RUTGERS
NEW BRUNSWICK

AI-Driven Drug Repurposing Using Multi-Modal Deep Learning & Graph Neural Networks

Manideep Pendyala, Ashwath Ramsundar

Introduction

Traditional Drug Development Challenges:

- **Time-intensive process:** 10-15 years from discovery to market [1].
- **Prohibitively expensive:** Average cost of $2.6 billion per new drug [2].
- **High failure rate:** > 90% of candidates fail in clinical trials
- **Limited return on investment:** Decreasing efficiency in pharma R&D [3].

Drug Repurposing :

- Finding new therapeutic uses for existing FDA-approved drugs
- Also known as : drug repositioning, drug reprofiling, drug redirecting
- Focuses on compounds with established safety profiles

Motivation

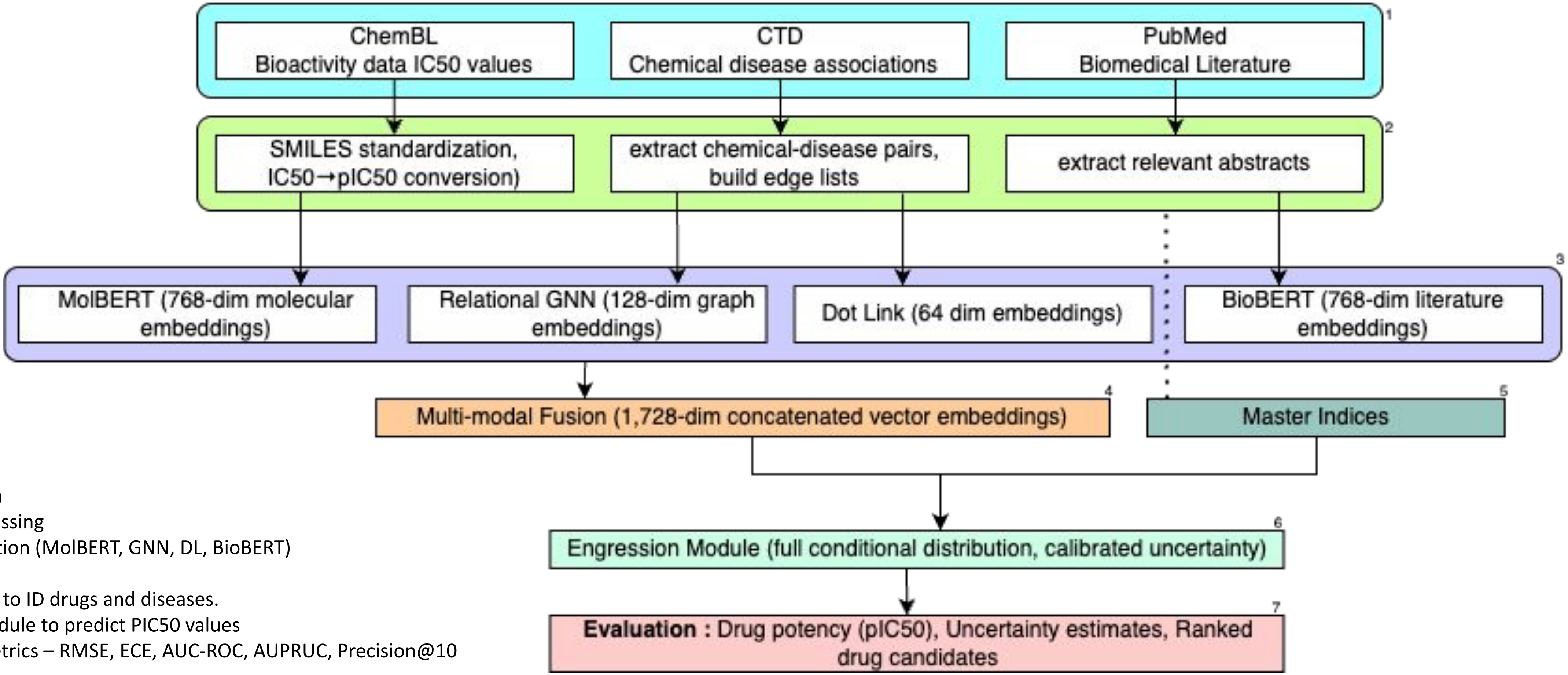
Advantages of Repurposing:

- **Reduced development time:** 3-12 years → 1-5 years [4].
- **Lower costs:** Up to 85% reduction compared to new drug development.
- **Decreased risk:** Known safety profiles and pharmacokinetics.
- **Established manufacturing protocols:** Streamlined production [5].

Examples of Successful Repurposing:

- Sildenafil: From angina treatment to erectile dysfunction (Viagra) [6].
- Thalidomide: From morning sickness to multiple myeloma treatment
- Metformin: From diabetes to emerging cancer applications [7].

Our Approach


1
ChemBL
Bioactivity data IC50 values
CTD
Chemical disease associations
PubMed
Biomedical Literature
2
SMILES standardization, IC50→pIC50 conversion)
extract chemical-disease pairs, build edge lists
extract relevant abstracts
3
MolBERT (768-dim molecular embeddings)
Relational GNN (128-dim graph embeddings)
Dot Link (64 dim embeddings)
BioBERT (768-dim literature embeddings)
4
Multi-modal Fusion (1,728-dim concatenated vector embeddings)
5
Master Indices
6
Engression Module (full conditional distribution, calibrated uncertainty)
7
Evaluation : Drug potency (pIC50), Uncertainty estimates, Ranked drug candidates

1. Data Collection
2. Data Pre-processing
3. Feature Extraction (MolBERT, GNN, DL, BioBERT)
4. Fusion layer
5. Master indices to ID drugs and diseases.
6. Engression module to predict PIC50 values
7. Evaluation (Metrics – RMSE, ECE, AUC-ROC, AUPRUC, Precision@10

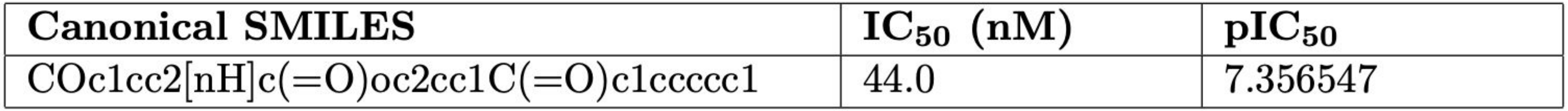
Fig 1 : Multi-modal, Uncertainty-aware Deep Learning Framework

RUTGERS

Datasets

ChEMBL (Chemical Database of Bioactive Molecules with Drug-Like Properties) dataset [8]:

- Bioactivity data, molecular structures, pharmacological annotations
- SMILES (**Simplified Molecular Input Line Entry System) -** A text-based notation that encodes a molecule’s **atoms, bonds, rings, and branches** as a compact string of characters.
- Bioactivity values (IC50, **Half-maximal inhibitory concentration**) - The concentration of a drug (or compound) required to inhibit a biological or biochemical function **by 50%**.
- 2.2 million compounds, 15 million activities, 13,000 targets.



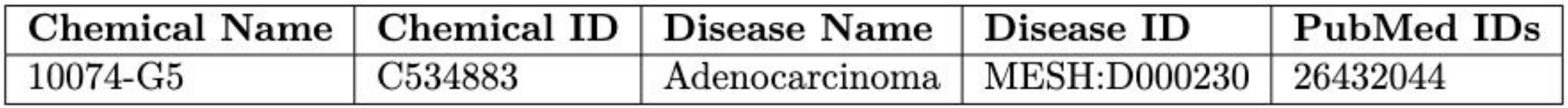
Canonical SMILES	**IC_{50} (nM)**	**pIC_{50}**
COc1cc2[nH]c(=O)oc2cc1C(=O)c1ccccc1	44.0	7.356547

Fig 2 : Example of a molecular bioactivity record showing SMILES, IC50, and pIC50 values.

Datasets

Comparative Toxicogenomics Database (CTD) [9]:

- **Content:** Chemical-disease-gene interactions, phenotypic outcomes
- **Scale:** 15,000 chemicals, 5,000 diseases, 120,000 associations
- **Our usage:** Knowledge graph construction for graph neural networks
- **Data format:** Chemical-disease pairs with identifiers



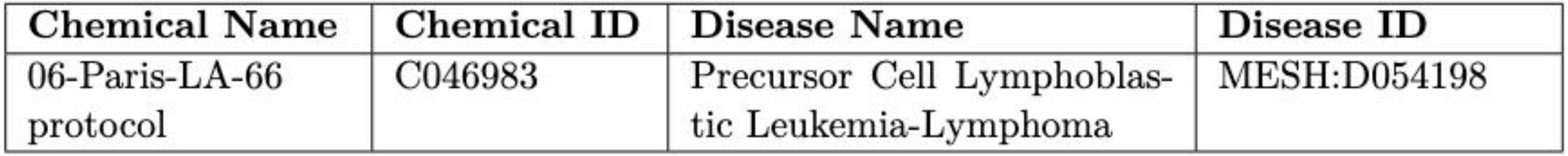
Chemical Name	Chemical ID	Disease Name	Disease ID	PubMed IDs
10074-G5	C534883	Adenocarcinoma	MESH:D000230	26432044

Fig 3 : Example of a chemical–disease interaction record from the CTD database

Datasets

PubMed Biomedical Literature [10]:

- Scientific abstracts, full-text articles, clinical reports
- 2,000 abstracts specifically relevant to drug repurposing
- Contextual information extraction via BioBERT
- Title, abstract, MESH terms, publication metadata



Chemical Name	Chemical ID	Disease Name	Disease ID
06-Paris-LA-66 protocol	C046983	Precursor Cell Lymphoblastic Leukemia-Lymphoma	MESH:D054198

Fig 4 : Example PubMed, extracted chemical–disease association for 06-Paris-LA-66

Master Indices

1. We created **master index dictionaries** to assign unique integer indices to each entity type - Chemicals, Diseases, Genes.
2. Instead of using raw string IDs (like C534883 or MESH:D000230), we mapped all identifiers to **global integer indices**.
3. Align chemical IDs across graph, text, molecular, and DotLink components.
4. The integer indices served as direct lookups into: GNN node embeddings, DotLink embedding tables, Fusion-layer concatenations.

Data Pre-processing

CTD Processing

- **Entity normalization:**
  - Mapped textual chemical/disease IDs to numeric indices
  - Resolved synonyms and alternative nomenclature
- **Knowledge graph construction:**
  - Created adjacency matrices (15K × 5K sparse matrices)
  - Generated positive chemical-disease edge lists (120K pairs)
- **Training preparation:**
  - 80/10/10 train/validation/test split at the relationship level
  - Negative sampling (1:3 ratio of positive to negative examples)

Data Pre-processing

ChEMBL Processing:

- SMILES standardization:
- Bioactivity transformation:
  - IC50 (nM) → pIC50 conversion (pIC50 = -log10(IC50 * 10^-9)

PubMed Processing:

- Corpus preparation:
  - Keyword-based retrieval (drug repurposing, repositioning)
  - Filtered for relevance and recency (past 5 years prioritized)
- Text preprocessing:
  - Entity recognition (drugs, diseases, genes)

GNN (Graph Neural Networks)

- Graph Neural Networks are machine learning models designed to work directly on **graph-structured data** — data made up of **node** (entities) and **edges** (relationships).
- R-GCNs don't treat all edges equally — they learn **separate weight matrices for each relatio type** to handle different kinds of interactions.
- **Why GNN? -** To capture chemical–disease–gene relationships from CTD graphs, including indirect, multi-hop links.

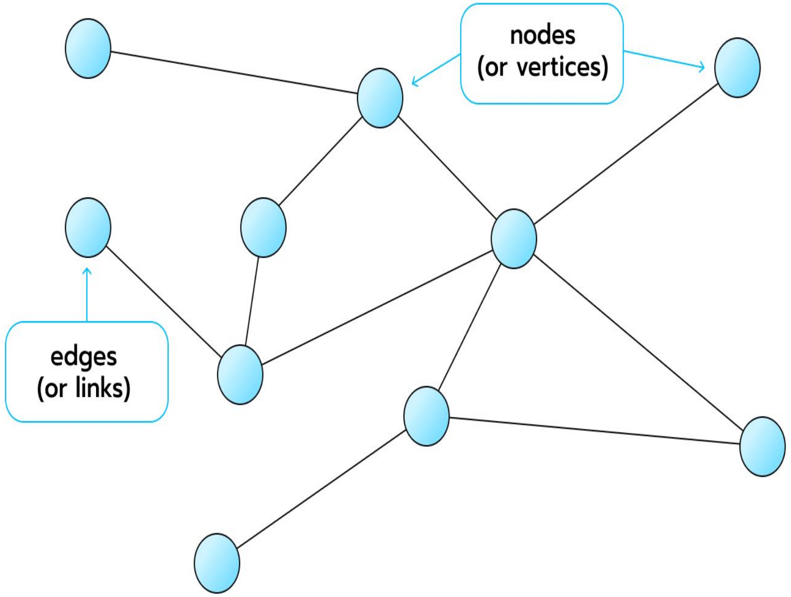


Fig 5 : GNN architecture [18]

GNN (Graph Neural Networks)

- **Inputs :** Edge lists from CTD; adjacency matrices; negative sampling.
- **Outputs :** 128-dim node embeddings (chemicals & diseases) capturing network topology.
- Provides structural knowledge to fusion layer, complementing MolBERT and BioBERT.
- Models indirect relationships and boosts predictions where molecular or text data alone is weak.

MolBERT (Molecular BERT)

- **Based on BERT (Bidirectional Encoder Representations from Transformers)** architecture. **We used pre-trained weights.**
- Encodes molecular structures (SMILES, ChEMBL) into 768-dim embeddings.
- BERT-based (12 layers, 768-dim), pretrained on 1.1M SMILES, mean-pooled outputs.
- Captures chemical grammar & activity; predicts potency from sequence; complements GNN for novel compounds.

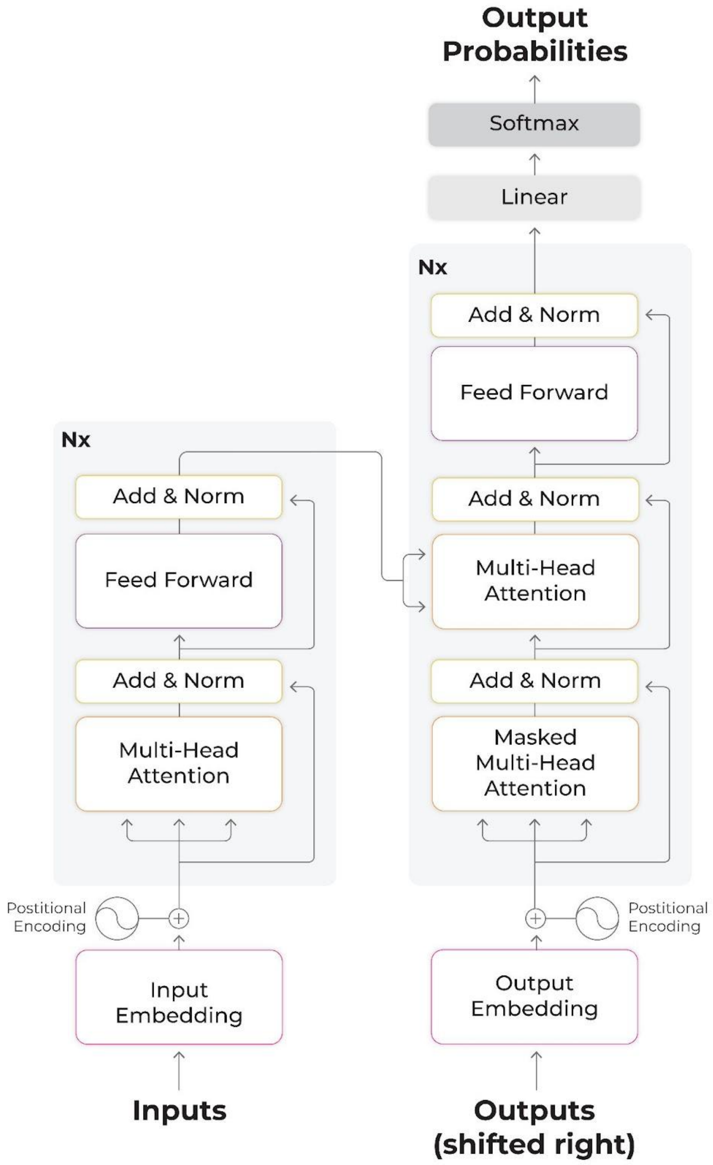


Fig 6 : BERT architecture [19]

BioBERT

- Encodes biomedical literature (PubMed abstracts) into 768-dim embeddings.
- 12-layer BioBERT (768-dim), pretrained on PubMed + PMC data.
- ~2,000 curated drug repurposing abstracts, focusing on drug–disease co-mentions.
- Extracts document-level embeddings via attention pooling; feeds into the fusion layer.
- Captures biomedical context, mechanisms, and trends; complements graph and molecular data; boosts predictions, especially for well-studied drugs.

Dot-Link

- Provides 64-dimensional embeddings of chemical–disease pairs from the CTD knowledge graph.
- **Architecture:** Matrix factorization over CTD pairs (1:3 pos–neg); embeds chemicals & diseases, combines via dot product.
- Complements the GNN by modeling broad, global co-occurrence patterns instead of just local relational structures.
- Acts as both a baseline predictor and an additional feature input for the multi-modal fusion layer.
- Especially valuable when molecular or text data are missing or weak, providing a foundational relational signal to strengthen predictions.

Fusion Methodology

- **Integration:** Concatenates embeddings → 128 (GNN) + 768 (MolBERT) + 768 (BioBERT) + 64 (DotLink) → 1728-dim
- **Reduction:** Two-layer MLP (1728 → 512 → 256) with ReLU.
- outputs compact 256-dim vector for downstream prediction, fuses multi-source knowledge into a unified representation.

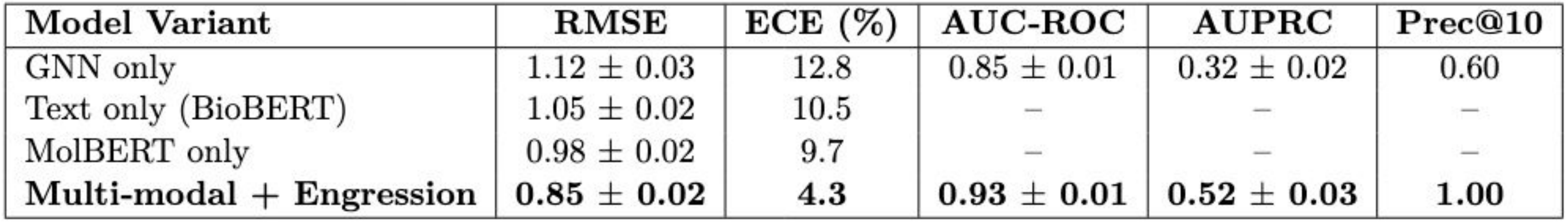
Engression

1. Engression goes beyond predicting just the mean by providing **full conditional distribution** P(y|x), uncertainty estimates, standard deviations, confidence intervals,
2. Learns an Energy function, where lower energy means higher likelihood.
3. **Training & Inference :** Uses score matching loss with noisy samples; at inference, integrates over the y-range to compute exceedance probabilities and asymmetric confidence intervals.
4. Improves **tail risk estimates** — critical for risk-aware decisions like drug candidate selection.

Results

Evaluation Metrics :

1. **RMSE (Root Mean Squared Error) :** Measures the average size of prediction errors; lower values mean better accuracy.
2. **ECE (Expected Calibration Error) :** Measures how well the predicted uncertainties match actual outcomes; lower values mean better-calibrated confidence.
3. **AUC-ROC (Area Under ROC Curve) :** Measures how well the model separates positive from negative cases; higher values mean better discrimination.
4. **AUPRC (Area Under Precision-Recall Curve) :** Measures the trade-off between precision and recall, especially on imbalanced data; higher values mean better positive detection.
5. **Precision@10 :** Measures the fraction of correct predictions in the top 10 ranked results; higher values mean better top-candidate selection.



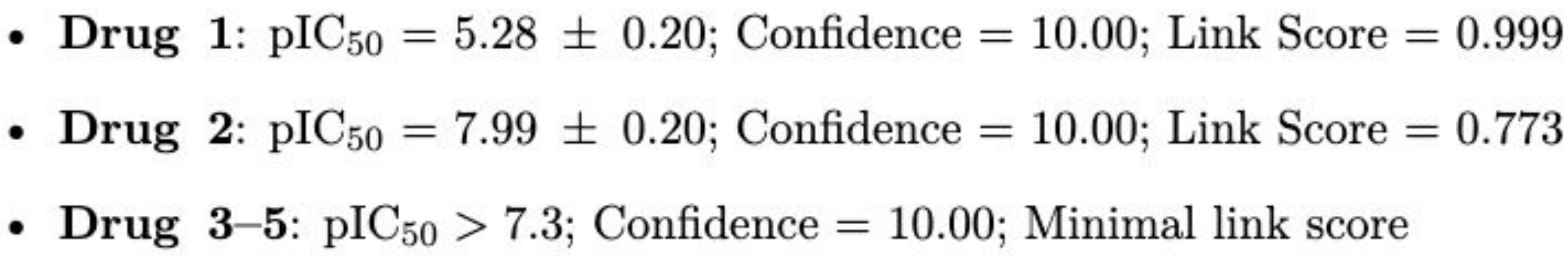
Model Variant	RMSE	ECE (%)	AUC-ROC	AUPRC	Prec@10
GNN only	1.12 ± 0.03	12.8	0.85 ± 0.01	0.32 ± 0.02	0.60
Text only (BioBERT)	1.05 ± 0.02	10.5	–	–	–
MolBERT only	0.98 ± 0.02	9.7	–	–	–
Multi-modal + Engression	$\mathbf{0.85 \pm 0.02}$	**4.3**	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.52 \pm 0.03}$	**1.00**

Results

- **Performance Gains :** 24% RMSE improvement over GNN-only, 13.3% boost over best single-modality (MolBERT), 66% reduction in calibration error (ECE).
- **Uncertainty Calibration :** Well-calibrated uncertainty with 4.3% ECE; Reliable confidence intervals for decision support.
- **Modality Contributions:** MolBERT → 40% gain (molecular structure); BioBERT → 25% gain (literature context); GNN → 35% gain (graph relationships).
- **Uncertainty Modeling Impact :** Engression module reduces calibration error by 33%.

Case study

1. **To evaluate real-world utility, we applied the model to disease, MESH : C000598644.**
2. **We identified five high-confidence drug candidates, all showing strong agreement across modalities**



- **Drug 1**: $\mathrm{pIC}_{50} = 5.28 \pm 0.20$; Confidence = 10.00; Link Score = 0.999
- **Drug 2**: $\mathrm{pIC}_{50} = 7.99 \pm 0.20$; Confidence = 10.00; Link Score = 0.773
- **Drug 3–5**: $\mathrm{pIC}_{50} > 7.3$; Confidence = 10.00; Minimal link score

- **Drug 1 :** Strong graph connectivity (0.999), Strong graph signal but low potency; **not recommended**.
- **Drug 2 :** Moderate graph connectivity (0.773), excellent **candidate**
- **Drugs 3–5 :** Lower graph (0.592–0.645), **Viable secondary candidates** with >90% probability of exceeding threshold.

Case study

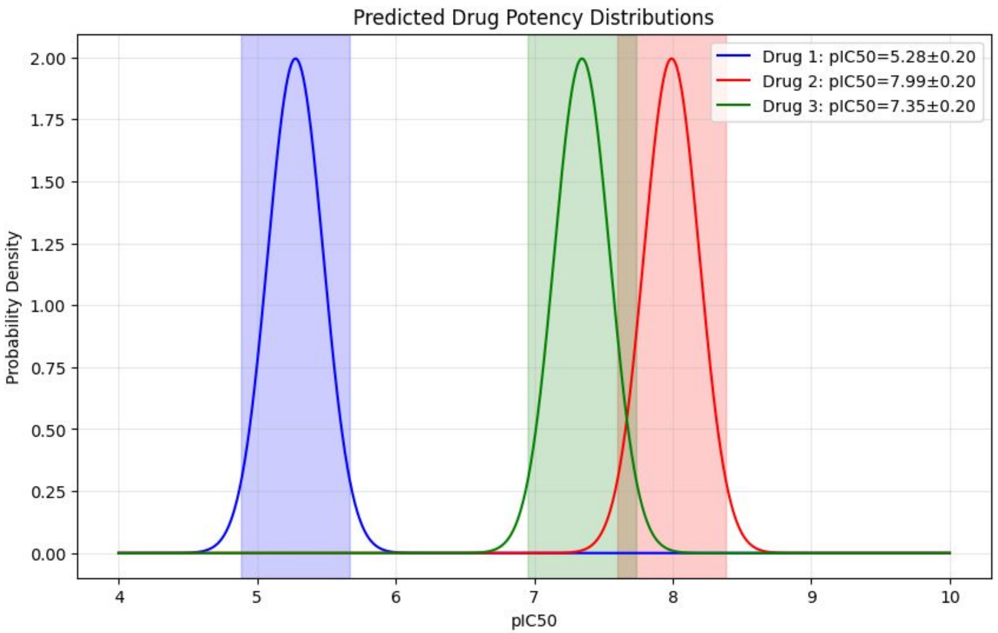


Fig 7 : Predicted drug potency distributions

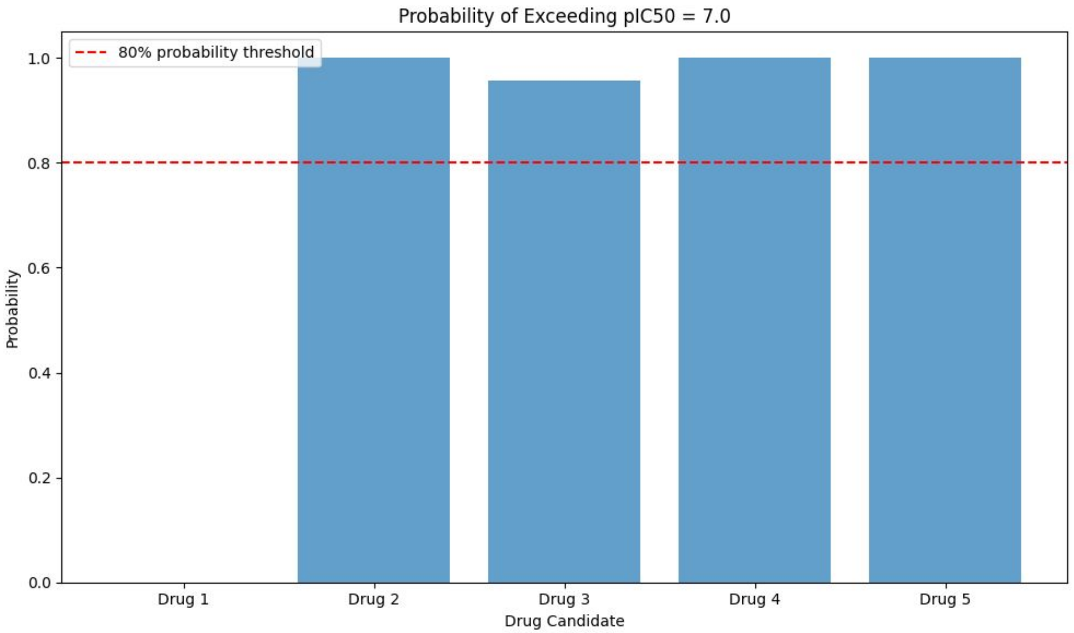


Fig 8 : Probability of exceeding pIC50 = 7.0 for each drug candidate

- **Potency Distribution :** Clear separation between Drug 1 and others, Tight confidence intervals, Drugs 2–5 mostly above therapeutic threshold.
- **Multi-modal Impact :** Avoided false positives from graph-only (Drug 1), Identified top candidate (Drug 2) via complementary signals.

Conclusion

- **Multi-modal integration:** Unified, end-to-end framework combining diverse biomedical data; +13.3% performance over best single modality.
- **Uncertainty quantification:** 66.4% reduction in calibration error; reliable confidence intervals.
- **Engression modeling:** Captures full output distribution; enables exceedance probabilities and robust candidate selection.
- **Interpretability:** Modality-level attribution; transparent, explainable predictions; builds expert trust.

Future Directions

1. Scaling the Knowledge Graph:

   1. Expansion to 500,000+ chemicals, 10,000+ diseases
   2. Integration with DrugBank, ClinicalTrials.gov, PubChem

2. Model Enhancement:

   3. Foundation models for chemical structures
   4. Incorporation of protein target information

References

1. https://lifesciences.n-side.com/blog/what-is-the-average-time-to-bring-a-drug-to-market-in-2022?utm_source=chatgpt.com
2. https://greenfieldchemical.com/2023/08/10/the-staggering-cost-of-drug-development-a-look-at-the-numbers/?utm_source=chatgpt.com
3. Fernald, K. D., Förster, P. C., Claassen, E., & Van de Burgwal, L. H. (2024). The pharmaceutical productivity gap – Incremental decline in R&D efficiency despite transient improvements. *Drug Discovery Today*, *29*(11), 104160. https://doi.org/10.1016/j.drudis.2024.104160
4. Rao, N., Poojari, T., Poojary, C. *et al.* Drug Repurposing: a Shortcut to New Biological Entities. *Pharm Chem J* **56**, 1203–1214 (2022). https://doi.org/10.1007/s11094-022-02778-w
5. https://www.drugpatentwatch.com/blog/drug-repurposing-an-overview/?srsltid=AfmBOorRjwjT2JGfrRpfXS_5sTBNZ1OF1iy6EbM3IZB3_twgeBSIEHEu&utm_source=chatgpt.com
6. https://pharmaceutical-journal.com/article/feature/repurposing-viagra-the-little-blue-pill-for-all-ills?utm_source=chatgpt.com
7. Hua, Y., Zheng, Y., Yao, Y. *et al.* Metformin and cancer hallmarks: shedding new lights on therapeutic repurposing. *J Transl Med* **21**, 403 (2023). https://doi.org/10.1186/s12967-023-04263-8
8. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012 Jan;40(Database issue):D1100-7. doi: 10.1093/nar/gkr777. Epub 2011 Sep 23. PMID: 21948594; PMCID: PMC3245175.
9. https://ctdbase.org/downloads/
10. https://pubmed.ncbi.nlm.nih.gov/
11. Fabian, B., Edler, C., et al. (2020). *Molecular representation learning with language models and domain-relevant auxiliary tasks*. arXiv:2011.13230.
12. Gal, Y., & Ghahramani, Z. (2016). *Dropout as a Bayesian approximation: Representing model uncertainty in deep learning*. ICML.
13. Kendall, A., & Gal, Y. (2017). *What uncertainties do we need in Bayesian deep learning for computer vision?* NeurIPS.
14. Kingma, D. P., & Ba, J. (2015). *Adam: A method for stochastic optimization*. ICLR.
15. Lee, J., Yoon, W., Kim, S., et al. (2020). *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*. Bioinformatics, 36(4), 1234–1240.
16. Loshchilov, I., & Hutter, F. (2017). *SGDR: Stochastic gradient descent with warm restarts*. ICLR.
17. Schlichtkrull, M., Kipf, T. N., Bloem, P., et al. (2018). *Modeling relational data with graph convolutional networks*. The Semantic Web.
18. https://www.avenga.com/magazine/graph-neural-networks-and-graph-convolutional-networks/
19. https://arize.com/blog-course/unleashing-bert-transformer-model-nlp/