Technical Report: BioGuard v1.0

BioGuard: Deep Learning Inference for Non-Linear DDI Prediction via Symmetric Multi-Modal Featurization

Abstract

Traditional structural similarity metrics are insufficient for predicting complex pharmacokinetic drug-drug interactions (DDIs). This report evaluates BioGuard, a deep learning framework utilizing symmetric ECFP4 and biophysical feature encoding to classify interaction potential. On a strict pair-disjoint test set, BioGuard achieved a Precision-Recall AUC (PR-AUC) of 0.819, compared with logistic regression at 0.816 and random forest at 0.588. Tanimoto similarity performed near-random with PR-AUC 0.169. The observed lift over the strongest linear baseline was modest (ΔPR-AUC 0.0027, approximately 0.34%). These results support the architecture's capacity to model non-linear interactions, while indicating that additive feature signal remains a dominant contributor under the current feature set.

Methodology

Featurization and Encoding

Input compounds are vectorized into $\mathbb{R}^{2064}$ using a hybrid feature set:

- Structural Topology: 2048-bit Extended-Connectivity Fingerprints (ECFP4) with a Morgan radius of 2.
- Biophysical Properties: 5 normalized scalars (Molecular Weight, LogP, TPSA, H-Donors, H-Acceptors).
- Enzyme Profile: A placeholder 11-dimensional vector reserved for CYP450 data (zero-padded in v1.0 inference).

To satisfy permutational invariance, where Interaction (A,B) is equivalent to Interaction (B,A), a symmetric pair encoding strategy was implemented.

For input vectors vA and vB, the network input $X \in \mathbb{R}^{6192}$ is constructed via element-wise operations:

$$X=[(vA+vB)\oplus \left| vA-vB \right| \oplus(vA\cdot vB)]$$

Network Architecture

The inference engine is a PyTorch-based Multi-Layer Perceptron (MLP). The architecture reduces the 6192-dimensional input through dense layers of 512 and 256 units, respectively. Non-linearity is introduced via ReLU activations, stabilized by Batch Normalization (1D) and regularized via Dropout (p=0.3) to prevent overfitting. The output layer yields a raw logit, which is calibrated via Isotonic Regression to provide a probability score [0,1].

Benchmarking Results

Model performance was assessed using a pair-disjoint split to eliminate data leakage (predicting interactions for unseen drug combinations). Due to class imbalance, PR-AUC was utilized as the primary success metric.

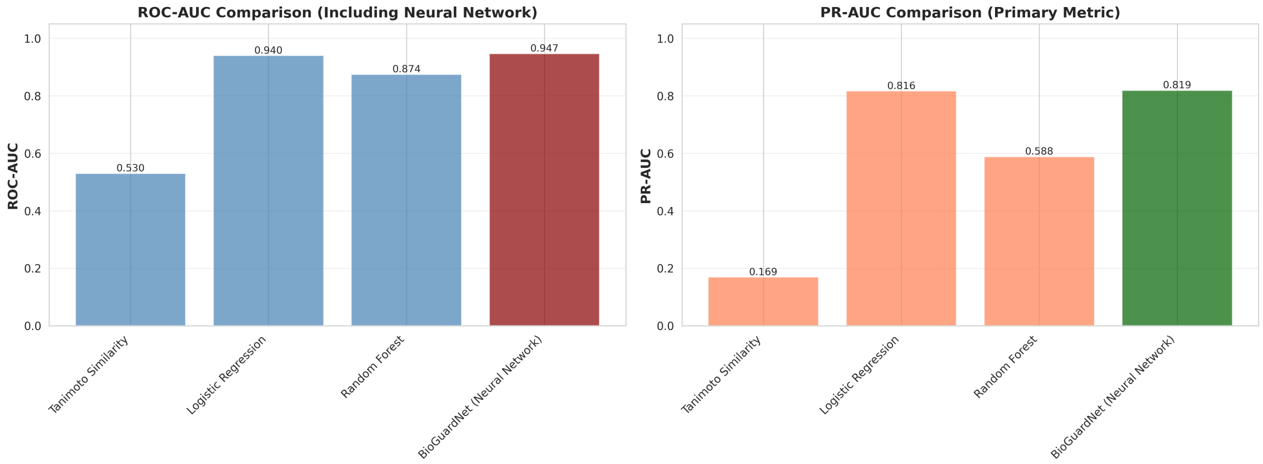

Figure 1. ROC-AUC and PR-AUC comparison across baselines and BioGuardNet on the pair-disjoint test split (n=83495).

Tanimoto Baseline (Structure-Only):

- PR-AUC: 0.169 | ROC-AUC: 0.530

Performance is near-random, confirming that structural homology alone is a poor predictor of pharmacokinetic interference.

Random Forest:

- PR-AUC: 0.588 | ROC-AUC: 0.874

Tree-based ensembles struggled with high-dimensional sparsity under the fingerprint-based representation.

Logistic Regression:

- PR-AUC: 0.816 | ROC-AUC: 0.940

The linear baseline performed strongly, indicating that additive features carry significant signal.

BioGuard (MLP):

- PR-AUC: 0.819 | ROC-AUC: 0.947
- F1 Score: 0.648

The MLP achieved a small improvement in PR-AUC over the linear baseline, demonstrating incremental gains from modeling non-linear dependencies in the interaction between structural motifs and biophysical properties.

Deployment

The system is deployed as a Python-based microservice. The inference backend is managed by FastAPI, utilizing a ProcessPoolExecutor to offload CPU-intensive RDKit featurization tasks and ensure thread safety for the PyTorch model. The end-user interface is rendered via Streamlit, allowing computational chemists to query candidate pairs and visualize risk probabilities in real-time.

Discussion and Limitations

While BioGuard v1.0 exhibits strong predictive power, limitations remain regarding mechanistic interpretability. The model currently relies on 2D structural proxies; the explicit metabolic features (CYP450 inhibition and induction) are currently zero-padded due to data sparsity in the screening phase. Consequently, the model predicts probability of interaction without identifying the specific metabolic pathway. Furthermore, while calibration improved reliability, high confidence scores in chemically sparse regions of the training manifold should be treated with caution. The modest performance delta over logistic regression suggests that further gains may require richer mechanistic features, improved label fidelity, or additional biological modalities. Future development will focus on integrating transcriptomic signatures to improve biological context.