Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

MULTIMODAL FOUNDATION MODEL FOR PHARMACEUTICAL KNOWLEDGE EXTRACTION FROM STRUCTURES, ASSAYS, PATENTS, AND REGULATORY TEXT

Paolo Ricci[1]*, Marco De Luca[1], Giulia Ferraro[2], Antonio Russo[1]

1. ***Department of AI-Based Drug Discovery, Faculty of Pharmacy, University of Naples Federico II, Naples, Italy.***
2. ***Department of Computational Pharmaceutical Systems, Faculty of Engineering, University of Bologna, Bologna, Italy.***

ARTICLE INFO

Received:
28 May 2025
Received in revised form:
12 September 2025
Accepted:
14 September 2025
Available online:
28 October 2025

Keywords: Multimodal foundation model, Pharmaceutical informatics, Chemical language model, Regulatory text mining, Patent knowledge extraction, Retrieval-augmented generation

ABSTRACT

Pharmaceutical research and development generates diverse knowledge spanning chemical structures, biological assay readouts, patent claims, and regulatory documents, yet these sources are typically curated, searched, and interpreted through separate workflows even when describing the same compound, target, or safety concern. No single system currently provides unified reasoning across structural chemistry, pharmacology, intellectual property, and regulatory evidence, forcing researchers to manually integrate information from multiple databases, document repositories, and expert interpretations. This article proposes a conceptual multimodal foundation model for pharmaceutical knowledge extraction that aligns molecules, assays, patents, and regulatory text within a shared representation space. The system architecture combines molecular encoders, assay-table encoders, document-text encoders, contrastive alignment modules, retrieval-augmented generation, and a conversational interface to enable evidence-grounded question answering across pharmaceutical data modalities. Such a model could assist medicinal chemists, pharmacologists, regulatory scientists, and competitive-intelligence teams in retrieving integrated answers that currently require separate searches, while also supporting drug repurposing, safety signal review, and patent landscape analysis by linking evidence across modalities. By facilitating cross-domain reasoning, a pharmaceutical multimodal foundation model could transform the synthesis of complex evidence into a routine and accessible capability.

To Cite This Article: Ricci P, De Luca M, Ferraro G, Russo A. Multimodal Foundation Model for Pharmaceutical Knowledge Extraction from Structures, Assays, Patents, and Regulatory Text. Pharmacophore. 2025;16(5):20-30. https://doi.org/10.51847/8PE7sUIyYy

Introduction

Modern drug discovery is increasingly data rich yet remains information poor because the most important signals are distributed across incompatible representations. A candidate molecule may appear as a SMILES string in PubChem [1], a curated bioactivity record in ChEMBL [2], and a drug-centric entry in DrugBank [3], while related machine-learning benchmarks treat chemical and biological information as structured prediction tasks rather than as integrated knowledge. MoleculeNet helped formalize molecular machine learning as a benchmark problem [4], but benchmark framing alone does not make assay evidence, patent claims, and regulatory safety language jointly queryable. The result is a pharmaceutical knowledge environment in which structure, activity, ownership, and safety are often connected only through manual expert synthesis.

Current pharmaceutical research workflows require scientists to move among chemistry databases, assay repositories, patent search systems, biomedical literature engines, and regulatory label collections. Deep learning systems such as DeepPurpose support drug–target interaction modeling [5], and learned molecular representations can improve property prediction from chemical structure [6], yet these tools generally do not explain how a predicted activity relates to a claim in a patent or an adverse-event statement in a label. Biomedical language models such as BioBERT enable domain-specific text mining [7], while scientific language models such as SciBERT extend transformer-based representations to scholarly writing [8], but they do not natively encode molecules, assay tables, and regulatory sections as a unified evidence graph. This fragmentation makes cross-domain questions difficult to answer at the scale of target families, drug classes, or therapeutic areas. **Figure 1** illustrates

Corresponding Author: Paolo Ricci; Department of AI-Based Drug Discovery, Faculty of Pharmacy, University of Naples Federico II, Naples, Italy. E-mail: paolo.ricci@gmail.com.

how the proposed foundation-model framework transforms fragmented pharmaceutical evidence sources into cross-domain reasoning paths that connect molecular structure.

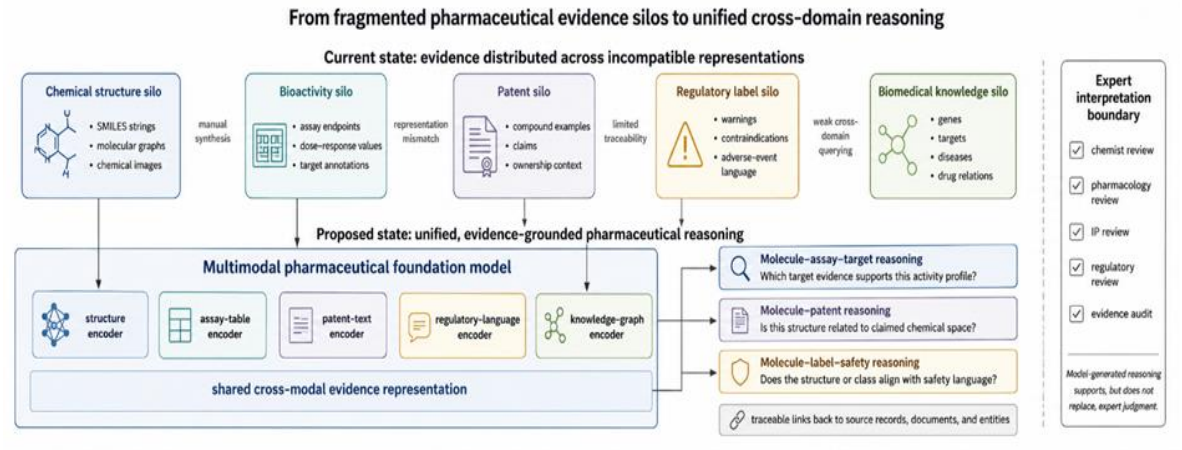

Figure 1. From Pharmaceutical Evidence Silos to Cross-Domain Foundation-Model Reasoning

Foundation-model methods have begun to connect molecular representations with biomedical and scientific language, creating a basis for broader multimodal pharmaceutical reasoning. SMILES Transformer demonstrated that molecular strings can be pre-trained as language-like sequences [9], MolGPT extended transformer decoding toward molecular generation [10], and Text2Mol showed that molecule retrieval can be driven by natural-language queries [11]. Later molecule–language translation work suggested that chemical structures and textual descriptions can be mapped across modalities [12], while structure–text systems such as KV-PLM and multimodal molecule structure–text models showed how biomedical text can be aligned with molecular information [13, 14]. These advances indicate that the core representational ingredients exist, but they remain narrower than the full pharmaceutical spectrum of structures, assays, patents, and regulatory documents.

The central thesis of this article is that a multimodal pharmaceutical foundation model could jointly encode chemical structures, assay tables, patent documents, and regulatory labels to support integrated reasoning over evidence that is currently hidden across silos. Such a system would not merely retrieve documents; it would align a molecule's graph or SMILES representation with assay endpoints, patent examples, label warnings, and knowledge-graph entities. Chemical image recognition systems such as DECIMER.ai could help recover structures from legacy documents [15], while regulatory adverse-event extraction resources and systems demonstrate that label text can be converted into structured evidence [16, 17]. Knowledge-graph resources such as PharmKG and precision-medicine knowledge graphs show how linked biomedical entities can support reasoning, but a learnable multimodal foundation model could make those links dynamic, queryable, and grounded in heterogeneous source evidence [18, 19].

Background

Pharmaceutical Data as Four Modalities

Pharmaceutical knowledge can be viewed as four interacting modalities: chemical structure, quantitative assay evidence, patent prose, and regulatory text. Chemical structure may be encoded through curated databases such as PubChem [1], ChEMBL [2], and DrugBank [3], while assay evidence often appears as heterogeneous endpoint-value-unit records that require interpretation in biological and experimental context. Patent documents contain claims, examples, Markush-like descriptions, and synthesis procedures, while regulatory labels contain structured sections on indications, contraindications, warnings, adverse reactions, and clinical pharmacology. A foundation model for pharmaceutical knowledge extraction must therefore preserve modality-specific meaning while also learning when a compound name, structure, assay result, claim, and safety statement refer to the same scientific object.

Foundation Models and Contrastive Multimodal Learning

Multimodal foundation models rely on separate encoders or shared transformer backbones that convert different data types into comparable representations. In the pharmaceutical setting, contrastive learning could bring together a molecule and its description, an assay and its target context, or a label section and its corresponding drug entity, extending the alignment logic demonstrated in molecule–text retrieval [11] and molecule–language translation [12]. KV-PLM showed that molecule structure and biomedical text can be bridged in a deep-learning system designed for cross-modal comprehension [13], while recent multimodal molecule structure–text modeling suggests that retrieval and editing can operate through a shared chemical-textual space [14]. These examples motivate a broader architecture in which patent paragraphs, regulatory sections, and assay rows are treated as additional views of pharmaceutical knowledge rather than as downstream annotations.

Chemical Language Models and Molecular Encoders

Molecular encoders provide the structural foundation for the proposed system because they transform chemical graphs, fingerprints, or strings into machine-readable representations. Learned molecular representation analysis has shown that graph-based methods can capture useful structure–property relationships [6], while SMILES Transformer framed molecular strings as pre-trainable sequences for low-data discovery settings [9]. MolGPT further demonstrated that transformer-decoder models can operate directly over molecular strings for generative chemistry [10], and the Molecular Transformer showed that chemical reaction information can be represented through transformer architectures [20]. A pharmaceutical multimodal foundation model would use these molecular encoders as one layer of a larger representational system that also handles assay measurements, patent language, and regulatory documents.

Tabular and Document Understanding for Pharmaceutical Text

Pharmaceutical evidence is frequently embedded in tables, scanned pages, figure panels, and semi-structured documents rather than in clean database fields. ChemDataExtractor 2.0 illustrates how scientific text can be converted into structured, ontology-linked records [21], while automated synthesis-action extraction demonstrates how experimental procedures can be parsed into machine-actionable steps [22]. Chemical image recognition systems such as DECIMER and its hand-drawn molecule image dataset show how document images can contribute molecular structures that are otherwise inaccessible to text-only NLP systems [23, 24]. Regulatory text mining also requires specialized document understanding, as shown by structured product label annotation for adverse drug reactions [16], adverse-event extraction evaluation from drug labels [17], and BERT-based classification of drug-induced liver injury risk from labeling documents [25].

Previous Efforts to Link Disparate Pharmaceutical Data Sources

Earlier integration efforts have often relied on curated databases, ontologies, and knowledge graphs to connect compounds, targets, diseases, and adverse events. PharmKG provides a dedicated biomedical knowledge-graph benchmark for data mining [18], while knowledge-graph completion has been applied conceptually to drug repurposing by linking biomedical entities through graph structure [26]. Precision-medicine knowledge graphs demonstrate how heterogeneous biomedical evidence can be organized into a unified resource for downstream analysis [19]. These systems are essential foundations, but a multimodal foundation model would go beyond fixed graph edges by learning representations from structures, tables, patent language, regulatory labels, and document images, allowing new associations to be retrieved and interrogated even before they have been manually encoded.

Model System Overview

High-Level Architecture

The proposed system would use a shared pharmaceutical representation backbone that accepts tokenized chemical structures, assay table rows, patent passages, and regulatory label sections. Molecules could be represented by graph encoders or SMILES transformers informed by molecular representation learning [6], SMILES pre-training [9], and transformer-based molecular generation [10]. Patent and regulatory text could be processed by biomedical and scientific transformer encoders derived from BioBERT [5] and SciBERT [8], while chemical-structure extraction from document images could be supported by DECIMER-style recognition pipelines [15]. A decoder head would then support retrieval-augmented generative question answering, with retrieved structures, assay rows, patent snippets, and label sections supplied as evidence rather than as unsupported model memory.

Figure 2 illustrates the proposed multimodal pharmaceutical foundation-model architecture, showing how chemical structures, assay tables, patent documents, and regulatory labels could be aligned into a shared representation space for retrieval-grounded cross-domain question answering.

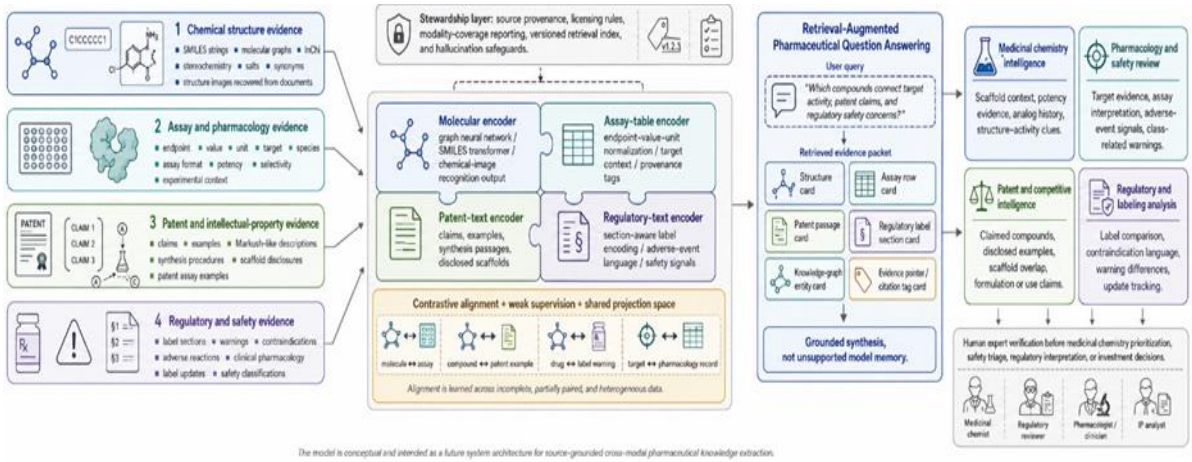

Figure 2. Multimodal Pharmaceutical Foundation Model for Cross-Domain Knowledge Extraction from Chemical Structures, Assay Evidence, Patent Documents, and Regulatory Text

Core Modalities and Their Encodings

The core encodings would treat molecules as graphs or SMILES strings, assays as normalized endpoint-value-unit sequences, patents as segmented legal and technical text, and regulatory labels as section-aware document sequences. Public chemical resources such as ChEMBL [2], PubChem [1], and DrugBank [3] would provide canonical identifiers and structure records, while drug–target modeling frameworks such as DeepPurpose indicate how compound and target representations can be combined for pharmacological tasks [5]. Patent examples and synthesis descriptions would benefit from automated chemical procedure extraction [22], and regulatory sections would require adverse-event-aware encoders informed by structured product label annotation [16]. All encodings would be projected into a common space so that a researcher could move from a structure to a relevant assay, from an adverse reaction to related compounds, or from a patent example to a regulatory warning.

Design Principles

The system should be modality-fair, evidence-grounded, robust to missing information, and usable through a natural-language interface. Modality fairness is important because text-rich patents or labels could otherwise dominate sparse but critical assay signals, while missing-modality handling is required because many investigational molecules lack complete regulatory or clinical records. Cross-modal molecule retrieval [11], molecule–language translation [12], and multimodal structure–text modeling [14] suggest that paired and weakly paired data can still support alignment when not every entity has all modalities available. Evidence grounding should rely on retrievable source units such as assay identifiers, patent paragraphs, document images, or label sections, reflecting the traceability requirements demonstrated in regulatory adverse-event extraction work [17, 27].

Data Modalities and Preprocessing

Chemical Structure Ingestion

Chemical structure ingestion would begin by parsing SMILES, InChI, names, salts, stereochemical variants, and synonyms from curated sources such as ChEMBL [2], DrugBank [3], and PubChem [1]. Standardization would harmonize tautomers, remove non-informative counterions when appropriate, preserve stereochemical information, and connect each normalized structure to identifiers used in assays, patents, and labels. Molecular pre-training approaches such as SMILES Transformer [9] and MolGPT [10] support the idea that string-based molecular representations can be used as foundation-model inputs, while graph-based molecular representation studies show why structural topology should also be preserved [6]. For patents and scanned literature, DECIMER-style chemical image recognition could recover structures that are present only as drawings, making document-derived chemistry available for multimodal alignment [15, 23].

Tabular Assay Data Extraction and Normalization

Assay preprocessing would convert heterogeneous bioactivity records into standardized machine-readable sequences that retain endpoint, value, unit, target, species, assay format, and measurement context. ChEMBL provides a natural foundation for curated bioactivity information [2], while MoleculeNet shows how molecular datasets can be organized for predictive tasks even when they originate from diverse experimental sources [4]. DeepPurpose illustrates the value of pairing drug and target representations for interaction prediction [5], but the proposed model would additionally preserve assay provenance and textual context so that retrieval remains auditable. Normalization would avoid forcing all assay results into a single potency scale when the underlying biological interpretation differs, instead encoding the measurement as structured evidence that can be compared, filtered, or retrieved.

Patent Document Processing

Patent processing would retrieve full-text patent documents, segment them into claims, descriptions, synthesis examples, and assay examples, and link mentioned compounds to canonical molecular structures. Automated extraction of synthesis actions from experimental procedures provides a precedent for turning chemistry prose into structured process information [22], while ChemDataExtractor 2.0 shows how scientific entities and relationships can populate machine-readable ontological forms [21]. Chemical image recognition through DECIMER and related datasets would support structure extraction from patent figures and legacy scans where molecules are shown as images rather than text [23, 24]. The resulting patent representation would allow a query to connect a compound or scaffold with claimed uses, example compounds, formulation language, and disclosed assay evidence without treating the patent as an unstructured text blob.

Regulatory Text and Structured Product Labels

Regulatory preprocessing would preserve the section structure of FDA Structured Product Labels, EMA Summary of Product Characteristics documents, and related regulatory texts so that retrieved evidence remains specific and verifiable. Annotated structured product label datasets for adverse drug reactions demonstrate how label sections can be converted into supervised extraction targets [16], and ADE Eval shows why evaluation of adverse-event extraction from labels requires domain-specific benchmarks [17]. BERT-based labeling studies for drug-induced liver injury risk indicate that regulatory text can support specialized safety classification when section language is modeled carefully [25], while LabelComp shows how artificial intelligence can identify adverse-event changes in FDA labeling [27]. A multimodal system would treat each label section as a source-grounded evidence unit linked to a product, ingredient, chemical structure, and pharmacological class.

Multimodal Architecture and Alignment

Modality Encoders and Shared Projection Space

The architecture would assign separate encoders to molecular graphs or SMILES, assay-table sequences, patent passages, document images, and regulatory-label sections before projecting their outputs into a shared latent space. Molecular encoders could draw on graph-based representation learning [6], SMILES pre-training [9], and transformer decoding over chemical strings [10], while text encoders could build on biomedical and scientific language models such as BioBERT [7] and SciBERT [8]. Document-image components would use chemical image recognition systems such as DECIMER.ai to convert structure depictions into molecular representations [15], and regulatory encoders would use label-aware signals from adverse-event extraction resources [16, 17]. In the shared projection space, similarity should indicate evidence relevance across modalities, so that a molecule, assay row, patent example, and warning section can be retrieved together when they support the same pharmaceutical question.

Table 1 defines how each pharmaceutical evidence modality contributes distinct representational content, preprocessing requirements, alignment signals, and retrieval value within the proposed multimodal foundation model.

Table 1. Cross-Modal Evidence Architecture for a Pharmaceutical Foundation Model

Evidence modality	Primary knowledge carried	Preprocessing and normalization logic	Encoder or representation strategy	Cross-modal alignment signal	Retrieval value for pharmaceutical reasoning	Main failure risk
Chemical structure	Molecular identity, scaffold, stereochemistry, salts, tautomers, analog relationships, chemical similarity, and structure-derived hypotheses	Standardize SMILES, InChI, salts, stereochemical variants, synonyms, and document-derived structure images; preserve identifiers from PubChem, ChEMBL, and DrugBank where available [1-3]	Molecular graph encoder, SMILES transformer, fingerprint embedding, or hybrid graph–sequence encoder informed by learned molecular representation and chemical language modeling approaches [6, 9, 10]	Compound identity, synonym matching, structure–description pairs, assay-linked compounds, patent examples, and regulatory active ingredients	Allows the system to move from a molecule or scaffold to assay records, patent examples, safety language, and related drug-class evidence	Structure ambiguity, salt or tautomer mismatch, image-recognition error, synonym collision, or loss of stereochemical specificity
Assay and pharmacology tables	Endpoint, potency, activity, target, species, assay format, value, unit, confidence, and biological context	Normalize endpoint–value–unit records without collapsing biologically different assays into a single artificial score; retain target, species, measurement type, and provenance [2, 4, 5]	Tabular sequence encoder, endpoint-aware transformer, target-context encoder, or structured evidence embedding	Molecule–assay pairs, target names, bioactivity records, endpoint semantics, and drug–target interaction context	Supports mechanistic and pharmacological interpretation by linking structures to observed activity rather than relying on text descriptions alone	False comparability across assays, unit inconsistency, sparse target evidence, noisy high-throughput measurements, or missing experimental context
Patent documents	Claims, examples, scaffold disclosures, synthesis procedures, formulation language, therapeutic-use statements, and competitive-intelligence context	Segment claims, descriptions, examples, synthesis passages, and assay examples; extract compounds from text and figures using chemical NLP and image-recognition pipelines [15, 21-24]	Patent-text encoder, claim-aware transformer, synthesis-procedure parser, structure-image recognition module, and document-section embedding	Compound mentions, disclosed examples, scaffold similarity, synthesis routes, assay examples, and patent-family metadata	Enables cross-domain answers that connect molecular or assay evidence with intellectual-property constraints and disclosed chemical space	Overinterpretation of claim scope, weak linkage between examples and claims, legal language ambiguity, OCR error, or incomplete patent-family coverage
Regulatory and safety text	Indications, contraindications, warnings, adverse reactions, clinical pharmacology, safety classifications, and label updates	Preserve section structure of labels and regulatory documents; extract adverse-event and safety language using section-aware regulatory NLP [17, 25, 27, 28]	Regulatory-text encoder, adverse-event extraction model, label-section classifier, and safety-signal embedding	Drug ingredient, product label, pharmacological class, adverse-event terms, contraindication language, and label-change signals	Allows safety, pharmacovigilance, and regulatory questions to be grounded in exact label sections rather than generic biomedical text	Hallucinated safety synthesis, section misclassification, failure to distinguish label language from inference, or missed label updates

Knowledge-graph entities and identifiers	Links among compounds, targets, diseases, pathways, adverse events, mechanisms, drug classes, and external database identifiers	Harmonize identifiers across DrugBank, PubChem, ChEMBL, PharmKG, and other biomedical knowledge resources [1-3, 18, 19, 26]	Entity embedding, graph neural network, knowledge-graph completion layer, or graph-augmented retrieval index	Shared identifiers, ontology links, compound–target–disease edges, adverse-event relationships, and biomedical co-occurrence	Provides a bridge across incomplete modalities and supports query expansion, entity disambiguation, and hypothesis generation	Propagation of curated-graph bias, outdated edges, incomplete coverage, or unsupported inference from weak graph links
Document images and scanned evidence	Molecules, tables, figures, chemical schemes, legacy patent drawings, scanned labels, and embedded experimental information	Use chemical image recognition, table extraction, OCR, and document-layout parsing while preserving source-page provenance [15, 23, 24]	Image-to-structure pipeline, layout-aware document encoder, table-recognition module, and image-derived evidence embedding	Figure–caption links, image-derived molecules, patent examples, scanned assay tables, and source-page anchors	Recovers pharmaceutical evidence that is unavailable in clean text or database fields	OCR error, incorrect structure recognition, table boundary error, low-quality scans, or missing context around extracted images
Natural-language user queries	Research intent, cross-domain question structure, constraints, target class, safety concern, scaffold interest, or regulatory-information need	Parse user query into entities, modality targets, filters, and evidence requirements; distinguish factual retrieval from inferential synthesis	Query encoder projected into the shared pharmaceutical embedding space	Query–evidence similarity, entity recognition, task intent, and modality-specific retrieval constraints	Enables conversational access to integrated pharmaceutical evidence across structures, assays, patents, and labels	Ambiguous query wording, unsupported inference request, missing evidence, or answer generation beyond retrieved source support

Contrastive Training Objectives

Contrastive objectives would align co-occurring or semantically equivalent information while separating unrelated evidence units. A molecule and its natural-language description could be aligned using principles demonstrated by Text2Mol [11] and molecule–language translation [12], while biomedical structure–text alignment could build from KV-PLM-style pre-training [13] and multimodal molecule structure–text modeling [14]. Patent examples, assay measurements, and label sections would provide weakly paired contexts, allowing the model to learn that a compound disclosed in a patent and the same active ingredient in a regulatory label are related even when their wording differs. The training objective would be conceptual rather than result-driven: it should encourage a semantically organized embedding space suitable for retrieval, evidence grounding, and downstream question answering.

Handling Unpaired and Partially Observed Data

Pharmaceutical records are often incomplete, so the model must handle entities that have structures and assays but no approved label, patents but no curated assay table, or regulatory text without full public experimental context. Weak supervision can exploit database identifiers, chemical names, synonyms, target names, and document co-occurrence signals from resources such as DrugBank [3], PubChem [1], ChEMBL [2], and PharmKG [18]. Knowledge-graph completion for drug repurposing shows how latent relationships can be inferred across incomplete biomedical graphs [26], and precision-medicine knowledge graphs show how heterogeneous sources can be integrated even when evidence density varies by entity [19]. During training, missing modalities would be masked rather than imputed as facts, while self-supervised reconstruction and retrieval tasks would help the system learn useful representations without hallucinating unavailable evidence.

Knowledge Retrieval and Question Answering

Retrieval-Augmented Generation Integration

For a user query, the system would first encode the question into the shared pharmaceutical embedding space and retrieve relevant molecules, assay records, patent passages, document-derived structures, and regulatory sections. Molecule–text retrieval methods such as Text2Mol show how natural language can be used to retrieve chemical structures [11], while multimodal molecule structure–text models suggest that aligned embeddings can support retrieval before generation [14]. The retrieved evidence would then be passed to a decoder language model that synthesizes an answer while preserving links to the underlying source units, rather than relying on unsupported parametric memory. This design would be especially important for regulatory and safety questions, where label-derived adverse-event evidence must remain traceable to the exact document section that supports it [16, 17].

Example Queries and Use Cases

A query such as “list CDK4/6 inhibitors with relevant patent claims and reported severe neutropenia in regulatory labeling” would require the model to connect target activity, chemical identity, patent scope, and adverse-event terminology. Drug and

target encoders such as those used in DeepPurpose illustrate how pharmacological relationships can be modeled computationally [5], while DrugBank provides drug-centric identifiers that can connect approved products with targets, mechanisms, and external references [3]. Regulatory label-mining systems for adverse-event extraction and label-change detection show why safety language must be interpreted as structured evidence rather than as generic text [17, 27]. A multimodal system could therefore retrieve compound-level evidence, patent passages, and label sections together, enabling a researcher to inspect the integrated answer rather than repeat separate searches manually.

Evidence Grounding and Citation

Evidence grounding would require every generated answer to identify the specific source units used in the response, such as a ChEMBL assay record, a patent example, a PubChem structure entry, or a regulatory label section. ChEMBL and PubChem provide complementary anchors for bioactivity and chemical identity [1, 2], while document-understanding tools such as ChemDataExtractor 2.0 can convert scientific text into structured evidence suitable for provenance tracking [21]. Regulatory adverse-event resources further show that source attribution is not optional when extracted statements may influence safety interpretation [16, 17]. The model should therefore expose evidence pointers as part of the answer interface, allowing the user to verify whether a claim was supported by assay data, patent text, regulatory labeling, or a combination of modalities.

Model Stewardship and Update Strategy

Continuous Data Ingestion and Model Refresh

A pharmaceutical foundation model would require continuous ingestion because assays, patents, labels, and knowledge graphs change over time. New chemical records from resources such as PubChem [1], updated bioactivity annotations from ChEMBL [2], and evolving drug-centric information from DrugBank [3] would need to be mapped into stable identifiers before model refresh. Regulatory updates would require section-aware comparison, and AI systems such as LabelComp demonstrate how changes in FDA labeling can be detected and organized for review [27]. Each model version should preserve its training snapshot, preprocessing rules, and retrieval index state so that prior answers can be reproduced and audited.

Bias and Completeness Monitoring

Bias monitoring would examine whether the system over-represents well-studied targets, approved products, English-language patents, text-rich labels, or therapeutic areas with stronger public curation. Knowledge-graph resources such as PharmKG show how biomedical entities can be linked across domains [18], but graph coverage can still vary by disease, drug class, and evidence type. Precision-medicine knowledge graphs similarly demonstrate the value of heterogeneous integration while highlighting the need to track provenance and coverage across sources [19]. The model should therefore report modality coverage for retrieved answers and flag cases where a response is driven mainly by one evidence channel, such as patent prose without supporting assay context or regulatory text without clear chemical linkage.

Integration Into Drug Discovery and Regulatory Workflows

Decision Support for Medicinal Chemists and Pharmacologists

In discovery workflows, the model could serve as an interactive knowledge assistant during compound design, target review, and competitive-intelligence discussions. Molecular representation learning can support structure-based reasoning [6], SMILES and transformer models can represent chemical series in language-like form [9, 10], and molecule–language alignment can connect structures to textual descriptions of function or phenotype [12, 13]. When combined with assay normalization, patent processing, and label retrieval, these representations could allow a chemist to ask whether a scaffold has public potency evidence, relevant intellectual-property constraints, or known class-related safety concerns. The system would not replace expert judgment, but it could reduce the cognitive burden of assembling evidence from separate scientific, chemical, patent, and regulatory sources.

Regulatory Intelligence and Label Harmonization

For regulatory intelligence, the model could compare label sections across drugs in the same class and identify differences in warnings, adverse reactions, contraindications, and clinical-use language. Structured product label annotation for adverse reactions provides a foundation for extracting safety-relevant information from labels [16], while BERT-based analysis of drug-induced liver injury language shows that specialized encoders can interpret regulatory text for risk-oriented tasks [25]. OnSIDES demonstrates how natural-language processing can extract adverse-event information from drug labels into a database-like form [28]. A multimodal extension would connect those label-derived signals to chemical structures, mechanisms, assay profiles, and patent landscapes, helping pharmacovigilance teams identify inconsistencies that merit expert review.

Evaluation Strategy

Cross-Modal Retrieval Accuracy

The evaluation strategy should test whether a query from one modality retrieves appropriate evidence from another modality, such as a label warning retrieving related molecules or a patent example retrieving associated assay evidence. MoleculeNet provides a precedent for benchmark-driven evaluation in molecular machine learning [4], while Text2Mol and molecule–

language translation work provide conceptual models for evaluating cross-modal retrieval between natural language and molecular structures [11, 12]. For the proposed system, retrieval should be assessed at the level of evidence relevance and provenance quality rather than as a stand-alone ranking exercise. Expert review would be needed to judge whether retrieved structures, assay rows, patent passages, and label sections are scientifically meaningful for the pharmaceutical question being asked.

Question-Answering Quality

Question-answering evaluation should use curated cross-modal questions that require evidence synthesis across chemical, assay, patent, and regulatory sources. Biomedical and scientific language models such as BioBERT and SciBERT provide useful baselines for text interpretation [7, 8], but the proposed task would require answers grounded in multiple modalities rather than biomedical prose alone. Regulatory extraction benchmarks such as ADE Eval demonstrate the need for domain-specific validation when answers involve adverse-event language [17], and label-comparison systems show why generated safety statements must be checked against source text [27]. Evaluation should therefore assess factual correctness, completeness, evidence grounding, and whether the answer clearly distinguishes retrieved evidence from model inference.

Table 2 consolidates the evaluation and governance requirements needed to distinguish reliable cross-modal pharmaceutical knowledge extraction from unsupported multimodal generation.

Table 2. Evaluation and Governance Framework for Source-Grounded Multimodal Pharmaceutical Question Answering

Evaluation or governance dimension	What must be tested	Recommended assessment design	Minimum evidence standard	Why it strengthens the conceptual model	Consequence if ignored
Cross-modal retrieval accuracy	Whether a query from one modality retrieves scientifically relevant evidence from another modality, such as a label warning retrieving related compounds or a patent example retrieving assay records	Curated benchmark of molecule-to-assay, assay-to-patent, patent-to-label, and label-to-structure retrieval tasks; include expert relevance grading rather than only ranking metrics	Retrieved items must include source identifiers, modality labels, and evidence-unit provenance	Converts the model from a generic embedding system into a testable pharmaceutical knowledge-retrieval architecture	The system may retrieve semantically similar but scientifically irrelevant evidence
Evidence-grounded answer quality	Whether generated responses are supported by retrieved structures, assay rows, patent passages, and regulatory sections	Expert-reviewed question-answering set requiring synthesis across at least two modalities; compare against text-only biomedical language-model baselines such as BioBERT and SciBERT [7, 8]	Each factual claim must be traceable to a retrieved source unit or explicitly labeled as inference	Ensures the decoder acts as a synthesis layer rather than unsupported model memory	Plausible hallucinations may be mistaken for verified pharmaceutical knowledge
Assay-context preservation	Whether potency, activity, endpoint, unit, species, target, and assay format remain interpretable after model encoding and retrieval	Challenge set with similar compounds but different assay formats, endpoints, species, or measurement units	Assay evidence must retain endpoint, value, unit, target, species, assay format, and provenance	Prevents the model from flattening heterogeneous bioactivity records into misleading similarity scores	Assay results may be compared incorrectly, producing false mechanistic or prioritization signals
Patent-scope discipline	Whether the system distinguishes disclosed examples, claimed compounds, synthesis procedures, and broader legal claim language	Patent-evidence review using segmented claims, examples, and descriptions; include legal or IP expert assessment for claim-boundary interpretation	Output must identify whether evidence came from a claim, example, description, synthesis passage, or assay disclosure	Protects the system from treating all patent text as equivalent scientific or legal evidence	Competitive-intelligence outputs may overstate freedom-to-operate risks or claimed chemical coverage
Regulatory-section fidelity	Whether warning, contraindication, adverse-reaction, indication, and clinical-pharmacology sections are retrieved and interpreted correctly	Section-aware regulatory extraction benchmark using structured product labels, ADE extraction resources, and label-change examples [16, 17, 25, 27, 28]	Safety-related statements must cite or point to the exact regulatory section from which they were derived	Supports safe use in pharmacovigilance, label harmonization, and regulatory-intelligence workflows	Generated safety summaries may blur label wording, inferred risk, and unsupported interpretation
Missing-modality robustness	Whether the model handles compounds with structures and assays but no label, patents with no curated assay table, or labels with incomplete public experimental context	Masked-modality evaluation in which one or more evidence channels are intentionally removed during retrieval and answering	The answer must report missing modalities and avoid imputing unavailable evidence as fact	Makes the conceptual model realistic for investigational compounds and incomplete public records	The system may hallucinate absent patents, assays, or regulatory evidence

Modality-balance monitoring	Whether abundant text evidence overwhelms sparse but important assay or structure evidence	Coverage reports showing how much each answer depends on structure, assay, patent, regulatory, and knowledge-graph channels	Every answer should disclose modality coverage and evidence imbalance when relevant	Adds transparency to cross-domain reasoning and helps experts judge confidence	Text-rich patents or labels may dominate sparse quantitative evidence
Versioning and update auditability	Whether answers can be reproduced after database, patent, label, or index updates	Maintain model-version, retrieval-index, preprocessing-rule, and source-snapshot metadata for each generated answer	Each answer should be linked to a model version and retrieval-index state	Supports reproducibility, regulatory audit, and longitudinal label or patent monitoring	Prior answers may become impossible to audit after data refresh
Licensing and access compliance	Whether source use respects database licenses, patent-document access rules, copyright limits, and institutional restrictions	Data-governance review before ingestion; classify sources by permissible training, indexing, display, and redistribution use	The system must document source availability, licensing status, and permitted downstream use	Addresses one of the manuscript's most important implementation constraints	A technically strong model may be unusable because of legal or institutional data restrictions
Human expert review boundary	Whether the model is used as decision support rather than an autonomous authority for chemistry, safety, regulatory, or IP decisions	Workflow simulation with medicinal chemists, pharmacologists, regulatory scientists, and IP analysts reviewing model outputs	Outputs influencing prioritization, safety triage, regulatory interpretation, or investment decisions must require expert verification	Aligns the model with high-stakes pharmaceutical practice and reduces automation risk	Users may overtrust generated synthesis in decisions requiring professional judgment

Impact on Workflow Efficiency

Workflow evaluation should examine how the model changes expert evidence synthesis rather than treating the system as an autonomous decision-maker. Knowledge-graph completion for drug repurposing shows how integrated biomedical evidence can support hypothesis generation [26], while precision-medicine knowledge graphs demonstrate how connected resources can help organize complex biomedical relationships [19, 29]. A model-augmented workflow could be compared conceptually with manual searching across chemical databases, patent systems, literature engines, and regulatory labels, with attention to the quality and traceability of the final answer. Any workflow study should preserve human review, especially when outputs influence medicinal chemistry prioritization, pharmacovigilance triage, or regulatory interpretation.

Limitations

Data Licensing, Copyright, and Access

Data access is a major limitation because high-value pharmaceutical evidence may be distributed across public databases, proprietary repositories, paywalled literature, patent services, and regulatory portals with different usage conditions. Public resources such as PubChem, ChEMBL, and DrugBank provide essential foundations [1-3], but they do not eliminate the legal and operational complexity of training on full-text documents or redistributing processed corpora. Patent and document-image processing can be aided by tools such as DECIMER.ai [15] and ChemDataExtractor 2.0 [21], yet access to source documents and derived annotations may still vary across jurisdictions and institutions. Reproducibility would therefore depend not only on model architecture but also on transparent documentation of licensing, preprocessing, source availability, and permissible downstream use.

Modality Imbalance and Hallucination Risks

A multimodal pharmaceutical model may over-rely on abundant text while underweighting sparse quantitative assay evidence or ambiguous chemical-structure links. Molecular encoders and chemical language models can represent structures effectively [6, 29], but they do not guarantee that a generated answer will respect assay uncertainty, patent claim boundaries, or regulatory wording. Regulatory NLP systems for adverse-event extraction and label-change analysis show that even focused safety tasks require careful validation and expert oversight [17, 27, 28]. If retrieval fails or evidence is incomplete, the decoder could produce plausible but incorrect synthesis, so safety-critical use should require source inspection, uncertainty communication, and human approval before any clinical, regulatory, or investment action.

Conclusion

A multimodal foundation model for pharmaceutical knowledge extraction would unify chemical structures, assay records, patent documents, and regulatory text into a single conversational knowledge environment. Its central contribution would be the ability to connect representations that are currently stored, searched, and interpreted separately. By aligning these modalities in a shared evidence space, the model could support questions that require structural, pharmacological, intellectual-property, and safety reasoning at the same time.

The strongest value of the proposed system would come from cross-modal reasoning and retrieval-grounded answers. Instead of asking researchers to move manually from a molecule to an assay table, from an assay table to a patent, and from a patent to a label, the system would retrieve the relevant evidence together. This could support use cases across discovery, development, competitive intelligence, regulatory review, and post-market surveillance.

Important challenges would remain before such a system could be trusted in practice. Data access, copyright, licensing, modality imbalance, incomplete evidence, and hallucination risk would all need careful governance. Domain experts would also need to validate generated answers before they are used for medicinal chemistry prioritization, regulatory interpretation, safety review, or clinical decision support.

A pre-competitive consortium could provide the best path toward building and maintaining this type of pharmaceutical foundation model. Shared benchmarks, transparent reporting, reproducible preprocessing, and source-grounded evaluation would help the field distinguish useful cross-modal reasoning from unsupported generation. With appropriate stewardship, a multimodal pharmaceutical knowledge model could become a trusted layer for navigating the world's distributed drug discovery and regulatory evidence.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47(D1):D1102-9.
2. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45(D1):D945-54.
3. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074-82.
4. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30.
5. Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J. DeepPurpose: a deep learning library for drug–target interaction prediction. Bioinformatics. 2020;36(22-23):5545-7.
6. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88.
7. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.
8. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019;3615-20.
9. Honda S, Shi S, Ueda HR. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. arXiv:1911.04738. 2019.
10. Bagal V, Aggarwal R, Vinod PK, Priyakumar UD. MolGPT: molecular generation using a transformer-decoder model. J Chem Inf Model. 2021;62(9):2064-76.
11. Edwards C, Zhai C, Ji H. Text2mol: Cross-modal molecule retrieval with natural language queries. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021;595-607.
12. Edwards C, Lai T, Ros K, Honke G, Cho K, Ji H. Translation between molecules and natural language. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022;375-413.
13. Zeng Z, Yao Y, Liu Z, Sun M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. Nat Commun. 2022;13(1):862.
14. Liu S, Nie W, Wang C, Lu J, Qiao Z, Liu L, et al. Multi-modal molecule structure–text model for text-based retrieval and editing. Nat Mach Intell. 2023;5(12):1447-57.
15. Rajan K, Brinkhaus HO, Agea MI, Zielesny A, Steinbeck C.decimer.ai : an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. Nat Commun. 2023;14(1):5045.
16. Demner-Fushman D, Shooshan SE, Rodriguez L, Aronson AR, Lang F, Rogers W, et al. A dataset of 200 structured product labels annotated for adverse drug reactions. Sci Data. 2018;5(1):180001.
17. Bayer S, Clark C, Dang O, Aberdeen J, Brajovic S, Swank K, et al. ADE eval: an evaluation of text processing systems for adverse event extraction from drug labels for pharmacovigilance. Drug Saf. 2021;44(1):83-94.
18. Zheng S, Rao J, Song Y, Zhang J, Xiao X, Fang EF, et al. PharmKG: a dedicated knowledge graph benchmark for biomedical data mining. Brief Bioinform. 2021;22(4):bbaa344.
19. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023;10(1):67.
20. Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci. 2019;5(9):1572-83.

21. Mavracic J, Court CJ, Isazawa T, Elliott SR, Cole JM. ChemDataExtractor 2.0: Autopopulated ontologies for materials science. J Chem Inf Model. 2021;61(9):4280-9.
22. Vaucher AC, Zipoli F, Geluykens J, Nair VH, Schwaller P, Laino T. Automated extraction of chemical synthesis actions from experimental procedures. Nat Commun. 2020;11(1):3601.
23. Rajan K, Zielesny A, Steinbeck C. DECIMER: towards deep learning for chemical image recognition. J Cheminform. 2020;12(1):65.
24. Brinkhaus HO, Zielesny A, Steinbeck C, Rajan K. DECIMER—hand-drawn molecule images dataset. J Cheminform. 2022;14(1):36.
25. Wu Y, Liu Z, Wu L, Chen M, Tong W. BERT-based natural language processing of drug labeling documents: a case study for classifying drug-induced liver injury risk. Front Artif Intell. 2021;4:729834.
26. Zhang R, Hristovski D, Schutte D, Kastrin A, Fiszman M, Kilicoglu H. Drug repurposing for COVID-19 via knowledge graph completion. J Biomed Inform. 2021;115:103696.
27. Neyarapally GA, Wu L, Xu J, Zhou EH, Dang O, Lee J, et al. Description and validation of a novel AI tool, LabelComp, for the identification of adverse event changes in FDA labeling. Drug Saf. 2024;47(12):1265-74.
28. Tanaka Y, Chen HY, Belloni P, Gisladottir U, Kefeli J, Patterson J, et al. OnSIDES database: Extracting adverse drug events from drug labels using natural language processing models. Med. 2025;6(7).
29. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife. 2017;6:e26726.