Enhanced Drug Toxicity Prediction via Reverse Transfer Learning and Graph-Based Visual Verification

Govind Bhatter[1], Animesh Shukla[1], Pratham Popatiya[1], Pratik Shah[1], and Jignesh Patel[1]*

Indian Institute of Information Technology Vadodara
Department of Computer Science and Engineering
{202311033,202311009,202311065}@diu.iiitvadodara.ac.in,
pratik@iiitvadodara.ac.in, *jignesh_patel@diu.iiitvadodara.ac.in*

Abstract. Predicting molecular toxicity and verifying drug identity in real-world scenarios is a fundamental challenge in ensuring pharmaceutical safety. This paper presents a combined framework addressing both issues through a two-stage pipeline process. First, we validate a reverse transfer learning module, demonstrating that a Graph Neural Network (GNN) pre-trained on high-level clinical phenotypes (SIDER 4.1 dataset) which captures rich molecular representations that are transferable to drug toxicity prediction (Tox21). Our base SIDER-transferred GINE model achieved a competitive ROC-AUC of 0.8125 on the Tox21 benchmark dataset. Second, to address the challenge of reading text from cylindrical medicine bottles prevalent in the current Indian market, we implemented an automated high-precision OCR pipeline. By employing a cascading deep learning strategy using EasyOCR combined with a heuristic smart-matching algorithm and an early-exit optimization, the visual system achieved an accuracy of 98.0% on a validation subset, demonstrating robustness against cylindrical distortion and specular reflection on medicine bottles.

Keywords: Graph Neural Networks, Transfer Learning, Tox21, OCR, Drug Verification, SIDER

1 Introduction

Drug discovery and pharmaceutical safety are highly regulated processes aimed at ensuring that medicines are both effective and safe for human use. Two persistent challenges in this pipeline are accurate molecular toxicity prediction during drug development and reliable drug identity verification in real-world usage scenarios.

* Corresponding author.

S. Sharma et al. (eds.), *Proceedings of the International Conference on Recent Advances in Intelligent and Sustainable Technologies (RAIST 2026)*, Atlantis Highlights in Intelligent Systems 17,
https://doi.org/10.2991/978-94-6239-707-1_29

Molecular Toxicity Prediction

One of the key questions in drug development is whether a molecular compound is toxic or not. In traditional computational models, molecular fingerprints are usually handcrafted; however, these 2D representations are not expressive enough to represent the properties of molecules, resulting in loss of information.

Graph neural networks (GNNs) overcome this problem by modeling molecules as graphs, where the atoms are treated as nodes and the chemical bonds as edges. This allows modeling of interactions between atoms, bonds, and topological properties of molecules.

To further improve the predictive capability, we adopt a transfer learning approach called *reverse transfer learning*, where a GNN is initially trained on high-level side-effect data from the Side Effect Resource Dataset (SIDER) to acquire transferable chemical common sense. The pre-trained model is then fine-tuned on the Tox21 dataset for molecular toxicity prediction.

In addition to molecular safety prediction, correct identification of pharmaceuticals in real-world scenarios is also of equal importance.

Visual Drug Verification via OCR

To build a comprehensive end-to-end safety solution, we introduce an optical character recognition (OCR) module to decode the information printed on the labels of medicine bottles. This is a challenging task because of the curved surface distortion, specular reflections from plastic or glass containers, and irregular orientations of the labels, which are commonly observed in real-world medicine bottles.

To address these challenges, we design a cascaded optimization framework that combines deep learning-based OCR with intelligent matching heuristics to efficiently and effectively detect medicines.

Contributions. The main contributions of this work are:

- A clinically guided transfer learning framework in which a graph neural network pre-trained on side-effect data (SIDER) is fine-tuned for molecular toxicity prediction on Tox21.
- A robust OCR-based visual verification pipeline for identifying drugs from cylindrical packaging using a cascading early-exit strategy and fuzzy matching.
- An integrated end-to-end pharmaceutical safety framework that bridges molecular prediction and real-world visual verification.

2 Methodology: Molecular Prediction

The proposed pipeline combines a GNN for molecular prediction with transfer learning strategies to leverage prior knowledge of drug side effects, along with a cascading optimization pipeline.

2.1 GNN Architecture

We adopt a six-layer GINEConv (Graph Isomorphism Network with Edge features) architecture, enhanced with residual connections to prevent information loss across layers. The architecture includes:

- Drug molecules are represented as graphs, where atoms are nodes and bonds between these atoms are edges of the graph.
- Node Features: Each atom is encoded using a 15-dimensional one-hot vector capturing its chemical properties, such as atomic symbol, degree (number of bonds), and implicit valence. The choice of 15 dimensions is motivated by grouping elements based on similar chemical properties, analogous to grouping elements in the periodic table.
- Edge Features: Each bond is represented using a four-dimensional one-hot vector encoding bond types (single, double, triple, aromatic).

The iterative update rule for node states is defined as follows:

$$\mathbf{x}_n^t = f_w \left(\mathbf{x}_n^{t-1}, \mathbf{l}_n, \sum_{m \in Ne(n)} \phi \left(\mathbf{x}_m^{t-1}, \mathbf{l}_m, \mathbf{e}_{m,n} \right) \right) \tag{1}$$

Here, $f_w(\cdot)$ denotes a learnable node update function parameterized by trainable weights w, which combines the previous node state, static node features, and aggregated messages. The function $\phi(\cdot)$ represents a message function that transforms neighboring node states and edge features before aggregation.

Here:

- $\mathbf{x}_n^t$ represents the state of node n at iteration t.
- $\mathbf{l}_n$ represents static node features (atom descriptors).
- $\mathbf{e}_{m,n}$ represents edge features describing the bond between atoms m and n.
- $Ne(n)$ denotes the set of neighboring atoms connected to node n.

This equation indicates that each node updates its state at every iteration by aggregating information from its neighboring nodes. Over multiple iterations, each node accumulates contextual information from the entire graph rather than only its local features. Consequently, after several iterations, the network produces a rich representation of the complete molecule. Finally, a global mean pooling layer aggregates all node states into a single graph-level embedding μ_G, which serves as the molecular fingerprint for downstream tasks.

2.2 Transfer Learning Strategy

The transfer learning strategy is applied on top of a previously trained GNN using the SIDER dataset:

i. Source Training (Side Effect Resource Dataset, SIDER): The model is first trained on the SIDER 4.1 dataset. This enables the network to learn molecular structures and recognize patterns between molecular features and observed side effects.

ii. Weight Transplantation: The learned weights of the embedding layers and the six GINEConv layers are frozen and transferred to the toxicity prediction task. This preserves the chemical intuition acquired during side-effect prediction.
iii. Targeted Fine-tuning (Tox21): A new classification head, implemented as a Multi-Layer Perceptron (MLP), is attached to the six-layer GNN. It is randomly initialized with an output size of 12, corresponding to the 12 toxicity assays in the Tox21 dataset. The model then learns to map molecular embeddings to toxicity outcomes.

This strategy allows the fine-tuned model to become specialized in toxicity prediction tasks. It can process a new compound as a molecular graph and output predictions across 12 toxicity categories, supporting the identification of potentially harmful drugs.

3 Methodology: Visual Verification Pipeline

Incorrect labeling can cause life-threatening medication errors. To tackle this problem, we created an advanced OCR module, the **Cascading Optimization Pipeline**, for text recognition that has the ability to overcome the challenges that traditional OCR modules face (Fig. 1).

3.1 The OCR Engine

We use the EasyOCR module, which is a deep learning engine that uses the following components:

- **CRAFT (Character Region Awareness for Text Detection):** This component is able to detect text regions even if they are curved or distorted.
- **ResNet + LSTM + CTC Network:** This architecture recognizes detected text by combining convolutional feature extraction (ResNet), sequential modeling (LSTM), and alignment-free decoding (CTC).

3.2 Cascading Optimization Strategy

We use an early exit strategy to avoid unnecessary computations while keeping the recognition accuracy intact and ensuring that new drugs/molecules are handled correctly:

i. Step 1: OCR is applied to the resized image. If a match is obtained from the drug database, the output is generated immediately.
ii. Step 2: If Step 1 is unsuccessful, the image is rotated by 90° clockwise and Step 1 is repeated.
iii. Step 3: If Step 2 is unsuccessful, the image is rotated by 270° to accommodate inverted bottles.

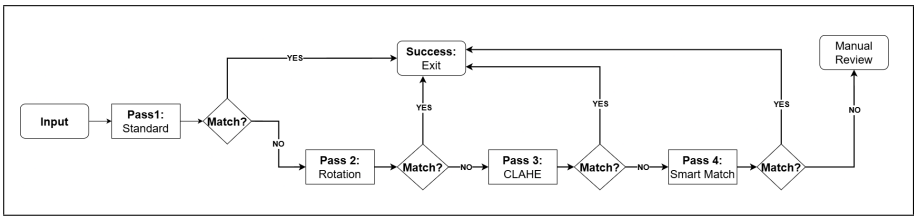


Fig. 1: The cascading "early exit" optimization pipeline for robust OCR on cylindrical medicine bottles.

iv. Step 4: If all the above steps are unsuccessful, CLAHE (Contrast Limited Adaptive Histogram Equalization) is used to remove glare and specular reflections, and OCR is reapplied.

This cascading process ensures that computationally intensive tasks are avoided unless absolutely necessary.

3.3 Smart Matching Algorithm

The output of the OCR module is noisy because similar words can easily be confused with each other. To overcome this problem, we propose:

- Text Tokenization: The extracted text is segmented into significant tokens, and frequent stopwords like "Tablet," "mg," and "Solution" are eliminated.
- Fuzzy Matching: Levenshtein distance, implemented using `SequenceMatcher`, is used with a stringent tolerance of 0.82 to accurately pick out the brand names despite subtle OCR errors.

This component explicitly enables reliable recognition of drug names.

4 Experimental Results

4.1 Dataset Description

Our framework utilizes two primary datasets to facilitate the reverse transfer learning process and to validate the visual pipeline.

SIDER 4.1 (Source Dataset for Pre-training) The Side Effect Resource (SIDER 4.1) is a comprehensive database containing information on marketed drugs and their recorded adverse drug reactions (ADRs). This dataset serves as the foundation for the model's "chemical intuition."

- Composition: It features 1,427 clinically approved small molecules mapped to 5,868 biological outcomes.

Table 1: Comparative Analysis of SIDER (Source) and Tox21 (Target) Datasets

Feature	SIDER 4.1	Tox21
Task Type	27 Multi-label Classification Tasks	12 Multi-label Classification Tasks
Dataset Size	Approximately 1,427 Molecules	Approximately 7,830 Molecules

- Hierarchy: Labels are organized according to the Medical Dictionary for Regulatory Activities (MedDRA), grouping symptoms into 27 System Organ Classes (SOC), such as hepatobiliary or cardiac disorders.
- Role: In the transfer learning pipeline, SIDER is utilized for Phase I pre-training. It allows the Graph Neural Network (GNN) to learn high-level clinical phenotypes and the systemic effects of various functional groups on the human body before moving to specific toxicity tasks.

Tox21 (Target Dataset for Fine-tuning) The Toxicology in the 21st Century (Tox21) dataset is a gold-standard benchmark in computational toxicology, originating from a federal collaboration between the EPA, NIH, and FDA. Unlike the clinical nature of SIDER, Tox21 focuses on the molecular "starting point" of toxicity.

- Composition: It consists of approximately 7,831 unique compounds evaluated via quantitative high-throughput screening (qHTS).
- Assay Panels: The compounds are screened against 12 different pathways, divided into two groups:
  i. Nuclear Receptor (NR) Panel: It comprises assays for estrogen (ER), androgen (AR), and glucocorticoid receptors.
  ii. Stress Response (SR) Panel: It comprises assays for p53 (DNA damage), heat shock response, and mitochondrial membrane potential.
- Role: This serves as the target dataset. Due to extreme class imbalance (often $< 10\%$ active molecules per assay), the pre-trained weights from SIDER are crucial for achieving high ROC-AUC scores on these sparse tasks.

4.2 Comparative Summary

Visual Verification Dataset Drug Packaging Dataset: To evaluate the OCR pipeline, we curated a subset of 100 high-resolution images of medicine bottles prevalent in the Indian market. These images represent real-world challenges, including cylindrical surface distortion, specular reflection, and multi-line pharmaceutical labeling.

4.3 Toxicity Prediction (SIDER $\rightarrow$ Tox21)

The transfer learning approach was validated on the Tox21 test set. The model demonstrated a strong ability to generalize from clinical phenotypes to molecular toxicity.

Table 2: Summary of Dataset Statistics

Dataset	Compounds	Tasks/Labels	Role
SIDER 4.1	1,427	2,060	Pre-training
Tox21	7,831	12	Fine-tuning
Drug Packaging	100 (Images)	N/A	OCR Evaluation

Table 3: Transfer Learning Performance on Tox21

Metric	Score	Context
Best Test ROC-AUC	**0.8125**	State-of-the-art competitive
Test AUPR (Macro)	**0.3074**	3.8× over random baseline

As shown in Table 3, the ROC-AUC score of 0.8125 outperforms traditional random forest models (≈ 0.76) and is competitive with deep learning benchmarks. The high AUPR is particularly significant due to the severe class imbalance in the target labels (often $< 10\%$ toxicity prevalence) in Tox21.

4.4 Baseline Training From Scratch.

A natural baseline for evaluating the benefit of the proposed transfer learning strategy would be training the same GINE-based architecture on the Tox21 dataset from random initialization. Prior studies on the Tox21 benchmark have reported that direct training of graph neural networks or conventional deep learning models (e.g., MLPs or CNNs operating on molecular fingerprints) typically yields ROC-AUC scores in the range of 0.74–0.79 under comparable settings, particularly in the presence of severe class imbalance.

In contrast, our proposed approach achieves a ROC-AUC of 0.8125, which is competitive with recent deep learning benchmarks and exceeds commonly reported non-transfer baselines. This performance gap suggests that pre-training on clinically grounded side-effect data enables the model to learn transferable chemical representations that are difficult to acquire when training on Tox21 alone.

4.5 Visual Pipeline Performance

The visual verification pipeline was evaluated on a randomized validation subset of 100 images from the Drug Packaging Dataset.

Qualitative Analysis of Logs The logs show the strength of the system in identifying brand names despite the presence of irrelevant text on medicine packaging. Some examples of the logs from the validation process are shown below:

Table 4: Automated Verification Pipeline Metrics

Metric	Value
Total Images Processed	100
Correct Identifications	98
Failed / Missed	2
Accuracy	98.00%

- Image 91 (Apulset Solution): The OCR identified "Apulset" and "Ondansetron USP". The system was able to successfully ignore the volume details ("50 ml") and correctly identify the brand name "Apulset".
- Image 98 (Complex Dosage): Ground truth: "Arbitel AM 5/80mg+5mg". Extracted: "Arbitel", "AM5/80", "Amlodipine". Result: Success. The OCR system was able to successfully link chemical ingredients and brand names despite being written on separate lines.
- Image 100 (Generic Detection): Ground truth: "Ardium 450mg". Extracted: "Ardium", "Micronised purified flavonoid...". Result: Success. Even with small print, the primary identifier was captured accurately.

5 Performance Analysis

5.1 Fine-Tuning Comparison

The initial training of the model on the SIDER 4.1 dataset provided a baseline "chemical intuition" for predicting 2,060 diverse clinical side effects. The pre-trained model was then transferred and fine-tuned for the 12 specific toxicity tasks of the Tox21 dataset. This is reflected in Table 5.

Table 5: Performance Comparison: Pre-training vs. Fine-tuning

Metric	SIDER Baseline	Tox21 Fine-tuned
Dataset	SIDER 4.1	Tox21
Tasks	2,060 (Side Effects)	12 (Toxicity Assays)
Test ROC-AUC	0.7149 ± 0.0132	**0.8125**
Test AUPR	0.2462 ± 0.0157	**0.3074**

The significant improvement is evident in the ROC-AUC value, which rises from around 0.71 in the source task to **0.8125** in the target task. This is despite the target task being from a different biological domain, and this serves as a strong indication for the proposed hypothesis.

5.2 Graphical Analysis

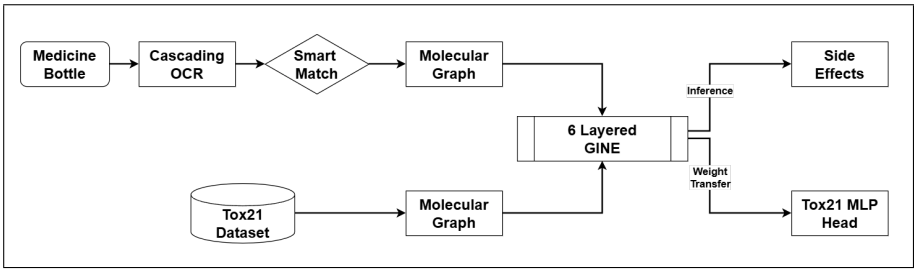


Fig. 2: End-to-end prediction pipeline illustrating the classification of toxic and non-toxic compounds.

Here, we provide further insights into the model's performance on the target Tox21 dataset through visual analysis.

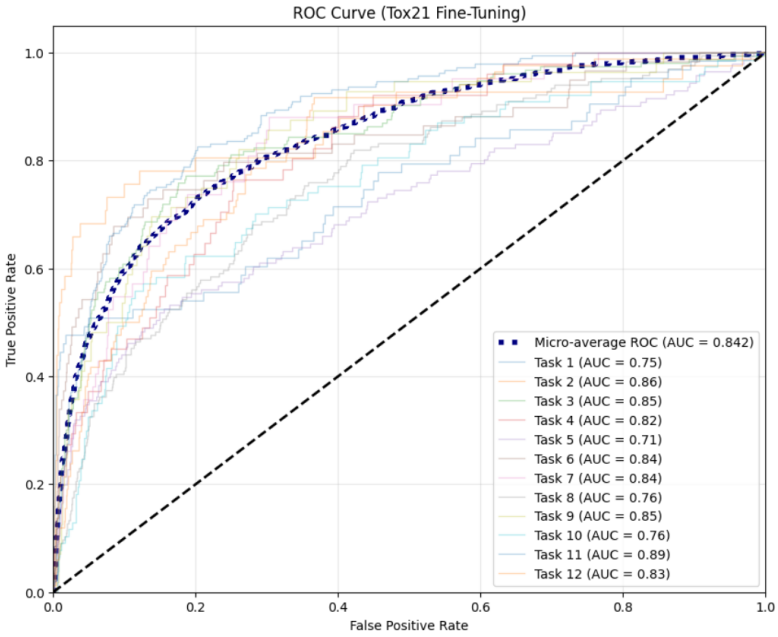
Figure 3 offers a qualitative view of the model's predictions, showcasing its ability to differentiate between toxic and non-toxic molecules within a sample subset.

Given the high class imbalance of the Tox21 dataset, the Precision–Recall (PR) curves shown in Figure 3 are particularly important. The model demonstrates a robust ability to maintain high precision at higher recall levels for the rare toxic class, achieving a macro-average AUPR of **0.3074**.

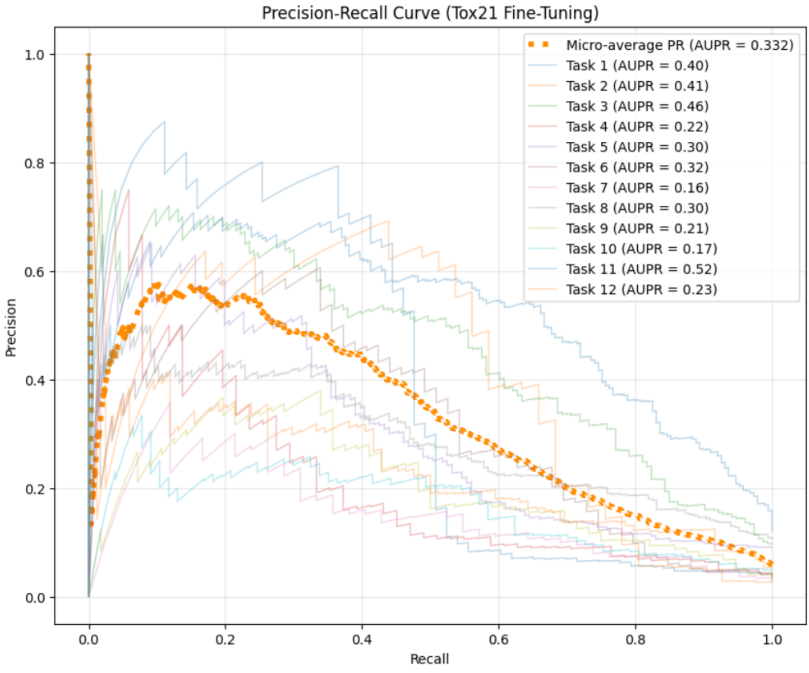
6 Conclusion

This study validates the reverse transfer learning hypothesis: pre-training on a human clinical dataset (SIDER) builds a robust "chemical intuition" that generalizes effectively to molecular toxicity prediction tasks (AUC 0.813), validating the use of the GINE residual architecture.

Furthermore, we bridge physical and digital gaps by combining EasyOCR with the cascading optimization pipeline and smart matching strategy, achieving 98% accuracy in drug identification from packaging. A key contributing factor is the smart matching logic; by tokenizing input text and applying fuzzy matching, the system effectively reduces discrepancies between raw OCR outputs and official database labels, demonstrating the effectiveness of hybrid AI systems in healthcare. While the proposed visual verification pipeline demonstrates strong performance on known drug packaging, it currently assumes access to a predefined drug database for brand-name matching. As a result, handling completely unseen, counterfeit, or non-standard drug labels remains a limitation of the present system and represents an important direction for future work.



(a) Receiver Operating Characteristic (ROC) curves for each of the 12 Tox21 assays.



(b) Precision–Recall (PR) curves for the 12 Tox21 assays.

Fig. 3: Performance evaluation curves for the proposed toxicity prediction model on the Tox21 benchmark.

References

1. Bongini, P., et al.: “A Deep Learning Approach to the Prediction of Drug Side-Effects on Molecular Graphs”. In: Proceedings of the 20th International Conference on Graph Modeling, pp. 45–58. Springer (2023)
2. M. Kuhn et al., “A side effect resource to capture phenotypic effects of drugs,” *Mol. Syst. Biol.*, vol. 6, 2010.
3. EasyOCR, `https://github.com/JaidedAI/EasyOCR`.
4. M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, “The SIDER database of drugs and side effects,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, Jan. 2016.
5. NIH, “Tox21 Data Challenge 2014,” `https://tripod.nih.gov/tox21/challenge/`.
6. P. Bongini, P. Faccioli, M. Ferri, and M. Bianchini, “Composite graph neural networks for molecular property prediction,” *Int. J. Mol. Sci.*, vol. 25, no. 12, p. 6583, 2024.
7. N. Pancino, P. Bongini, F. Scarselli, and M. Bianchini, “GNNkeras: A Keras–based library for graph neural networks and homogeneous and heterogeneous graph processing,” *SoftwareX*, vol. 18, p. 101061, 2022.