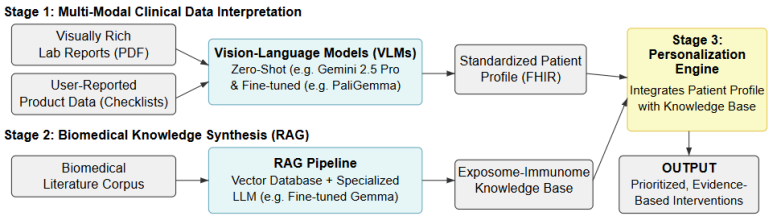
The Exposome Interpreter as a Multi-modal Framework for Autoimmune Trigger Identification

The prevalence of autoimmune diseases is rising, yet personalized treatment remains elusive due to a significant gap between the established knowledge that environmental exposures (exposome) contribute up to 70% of disease risk and the clinical ability to identify patient-specific triggers [1]. This translational challenge is rooted in two fundamental information processing bottlenecks: a clinical data standardization bottleneck, where crucial patient biomarker and exposure data are fragmented across visually complex, semi-structured documents like PDF lab reports, and a biomedical knowledge synthesis bottleneck, where understanding the immunotoxic effects of thousands of chemicals requires synthesizing a vast and rapidly evolving body of scientific literature. Collectively, these bottlenecks preclude clinicians from manually reconciling a patient's unique immunological profile against the dynamic body of knowledge concerning environmental chemicals and their biological impacts. To address this challenge, we introduce the Exposome Interpreter, a multi-modal framework that infers evidence-based links between a patient's environmental exposures and their specific immunological dysregulation.

The Exposome Interpreter [2] is implemented through a modular, three-stage architecture. The first stage, Multi-modal Data Interpretation, formulates structured information extraction from heterogeneous clinical documents (reports, imagery) as a prediction task. We propose a hybrid VLM (Vision-Language Model) architecture that employs a large generalist model to generate programmatic weak supervision, which is then used to efficiently fine-tune a smaller, specialized model for high-precision key-value extraction. The second stage, Biomedical Knowledge Synthesis, constructs a dynamic knowledge base using a Retrieval-Augmented Generation (RAG) pipeline. This pipeline is optimized with specialized biomedical embeddings for high-fidelity retrieval from scientific literature, while a fine-tuned LLM performs multi-document synthesis to generate evidence-grounded causal hypotheses. Finally, the Personalization Engine frames trigger identification as a personalized ranking problem. It learns to score and rank potential exposome-biomarker links by integrating the structured patient profile with the synthesized knowledge, thereby prioritizing the most probable causal factors. The end-to-end performance is tested on link prediction in simulated clinical scenarios.



The efficacy of our framework is validated through an evaluation strategy designed to address the significant challenge of data scarcity and compliance in healthcare. We employ a dual-data approach: while initial development is grounded in a small, fully anonymized set of real-world clinical reports from autoimmune patients, this internal data is not HIPAA compliant for public sharing. Therefore, to ensure reproducible validation, we test the generalizability of our VLM-based parsing engine on a publicly available collection of 426 medical reports [3]. On this public dataset, our hybrid VLM strategy proved effective: a zero-shot Gemini 2.5 Flash model achieved a macro-averaged Precision of 0.88 and Recall of 0.86, which improved to 0.91 Precision and 0.89 Recall with our fine-tuned PaliGemma model. This demonstrates the robustness of our approach even in a data-constrained environment. As this is a work in progress, an evaluation of the Stage 2 RAG pipeline is planned. Future work will assess the retrieval component using metrics like Recall@k, Mean Reciprocal Rank (MRR), and the generation component for Faithfulness (factual consistency) and Answer Relevance. Overall, the Exposome Interpreter presents a viable path to integrating the exposome into clinical care, offering a scalable solution to transform complex health data into actionable, personalized guidance.

[1] Vojdani, A., Pollard, K. M., & Campbell, A. W. (2014). Environmental triggers and autoimmunity. Autoimmune diseases, 2014

[2] Exposome Interpreter Framework. https://valencewellbeing.com

[3] The dataset of 426 diverse reports is available at https://www.kaggle.com/datasets/dikshaasinghhh/bajaj-213