Technical Dossier: BioGuard v2.2

Role: OOD Drug-Drug Interaction (DDI) Inference Engine
Status: Dockerized / Production-Ready (<200ms Latency)

1. Executive Summary: The Generalization Gap

Current SOTA baselines (e.g., DeepDDS, CASTER, FG-DDI) report ROC-AUCs >0.90 by utilizing random, or ID-disjoint data splits. This introduces structural leakage, where models memorize training scaffolds rather than learning interaction mechanisms.

BioGuard v2.2 is designed to audit and correct this failure mode. By enforcing strict **Bemis-Murcko Scaffold-Disjoint** evaluation and injecting CYP450 Metabolic Priors, BioGuard recovers signal in the Out-of-Distribution (OOD) regime that pure structural models miss.

---

2. Hybrid Architecture

The system employs a Dual-Track Consensus strategy to balance high Precision with high Recall.

Track A: LightGBM

- **Input:** 1024-bit ECFP4 Fingerprints + 30-dim Explicit CYP Vector (ChEMBL).
- **Logic:** Gradient-based One-Side Sampling (GOSS).
- **Role:** High-fidelity filtration of Easy Negatives based on rigid structural alerts.
- **Performance:** Achieves SOTA Precision (**0.43 PR-AUC**) on the Scaffold split.

Track B: BioGuard GATv2

- **Input:** Molecular Graph + Metabolic Node Features.
- **Pre-Training: Self-Supervised Learning (SSL)** via Masked Atom Prediction (Acc: 81%) to initialize weights with chemical valency intuition before DDI fine-tuning.
- **Architecture:** Siamese GATv2 with Multi-Head Attention ($k = 4$) employing specific fusion and readout strategies:
- **Metabolic Fusion:** The explicit enzyme vector (E) is concatenated to the graph embedding (G) to form the arm vector:

$$v_{arm} = [G \oplus E]$$

- **Symmetric Interaction Head:** Ensures permutation invariance ($f(A, B) = f(B, A)$) via:

$$X = [(v_A + v_B) \oplus |v_A - v_B| \oplus (v_A \cdot v_B)]$$

- **Role: Maximal Recall (0.70).** Propagates metabolic risk through the graph topology even when the scaffold is novel.

---

3. OOD Benchmarks (Scaffold-Disjoint)

Evaluation performed on a held-out test set of entirely unseen molecular scaffolds to simulate NCE discovery.

Model	Architecture	ROC-AUC	PR-AUC	Recall (Sensitivity)	Verdict
LightGBM (Hybrid)	Gradient Boosting	**0.71**	**0.43**	0.66	**Precision Baseline.** Best for active screening.
BioGuard GATv2	Pre-trained GNN	0.66	0.39	**0.70**	**Safety Filter.** Catches 70% of toxic events in OOD space.
Naive GAT	Graph Neural Net	0.64	0.37	0.27	*Failed Control (Lacks Biological Context).*

Analysis:
The Hybrid Ensemble provides a **50% Signal Lift** over the random screening baseline (0.28). While Trees dominate Precision on known substructures, the GATv2 is essential for catching "Structural Outliers"—compounds that look safe structurally but carry metabolic liability.

---

4. Engineering & Deployment

- **Containerization:** Fully Dockerized (`bioguard_app:latest`) for zero-config reproducibility.
- **Inference Latency:** <200ms per pair (CPU-optimized for high-throughput screening).
- **Data Pipeline:** Custom ETL to map metabolic hubs (CYP3A4, 2D6, 2C9, 2C19, 1A2) from ChEMBL to graph node features.

5. Limitations & Roadmap

1. **ChEMBL Sparsity:** NCEs with no known metabolic assay data are treated as zero-vectors (degraded mode).
    - *Fix:* Integration of an upstream CYP-prediction module to impute missing priors.
2. **Latent Leakage:** Murcko scaffolds are a proxy for disjointness, but UMAP analysis suggests latent overlap may persist.
    - *Fix:* Roadmap includes UMAP-based cluster splitting for V3.0.