

Effective remediation begins with correctly identifying and delineating the source of contamination. Across Alberta and beyond, contaminated sites from upstream oil and gas operations to industrial facilities and legacy disposal areas generate large, multivariate chemical datasets alongside decades of historical records. When sources are mischaracterized or co-mingled impacts are not separated, the result is over-scoped remediation, unnecessary monitoring, escalating cost, and delayed closure. We present two complementary AI frameworks: a document intelligence system and an interactive visualization platform. These frameworks improve speed, accuracy, and defensibility of source identification, and in turn the remediation and closure decisions that follow. The approach has been applied across contaminants relevant to Alberta practitioners, including PCDD/Fs, PCBs, PFAS, PAHs, petroleum hydrocarbons, metals, and salinity (chloride) impacts.

The first framework synthesizes large collections of site-specific historical documents—investigation reports, analytical data packages, operational records, and regulatory correspondence. Built on large language model (LLM) based retrieval-augmented generation (RAG), it enables structured natural-language querying across indexed document corpora. Its key innovation is the concept of knowledge nodes: curated, domain-specific reference sets such as keystone peer-reviewed literature or research summaries encoding current scientific consensus. These act as guardrails within the AI system. By anchoring responses to vetted, citation-traceable knowledge nodes, the system keeps outputs within a defensible scientific boundary and reduces unsupported inferences. For remediation programs, site history, potential source areas, and supporting science can be assembled rapidly to focus the investigation and support regulatory submissions.

The second framework targets pattern recognition within high-dimensional chemical datasets. Our interactive visual data portal integrates Uniform Manifold Approximation and Projection (UMAP), hierarchical cluster analysis (HCA), chemical fingerprint visualization, and geospatial mapping in a single environment. UMAP projects multivariate contaminant profiles into low-dimensional embeddings, revealing clusters that share a common chemical signature while highlighting transitional samples where sources mix; HCA adds complementary grouping, and mapping places clusters in the site's geography. The portal is fully interactive: an expert can select a cluster in UMAP space and instantly view the matching fingerprints, dendrogram membership, and geographic distribution or select by location and examine the chemistry. This enables rapid separation of site-related impacts from background, delineation of distinct source areas, and focusing of remediation where it matters.

Applied across source-identification projects, these frameworks deliver measurable gains in throughput and transparency. The knowledge-node architecture keeps insights anchored to credible reference material, while the visual portal produces source identification that is both rigorous and readily communicable to regulators, clients, and communities supporting faster, lower-cost paths to closure. Case studies will demonstrate the workflow in practice, illustrating how machine learning and AI are advancing environmental forensic practice and the remediation decisions it informs.