Darlyne AI: A Deterministic, Structural Retrieval Engine for PGx Exploration

Demesvar Destin
demesvar@destingenomics.com
www.darlyne.ai

Abstract. The exponential growth of pharmacogenomic (PGx) data has created a heavily siloed landscape, forcing researchers to manually synthesize findings across disparate databases. While artificial intelligence and Retrieval-Augmented Generation (RAG) offer potential solutions to this fragmentation, current probabilistic models suffer from severe limitations in biomedical contexts, primarily "hallucinations" and inaccurate source retrieval. We propose a solution to the PGx data fragmentation problem using a deterministic, structural retrieval architecture. By aggregating proven clinical topologies into a unified PostgreSQL matrix and restricting the Large Language Model (LLM) strictly to these verified structural edges, we eliminate probabilistic guessing. The resulting platform, Darlyne, serves as a centralized, hallucination-free knowledge graph that democratizes precision medicine research.

1. Introduction

Modern biomedical research relies almost exclusively on institutional databases serving as trusted repositories for genetic and pharmacological data. Platforms like GeneCards, OpenTargets, PharmGKB, and OpenFDA provide rigorous, peer-reviewed insights. However, the system suffers from the inherent weaknesses of data fragmentation. To map a single adverse drug event to a specific genetic variant, a researcher must cross-reference multiple independent interfaces, normalize the differing terminologies, and manually synthesize the results.

This multi-platform reliance creates significant friction. It slows the pace of academic research and increases the cognitive load on clinical scientists. What is needed is a unified exploration system that aggregates these trusted sources into a single conversational interface.

Recent advancements in Large Language Models have attempted to solve this issue. Unfortunately, generalized LLMs are inherently probabilistic. They predict the next most likely word rather than retrieving absolute facts, making them dangerously unreliable for clinical

research. To counteract this, engineers developed Retrieval-Augmented Generation.

2. The Limitations of Standard Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a framework that improves LLM accuracy by fetching external documents before generating a response (Lewis et al., 2020). In a standard biomedical RAG pipeline, vast amounts of unstructured text (like PDF research papers or medical journals) are converted into vector embeddings. When a user asks a question, the system performs a "similarity search" to find paragraphs that mathematically resemble the user's query, which the LLM then reads to formulate an answer.

While standard RAG is highly effective for general enterprise data, it is fundamentally flawed for deep pharmacogenomics.

1. **Probabilistic Matching:** Vector searches rely on semantic similarity. If a user queries the interactions of "CDH13", a standard RAG system might retrieve articles about genes that simply sound similar or appear in similar contexts, missing the exact pharmacological targeting data entirely.
2. **Unverified Extraction:** Standard RAG cannot differentiate between a proven structural interaction and a highly debated hypothesis mentioned in a single obscure paper.
3. **Black Box Retrieval:** The user has no guarantee that the LLM successfully retrieved all available data, leading to incomplete or skewed clinical summaries.

3. The Proposed Solution: Deterministic Structural Injection

We propose an improved PGx exploration system that entirely abandons probabilistic vector search in favor of deterministic structural retrieval. Instead of searching through unstructured PDFs, Darlyne operates on a heavily curated, relational knowledge graph. The Darlyne architecture requires minimal unstructured searching. The steps to run the network are as follows:

1. **Matrix Aggregation:** Millions of clinically validated topological edges (gene-gene interactions, pathogenic links, SNP overlaps) are harvested directly from OpenTargets, PharmGKB, OpenFDA, Ensembl, MyGene.info, PubChem, and DGIdb. These are vaulted as strict relational rows in a PostgreSQL database.

2. **Entity Classification:** When a user submits a natural language query, a routing layer extracts the precise clinical targets (e.g., specific drugs, genes, or rsIDs).
3. **Structural Retrieval:** The system queries the PostgreSQL matrix for explicit, mathematically weighted connections tied to those exact entities.
4. **Context Injection:** The raw, irrefutable matrix data (along with verified biological metadata) is injected directly into the LLM's context window.
5. **Restricted Synthesis:** The LLM is strictly instructed by the system prompt to synthesize only the provided matrix edges. It acts purely as a translation engine, converting complex database rows into highly readable, categorized clinical profiles.

If the database contains no proven structural edge for a query, the AI clearly flags any subsequent response as a "Literature Fallback" hypothesis. This guarantees that the core outputs remain deterministic, reproducible, and deeply grounded in proven science.

4. Proof of Concept: CDH13 Pharmacological Targeting

To demonstrate the limitations of probabilistic retrieval and the superiority of structural injection, we conducted a comparative case study using the query: *"What drugs target CDH13?"*

- **Generalized LLM Performance (Google AI)**
  When queried, Google's AI overview relied on surface-level internet scraping. It returned generalized therapeutic concepts such as "Epigenetic Drugs" and "Cancer Sensitizers." It failed entirely to identify the specific, clinically approved pharmacological agents that directly target the CDH13 gene.
- **Standard Probabilistic RAG Performance (Perplexity AI)**
  Perplexity successfully identified that a curated database (GeneCards) contained exactly 12 drugs associated with CDH13. However, because standard RAG operates as an external text scraper, it was blocked by the database's user interface architecture. The system explicitly stated: *"I couldn't verify the full list of 12 drugs from the available information... the visible snippet does not expose the full list."* The probabilistic agent knew the data existed but lacked the structural access to retrieve it.
- **Deterministic Structural Retrieval (Darlyne AI)**

Because Darlyne inherently houses the deep structural data within its own relational matrix, it effortlessly bypassed the limitations of web scraping. Darlyne instantly retrieved and formatted the exact pharmacological agents, actively categorizing them by mechanism:

- ➢ **Monoamine Vesicular Transporter Inhibitors:** Tetrabenazine, Deutetrabenazine
- ➢ **Serotonergic Agents:** Ketanserin
- ➢ **Sympathomimetic Agents:** Dextroamphetamine Sulfate, Reserpine

By eliminating the probabilistic middle-step of semantic search, Darlyne achieves a level of accuracy and data retrieval that standard LLM pipelines currently do not match.

5. Industry Applications and Democratization

The Darlyne architecture is designed to accelerate workflows across the precision medicine landscape.

- ❖ **Pharmaceutical Research:** Drug developers can instantly map out the downstream targets of a novel compound, identify potential off-target toxicities early in the pipeline, and discover repurposing opportunities for existing FDA-approved drugs based on shared genetic pathways.
- ❖ **Academic Advancement:** Doctoral students and independent researchers can bypass weeks of manual literature review and multi-site navigation. By querying specific SNPs or disease phenotypes, researchers receive instant, highly cited topological maps to support dissertation writing and grant proposals.
- ❖ **Clinical Investigation:** Pharmacists and clinical geneticists can utilize the platform to investigate the molecular mechanisms behind rare adverse drug events, aiding in the interpretation of complex patient pharmacogenomic profiles.

Crucially, current advanced bioinformatics platforms (such as Causaly or Tempus) gatekeep their insights behind prohibitive enterprise licensing fees. Darlyne is engineered to democratize this technology, providing enterprise-grade, deterministic architecture completely free of charge to the independent research community.

6. Conclusion

We have proposed a system for pharmacogenomic exploration that does not rely on probabilistic guessing. We started with the RAG framework, which provides a method for external data injection but is incomplete without a way to guarantee the accuracy and completeness of the retrieved medical data. To solve this, we created a deterministic architecture utilizing a unified knowledge graph populated exclusively by proven clinical topologies. The resulting platform provides researchers with a highly efficient, conversational alternative to multi-site browsing. Users receive hallucination-free insights grounded in the industry's most rigorous databases, advancing the accessibility and speed of precision medicine research.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Yih, W. T. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
2. Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., ... & Dunham, I. (2021). Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic acids research*, 49(D1), D1077-D1084.
3. Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., ... & Klein, T. E. (2012). Pharmacogenomics knowledge for personalized medicine. *Clinical pharmacology & therapeutics*, 92(4), 414-417.