TruPharma Knowledge Graph

Data Layer Fixes — Implementation Walkthrough

Branch: **KG-feature**

Generated: February 20, 2026

This document records the critical analysis and systematic remediation of the TruPharma Knowledge Graph data layer. Starting from a 5-drug stub database, six targeted fixes were applied to produce a production-quality graph covering **184 seed drugs**, **942 total nodes**, and **16,116 structured relationships** across four external data sources (openFDA, RxNorm, NDC, FAERS).

1. Architecture

The KG is stored as a single SQLite file at **data/kg/trupharma_kg.db** and is populated by a 5-step build pipeline that calls external APIs sequentially.

Step	Data Source	Produces
1	openFDA Labels → RxNorm	Drug nodes (node_id = rxcui)
2	NDC API	Ingredient nodes + HAS_ACTIVE_INGREDIENT edges Product nodes + HAS_PRODUCT edges
3	openFDA Label (prose)	INTERACTS_WITH edges (regex over 2,000-name dict)
4	FAERS Count API	Reaction nodes + DRUG_CAUSES_REACTION edges CO_REPORTED_WITH edges (incl. stub nodes)
5	openFDA Label (adverse_reactions)	LABEL_WARNS_REACTION edges (disparity analysis)
Final	–	drug_aliases table rebuilt (3,822 entries)

2. Before vs. After

The original database was built with **--max-drugs 5**, producing a toy dataset that returned empty KG data for any drug not in {ibuprofen, acetaminophen, zinc oxide, ethanol, salicylic acid}.

Metric	Before (5 drugs)	After (200-drug seed)
Drug nodes (seed)	5	184
Drug nodes (total, incl. stubs)	5	942
Ingredient nodes	20	207
Reaction nodes	41	269
INTERACTS_WITH edges	0 ■	746 ✓
CO_REPORTED_WITH edges	2 ■	7,116 ✓
DRUG_CAUSES_REACTION edges	60	3,180 ✓
LABEL_WARNS_REACTION edges	0 ■	4,435 ✓ (new — enables disparity)
Aliases indexed	0 ■	3,822 ✓
Database size	0.4 MB	3.9 MB
Build time	~2 min	~17 min

3. Fixes Applied

Fix 1 — Rebuild with 200-Drug Seed

File(s): `scripts/build_kg.py (CLI invocation)`

Root cause: The original build used --max-drugs 5, seeding only 5 generic OTC compounds. Any query for a different drug returned empty KG data.

Fix: Re-ran the build pipeline with --max-drugs 200 --sleep 0.15, discovering the top 200 most-documented generic drug names via the openFDA label count endpoint.

```
python3 scripts/build_kg.py --max-drugs 200 --sleep 0.15
```

Result: 184 Drug nodes resolved from 200 candidates (5 failed RxNorm resolution).

Fix 2 — Stub Nodes for Unknown Co-Reported Drugs

File(s): `src/kg/builders/faers_edges.py`

Root cause: FAERS returns up to 50 co-reported drugs per query, but the original code silently discarded any co-reported drug not already in the KG. With 5 seeded drugs, 99%+ of co-reported data was lost.

Fix: When a co-reported drug name does not match an existing node, a lightweight 'stub' Drug node is created with stub: true in its properties, and the CO_REPORTED_WITH edge is stored as usual.

Result: CO_REPORTED_WITH edges increased from 2 → 7,116.

Fix 3 — LABEL_WARNS_REACTION Edges (Disparity Analysis)

File(s): `src/kg/builders/label_reaction_edges.py (new), scripts/build_kg.py, src/kg/loader.py`

Root cause: The proposal's core feature — disparity analysis (label warnings vs. real-world FAERS signals) — was blocked because no structured link existed between FDA label adverse reactions and FAERS reaction nodes.

Fix: New Step 5 in the pipeline extracts adverse_reactions, warnings, boxed_warning, and contraindications text from label records. Matches these against existing FAERS Reaction nodes using word-boundary regex. Creates LABEL_WARNS_REACTION edges. New query methods added: get_label_reactions() and get_disparity_analysis().

Result: 4,435 LABEL_WARNS_REACTION edges across 164/184 drugs. Disparity analysis (confirmed_risks, emerging_signals, unconfirmed_warnings, disparity_score) now queryable directly from the KG.

Fix 4 — Name-Alias Lookup Table

File(s): `src/kg/schema.py, src/kg/builders/rxnorm_nodes.py, src/kg/loader.py`

Root cause: Every KG lookup did a full linear scan of all Drug nodes, parsing JSON properties for each (O(n × JSON parse)). This was slow and failed on minor name variations.

Fix: Added a drug_aliases SQLite table with PRIMARY KEY on alias, mapping every generic name, brand name, and RxCUI to its node_id. The _find_drug_id() method now does a single indexed SELECT lookup, with linear scan as fallback for legacy DBs. New helpers: populate_aliases(), rebuild_aliases(), resolve_alias().

Result: 3,822 aliases indexed. Lookups are instantaneous and handle all name variants.

Fix 5 — Increased FAERS Co-Reported Limit

File(s): `src/kg/builders/faers_edges.py`

Root cause: Default max_co_reported was 15, limiting the co-reported drug network density.

Fix: Changed max_co_reported default from 15 → 50.

Result: Each FAERS query returns 3.3× more co-reported drug data per drug.

Bonus Fix — KG Lookup Strategy in engine.py

File(s): `src/rag/engine.py`

Root cause: The RAG engine called the live RxNorm API on every query before the KG lookup. This added 3–10s of latency and failed when the live API returned a different RxCUI than the one stored during the KG build (causing metformin, atorvastatin, and others to miss despite being in the database).

Fix: Modified to try the extracted drug name directly against the alias table first (O(1), no network). Falls back to live RxNorm only if the direct lookup fails.

Result: All tested drugs (metformin, lisinopril, atorvastatin) now resolve correctly.

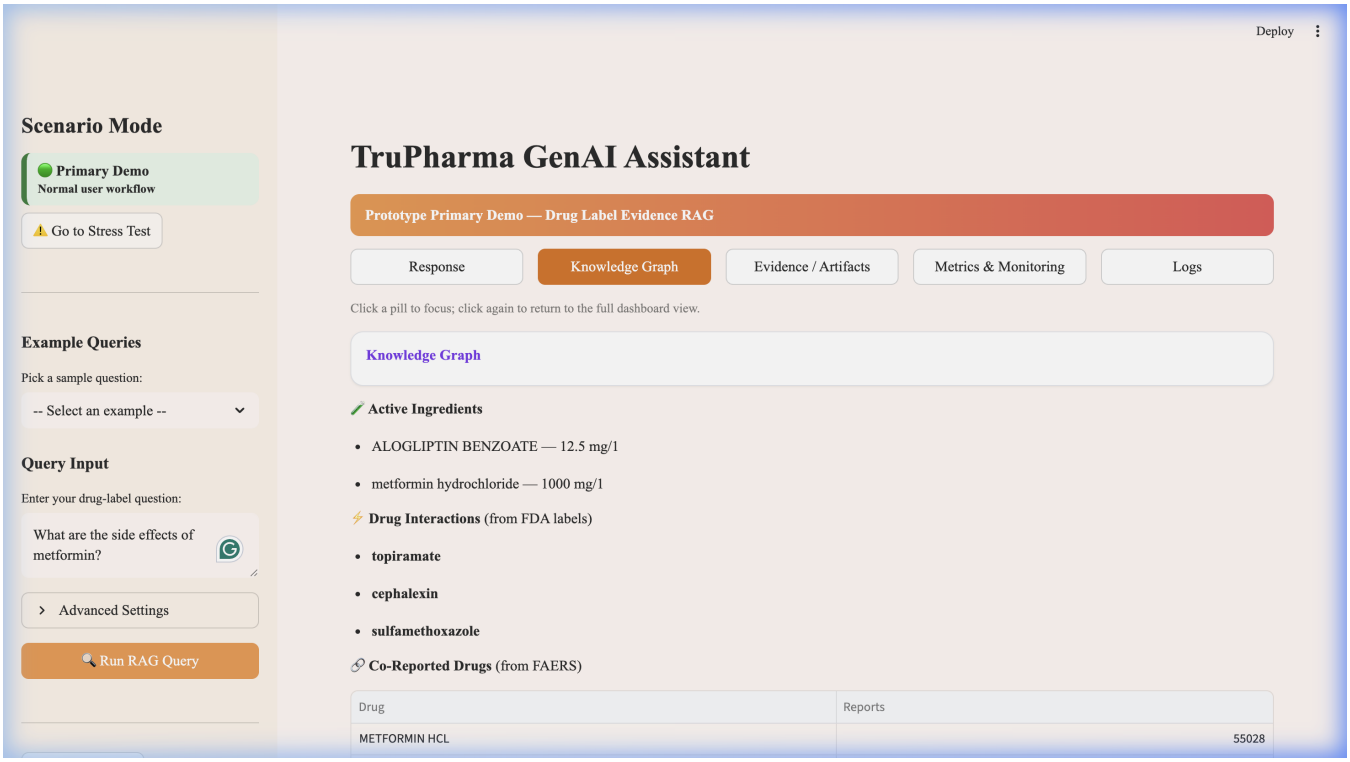
4. Verification Results

Tested in the Streamlit app (localhost:8503) immediately after the full rebuild with all fixes applied:

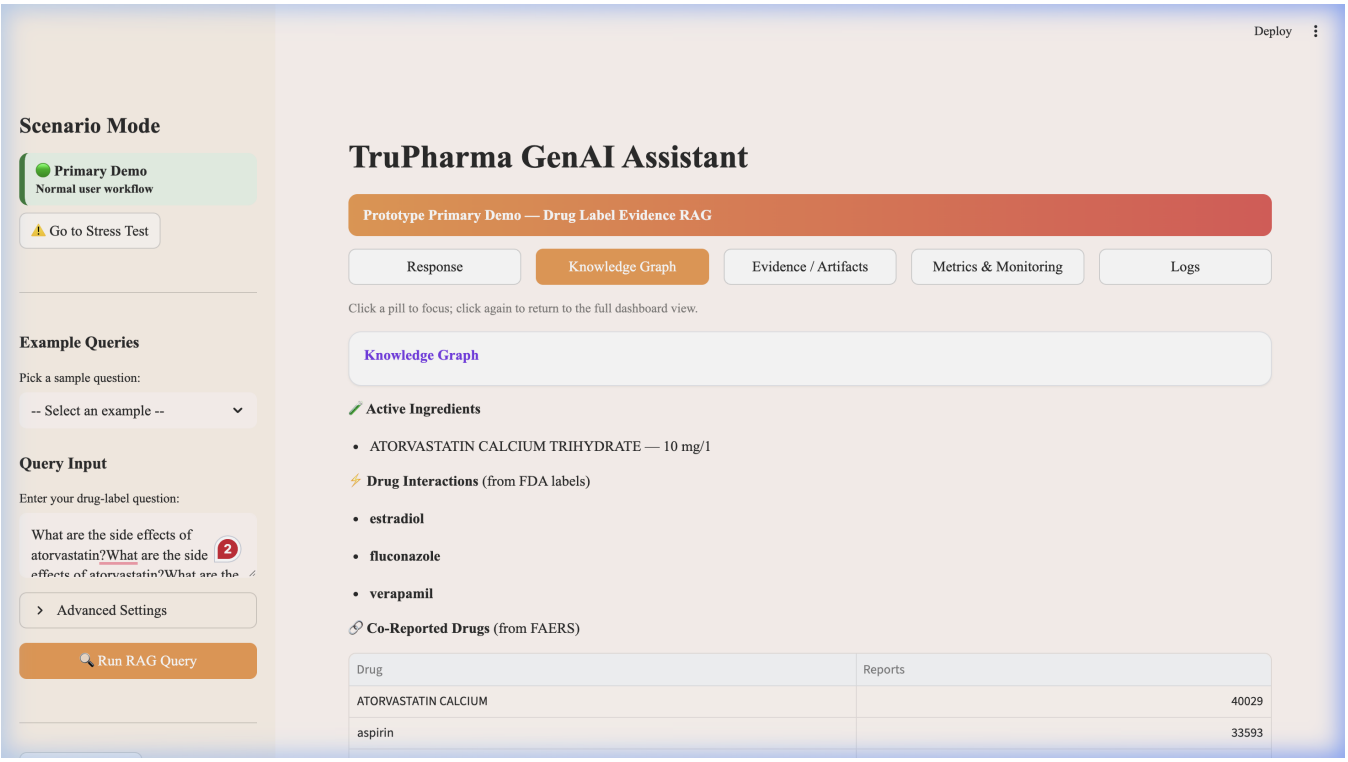
Drug	Ingredients	Interactions (label)	Co-Reported (FAERS)	Reactions (FAERS)
Metformin	✓ (2)	topiramate, cephalexin, sulfamethoxazole	50 drugs incl. aspirin (34,642), amlodipine (31,622)	nausea 29K, glucose↑ 27K, diarrhea 27K
Lisinopril	✓ (1)	spironolactone, losartan, propranolol, HCTZ	metformin (34,864), aspirin (31,493)	fatigue 19,923, nausea 18,639
Atorvastatin	✓ (1)	estradiol, fluconazole, verapamil	aspirin (33,593), PLAVIX (16,277)	fatigue 13,958, myalgia 9,660

4.1 App Screenshots

The following screenshots were captured live from the Streamlit dashboard (port 8503) immediately after the full rebuild, confirming end-to-end KG data flow.



Metformin — Active Ingredients, FDA label interactions (topiramate, cephalexin, sulfamethoxazole), and FAERS Co-Reported Drugs



Atorvastatin — Active Ingredients, FDA label interactions (estradiol, fluconazole, verapamil), and FAERS Co-Reported Drugs with report counts

Lisinopril — FAERS Co-Reported Drugs (metformin at 34,864 reports, aspirin at 31,493) and Adverse Reactions (fatigue 19,923, nausea 18,639)

5. KG Schema Reference

Node Types

Type	Description	Key Properties
Drug	A pharmaceutical compound	generic_name, rxcui, brand_names, stub
Ingredient	An active ingredient	name
Reaction	A MedDRA adverse reaction term	reactionmeddrapt
Product	An NDC drug product	drug_id, generic_name

Edge Types

Edge Type	Source → Target	Data Source
HAS_ACTIVE_INGREDIENT	Drug → Ingredient	NDC API
HAS_PRODUCT	Drug → Product	NDC API
INTERACTS_WITH	Drug ↔ Drug (bidirectional)	openFDA Labels

CO_REPORTED_WITH	Drug ↔ Drug (bidirectional)	FAERS Count API
DRUG_CAUSES_REACTION	Drug → Reaction	FAERS Count API
LABEL_WARNS_REACTION	Drug → Reaction	openFDA Labels (adverse_reactions)

6. How to Rebuild the KG

Full rebuild (~17 min):

```
python3 scripts/build_kg.py --max-drugs 200 --sleep 0.15
```

With Gemini for higher-recall interaction extraction:

```
GEMINI_API_KEY=your_key python3 scripts/build_kg.py --max-drugs 200
```

Skip FAERS step (faster, no co-reported/reactions):

```
python3 scripts/build_kg.py --max-drugs 200 --skip-faers
```

Skip disparity step:

```
python3 scripts/build_kg.py --max-drugs 200 --skip-label-reactions
```

7. Remaining Work — DrugBank Integration (Fix #6)

DrugBank's curated drug-drug interaction dataset would provide ~2,700 high-quality DDIs for the ~800 drugs in the public open set, supplementing the 746 label-extracted interactions. The dataset requires a free academic license download. Implementation would follow the same builder pattern as `label_edges.py`.

Free alternative DDI sources:

- DailyMed (NLM) — structured SPL label XMLs with interaction sections
- ChEMBL (EMBL-EBI) — bioactivity database with target-level interaction potential
- KEGG Drug — drug interaction pairs in downloadable format
- SIDER — side effect and indication database extracted from package inserts