Pharmacophore

ISSN-2229-5402

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

AI LITERATURE SURVEILLANCE SYSTEM FOR EMERGING FORMULATION TECHNOLOGIES USING SEMANTIC SEARCH AND PATENT LINKAGE

Diego Morales[1]*, Andres Gutierrez[1], Lucia Navarro[2], Pablo Rios[1]

1. ***Department of Computational Pharmaceutical Sciences, Faculty of Pharmacy, University of Lima, Lima, Peru.***
2. ***Department of Intelligent Drug Engineering, Faculty of Medicine, Pontifical Catholic University of Peru, Lima, Peru.***

ARTICLE INFO

Received:
28 February 2026
Received in revised form:
24 June 2026
Accepted:
25 June 2026
Available online:
28 June 2026

Keywords: Artificial intelligence, Pharmaceutical formulation, Semantic search, Patent landscaping, Technology surveillance, Innovation intelligence

ABSTRACT

Pharmaceutical formulation technology evolves rapidly, with critical innovations often appearing in dispersed publications and patent documents before they become mainstream. An integrated surveillance system is needed to monitor both scientific and intellectual-property sources for early evidence of emerging platforms. Competitive intelligence and R&D planning often rely on manual literature reviews and separate patent searches. This separation can obscure the connection between early scientific reports and the patent activity that signals a technology's development trajectory. This article proposes an AI literature surveillance system that continuously ingests scientific publications and global patent documents. The system uses semantic search to interpret formulation-specific queries and links publications to related patents to reveal technology evolution over time. The proposed framework includes a dual-stream ingestion pipeline for literature and patents, a semantic search engine, a patent-literature linkage module, a technology-trend detection engine, and an analyst dashboard. These components are designed to support continuous monitoring rather than one-time retrospective review. The system would help formulation scientists detect nascent technologies such as new lipid nanoparticle compositions, printable excipient systems, continuous-manufacturing approaches, or long-acting delivery platforms. By linking technical evidence to intellectual-property activity, the system could support earlier assessment of maturity, ownership, and competitive relevance. AI-driven surveillance of literature and patents could shift formulation innovation from reactive awareness to proactive strategic intelligence. Such systems should be evaluated prospectively in real pharmaceutical innovation intelligence workflows.

To Cite This Article: Morales D, Gutierrez A, Navarro L, Rios P. AI Literature Surveillance System for Emerging Formulation Technologies Using Semantic Search and Patent Linkage. Pharmacophore. 2026;17(3):132-141. https://doi.org/10.51847/rgDsedVJdS

Introduction

Pharmaceutical formulation innovation increasingly develops across fragmented technical sources, including journal articles, preprints, conference outputs, patent publications, and regulatory-facing technical disclosures. Recent formulation platforms such as lipid nanoparticle delivery, three-dimensional printing, and continuous manufacturing show how scientific and engineering signals can emerge across different communication channels before they consolidate into recognizable product strategies [1]. Lipid nanoparticle systems for mRNA delivery illustrate the pace at which formulation knowledge can move from mechanistic research to platform-level industrial relevance [2]. Continuous monitoring is therefore essential because early signals may be distributed across formulation science, materials science, drug delivery, manufacturing, and intellectual-property documents.

Current competitive intelligence practice often depends on periodic keyword searches, analyst-curated watchlists, and manual categorization of retrieved records. These approaches can miss semantically related concepts when authors and patent applicants use different terms for similar formulation mechanisms, excipient functions, manufacturing processes, or delivery profiles [3]. Automated patent landscaping has shown that structured analysis of patent text can support technical mapping, but the formulation domain still requires systems that connect patents with scientific evidence rather than treating them as

Corresponding Author: Diego Morales; Department of Computational Pharmaceutical Sciences, Faculty of Pharmacy, University of Lima, Lima, Peru. E-mail: morales@gmail.com.

isolated information streams [4]. Manual review remains valuable, but it is difficult to scale when formulation technologies evolve through subtle combinations of materials, processing conditions, and therapeutic contexts.

The maturation of biomedical language models and scientific-document representation learning makes a continuous AI surveillance framework technically feasible. BioBERT and related biomedical transformer models demonstrate how domain-specific language representations can improve biomedical text mining by encoding scientific meaning beyond surface keywords [5]. Broader biomedical language models such as BioGPT and domain-specific pretraining frameworks further support natural-language understanding over complex scientific statements [6, 7]. Scientific representation models such as SciBERT and SPECTER also enable document-level retrieval based on scientific similarity, which is central to surveillance systems that must connect related publications even when terminology differs [8, 9].

This article proposes an AI literature surveillance system for emerging formulation technologies that combines semantic search over scientific literature with automated patent linkage. The system is conceptualized as an AI Systems/Frameworks contribution rather than an experimental performance report, so it emphasizes architecture, data flow, evaluation strategy, and governance rather than numerical results. Patent analytics methods, including BERT-based classification and deep-learning approaches to patent analysis, provide a foundation for linking formulation concepts to intellectual-property documents [4, 10]. By integrating semantic retrieval, patent-publication linkage, and analyst-facing dashboards, the proposed system would support R&D and strategy teams in detecting, interpreting, and prioritizing early technology signals.

Table 1 presents the architecture-to-intelligence translation logic through which the proposed surveillance system converts fragmented scientific and patent evidence into formulation-specific strategic intelligence.

Table 1. Architecture-to-Intelligence Translation Map for an AI Formulation Technology Surveillance System

System layer	Primary evidence handled	Analytical transformation added by the system	Formulation-specific intelligence produced	Why this adds value beyond manual search
Dual-stream evidence ingestion	Scientific articles, preprints, conference abstracts, patent applications, granted patents, claims, descriptions, examples, priority data, assignee information	Converts fragmented scientific and intellectual-property sources into synchronized surveillance streams	A continuously refreshed evidence base covering both research emergence and protection activity	Prevents literature review and patent search from remaining separate, delayed, or inconsistently monitored activities
Document normalization and segmentation	Titles, abstracts, full text, patent claims, patent descriptions, citation fields, classifications, author and assignee metadata	Cleans, deduplicates, segments, and enriches heterogeneous documents into comparable analytical units	Searchable passages linked to formulation entities, claims, dates, organizations, and evidence sources	Enables fine-grained retrieval of dosage-form mechanisms, excipient functions, manufacturing parameters, and claim language
Domain-adapted semantic indexing	Formulation terminology, biomedical language, patent language, delivery-route concepts, manufacturing terms	Embeds documents and passages by conceptual meaning rather than surface keywords	Retrieval space capable of finding related technologies despite vocabulary differences across papers and patents	Reduces missed signals caused by synonymy, applicant-specific language, or cross-disciplinary terminology
Natural-language query interface	Analyst questions about formulation platforms, materials, excipients, routes, release mechanisms, or processes	Converts formulation-specific questions into semantic searches with metadata filtering and re-ranking	Relevant documents, passages, patent claims, and source-grounded summaries aligned with user intent	Allows formulation scientists and strategy teams to ask surveillance questions without constructing complex Boolean strings
Publication–patent linkage module	Scientific citations, patent citations, shared technical language, claim overlap, assignee and inventor metadata	Creates explicit edges between publications and patents using citation, semantic, and metadata relationships	Linked publication–patent pairs showing how scientific concepts connect to protected inventions	Reveals whether an emerging scientific idea is associated with ownership, claims, filings, or competitive development
Technology clustering and taxonomy mapping	Linked evidence pairs, formulation entities, patent classifications, citation relationships, semantic similarity clusters	Groups documents into formulation themes and maps them to dosage form, delivery route, manufacturing method, excipient class, and therapeutic context	Interpretable technology clusters such as lipid nanoparticles, 3D-printed dosage forms, continuous manufacturing, or long-acting depots	Moves surveillance from isolated records toward structured understanding of technology families and platform evolution
Trend detection and maturity interpretation	Time-stamped publications, patent filings, assignee activity, citation patterns, convergence signals	Tracks weak signals, growth patterns, convergence, and strategic activity across evidence streams	Qualitative maturity labels such as early, consolidating, strategically active, crowded, or uncertain	Supports earlier discussion of whether a formulation technology is becoming technically and competitively relevant

Analyst dashboard and intelligence briefs	Linked records, cluster maps, evidence summaries, alerts, uncertainty indicators, source links	Converts system outputs into reviewable decision-support artifacts	Evidence dossiers, alert histories, watchlists, patent-landscape views, and technology briefs	Embeds surveillance outputs into R&D, patent strategy, competitive intelligence, and portfolio-planning workflows
Governance and evaluation layer	Source provenance, coverage warnings, expert feedback, retrospective validation cases, prospective pilot use	Applies traceability, uncertainty labeling, false-positive control, and expert review	Trustworthy, auditable intelligence outputs that remain anchored to source evidence	Protects against unsupported automated conclusions and reinforces the system as expert decision support

Background

Formulation Technology Innovation and the Need for Surveillance

Formulation technology innovation often begins as dispersed scientific work before it becomes visible as an organized platform. Three-dimensional printed drug products, for example, were described as a new chapter in pharmaceutical manufacturing because they combine formulation design, process control, and individualized dosage-form concepts [1]. Later reviews of three-dimensional printing in pharmaceuticals extended this view from drug development to frontline care, showing how early technical publications can signal broader changes in manufacturing and delivery models [11]. Lipid nanoparticle mRNA systems similarly demonstrate how advances in lipid composition, particle structure, and stability can become central to a major formulation platform [2]. Continuous manufacturing of oral solid dosage forms also shows how process technologies can progress through the literature before wider industrial adoption.

Semantic Search and Natural-Language Processing for Scientific Literature

Semantic search differs from keyword retrieval because it represents documents and queries by meaning, allowing related concepts to be retrieved even when they use different surface terms. In biomedical literature, BioBERT showed that transformer-based pretraining can encode biomedical terminology and improve text-mining tasks relevant to scientific discovery [5]. Domain-specific pretraining approaches such as PubMedBERT further support retrieval and question answering in biomedical contexts by adapting language models to the structure and vocabulary of biomedical text [6]. BioGPT extends these ideas toward generative biomedical text mining, while SciBERT and SPECTER support scientific-document understanding and citation-informed similarity search [7-9]. These models provide the conceptual basis for a formulation-focused search layer that can interpret natural-language questions about dosage forms, excipients, delivery systems, and manufacturing technologies.

Patent Analytics for Formulation Technology

Patent analytics is central to formulation surveillance because patent claims often describe compositions, excipient combinations, manufacturing parameters, and delivery mechanisms before they appear in product-facing summaries. Automated patent landscaping has demonstrated how machine learning can organize large patent corpora into interpretable technology maps [3]. BERT-based patent classification and deep-learning surveys for patent analysis show that transformer and neural retrieval approaches can support more nuanced classification of technical documents than traditional keyword methods [4, 10]. Pharmaceutical patent enrichment methods and drug-discovery-oriented patent landscaping further illustrate how patent text can be converted into structured intelligence for life-science decision-making [12, 13]. For formulation technologies, such methods would need to emphasize claim language, dosage-form terminology, chemical entities, delivery routes, and assignee behavior.

AI-Based Technology Forecasting and Innovation Detection

AI-based technology forecasting aims to detect weak signals, convergence patterns, and emerging clusters before they become obvious in mainstream scientific or commercial discourse. Supervised patent-analysis approaches have been proposed for forecasting emerging technologies by learning patterns in patent information [14]. Machine-learning-based semantic analysis can also support the identification of new technology convergence, which is relevant when formulation advances arise from combinations of materials science, device engineering, and pharmaceutical manufacturing [15]. Link-prediction methods for technological convergence and graph-convolutional approaches to patent text further suggest that future formulation opportunities may be inferred from evolving relationships among technologies [16, 17]. Reviews of technological forecasting from a complex-systems perspective reinforce the need to treat innovation as a dynamic stream rather than as a static set of documents [18].

Existing Literature-Surveillance and Patent-Linking Tools and Their Gaps

Existing biomedical information platforms support literature search, database integration, and drug-target intelligence, but they are not typically designed as formulation-specific surveillance systems that connect publications to patent trajectories. Systematic biomedical knowledge integration has shown how heterogeneous biomedical evidence can be organized to support prioritization and discovery [19]. DrugBank, ChEMBL, and Open Targets demonstrate the value of structured biomedical

knowledge resources, although their primary emphasis differs from formulation technology surveillance and patent-literature linkage [20-22]. Pandemic-scale retrieval benchmarks such as TREC-COVID illustrate how urgent biomedical questions can motivate specialized retrieval collections and evaluation workflows [23]. A formulation surveillance system would extend these ideas by integrating scientific articles, preprints, and patents around formulation concepts rather than diseases, targets, or compounds alone.

System Architecture Overview

High-Level Design

The proposed system has three main layers: a data-acquisition layer, an intelligence layer, and a presentation layer. The data-acquisition layer continuously harvests scientific publications and patent documents, while the intelligence layer performs semantic indexing, natural-language retrieval, patent linkage, clustering, and trend detection. Scientific-document models such as SPECTER support document-level similarity analysis, and patent-specific transformer approaches support technical classification of intellectual-property records 4, 9]. The presentation layer then converts these linked outputs into analyst-facing dashboards, alerts, and evidence dossiers that can be reviewed by formulation scientists and competitive intelligence teams.

Figure 1 illustrates the proposed AI surveillance architecture in which continuous scientific-literature ingestion, patent-document analysis, semantic search, publication–patent linkage, technology mapping, and analyst-facing intelligence briefs operate as an integrated decision-support system for emerging formulation technologies.

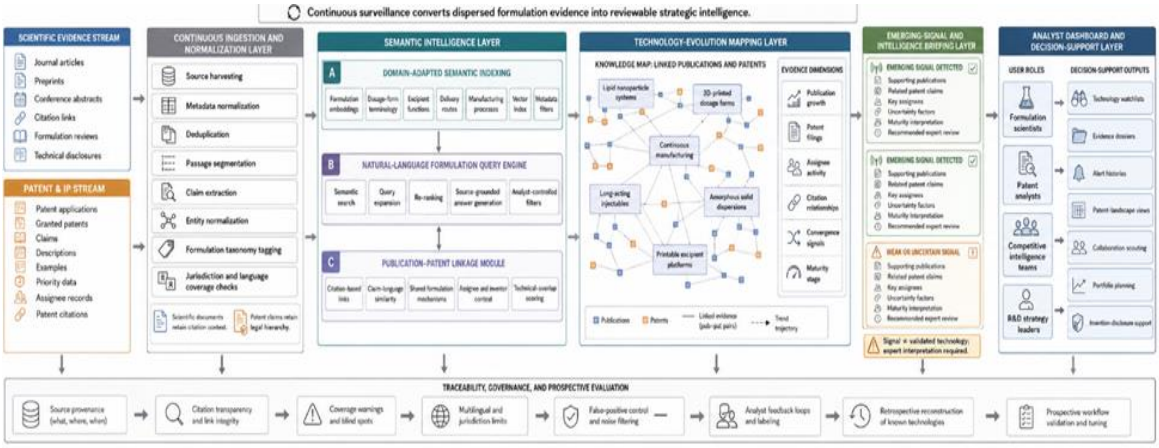

Figure 1. AI Literature Surveillance Architecture for Emerging Formulation Technologies Linking Scientific Evidence, Patent Activity, and Analyst Decision Support

Core Data Ingested and Outputs

The system ingests full-text scientific articles, preprints, conference abstracts, patent publications, patent grants, patent claims, patent descriptions, bibliographic metadata, assignee information, and citation relationships. Patent analytics work has emphasized that claims and classifications are especially important for understanding the protected technical scope of inventions [3, 10]. The expected outputs are structured technology records that connect key publications, related patents, probable formulation themes, assignees, research groups, and supporting evidence. These outputs should be treated as decision-support artifacts rather than automated conclusions, because biomedical knowledge systems require traceable evidence and expert interpretation [19].

Design Principles

The design principles are continuity, domain awareness, transparency, and workflow alignment. Continuity means that the system monitors new documents as they appear rather than relying only on scheduled manual reviews. Domain awareness requires language models and taxonomies that recognize formulation-specific terms, such as ionizable lipids, polymer matrices, amorphous dispersions, printable binders, depot systems, and process analytical technologies, building on domain-specific biomedical pretraining approaches [5, 6]. Transparency requires every generated insight to preserve links to source publications and patents, consistent with the evidence-oriented design of biomedical databases and knowledge platforms [20-22].

Data Ingestion and Preprocessing

Scientific Literature Ingestion

Scientific literature ingestion would continuously collect records from biomedical databases, preprint servers, publisher feeds, and citation indexes. Each record would be normalized to capture title, abstract, full text when available, authors, affiliations, publication date, journal, references, and cited-by relationships. Biomedical retrieval work such as TREC-COVID shows that rapid and structured literature collection can support specialized search tasks when the corpus is aligned with domain needs

[23]. For formulation surveillance, ingestion would prioritize drug delivery, pharmaceutics, manufacturing, materials science, and biotechnology sources so that early technical signals are not missed because they appear outside traditional formulation journals.

Patent Ingestion and Claim Extraction

Patent ingestion would retrieve publications and grants from major patent jurisdictions and parse bibliographic fields, claims, descriptions, examples, applicants, inventors, classifications, priority data, and citation links. Automated patent landscaping provides a basis for transforming unstructured patent text into analyzable technology spaces [3]. BERT-based patent classification and broader deep-learning methods for patent analysis can support automated categorization of patents into formulation-relevant domains such as drug delivery, dosage forms, excipients, manufacturing processes, and device-enabled administration [4, 10]. Pharmaceutical patent enrichment tools also show how patent text can be processed for life-science intelligence, although a formulation-specific system would need additional extraction rules for dosage-form composition, release mechanisms, and process parameters [12].

Normalization and Indexing

After ingestion, scientific and patent documents would be cleaned, deduplicated, segmented into semantically coherent passages, and enriched with normalized metadata. Each passage would be embedded with a domain-adapted language model, following the general logic of biomedical and scientific representation learning in BioBERT, PubMedBERT, SciBERT, and SPECTER [5, 6, 8, 9]. Patent passages would also retain claim hierarchy, classification codes, assignee fields, and priority relationships so that retrieval can combine semantic similarity with legal and technical metadata [4, 10]. The indexed corpus would support filtering by time period, technology class, organization, formulation type, route of administration, and evidence source.

Semantic Search and Query Interface

Semantic Query Understanding

The query interface would allow users to ask natural-language questions about emerging formulation technologies rather than construct complex Boolean strings. A formulation scientist might ask about sustained-release injectables, lipid nanoparticle compositions, printable oral dosage forms, or continuous-manufacturing methods, and the system would map the query to semantically related passages across literature and patents. Biomedical language models such as BioBERT and BioGPT support this kind of meaning-based interpretation because they encode domain-specific biomedical context rather than treating terms as isolated keywords [5, 7]. Scientific-document models such as SciBERT and SPECTER further help retrieve related publications that share conceptual content even when they differ in terminology or disciplinary framing [8, 9].

Domain-Adapted Retrieval and Ranking

The retrieval pipeline would combine vector search, metadata filtering, and re-ranking to prioritize records that are technically relevant to formulation innovation. Domain-specific pretraining provides a foundation for adapting retrieval to biomedical vocabulary, while patent-specific transformer classification supports the interpretation of specialized claim and description language [4, 6]. Deep-learning methods for patent analysis can help distinguish patents that merely mention a formulation term from those that actually claim a relevant dosage form, delivery system, or manufacturing process [10]. Technology-clustering approaches based on patent landscapes and semantic similarity would further allow the system to surface groups of related documents rather than isolated records [24-27].

Answer Generation

Answer generation would use retrieved publications and patent passages to produce concise, source-grounded summaries for analysts. The language model would not invent unsupported conclusions; instead, it would explain what the retrieved evidence suggests, identify relevant publications and patents, and provide links for expert review. Biomedical generative models such as BioGPT show how language models can synthesize biomedical text, but surveillance use requires strict grounding in retrieved sources and transparent citation behavior [7]. The resulting interface would allow analysts to move from a high-level answer to underlying evidence, patent claims, publication abstracts, and technology-cluster maps, supporting decision-making without replacing expert judgment.

Patent Linkage and Technology Mapping

Linking Publications to Patents

The patent-linkage module would create explicit edges between scientific publications and patent documents when the publication is cited by the patent, shares unusually similar technical language, or describes formulation mechanisms that correspond to claimed inventions. Patent enrichment methods in pharmaceutical research show how patent documents can be processed as structured sources of drug-discovery intelligence rather than treated only as legal records [12]. Automated patent landscaping and BERT-based patent classification provide complementary mechanisms for connecting scientific terminology with claim language, especially when the same formulation concept is described differently across journals and patents [3, 4].

The resulting technology-evolution graph would allow analysts to trace how a scientific observation, excipient platform, delivery mechanism, or manufacturing process becomes associated with intellectual-property activity.

Technology Clustering and Taxonomy Mapping

Technology clustering would group linked publication-patent pairs into formulation themes such as lipid nanoparticle mRNA delivery, three-dimensional printed polypills, amorphous solid dispersions, continuous oral solid manufacturing, or long-acting injectable depots. Automated patent-landscape methods can organize large patent collections into interpretable technical spaces, while patent-domain transformer models can improve the classification of specialized technological language [3, 25]. Topic modeling and semantic patent analytics can also support automatic cluster labeling when related documents share formulation terms, materials, routes of administration, or process characteristics [26, 27]. These clusters would be mapped to a formulation taxonomy so that analysts can compare emerging themes across dosage form, delivery route, manufacturing method, excipient class, and therapeutic context.

Technology Maturity and Competitive Assessment

For each technology cluster, the system would examine the relative ordering of publications, patent filings, assignee activity, inventor networks, and citation relationships to support a qualitative maturity assessment. Machine-learning approaches for technology forecasting show that patent streams can be interpreted as signals of emerging technological direction rather than as static archives [14]. Technology-convergence methods using semantic analysis and link prediction further suggest that formulation maturity may be inferred from connections among materials science, process engineering, drug delivery, and therapeutic application areas [15, 16]. The system would therefore describe clusters as early, consolidating, or strategically active in conceptual terms, while identifying relevant academic groups, companies, patent assignees, and document evidence for expert review.

Trend Detection And Innovation Intelligence Engine

Emerging Technology Signal Detection

The trend-detection engine would monitor the literature and patent streams for signals such as new formulation terms, increased co-occurrence of previously separate technologies, repeated appearance of new excipient-function combinations, or growth in related patent claims. Graph-based patent text methods show how technology convergence can be detected from relationships among technical concepts rather than only from individual document counts [17]. Complex-systems perspectives on technology forecasting also support monitoring innovation as a dynamic interaction among actors, documents, and technical domains [18]. In this framework, a signal would not be treated as proof of future success, but as an analytically useful prompt for expert investigation.

Alert Generation and Intelligence Briefing

When the system identifies a plausible emerging formulation signal, it would generate an intelligence brief summarizing the technology concept, supporting publications, related patent documents, key assignees, and uncertainty factors. Biomedical knowledge integration studies show the value of connecting heterogeneous evidence into structured decision-support views rather than presenting isolated search results [19]. DrugBank, ChEMBL, and Open Targets illustrate how curated biomedical resources can support structured navigation across entities, mechanisms, and evidence sources, although formulation surveillance would require a stronger emphasis on dosage forms, delivery platforms, and patent claims [20-22]. The brief would be routed to subscribed analysts as a reviewable evidence package rather than as an automated recommendation.

User Dashboard and Decision Support

Interactive Dashboard for Technology Surveillance

The dashboard would allow users to explore formulation technology clusters, inspect linked publication-patent pairs, filter by date, organization, formulation type, patent assignee, delivery route, and evidence source, and configure alerts for topics of interest. Scientific retrieval benchmarks such as TREC-COVID demonstrate that specialized interfaces and curated retrieval tasks can support rapid navigation of complex biomedical literature [23]. Patent-landscape visualization methods similarly show how technical collections can be organized for analyst interpretation rather than simple document listing [3, 24]. A formulation-focused dashboard would therefore combine semantic search, evidence traceability, cluster maps, and alert histories in a single decision-support environment.

Integration with R&D and Strategy Workflows

The system's intelligence briefs would be integrated into formulation R&D, competitive intelligence, business development, due diligence, and patent-strategy workflows. Pharmaceutical patent landscaping methods show how patent-derived intelligence can inform life-science decision-making when documents are interpreted through a domain-specific lens [13]. Technology-forecasting and convergence analysis can support early discussion of whether a formulation theme is technically maturing, strategically crowded, or still exploratory [14-16]. In practice, the system would provide structured evidence for human review during portfolio planning, technical scouting, collaboration assessment, and invention-disclosure preparation.

Evaluation Strategy

Retrospective Detection of Known Technologies

The system should first be evaluated retrospectively using well-known formulation technologies whose scientific and patent histories can be reconstructed without claiming prospective predictive success. Lipid nanoparticle mRNA delivery, three-dimensional printed drug products, and continuous manufacturing provide suitable conceptual examples because their literature traces include technical evolution across composition, process design, stability, and dosage-form development [1, 2, 11]. The evaluation would ask whether the system could detect coherent clusters of relevant documents and link them to patent activity in a way that expert reviewers find meaningful. This assessment should emphasize qualitative interpretability, evidence completeness, and traceability rather than unverified numerical performance claims.

Semantic Search Relevance and Patent-Linkage Accuracy

Semantic search relevance should be evaluated by formulation experts who judge whether retrieved publications and patents answer domain-specific questions about dosage form, excipient function, release mechanism, delivery route, or manufacturing process. Biomedical and scientific language models such as BioBERT, PubMedBERT, SciBERT, SPECTER, and BioGPT provide the conceptual basis for retrieval and answer generation, but their usefulness in formulation surveillance must be judged against expert expectations [5-9]. Patent-linkage accuracy should be assessed by comparing system-proposed publication-patent connections with expert-curated examples based on citations, shared claims, and technical overlap [3, 4, 10, 12]. The goal would be to determine whether the system supports reliable expert review, not to replace expert interpretation.

Table 2 defines the evaluation and governance criteria needed to determine whether AI-driven formulation surveillance produces reliable, traceable, and workflow-relevant intelligence rather than unsupported automated trend claims.

Table 2. Evaluation and Governance Framework for Reliable AI-Driven Formulation Technology Surveillance

Evaluation domain	Core question	Suggested assessment approach	Evidence of success	Key risk if not evaluated	Governance control
Coverage adequacy	Does the system ingest the right scientific, patent, and technical sources for formulation surveillance?	Compare ingested sources against expert-curated lists of formulation journals, patent jurisdictions, preprint servers, conference sources, and technical domains	Important sources are represented; missing sources are documented; coverage warnings are visible to users	Early formulation signals may be missed because they appear outside the monitored corpus	Source coverage register, database audit, jurisdiction and language warnings
Semantic search relevance	Do retrieved records answer formulation-specific questions rather than merely match terminology?	Expert review of retrieved publications and patents for queries about dosage forms, excipient functions, release mechanisms, routes, and manufacturing processes	Retrieved records are technically relevant, diverse, and ranked in a way that supports expert review	Users may receive semantically broad but practically irrelevant results	Expert-labeled relevance sets, re-ranking review, query-feedback mechanism
Patent-claim interpretation	Does the system distinguish patents that truly claim a formulation technology from patents that only mention it?	Compare system-identified claim relevance with patent analyst judgments across selected technology themes	Claim-bearing patents are separated from incidental references; claim sections remain traceable	Competitive intelligence may overstate ownership, crowding, or technical protection	Claim-level extraction, legal-context labels, patent-professional review
Publication–patent linkage validity	Are proposed links between scientific literature and patents technically meaningful?	Evaluate links using citations, shared mechanisms, semantic overlap, assignee or inventor continuity, and expert-curated examples	Links are explainable and supported by visible evidence rather than opaque similarity scores	False links may imply nonexistent technology translation or ownership pathways	Link-type labeling, confidence bands, source-to-source traceability
Technology-cluster coherence	Do clusters represent interpretable formulation themes?	Have formulation experts judge whether grouped records share dosage-form, material, mechanism, route, or manufacturing logic	Clusters are conceptually coherent and can be named by experts	Noisy clusters may confuse surveillance users or hide important subthemes	Taxonomy mapping, cluster-editing tools, manual merge/split controls
Emerging-signal reliability	Are alerts useful prompts for expert review rather than noisy trend claims?	Review generated alerts for novelty, relevance, supporting evidence, uncertainty, and decision usefulness	Alerts surface plausible weak signals with clear evidence and uncertainty factors	Excessive or unsupported alerts may create signal fatigue and reduce trust	Alert thresholds, sensitivity settings, uncertainty labels, analyst feedback
Maturity and competitive interpretation	Does the system support careful qualitative assessment of technology development?	Compare maturity narratives against known histories of technologies such as lipid nanoparticles, 3D-printed	Maturity labels are evidence-based, cautious, and understandable to R&D and strategy users	The system may imply predictive certainty or commercial significance without sufficient evidence	"Signal not validation" labeling, source-backed maturity rationale, expert approval

		dosage forms, and continuous manufacturing			
Workflow adoption	Does the system fit real formulation, patent, competitive-intelligence, and R&D planning workflows?	Pilot use with formulation scientists, patent analysts, competitive intelligence teams, and R&D managers	Users report improved question framing, evidence discovery, cross-functional discussion, and decision preparation	Technically strong outputs may fail if they do not match decision processes	Role-specific dashboards, evidence dossiers, review queues, user-training feedback
Transparency and auditability	Can users trace every answer, alert, and cluster back to source documents?	Audit generated summaries and alerts for source links, evidence provenance, and reproducibility	Every insight includes source records, patent claims, dates, and explanation of linkage logic	Unsupported generation may undermine trust or create compliance concerns	Source-grounded generation, citation display, audit logs, restricted unsourced synthesis
Prospective intelligence value	Does the system improve real-time surveillance beyond retrospective reconstruction?	Deploy pilot surveillance in active technology areas and evaluate whether outputs inform meetings, scouting, patent strategy, or portfolio discussions	Intelligence briefs contribute to documented expert review, technology prioritization, or strategic decisions	The system may appear useful retrospectively but fail in live innovation settings	Prospective validation protocol, decision-impact tracking, periodic governance review

User Adoption and Intelligence Value

User adoption should be evaluated through pilot use by formulation scientists, patent analysts, competitive intelligence specialists, and R&D managers. The evaluation would examine whether the system helps users frame better surveillance questions, discover relevant documents, understand patent landscapes, and communicate technology signals across functions [13, 19]. Knowledge resources such as DrugBank, ChEMBL, and Open Targets show that biomedical platforms become valuable when they align structured data with practical decision workflows [20-22]. For the proposed system, intelligence value should therefore be assessed through user feedback, case review, and decision relevance rather than through unsupported claims of automated discovery performance.

Limitations

Language and Database Coverage

The system would be limited by the scope, accessibility, and language coverage of the scientific and patent databases it ingests. Major biomedical and patent sources may underrepresent formulation signals that appear in regional journals, non-English filings, conference-only disclosures, proprietary reports, or regulatory documents not available for automated ingestion [3, 23]. Patent analytics methods also depend on the quality of claim extraction, classification metadata, and jurisdiction-specific publication practices, which can vary across patent offices [4, 10]. As a result, the system should present coverage warnings and allow analysts to supplement automated monitoring with targeted manual review.

Early-Stage Ambiguity and Signal Overload

Early-stage formulation signals are inherently ambiguous because increased publication or patent activity does not necessarily indicate technical feasibility, regulatory viability, manufacturability, or commercial importance. Technology-forecasting research emphasizes that emerging technologies should be interpreted as evolving systems of evidence rather than as deterministic predictions [14, 18]. Semantic clustering and patent-landscape methods may also generate overly broad or noisy groups when terminology overlaps across unrelated formulation contexts [17, 24, 26]. The system should therefore include analyst controls for alert sensitivity, evidence thresholds, and manual feedback so that surveillance remains useful rather than overwhelming.

Conclusion

The proposed AI literature surveillance system provides a conceptual framework for monitoring emerging pharmaceutical formulation technologies across scientific publications and patent documents. By combining continuous ingestion, semantic search, patent linkage, trend detection, and analyst-facing dashboards, it would support earlier recognition of formulation platforms that are still distributed across fragmented evidence sources. The system is designed as decision support rather than automated prediction. Its central purpose is to help experts interpret technical signals with better context, traceability, and continuity.

The main strength of the framework is its integration of scientific discovery signals with intellectual-property evidence. Semantic search would allow users to ask formulation-specific questions in natural language, while patent linkage would show whether related technical concepts are being protected, claimed, or developed by identifiable organizations. Technology maps and intelligence briefs would help translate retrieved documents into strategic views that can be used by R&D, competitive intelligence, and patent teams. This integration could make formulation surveillance more proactive, structured, and transparent.

Important challenges remain before such a system could be adopted in industry. Data coverage, multilingual patent interpretation, full-text access, entity normalization, and jurisdictional differences in patent documentation would all affect the reliability of surveillance outputs. False-positive alerts would require careful management because weak signals are not equivalent to validated technologies. Prospective validation in real innovation intelligence workflows would be necessary to determine whether the system provides actionable value.

Pilot implementations should therefore be developed within pharmaceutical innovation intelligence functions, where formulation scientists, patent professionals, and strategy teams can jointly evaluate the system. These pilots should focus on evidence quality, user trust, workflow fit, and the ability to support better technology discussions. Community benchmark datasets for formulation technology surveillance would also help compare future systems in a transparent way. Over time, AI-driven literature and patent surveillance could become a core infrastructure for strategic formulation innovation.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Norman J, Madurawe RD, Moore CM, Khan MA, Khairuzzaman A. A new chapter in pharmaceutical manufacturing: 3D-printed drug products. Adv Drug Deliv Rev. 2017;108:39-50.
2. Hou X, Zaks T, Langer R, Dong Y. Lipid nanoparticles for mRNA delivery. Nat Rev Mater. 2021;6(12):1078-94.
3. Abood A, Feltenberger D. Automated patent landscaping. Artif Intell Law. 2018;26(2):103-25.
4. Lee JS, Hsiang J. Patent classification by fine-tuning BERT language model. World Pat Inf. 2020 Jun 1;61:101965.
5. 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.
6. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. 2021;3(1):1-23.
7. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23(6):bbac409.
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. Cohan A, Feldman S, Beltagy I, Downey D, Weld DS. Specter: Document-level representation learning using citation-informed transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020;2270-82.
10. Krestel R, Chikkamath R, Hewel C, Risch J. A survey on deep learning for patent analysis. World Pat Inf. 2021;65:102035.
11. Trenfield SJ, Awad A, Goyanes A, Gaisford S, Basit AW. 3D printing pharmaceuticals: drug development to frontline care. Trends Pharmacol Sci. 2018;39(5):440-51.
12. Gadiya Y, Zaliani A, Gribbon P, Hofmann-Apitius M. PEMT: a patent enrichment tool for drug discovery. Bioinformatics. 2023;39(1):btac716.
13. Gadiya Y, Gribbon P, Hofmann-Apitius M, Zaliani A. Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective. Artif Intell Life Sci. 2023;3:100069.
14. Kyebambe MN, Cheng G, Huang Y, He C, Zhang Z. Forecasting emerging technologies: A supervised learning approach through patent analysis. Technol Forecast Soc Change. 2017;125:236-44.
15. San Kim T, Sohn SY. Machine-learning-based deep semantic analysis approach for forecasting new technology convergence. Technol Forecast Soc Change. 2020;157:120095.
16. Cho JH, Lee J, Sohn SY. Predicting future technological convergence patterns based on machine learning using link prediction. Scientometrics. 2021;126(7):5413-29.
17. Zhu C, Motohashi K. Identifying the technology convergence using patent text information: A graph convolutional networks (GCN)-based approach. Technol Forecast Soc Change. 2022;176:121477.
18. Feng L, Wang Q, Wang J, Lin KY. A review of technological forecasting from the perspective of complex systems. Entropy. 2022;24(6):787.
19. 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.
20. 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.

21. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 2019;47(D1):D930-40.
22. Ochoa D, Hercules A, Carmona M, Suveges D, Baker J, Malangone C, et al. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res. 2023;51(D1):D1353-9.
23. Voorhees E, Alam T, Bedrick S, Demner-Fushman D, Hersh WR, Lo K, et al. TREC-COVID: constructing a pandemic information retrieval test collection. In: ACM SIGIR Forum. 2021;54(1):1-12.
24. Antonin B, Cyril V. Identifying technology clusters based on automated patent landscaping. PLoS One. 2023;18(12):e0295587.
25. Maehara Y, Kuku A, Osabe Y. Macro analysis of decarbonization-related patent technologies by patent domain-specific BERT. World Pat Inf. 2022;69:102112.
26. Yang H, Chen S. Industry 5.0: Life-Cycle mapping of sustainable technologies using BERTopic-Driven patent analytics. World Pat Inf. 2025;83:102406.
27. Lee JS. Generating patent claims with semantic novelty. World Pat Inf. 2025;83:102404.