PORTFOLIO CASE STUDY

Nexus Knowledge Graph: turning biomedical literature into inspectable hypotheses

A cross-domain correlation analysis pipeline for moving from messy biomedical papers to structured evidence, graph-searchable mechanisms, and testable research hypotheses.

PYTHON · PUBMED API · SQLITE STAGING · LLM EXTRACTION · NEO4J · CYPHER DISCOVERY

The Problem

Biomedical literature is not merely large. It is fragmented across vocabularies, model organisms, disease areas, paper formats, and levels of evidence. A mechanism may appear in one field as a microbial metabolite, in another as an immune marker, and in a third as a mitochondrial or circadian signal. The valuable part is often the bridge between those domains, not the isolated paper.

Nexus was built to make those bridges easier to inspect. Instead of treating papers as a pile of PDFs, it turns them into structured records and graph relationships so that contradictions, convergence points, organism-metabolite chains, and underexplored mechanisms can be queried directly.

What I Built

Nexus is a private research codebase and the graph itself is not public. This portfolio write-up describes the system at a high level: candidate discovery, abstract scoring, full-text acquisition, schema-guided extraction, adversarial verification, normalization, graph ingestion, and downstream hypothesis generation.

SEARCH
PubMed keyword taxonomies and domain-specific query sets.

SCORE
LLM relevance scoring to narrow large candidate sets.

EXTRACT
Structured JSON records from abstracts and full texts.

GRAPH
Neo4j entities, relationships, paper links, and evidence counts.

DISCOVER
Cypher pattern passes and LLM-assisted hypothesis assessment.

Scale

The current Neo4j graph contains six ingested research domains: lupus/microbiome, CRS/migraine, male fertility, mitochondrial biology, psychiatric medications, and circadian immune biology. Additional staging corpora were used while exploring serotonin/psychedelic and vagal/neuroimmune angles.

These domains were not chosen as a generic breadth exercise. They were selected because there is plausible mechanistic overlap between groups of them: gut metabolites and immune signalling, mitochondrial stress and inflammation, circadian rhythm and immune timing, neuroimmune pathways, medication effects, reproductive biology, and systemic inflammatory disease.

~10,000 keyword-matched papers	**144,087** biomedical entity nodes	**181,131** semantic entity relationships
225,698 evidence-counted observations	**210,168** paper-to-entity source links	**6** research domains ingested into the graph

Figures are from the live local Neo4j graph status output. Semantic relationships exclude paper mention edges.

Corpus Collection

The acquisition layer uses keyword-driven PubMed API searches, with domain-specific query sets for lupus/microbiome, mitochondrial biology, circadian immune biology, CRS/migraine, serotonin and psychedelic mechanisms, vagal neuroimmune mechanisms, male fertility, and psychiatric medication effects.

ACQUISITION DESIGN

- Search terms were organized as taxonomies, not single keywords, so the pipeline could capture different names for the same biological idea.
- Large candidate sets were staged in SQLite, deduplicated across related domains, and advanced through resumable statuses.
- New mechanisms, organisms, metabolites, interventions, and adverse-effect targets were fed back into later query passes.
- The pipeline supported full-text acquisition where available, while retaining abstract-only records where necessary.

Extraction

Nexus uses modular extraction schemas rather than one generic prompt. Every paper can produce a universal biomedical core, while domain extensions capture the details that matter for a specific question: taxa and metabolites for microbiome work, mitochondrial pathways for immunometabolism, medication effects for psychiatric drug research, reproductive endpoints for male fertility, and circadian markers for sleep-immune biology.

The key output is a standardized JSON record. For non-technical readers, that means each dense paper is converted into a consistent data card: what was studied, in what population or model, what changed, what did not change, what mechanism was proposed, what evidence supported it, and which source paper the claim came from.

QUALITY CONTROLS

- Schema-guided extraction required explicit fields rather than freeform summaries.
- Null and negative findings were treated as useful data, reducing pressure to invent signal where a paper was silent.
- Consensus and adversarial verification passes checked structured outputs against source text.
- Quality gates excluded blocked records before graph ingestion.
- Entity normalization merged synonyms so the graph did not split one concept across many spellings.

Graph Design

The graph separates papers from biomedical entities. Papers anchor the provenance; entity nodes represent compounds, cell types, organisms, pathways, metabolites, diseases, organs, outcomes, dietary factors, and medications. Relationships capture direction and mechanism: increases, decreases, no effect, associated with, enriched in, depleted in, produces, and mechanism.

This design lets a question become a graph query. Instead of asking an LLM to "think about lupus and the gut," Nexus can ask: which organisms produce metabolites that connect to SLE through immune markers, where is evidence contradictory, what nodes bridge multiple domains, and which apparently indirect paths have enough support to justify a closer literature review?

Discovery Layer

The discovery scripts run Cypher pattern passes over the graph, then send selected candidates to an LLM assessment pass for novelty, plausibility, mechanism, and suggested follow-up. This was used for general SLE/gut analysis and for targeted CRS/migraine questions involving gut biology, mitochondrial dysfunction, circadian immune rhythm, mast cells, serotonin, and neuroinflammation.

PATTERN TYPES

- Contradictory evidence pairs.
- Organism-to-metabolite-to-disease chains.
- Cross-domain bridge nodes.
- High-connectivity hub mechanisms.
- Research gaps near a selected entity.
- Shared medication mechanisms.

Example Output

One downstream pass converted the SLE/gut-health analysis into a structured personal research report, supplement rationale, and shareable timing protocol. It organized candidate mechanisms around SCFAs, bile acids, tryptophan routing, butyrate and valerate production, mitochondrial ROS, circadian disruption, gut barrier integrity, and immune balance.

The practical result was not just an interesting graph. It helped surface a real supplement protocol that I could reason about, sequence, and test carefully, with a noticeable positive impact from the resulting stack.

That output is best understood as personal hypothesis generation rather than universal clinical guidance. It demonstrates the value of the pipeline: moving from thousands of papers to a map of plausible mechanisms, candidate interventions, caveats, and follow-up questions that can be reviewed with appropriate medical or research expertise.

Why It Matters

The interesting part of Nexus is not that it processed a large number of papers. The interesting part is that it created an inspectable bridge between literature search, structured extraction, provenance, graph reasoning, and human review. The system is designed to make speculative connections more disciplined: every claim should have a source, every relationship should be queryable, and every generated hypothesis should remain attached to its supporting evidence and uncertainty.

Ownership

I built Nexus end to end, using LLMs as both part of the research pipeline and as development assistance while designing, implementing, testing, and iterating the system.

- Designed the cross-domain research architecture and domain taxonomy strategy.
- Built the Python collection, scoring, extraction, enrichment, verification, and ingestion tooling.
- Created modular Pydantic schemas for core biomedical evidence and domain-specific extensions.
- Implemented SQLite staging databases, resumable pipeline steps, and Neo4j graph ingestion.
- Wrote Cypher-backed discovery scripts for contradictions, bridge mechanisms, hubs, gaps, and ranked candidate hypotheses.
- Translated graph outputs into research reports and public-safe portfolio summaries.

Prepared as a public-safe portfolio case study. Source materials include the Nexus technical specification, local source tree, graph status output, discovery reports, SLE research report, and supplementation protocol. This document excludes raw private database contents and is not medical advice.