Hepa-RAFT: Retrieval-Augmented Virtual Hepatocyte Responses for Hepatotoxicity Prediction

Anonymous Authors[1]

Abstract

Generative and agentic systems for therapeutic design need inexpensive biological feedback before prioritizing compounds for synthesis or wet-lab testing. Drug-induced liver injury (DILI) is a high-impact case: structure-only predictors capture chemical liabilities but miss cellular response, whereas measured transcriptomic assays provide mechanism but are too costly for early-stage screening. We present Hepa-RAFT, a Hepatocyte Retrieval-Augmented Framework for Toxicogenomics that provides a structure-conditioned virtual hepatocyte-response readout: it maps a query SMILES to predicted drug-induced hepatocyte expression and uses that model-derived state as an intermediate retrieval key. The framework has two task-adaptive heads sharing one FiLM-conditioned expression predictor: an expression head corrects predicted profiles by retrieving training drugs with similar predicted perturbation, and a DILI head performs similarity-weighted voting over a small labeled reference set using chemical, predicted-transcriptomic, pharmacokinetic, and drug–target channels. This yields a biological design principle: retrieval should be matched to the phenotype being queried, with perturbation correction favoring a focused predicted-response geometry and DILI liability favoring broader evidence aggregation. Across held-out scaffold splits, the end-to-end retrieval-augmented Hepa-RAFT pipeline reaches Pearson correlation 0.631 on the top 20 perturbation-responsive genes; on external DILImap and DILIst benchmarks, Hepa-RAFT reaches AUROC 0.713 and 0.637 without retraining. Because query-drug expression and labels are never accessed, Hepa-RAFT is best interpreted as a lightweight, biologically grounded virtual-response module for hepatocyte-intrinsic DILI liability, with training-overlap-controlled external comparisons and overlap audits reported for available competitor training sets.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

1. Introduction

Drug-induced liver injury (DILI) accounts for more than half of acute liver failure cases in the United States—primarily acetaminophen overdose (∼39%) together with idiosyncratic drug reactions (∼13%) (Ostapowicz et al., 2002)—and was the single most common cause of post-marketing drug withdrawal in a systematic review of 462 products (hepatotoxicity: 18%) (Onakpoya et al., 2016). Identifying DILI liability early in drug development could prevent costly late-stage failures and expensive safety studies (Watkins, 2011), yet reliable prediction remains an open challenge.

Current computational DILI predictors fall into three categories. Structure-only methods such as GATNN (Wibowo et al., 2025) and classical fingerprint baselines capture chemical liability signals but miss the cellular response that makes liver injury a biological phenotype rather than a purely structural property. Proxy-augmented methods like DILIPredictor (Seal et al., 2024) combine chemical descriptors with predicted in vitro and in vivo proxy DILI scores; such methods can be highly effective but rely on upstream proxy models and assay-derived resources. Expression-based methods leverage measured transcriptomic profiles to capture cellular response but require new gene-expression experiments for query compounds, limiting their use inside early generative-design loops.

This leaves a gap central to generative biology: can a model provide a queryable, structure-derived esti-

mate of hepatocyte response, and can that predicted response guide drug-safety decisions without measured query assays? We address this by predicting drug-induced hepatocyte expression computationally and using the predicted biological state as a retrieval key over a small labeled reference set; the response is not a new wet-lab measurement, but a transparent intermediate hypothesis used to organize retrieval. The resulting system is generative-compatible rather than a molecule generator itself: it can be placed downstream of any generator that proposes valid SMILES and can return a hepatocyte-response estimate before synthesis. Unlike endpoint-only toxicity classifiers, Hepa-RAFT exposes an intermediate predicted hepatocyte response, allowing candidate molecules to be ranked not only by a scalar liability score but also by the transcriptomic state supporting that score.

Our retrieval-first design follows work on retrieval-augmented and few-shot drug discovery. Schimunek et al. (Schimunek et al., 2023) (MHNfs, ICLR 2023) showed that context-enriched molecular representations improve few-shot drug-discovery tasks. We operate in an even smaller DILI-label regime (n=165), so we keep the learned component focused on hepatocyte-response prediction and use non-parametric retrieval for the small-label downstream tasks.

We introduce Hepa-RAFT (Hepatocyte Retrieval-Augmented Framework for Toxicogenomics), a task-adaptive retrieval framework with two heads sharing a single FiLM-decoder backbone:

1. Virtual hepatocyte-response head. A FiLM-conditioned neural network maps SMILES inputs to drug-induced expression profiles over 5,000 hepatocyte genes. The profile is refined by similarity-weighted residuals from training drugs whose predicted perturbation is most similar to the query's. Ablation shows that this task is best served by predicted-delta similarity alone.

2. DILI-liability head. DILI liability is scored by similarity-weighted voting over the same labeled reference set. Here retrieval combines four evidence channels—chemical substructure (Morgan), predicted transcriptomic state, pharmacokinetic properties (ADMET-AI), and drug–target profiles (DGIdb)—because hepatotoxicity arises from multiple partly independent biological mechanisms (Aleo et al., 2014; Uetrecht, 2019).

In summary, our contributions are: (i) a SMILES-queryable, structure-conditioned hepatocyte-response readout producing 5,000-gene predicted expression profiles; (ii) retrieval correction that improves held-out perturbation prediction without adding learned parameters; (iii) a task-adaptive retrieval design showing that expression prediction and DILI liability prefer different biological similarity geometries; and (iv) external DILI evaluation with explicit training-overlap audits.

For external comparisons, we report training-overlap-controlled subsets whenever competitor training lists are available and separately audit benchmark-level and feature-level overlap for our own pretraining corpus (Appendix A). This audit is motivated by standardized molecular benchmarking practice (Wu et al., 2018; Huang et al., 2022) while avoiding a stronger claim than the available training-set disclosures support.

2. Method

2.1. Framework Overview

Hepa-RAFT consists of four functional blocks (Figure 1): (i) a frozen pre-trained chemical-language drug encoder (MoLFormer; Section 2.2), (ii) a FiLM-conditioned expression decoder shared by both retrieval heads (Section 2.2), (iii) a fixed reference set of n=165 labeled training drugs against which retrieval is performed (Section 2.3), and (iv) two retrieval heads with task-specific similarity functions: an expression-correction head (Section 2.4, single-channel retrieval on predicted-delta similarity) and a DILI-liability retrieval head (Section 2.5, multi-channel retrieval combining four similarity measures).

The architecture is designed for unseen-drug inference. Operationally, the required query input is a SMILES string: the drug encoder maps SMILES to a fixed-dimensional chemical representation, the FiLM decoder produces a structure-derived predicted hepatocyte expression profile, and the retrieval heads aggregate similarity-weighted information from the reference set without ever accessing the query drug's measured expression or DILI label. The DILI-liability retrieval head can additionally use target-channel evidence when a DGIdb match is resolved; otherwise it falls back to the available SMILES-derived and predicted-expression channels. Both heads are non-parametric in the retrieval step (no per-drug training). DILI-label supervision and the retrieval memory are restricted to the 165-drug benchmark, while the FiLM expression decoder is trained on the hepatocyte-expression corpus described in Section 3 and Appendix A; consequently both heads are directly applicable to new drugs without retraining or wet-lab measurements of the query.

2.2. Drug Encoder and FiLM-Conditioned Expression Decoder

Drug encoder (MoLFormer). For input drug q with SMILES string s_q, we obtain a chemical embedding

$$\mathbf{d}_q = f_{\text{MoLFormer}}(s_q) \in \mathbb{R}^{768} \tag{1}$$

using the publicly released MoLFormer (Ross et al., 2022) chemical-language model, kept frozen throughout. MoLFormer was pre-trained on ∼1B SMILES via masked language modeling, providing a chemically meaningful representation that generalizes to drugs outside the 165-drug reference set: any new SMILES with valid tokens yields a usable embedding without further training. Auxiliary inputs to the decoder are the cell type $\mathbf{c} \in \{0, 1\}$ (PHH vs. HepaRG) and the dose $\delta \in \mathbb{R}^+$, fixed to the predefined inference condition (HepaRG, 5 μM) for all retrieval operations (Section 3).

FiLM expression decoder. The decoder predicts a drug's hepatocyte expression response from $(\mathbf{d}_q, \mathbf{c}, \delta)$ via Feature-wise Linear Modulation (FiLM) (Perez et al., 2018). The drug embedding produces per-gene conditioning parameters

$$(\boldsymbol{\gamma}_g, \boldsymbol{\beta}_g) = f_{\text{cond}}\big(\mathbf{d}_q, \mathbf{c}, \delta;\, g\big), \qquad g = 1, \ldots, 5{,}000, \tag{2}$$

which modulate gene-level hidden states $\mathbf{h}_g$:

$$\mathbf{h}'_g = \boldsymbol{\gamma}_g \odot \mathbf{h}_g + \boldsymbol{\beta}_g. \tag{3}$$

A heteroscedastic Gaussian NLL output head (Nix & Weigend, 1994; Kendall & Gal, 2017) produces a per-gene mean $\hat{\boldsymbol{\mu}}_q \in \mathbb{R}^{5000}$ and log-variance $\log \hat{\boldsymbol{\sigma}}_q^2$. The per-gene log-variance branch is intended to down-weight noisy genes during training under the small-data, high-dimensional regime (n=165 drugs × 5,000 genes). We denote the mean output for drug q at the predefined HepaRG, 5 μM inference condition as

$$\hat{\mathbf{y}}_q^{\text{pred}} \;\equiv\; \hat{\boldsymbol{\mu}}_q \;\in\; \mathbb{R}^{5000}, \tag{4}$$

and at DMSO control as $\hat{\mathbf{y}}_{\text{DMSO}}^{\text{pred}}$, obtained by passing the same vehicle-control representation used during training through the decoder under the same cell-type condition. Implementation hyperparameters appear in Appendix B.

Trained once, applies to any SMILES. The encoder is frozen (MoLFormer pre-trained weights, used as-is). The FiLM decoder is trained once on TG-GATEs and ToxCast HTTr hepatocyte profiles (Section 3) and then frozen for all retrieval operations downstream. Inference for a new query drug requires no per-drug fine-tuning, no measured expression of the query, and no query label. The required input is a SMILES string; optional target evidence is used only when the query can be resolved in DGIdb, with a predefined fallback when it cannot. This property permits both predicted-expression correction and predicted DILI-liability scoring to operate as a computational screening module.

2.3. Reference Database

Both heads retrieve against the same fixed database of n=165 labeled training drugs. For each training drug i, the database stores: (i) the observed hepatocyte expression profile $\mathbf{y}_i^{\text{real}} \in \mathbb{R}^{5000}$ (TG-GATEs and ToxCast HTTr; Section 3); (ii) the binary DILI label $\text{DILI}_i \in \{0, 1\}$; (iii) precomputed similarity features used by the DILI-liability retrieval head—Morgan fingerprint $\mathbf{m}_i \in \{0, 1\}^{2048}$, ADMET-AI vector $\mathbf{a}_i \in \mathbb{R}^{16}$, and binary DGIdb target profile $\mathbf{t}_i \in \{0, 1\}^{1307}$; and (iv) the raw FiLM-decoder output $\hat{\mathbf{y}}_i^{\text{pred}}$ from the frozen checkpoint, used by both heads. External query drugs are never inserted into the database, and their measured expression and labels are never consulted at inference. The two heads define their own similarity functions over this database (Sections 2.4 and 2.5).

2.4. Expression Prediction via Predicted-Delta Retrieval Correction

The expression-correction head refines the FiLM decoder's raw expression prediction using similarity-weighted residuals from the k=7 training drugs whose predicted perturbation is most similar to the query's. Retrieval uses a single similarity channel: the Pearson correlation between predicted deltas. Among the channels available, this is the one that directly answers the task-relevant question—which training drugs does the model think behave most similarly to this query? Adding chemical, pharmacokinetic, or target-based channels to this retrieval step did not improve performance and slightly degraded it in our ablation (Section 4.3), so the expression-correction head is single-channel by design.

Predicted delta. The predicted delta for query drug q,

$$\hat{\boldsymbol{\delta}}_q \;=\; \hat{\mathbf{y}}_q^{\text{pred}} - \hat{\mathbf{y}}_{\text{DMSO}}^{\text{pred}}, \tag{5}$$

isolates the drug-specific perturbation signal from the baseline cell state. It is a pure function of the trained decoder and uses no observed expression measurement.

Retrieval similarity. The expression-correction similarity is the Pearson correlation between predicted deltas:

$$S^{(1)}(q, i) \;=\; \text{Pearson}(\hat{\boldsymbol{\delta}}_q, \hat{\boldsymbol{\delta}}_i). \tag{6}$$

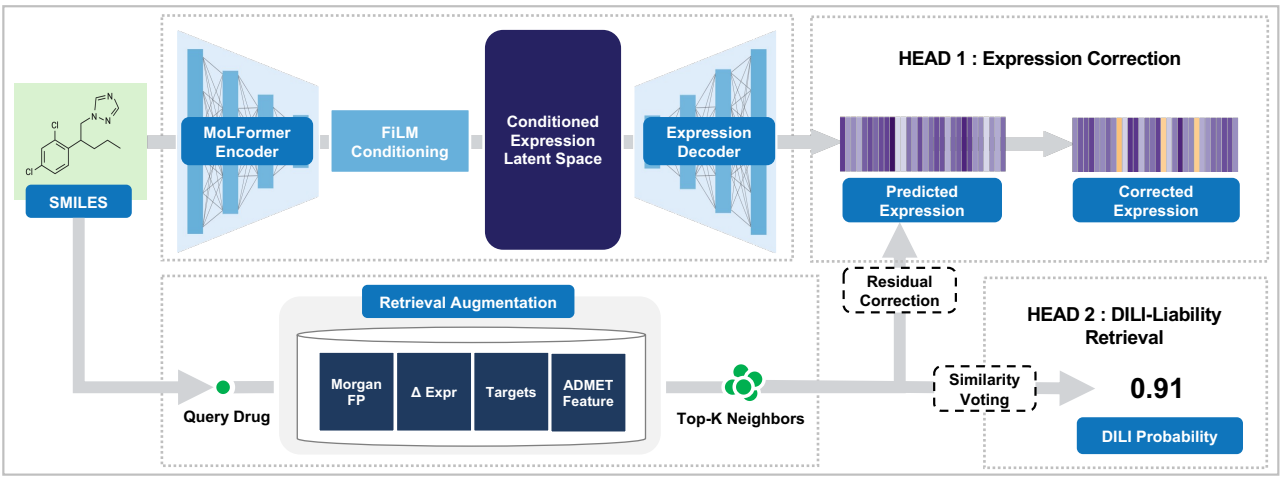

Figure 1. Hepa-RAFT pipeline. A query SMILES is encoded by a frozen MoLFormer drug encoder and passed through the FiLM expression decoder to produce raw predicted hepatocyte expression. Retrieval augmentation then uses a fixed training-drug reference memory to identify top-k neighbors. The expression-correction head retrieves neighbor residuals to produce corrected expression (Section 2.4), whereas the DILI-liability retrieval head performs similarity-weighted voting over retrieved labels to estimate DILI probability (Section 2.5). Both heads operate at inference without measured query expression or query labels.

This is computed entirely in predicted space: both vectors come from the FiLM decoder, and the similarity computation never uses measured expression of the query or its neighbors at retrieval time.

Retrieval correction. Let $\mathcal{N}_k^{(1)}(q)$ denote the k=7 training drugs with the largest similarity $S^{(1)}(q,\cdot)$. For each retrieved neighbor i we form the training-side residual $\mathbf{r}_i = \mathbf{y}_i^{\text{real}} - \hat{\mathbf{y}}_i^{\text{pred}}$, and aggregate:

$$\hat{\mathbf{y}}_q^{\text{corrected}} = \hat{\mathbf{y}}_q^{\text{pred}} + \sum_{i \in \mathcal{N}_k^{(1)}(q)} w_i \, \mathbf{r}_i, \tag{7}$$

$$w_i = \frac{S^{(1)}(q,i)}{\sum_{j \in \mathcal{N}_k^{(1)}(q)} S^{(1)}(q,j)}. \tag{8}$$

Similarity weighting lets close neighbors dominate the correction, so the aggregate remains meaningful when only a few training drugs share the query's mechanism. Observed expression $\mathbf{y}_i^{\text{real}}$ enters only through training-side residuals; the query drug's measured expression is never used at retrieval or aggregation. The expression-correction head therefore generalizes to any new drug for which only a SMILES string is available. Empirically, this step lifts $\text{Pearson}_{\Delta\text{DE20}}$ from 0.482 (FiLM decoder alone) to 0.631 at k=7 under multi-seed 5-fold scaffold CV (+0.149 absolute, +30.9% relative), adding no learned parameters over the backbone.

2.5. DILI Classification via Multi-Channel Retrieval Voting

The DILI-liability retrieval head produces a liability score for a new drug by similarity-weighted soft voting over the labels of the k=10 most similar training drugs. Unlike expression correction, estimating DILI liability from molecular structure does not have a single similarity channel that captures the relevant evidence. DILI liability arises from multiple, partly independent mechanisms: chemical reactivity (e.g., reactive metabolite formation), transcriptomic perturbation (mitochondrial stress, oxidative damage), pharmacokinetic and exposure-related properties, and drug–target interactions such as BSEP inhibition and mitochondrial dysfunction-related targets (Aleo et al., 2014; Uetrecht, 2019). The DILI-liability retrieval head therefore combines four similarity channels that span this evidence space. In internal multi-seed CV, adding predicted expression to Morgan similarity gives the largest gain, with smaller additional gains from ADMET and target evidence (Section 4.3); external behavior remains mechanism-dependent (Section 5).

Similarity channels.

- M (Morgan). Tanimoto similarity on 2,048-bit Morgan fingerprints (Rogers & Hahn, 2010), capturing chemical substructure: $S_M(q,i) = |\mathbf{m}_q \cap \mathbf{m}_i| / |\mathbf{m}_q \cup \mathbf{m}_i|$.

- P (Predicted expression). Pearson correlation between raw predicted expression profiles: $S_P(q,i) = \text{Pearson}(\hat{\mathbf{y}}_q^{\text{pred}}, \hat{\mathbf{y}}_i^{\text{pred}})$. Note that DILI-liability voting uses the raw decoder output rather than the predicted delta of Equation (5): for DILI voting the absolute predicted profile (which encodes both baseline cell state and perturbation) is more discriminative than perturbation direction alone.
- A (ADMET). Cosine similarity on our curated 16-dimensional subset of ADMET-AI (Swanson et al., 2024) predictions, capturing ADMET/property evidence.
- T (Target). Jaccard similarity on binary drug–target profiles from our processed DGIdb (Freshour et al., 2021) snapshot (1,307 targets; 158/165 reference drugs covered).

For a query drug q, $\mathbf{m}_q$ is computed via RDKit, $\mathbf{a}_q$ via ADMET-AI, $\mathbf{t}_q$ from DGIdb when available (drugs lacking a match fall back to a T-free combination), and $\hat{\mathbf{y}}_q^{\text{pred}}$ via the FiLM decoder of Section 2.2.

Channel combination. The combined similarity is the mean over channels available for both drugs:

$$S^{(2)}(q,i) \;=\; \frac{1}{|C_{qi}|} \sum_{c \in \{M,P,A,T\} \cap C_{qi}} S_c(q,i), \qquad (9)$$

where $C_{qi} \subseteq \{M,P,A,T\}$ excludes channels for which either drug lacks the corresponding feature.

Voting. Let $\mathcal{N}_k^{(2)}(q)$ denote the top-k=10 training drugs under $S^{(2)}(q,\cdot)$, with similarities clipped to $[0,\infty)$. The DILI probability for query drug q is

$$\begin{aligned} P(\text{DILI}^+ \mid q) &= \sum_{i \in \mathcal{N}_k^{(2)}(q)} w_i \mathbb{1}[\text{DILI}_i = 1], \\ w_i &= \frac{S^{(2)}(q,i)}{\sum_{j \in \mathcal{N}_k^{(2)}(q)} S^{(2)}(q,j)}. \end{aligned} \qquad (10)$$

When $\sum_j S^{(2)}(q,j) = 0$ we fall back to the training-set DILI prevalence. Weighted voting lets close neighbors dominate the prediction while still drawing on the full k-neighbor context. The training-drug labels DILI_i are the only label information consumed; the query drug's label is never accessed at inference. The DILI-liability retrieval head therefore returns a score for any new drug with a valid SMILES and the available optional side information, just as the expression-correction head returns a predicted response. Because both heads are retrieval-based, each prediction can be accompanied by nearest reference drugs, similarity scores, and active evidence channels, giving users a direct audit trail for whether a liability score is driven by chemical similarity, predicted transcriptomic similarity, ADMET resemblance, target overlap, or their combination.

3. Experimental Setup

Training data and platform. We train the FiLM expression decoder on hepatocyte gene expression profiles from Open TG-GATEs (Igarashi et al., 2015) (primary human hepatocytes; Affymetrix profiles, RMA-normalized in our preprocessing) and EPA ToxCast high-throughput transcriptomics (HTTr) profiles (Richard et al., 2016; Harrill et al., 2021) (including HepaRG TempO-Seq targeted transcriptomic profiles, $\log_2$-CPM in our preprocessing). The full decoder pretraining corpus contains 1,284 unique InChIKeys (feature-level external overlap audited in Appendix A). The labeled retrieval benchmark is the subset with hepatocyte DEG measurements and binary DILI labels derived from a consensus across public DILI resources including DILIrank (Chen et al., 2016), LiverTox (Hoofnagle et al., 2013), DILIst (Thakkar et al., 2020), and DILImap (Bergen et al., 2025). This benchmark contains 165 drugs with 71.5% DILI^+ prevalence; we therefore report AUPRC lift relative to prevalence alongside AUROC and AUPRC. The gene vocabulary spans 5,000 genes selected from control-condition expression variance on the probe/transcript intersection of the two platforms.

External inference conditions. For every external query drug we run a single forward pass through the shared FiLM decoder at the predefined inference condition: cell type = HepaRG and dose = 5 μM. These conditions are fixed a priori as the standardized virtual-response condition used throughout the harmonized hepatocyte-response benchmark, not as a patient exposure estimate, and are applied uniformly across DILImap and DILIst inference.

Evaluation protocol. Internal evaluation uses repeated 5-fold Murcko-scaffold cross-validation (Bemis & Murcko, 1996) with out-of-fold expression predictions, so every held-out drug's predicted response comes from a decoder checkpoint that did not see it during training. During internal DILI evaluation, the retrieval memory is restricted to scaffold-training drugs for each held-out query, so the query drug and held-out test drugs cannot be retrieved. Main-result ± bars report between-split standard deviations across 5 split seeds; detailed

fold statistics are reported in Appendix B. For the expression-correction head, we report Pearson and Spearman correlation on predicted vs. true expression deltas for the top 20 perturbation-responsive genes per drug (Pearson/Spearman$_{\Delta\text{DE20}}$), direction accuracy (Dir$_{\text{DE20}}$), and whole-transcriptome correlations (Pearson/Spearman$_{\Delta\text{all}}$). For the DILI-liability retrieval head, we report AUROC, AUPRC, and AUPRC lift (AUPRC minus DILI prevalence)—a prevalence-aware ranking metric relevant under the heterogeneous prevalences of internal (71.5%) and external (26.3–59.1%) sets. We select k using internal CV only, fixing k=7 for expression correction and k=10 for DILI liability before external evaluation.

External validation datasets. We evaluate on two external datasets: a DILImap-derived compound-label set from a primary human hepatocyte SMART-Seq toxicogenomics resource (Bergen et al., 2025) (n=893 after deduplication, 235 DILI$^+$, prevalence 26.3%) and a DILIst-derived compound-label set (Thakkar et al., 2020) (n=1,077 after standardization, 636 DILI$^+$, prevalence 59.1%). After InChIKey deduplication against our 165-drug benchmark, both external sets have 0% benchmark-level drug overlap. DGIdb target profiles are used where available; drugs without DGIdb matches fall back to the $\{M, P, A\}$ channel set per Equation (9). Because external evaluation uses a single frozen checkpoint and fixed retrieval configuration, Table 2 reports point estimates; a fold-checkpoint sensitivity analysis is summarized in Appendix B.

Baselines and external overlap control. We compare against DILIPredictor (Seal et al., 2024), ADMET-AI (Swanson et al., 2024), GATNN (Wibowo et al., 2025), and a Morgan fingerprint Random Forest baseline trained on the 165-drug benchmark. For comparators with disclosed training-set drug lists, we evaluate only external drugs absent from the corresponding training set; subset sizes are shown in Table 2. Morgan FP RF and Hepa-RAFT are evaluated on the full external benchmarks for label-level evaluation because their DILI supervision is restricted to the 165-drug benchmark. Hepa-RAFT's FiLM-decoder feature-level pretraining overlap is audited separately in Appendix A, together with all training-set provenance.

4. Results

4.1. Expression Prediction

Table 1 shows that retrieval-augmented correction substantially improves perturbation prediction. Under repeated 5-fold scaffold cross-validation, predicted-delta retrieval correction (k=7) achieves Pearson$_{\Delta\text{DE20}}$ of 0.631±0.033, a 30.9% improvement over the raw FiLM decoder (0.482) and 87% over the fold-wise training-set mean baseline (0.338). The expression-correction head produces a virtual hepatocyte response for new drugs, while the DILI retrieval head uses the shared decoder's raw predicted expression as one of its retrieval channels. Direction accuracy reaches 0.829, correctly predicting up- or down-regulation for 82.9% of the most perturbation-responsive genes per drug. The gain is not confined to DE20: retrieval-corrected expression remains stable on broader DE50 and DE100 panels (Table 5).

4.2. DILI Classification

Internal performance. With the four-channel Morgan + predicted-expression + ADMET-AI + target-profile setting at k=10, Hepa-RAFT reports internal AUROC 0.779±0.022 and AUPRC 0.876±0.014 under multi-seed 5-fold scaffold cross-validation (5 split seeds; prevalence 71.5%; AUPRC lift +0.161 over prevalence). For reference, the Morgan fingerprint Random Forest baseline reports 0.523±0.080 / 0.711±0.048 (lift −0.004) under the same protocol.

External DILI classification. Table 2 presents external DILI classification performance for Hepa-RAFT and the comparator methods. Hepa-RAFT reports DILImap AUROC 0.713 (AUPRC 0.439, lift +0.176 over the 26.3% prevalence) and DILIst AUROC 0.637 (AUPRC 0.697, lift +0.106 over the 59.1% prevalence) without measured query expression or query labels at inference. Viewed against disclosed DILI-label supervision (Figure 2), and recognizing that comparator methods are evaluated on different overlap-filtered subsets, Hepa-RAFT uses a 165-drug DILI-label retrieval memory yet remains competitive on DILImap with larger supervised DILI models after overlap filtering (0.713 AUROC vs. 0.740 for DILIPredictor and 0.684 for GATNN). On DILIst, larger supervised DILI models perform better, highlighting the complementary regime targeted by Hepa-RAFT: small-label, virtual-response-based retrieval rather than label-rich endpoint classification.

Our own pretrain corpus. We audit feature-level overlap between the FiLM-decoder pretraining corpus and the external benchmarks in Appendix A. This overlap is distinct from label-level evaluation: the DILI-liability retrieval head retrieves only over the 165-drug benchmark and uses only the training-side labels of those 165 drugs at external inference; external drug labels are never consulted by the voting step.

Table 1. Expression-correction head performance on held-out drugs. All rows use the same multi-seed 5-fold scaffold CV regime (5 split seeds × 5 folds = 25 expression-decoder retrainings; ± = between-split std, except the fold-invariant Train mean row which carries eval-seed std). Pear/Spear$_{\Delta DE20}$: Pearson/Spearman correlation of predicted vs. true expression delta on top 20 perturbation-responsive genes; Dir: fraction with correct up/down direction; Pear/Spear$_{\Delta all}$: whole-transcriptome correlations on delta.

Method	Pear$_{\Delta DE20}$	Spear$_{\Delta DE20}$	Dir$_{DE20}$	Pear$_{\Delta all}$	Spear$_{\Delta all}$
Fold train mean expr.	0.338±0.044	0.269±0.043	0.539±0.064	0.132±0.026	0.115±0.019
FiLM decoder only	0.482±0.076	0.393±0.051	0.775±0.042	0.248±0.095	0.257±0.082
Hepa-RAFT P_Δ	0.631±0.033	0.501±0.024	0.829±0.045	0.388±0.024	0.365±0.031

Table 2. External DILI classification with per-method training-overlap control. Each method is evaluated on the subset of external drugs not present in its training-set disclosure; subset size n is reported per method. Morgan FP RF is trained only on the 165-drug benchmark. Hepa-RAFT's DILI-label supervision and retrieval memory are restricted to the same 165-drug benchmark; its FiLM-decoder pretraining overlap is audited separately in Appendix A. These two methods are evaluated on the entire external benchmark where valid predictions are available. DILImap prevalence 26.3%, DILIst prevalence 59.1%. AUPRC lift is AUPRC minus the prevalence on the corresponding evaluation set.

	DILImap			DILIst		
Method	n	AUROC	AUPRC	n	AUROC	AUPRC
Morgan FP RF (165 train)	893	0.602	0.340	1,077	0.612	0.656
ADMET-AI	768	0.560	0.258	745	0.574	0.660
GATNN	276	0.684	0.296	539	0.792	0.755
DILIPredictor	684	0.740	0.471	357	0.824	0.845
Hepa-RAFT (ours)	893	0.713	0.439	1,077	0.637	0.697
AUPRC lift vs. prevalence		+0.176			+0.106	

The external comparison is contextualized by disclosed DILI-label training size in Figure 2.

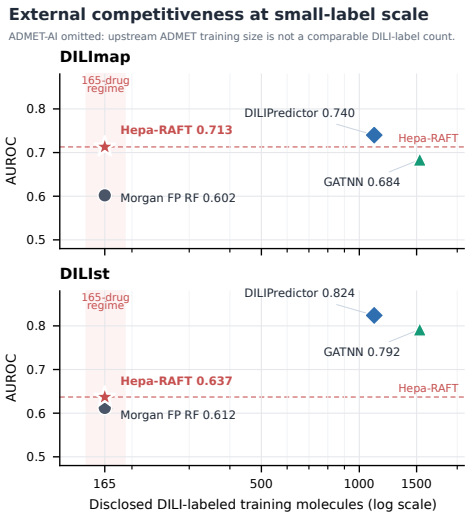

Figure 2. External AUROC by disclosed DILI-labeled training-set size. Dashed red lines mark Hepa-RAFT's AUROC; the x-axis counts DILI-label supervision only. Comparator disclosure details and overlap audits are reported in Table 2 and Appendix A.

4.3. Channel Ablation: Task Complexity Asymmetry

Expression correction: a single channel suffices. Across all 15 non-empty subsets of the four retrieval channels tested at k=7, the maximum Pearson$_{\Delta DE20}$ is achieved by predicted-delta similarity alone (P_Δ; 0.631±0.033). Every other subset is at or below this value. Adding chemical structure (M) is approximately neutral; adding ADMET-AI properties (A) costs −0.015; adding drug–target profiles (T) costs −0.029. Removing predicted-expression similarity entirely drops performance below the FiLM decoder's own baseline of 0.482, indicating that the retrieval correction's signal lives almost entirely in the predicted-perturbation channel. The expression-correction head therefore uses P_Δ (Section 2.4): a single channel that directly answers the task-relevant question.

DILI classification: internal multi-channel gains. For DILI classification under internal CV, predicted expression alone reaches AUROC 0.726, already close to M+P (0.730) and substantially above M-only retrieval (0.555). Adding ADMET-AI properties and drug–target profiles increases AUROC to 0.779 at k=10. The four channels capture complementary evidence sources (chemistry, transcriptomic stress, ADMET/property evidence, target evidence), so the DILI-liability retrieval head uses $\{M, P_{\text{raw}}, A, T\}$ (Section 2.5).

Task-dependent retrieval geometry. The expression-correction task is concentrated: predicting which genes a drug perturbs and by how much. The model's own predicted-delta vector is the natural similarity for retrieving neighbors with correlated systematic errors, and other channels add noise to that specific objective. The DILI-liability task is composite: hepatotoxicity depends on multiple, partly independent mechanisms, so retrieval over a single channel leaves signal on the table even when predicted expression is the strongest individual channel.

Table 3. Per-head channel ablation under multi-seed 5-fold scaffold CV (5 split seeds × 10 eval seeds, n=165, prevalence 71.5%; ± = between-split std). The expression-correction head (left columns) reports $\text{Pearson}_{\Delta\text{DE20}}$ at k=7; the selected predicted-delta row (P) is bolded. The DILI-liability retrieval head (right columns) reports DILI AUROC at k=10; the selected four-channel row is bolded. Channels: M = Morgan fingerprint, P = predicted-expression similarity (predicted-delta similarity for expression correction; raw predicted-expression similarity for DILI liability), A = ADMET-AI properties, T = DGIdb target profiles. Δ columns are differences vs. each head's selected row. "Decoder" is the FiLM decoder evaluated without retrieval.

	Expression correction		DILI classification	
Channels	$\text{Pear}_{\Delta\text{DE20}}$	Δ	AUROC	Δ
Decoder only	0.482±0.076	−0.149	—	—
M only	0.441±0.075	−0.190	0.555±0.000	−0.225
P only	0.631±0.033	—	0.726±0.023	−0.053
M+P	0.631±0.039	−0.001	0.730±0.015	−0.049
M+P+A	0.616±0.049	−0.015	0.754±0.032	−0.025
M+P+A+T	0.602±0.050	−0.029	0.779±0.022	—

5. Discussion

Hepa-RAFT reframes DILI prediction as retrieval over a virtual biological response rather than as a purely discriminative small-label classifier. The main design lesson is task asymmetry: expression correction is best performed in predicted perturbation space, whereas DILI liability benefits from combining predicted transcriptomics with chemical, pharmacokinetic, and target evidence. Development checks found higher-capacity heads less robust externally at this label size (Appendix D), motivating the retrieval-first design. This choice also makes the DILI head easy to update: adding new labeled reference drugs changes the retrieval memory without retraining a high-capacity classifier.

Role of predicted expression. The predicted-expression component is useful here primarily as a retrieval geometry, not as proof that every individual gene prediction is mechanistically validated. It contributes because predicted-delta retrieval improves perturbation prediction (Table 1) and predicted-expression similarity is much stronger than Morgan-only retrieval for internal DILI classification (Table 3). The DILIst result nevertheless shows that it is not a universal replacement for larger endpoint-supervised models; gene-level biological claims are therefore limited to the evaluated perturbation-responsive panels.

Limitations and scope. The 165-drug label regime remains an important constraint, but it is also the regime Hepa-RAFT is designed for: learning is concentrated in hepatocyte-response prediction, while the DILI endpoint is handled by an updatable retrieval memory. The external tables and training-size view in Table 2 and Figure 2 are contextual rather than a controlled head-to-head scaling law, because comparator resources differ in label provenance, auxiliary supervision, and overlap-filtered subsets. Hepa-RAFT estimates hepatocyte-intrinsic liability rather than total clinical DILI risk; it is not designed to model host immune or HLA-gated risk factors. Finally, the external benchmarks use related public DILI curation protocols; drug-level overlap after standardized-SMILES deduplication is zero, but external AUROC should be interpreted as unseen-drug generalization under related label definitions rather than as fully independent clinical validation. These scope limits motivate future work on mechanism-stratified labels and adaptive channel weighting.

Broader impact. Hepa-RAFT provides a SMILES-queryable virtual hepatocyte-response module that could be used inside generative or agentic drug-design workflows to triage compounds before experimental testing, potentially reducing animal testing and accelerating early-stage safety screening. False negatives, however, could create unwarranted confidence in unsafe compounds, so predictions should complement—not replace—experimental safety evaluation and clinical judgment.

References

Aleo, M. D., Luo, Y., Swiss, R., Bonin, P. D., Potter, D. M., and Will, Y. Human drug-induced liver injury severity is highly associated with dual inhibition of liver mitochondrial function and bile salt export pump. Hepatology, 60(3):1015–1022, 2014. doi: 10.1002/hep.27206.

Bemis, G. W. and Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. Journal of

Medicinal Chemistry, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928.

Bergen, V., Kodella, K., Srikrishnan, S., Barrandon, O., Anderson, S., Rogers-Grazado, M., Fowler, C., Beyene, H., Robichaud, N., Fulton, T., Lapchyk, N., Cortes, M., Plugis, N., Goddeeris, M., and Zamanighomi, M. A large-scale human toxicogenomics resource for drug-induced liver injury prediction. Nature Communications, 2025. doi: 10.1038/s41467-025-65690-3. Dataset and code: https://github.com/Cellarity/DILImap. Primary human hepatocyte SMART-Seq profiles with multi-concentration labels.

Chen, M., Suzuki, A., Thakkar, S., Yu, K., Hu, C., and Tong, W. DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. Drug Discovery Today, 21(4): 648–653, 2016. doi: 10.1016/j.drudis.2016.02.015.

Freshour, S. L., Kiwala, S., Cotto, K. C., Coffman, A. C., McMichael, J. F., Song, J., Griffith, M., Griffith, O., and Wagner, A. Integration of the drug–gene interaction database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Research, 49(D1): D1144–D1151, 2021. doi: 10.1093/nar/gkaa1084.

Harrill, J. A., Everett, L. J., Haggard, D. E., Sheffield, T., Bundy, J. L., Willis, C. M., Thomas, R. S., Shah, I., and Judson, R. S. High-throughput transcriptomics platform for screening environmental chemicals. Toxicological Sciences, 181(1):68–89, 2021. doi: 10.1093/toxsci/kfab009.

Hoofnagle, J. H., Serrano, J., Knoben, J. E., and Navarro, V. J. LiverTox: a website on drug-induced liver injury. Hepatology, 57(3):873–874, 2013. doi: 10.1002/hep.26175.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Artificial intelligence foundation for therapeutic science. Nature Chemical Biology, 18:1033–1036, 2022. doi: 10.1038/s41589-022-01131-2.

Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., and Yamada, H. Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Research, 43(D1):D921–D927, 2015. doi: 10.1093/nar/gku955.

Kendall, A. and Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems, volume 30, pp. 5574–5584, 2017.

Nix, D. A. and Weigend, A. S. Estimating the mean and variance of the target probability distribution. In Proceedings of 1994 IEEE International Conference on Neural Networks, volume 1, pp. 55–60. IEEE, 1994. doi: 10.1109/ICNN.1994.374138.

Onakpoya, I. J., Heneghan, C. J., and Aronson, J. K. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. BMC Medicine, 14: 10, 2016. doi: 10.1186/s12916-016-0553-2.

Ostapowicz, G., Fontana, R. J., Schiødt, F. V., Larson, A., Davern, T. J., Han, S. H. B., McCashland, T. M., Shakil, A. O., Hay, J. E., Hynan, L., Crippin, J. S., Blei, A. T., Samuel, G., Reisch, J., and Lee, W. M. Results of a prospective study of acute liver failure at 17 tertiary care centers in the United States. Annals of Internal Medicine, 137(12):947–954, 2002. doi: 10.7326/0003-4819-137-12-200212170-00007.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. FiLM: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

Richard, A. M., Judson, R. S., Houck, K. A., Grulke, C. M., Volarath, P., Thillainadarajah, I., Yang, C., Rathman, J., Martin, M. T., Wambaugh, J. F., et al. ToxCast chemical landscape: paving the road to 21st century toxicology. Chemical Research in Toxicology, 29(8):1225–1251, 2016. doi: 10.1021/acs.chemrestox.6b00135.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence, 4:1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Schimunek, J., Seidl, P., Friedrich, L., Kuhn, D., Rippmann, F., Hochreiter, S., and Klambauer, G. Context-enriched molecule representations improve few-shot drug discovery. In International Conference on Learning Representations (ICLR), 2023. OpenReview forum XrMWUuEevr; arXiv:2305.09481.

Seal, S., Trapotsi, M.-A., Spjuth, O., Singh, S., Carreras-Puigvert, J., Greene, N., Bender, A., and Carpenter, A. E. Improved detection of drug-induced liver injury by integrating predicted in vivo

and in vitro data. Chemical Research in Toxicology, 37(7):1290–1305, 2024. doi: 10.1021/acs.chemrestox.4c00015.

Swanson, K., Walther, P., Leitz, J., Mukherjee, S., Wu, J. C., Shivnaraine, R. V., and Zou, J. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. Bioinformatics, 40(7):btae416, 2024. doi: 10.1093/bioinformatics/btae416.

Thakkar, S., Li, T., Liu, Z., Wu, L., Roberts, R., and Tong, W. Drug-induced liver injury severity and toxicity (DILIst): binary classification of 1279 drugs by human hepatotoxicity. Drug Discovery Today, 25(1):201–208, 2020. doi: 10.1016/j.drudis.2019.09.022.

Uetrecht, J. Mechanisms of idiosyncratic drug-induced liver injury. In Drug-Induced Liver Injury, volume 85 of Advances in Pharmacology, pp. 133–163. Academic Press, 2019. doi: 10.1016/bs.apha.2018.12.001.

Watkins, P. B. Drug safety sciences and the bottleneck in drug development. Clinical Pharmacology & Therapeutics, 89(6):788–790, 2011. doi: 10.1038/clpt.2011.63.

Wibowo, A. S., Chong, K. T., and Tayara, H. Enhancing DILI toxicity prediction through integrated graph attention (GATNN) and dense neural networks (DNN). Toxicology, 514:154108, 2025. doi: 10.1016/j.tox.2025.154108.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. MoleculeNet: a benchmark for molecular machine learning. Chemical Science, 9:513–530, 2018. doi: 10.1039/C7SC02664A.

A. Per-Method Training-Set Sources

Table 4 lists the training-set disclosures used to exclude overlapping external drugs for Table 2. Matching uses standardized InChIKey first and standardized SMILES as fallback; reconstruction assumptions are stated when source lists are incomplete.

Table 4. Comparator training-set sources and reconstruction notes. Matching key: standardized InChIKey, with standardized SMILES fallback. The FiLM-decoder row reports feature-level pretraining overlap against the external sets.

Method / corpus	Training list source	Reconstruction notes
Published DILIPredictor training list	Seal et al. (2024) Supp. Table S1	Published 888-drug training list.
DILIPredictor disclosure universe	Seal et al. (2024) published training/held-out/resource lists	Reconstructed 1,111-drug conservative disclosure universe for the matched subsets in Table 2 and the training-size axis in Figure 2.
ADMET-AI (zero-shot)	Swanson et al. (2024) DILI training set	Corpus enumerated in the public release; Table 2 reports external performance after removing drugs overlapping this disclosed training set.
GATNN	Wibowo et al. (2025) preprocessing files (1,534 unique SMILES)	Training data fully enumerated in the released artifacts; Table 2 reports external performance after removing drugs overlapping this disclosed training set.
Morgan FP RF (ours)	Our 165-drug benchmark	0% overlap by construction.
Ours (FiLM pretraining)	1,284 InChIKeys	Self-disclosure: 4/893 DILImap and 90/1,070 DILIst SMILES appeared in FiLM pretraining (1,070 = DILIst entries with standardized InChIKeys; 7/1,077 lack standardized InChIKey). External labels are never used by the DILI-liability k-NN vote (Section 4.2).

B. Implementation Hyperparameters

FiLM-decoder configuration, shared by the final checkpoint and all 25 scaffold-CV retrainings: 5,000 control-HVG genes on the TG-GATEs/ToxCast intersection; frozen 768-d MoLFormer encoder; hidden dimension 64; attention-weighted FiLM ("attn_film", per-gene drug attention); Gaussian negative log-likelihood with mean/log-variance heads; Adam, learning rate 10^{-3}, batch size 32, maximum 200 epochs, early stopping patience 45; cell-type vocabulary $\{\text{HepaRG}, \text{PHH}\}$; training dose range 0.01–100 μM. External inference uses HepaRG and 5 μM (Section 3). Both retrieval heads are deterministic and parameter-free given frozen FiLM outputs plus Morgan, ADMET-AI, and DGIdb features. Internal scaffold folds contain 32–34 drugs, DILI$^+$ prevalence 0.706–0.727, and 13–29 unique Murcko scaffolds.

C. Algorithm

Algorithm 1 Hepa-RAFT Inference

Input: Query drug SMILES q, training drug database $\mathcal{D}$
Output: Predicted expression $\hat{\mathbf{y}}_q$, DILI probability p_q

// Expression-correction head
$\mathbf{d}_q \leftarrow \text{MoLFormer}(q)$ Drug embedding
$\hat{\mathbf{y}}_q^{\text{pred}} \leftarrow \text{FiLM decoder}(\mathbf{d}_q; \text{HepaRG}, 5\,\mu\text{M})$ Raw prediction
$\hat{\boldsymbol{\delta}}_q \leftarrow \hat{\mathbf{y}}_q^{\text{pred}} - \hat{\mathbf{y}}_{\text{DMSO}}^{\text{pred}}$ Predicted delta
$\mathcal{N}_1 \leftarrow k\text{-NN}(q, \mathcal{D}, S^{(1)}, k{=}7)$ Expression neighbors
$\hat{\mathbf{y}}_q \leftarrow \hat{\mathbf{y}}_q^{\text{pred}} + \sum_{i \in \mathcal{N}_1} w_i \mathbf{r}_i$ Residual correction

// DILI-liability retrieval head
$\mathcal{N}_2 \leftarrow k\text{-NN}(q, \mathcal{D}, S^{(2)}, k{=}10)$ DILI neighbors using M/raw-P/A/T channels
$p_q \leftarrow \sum_{i \in \mathcal{N}_2} w_i \mathbb{1}[\text{DILI}_i{=}1], \quad w_i = S_{q,i} / \sum_j S_{q,j}$ Weighted soft vote

return $\hat{\mathbf{y}}_q$, p_q

D. Internal Development Checks

During development, we also evaluated higher-capacity DILI heads on FiLM-derived representations, including neural classifiers, boosted trees, and representation-learning variants. These diagnostics often improved or matched internal scores but did not improve external generalization at the available 165-drug DILI-label scale. We therefore use them only to motivate the final retrieval-first design rather than as controlled benchmark comparisons.

E. Additional Results

Perturbation prediction across gene panel sizes. Table 5 shows that retrieval-augmented correction is robust across different numbers of differentially expressed genes, maintaining high performance from DE20 to DE100.

Table 5. Expression-correction performance by gene panel size (predicted-delta retrieval, k=7).

Gene panel	Pear_Δ	Direction
DE20	0.631±0.033	0.829
DE50	0.635±0.034	0.817
DE100	0.628±0.034	0.798