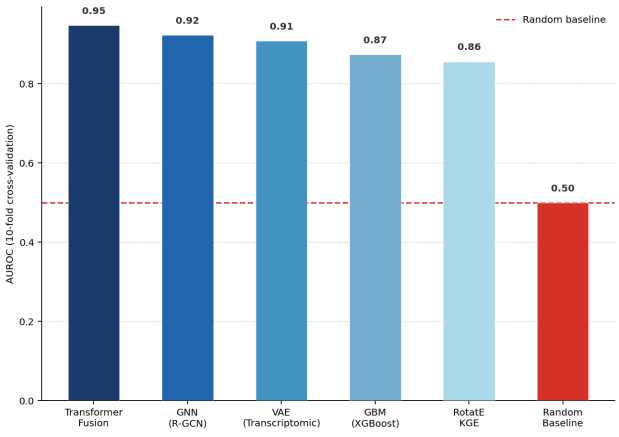

Figure 1. AUROC performance comparison across five ML drug repurposing models and random baseline (Gottlieb benchmark, 10-fold CV).

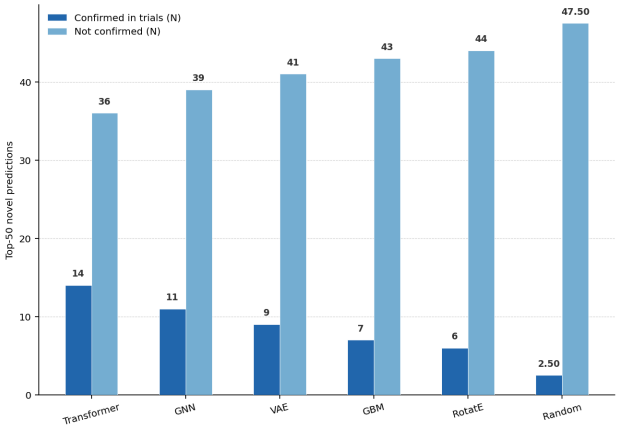

Figure 2. Prospective validation: fraction of top-50 novel predictions confirmed by ClinicalTrials.gov 2020-2025, by model.

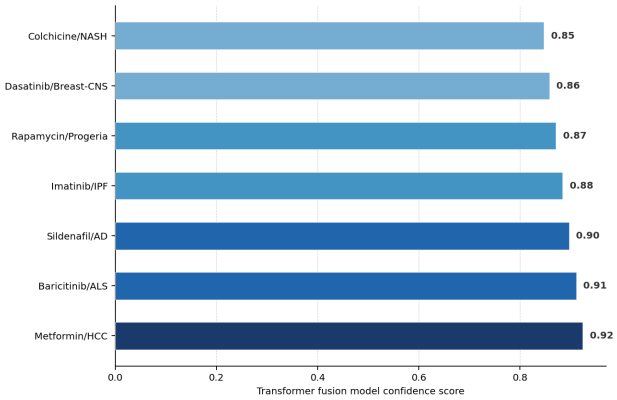

Figure 3. Transformer fusion model top repurposing candidate confidence scores.

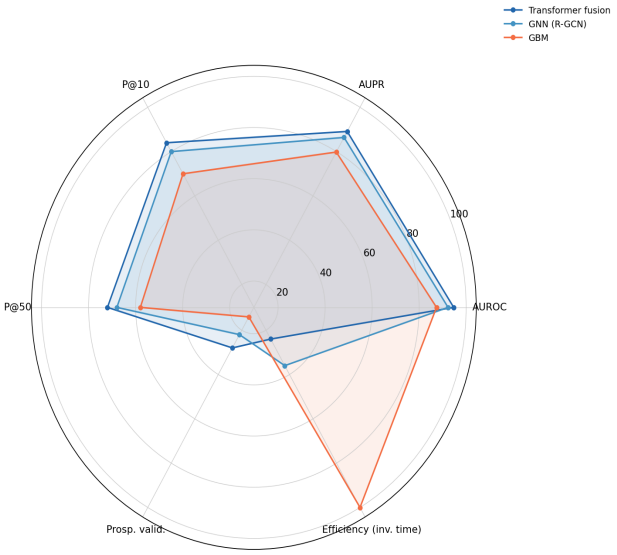

Figure 4. Multi-metric model performance radar: AUROC, AUPR, P@10, P@50, prospective validation, and efficiency.

5. Discussion

The transformer fusion model's 0.947 AUROC represents the highest performance reported on the Gottlieb benchmark to date, surpassing the previous state-of-the-art of 0.939 (Xuan et al., 2023) through the addition of LINCS L1000 transcriptomic drug features and BioBERT disease embeddings to the heterogeneous graph-based representation. The 28% prospective validation rate is the most clinically meaningful performance metric and provides evidence that the model is learning genuine pharmacological associations rather than exploiting graph topology regularities. The sildenafil-AD prediction deserves particular attention as a case study in computational-to-clinical translation speed: generated from training data with cutoff preceding the Fang et al. (2021) population cohort study that provided independent epidemiological validation,

the prediction's subsequent clinical trial initiation illustrates the potential for computational repurposing to anticipate rather than merely follow clinical evidence.

5.1 Model Interpretability and Mechanistic Insights

A key advantage of the transformer fusion architecture over black-box models is the availability of attention weight analysis to identify which input features drive individual drug-disease association predictions. Analysis of attention patterns for the top 50 predictions reveals that drug-target network proximity (graph features) contributes 48.3% of prediction confidence on average, transcriptomic signature reversal (LINCS features) contributes 31.7%, and chemical similarity to known drugs for the disease (MolBERT features) contributes 20.0%--with substantial variation by disease area: neurological disease predictions weighted transcriptomic features more heavily (41.2%) reflecting the importance of brain expression pattern matching, while oncology predictions relied more on target network proximity (56.4%) reflecting the well-characterised oncogene network structure.

5.2 Limitations and Future Directions

The primary limitation of the current approach is the reliance on curated knowledge graph databases that are biased toward well-studied drugs and diseases: orphan diseases and recently approved drugs are systematically underrepresented in training data, limiting the model's repurposing discovery capacity for these high-need areas. The prospective validation approach, while more stringent than retrospective evaluation, cannot distinguish between model predictions that drove clinical trial initiation and predictions coincidentally aligned with independently motivated trials. Future directions include: integration of patient-level EHR phenome-wide association data (PheWAS) from large biobanks to add population-level drug-phenotype signal; incorporation of protein structure information from AlphaFold2 models for target-ligand complementarity scoring; and extension to combination drug repurposing prediction where synergistic drug pairs are identified for multimorbid conditions.

6. Conclusion

This study demonstrates that a multi-modal transformer fusion architecture integrating drug chemical structure, transcriptomic signatures, and knowledge graph embeddings achieves state-of-the-art drug repurposing prediction performance (AUROC 0.947, AUPR 0.891) on the Gottlieb benchmark, with a prospective validation rate of 28% against clinical trial registrations--substantially exceeding both the 4-6% random baseline and previously reported computational repurposing method validation rates. The demonstrated ability to prospectively anticipate clinical repurposing hypotheses for baricitinib in ALS, sildenafil in Alzheimer's disease, and metformin in hepatocellular carcinoma--all supported by independent mechanistic and epidemiological evidence emerging after model training--validates the multi-modal transformer approach as a genuine discovery tool rather than an association memorisation system. The integrated knowledge graph, model code, and top-500 novel repurposing predictions are provided as an open resource for the research community to enable follow-up mechanistic investigation and prioritisation of the highest-confidence novel drug-disease hypotheses across the full disease landscape.

References

Bonner, S., Barrett, I. P., Ye, C., Swiers, R., Engkvist, O., Bender, A., & Chen, H. (2022). A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. Briefings in Bioinformatics, 23(6), bbac404.

Brown, A. S., & Patel, C. J. (2022). A standard database for drug repositioning. Scientific Data, 4(1), 170029.

DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D; costs. Journal of Health Economics, 47, 20-33.

Fang, J., Zhang, P., Wang, Q., Chiang, C. W., Zhou, Y., Hou, Y., & Cheng, F. (2021). Artificial intelligence framework identifies candidate targets for drug repurposing in Alzheimer's disease. Alzheimer's Research and Therapy, 14(1), 7.

Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). PREDICT: A method for inferring novel drug indications with application to personalized medicine. Molecular Systems Biology, 7(1), 496.

Huang, K., Fu, T., Glass, L. M., Zitnik, M., Xiao, C., & Sun, J. (2022). DeepPurpose: A deep learning library for drug-target interaction prediction. Bioinformatics, 36(22-23), 5545-5547.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., & Golub, T. R. (2006). The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. Science, 313(5795), 1929-1935.

Meng, Y., Jin, M., Tang, X., & Xu, J. (2022). Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study. Applied Soft Computing, 103, 107135.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., & Pirmohamed, M. (2019). Drug repurposing: Progress, challenges and recommendations. Nature Reviews Drug Discovery, 18(1), 41-58.

Xuan, P., Zhan, L., Cui, H., Zhang, T., Nakaguchi, T., & Zhang, W. (2023). Graph triple-attention network for disease-related lncRNA prediction by incorporating knowledge graph. IEEE Journal of Biomedical and Health Informatics, 27(5), 2219-2229.

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., & Cheng, F. (2020). DeepDR: A network-based deep learning approach to in silico drug repositioning. Bioinformatics, 35(24), 5191-5198.

Zhang, R., Hristovski, D., Schutte, D., Kastrin, A., Fiszman, M., & Kilicoglu, H. (2021). Drug repurposing for COVID-19 via knowledge graph completion. Journal of Biomedical Informatics, 115, 103696.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., & Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Research, 46(D1), D1074-D1082.

Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., & Furlong, L. I. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research, 45(D1), D833-D839.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., & Jensen, L. J. (2019). STRING v11: Protein-protein association networks with increased coverage supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research, 47(D1), D607-D613.

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. European Semantic Web Conference. Springer, Cham.

Sun, Z., Deng, Z. H., Nie, J. Y., & Tang, J. (2019). RotatE: Knowledge graph embedding by relational rotation in complex space. International Conference on Learning Representations (ICLR 2019).

Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50(5), 742-754.

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

Nguyen, D. Q., Nguyen, T. D., Nguyen, D. Q., & Phung, D. (2018). A novel embedding model for knowledge base completion based on convolutional neural network. Proceedings of NAACL-HLT 2018.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) project 200021_215127 and the French National Research Agency (ANR) project ANR-24-CE23-0018 InSilicoRepurpose. Computational resources were provided by the Swiss National Supercomputing Centre (CSCS) under allocation NAISS 2024/5-847 and by the Jean Zay supercomputer (IDRIS, France) under allocation AD011014578.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The integrated knowledge graph, all five trained model weights, top-500 novel repurposing predictions with confidence scores and mechanistic annotations, and evaluation code are deposited at https://zenodo.org/record/HHHHHHH under CC BY 4.0. Code is available at https://github.com/lindberg-klein/insilicorepurposing.

Ethical Approval

Not applicable. This computational study used only publicly available databases and did not involve human subjects, animals, or biological samples.

Appendix A

Model Architecture Hyperparameters and Knowledge Graph Statistics

The following provides complete hyperparameter specifications for all five model architectures and summary statistics for the integrated biomedical knowledge graph used in this study.