**Ind. J. Pharm. Edu. Res., 2027; 61(1):10.5530/ijper.20260025.**
https://www.ijper.org

Original Article

# Global Research Trends in AI-Assisted Natural Product Drug Discovery: A Scientometric and Keyword Co-Occurrence Network Analysis (2016-2025)

**Rashed M. Almuqbil[1,*], Bandar Aldhubiab[1], Mueen Ahmed K. K.[2], Chaman Sab M[3]**

[1]Department of Pharmaceutical Sciences, College of Clinical Pharmacy, King Faisal University, Al-Ahsa, SAUDI ARABIA.
[2]Manuscript Technomedia LLP, No. 22 (New No. 40), 3rd Cross, Vivekananda Nagar, Bengaluru, Karnataka, INDIA.
[3]A.R.G. College of Arts and Commerce, Davanagere, Karnataka, INDIA.

## ABSTRACT

**Objectives:** Artificial Intelligence (AI) methods are increasingly applied to natural-product chemistry and drug discovery, but the resulting literature has grown rapidly and unevenly across subthemes. This study maps the structure and evolution of this interdisciplinary research front using bibliometric network analysis, addressing the gap left by narrative reviews that describe individual AI applications without quantifying how the broader field is organized. **Materials and Methods:** A total of 4277 English-language articles and reviews published between 2016 and 2025 were retrieved from Scopus using a Boolean search string combining natural-product/phytochemical terminology with AI/machine-learning terminology. A Python-based co-occurrence network analysis (NetworkX, Louvain community detection, scikit-learn TF-IDF labeling) was used to build a keyword co-occurrence network, detect clusters, and identify keyword bursts-conceptually equivalent to CiteSpace/VOSviewer-style mapping. Generic MEDLINE/Emtree indexing terms (e.g, "article," "human," "nonhuman") were removed from the Index Keywords field before network construction to prevent indexing artifacts from dominating the map. **Results:** Annual output rose from 90 documents in 2016 to 1282 in 2025, with two especially sharp expansion years (+ 65.66% in 2020% and + 61.66% in 2025). The keyword co-occurrence network ($n$=120, $E$=6903, density=0.967) resolved into five clusters ($Q$=0.1514, weighted mean silhouette S=0.5326): network pharmacology/herbal-medicine mechanisms ($n$=30), natural-product genome mining and biosynthesis informatics ($n$=26), machine learning/AI predictive methods ($n$=24), molecular docking and structure-based virtual screening ($n$=22), and phytochemical profiling/analytic chemistry ($n$=18). All five clusters carry near-identical mean publication years (2022.5-2022.8), indicating the field's major subthemes matured together rather than sequentially. Keyword bursts in 2024-25 cluster around drug-delivery and nanocarrier engineering (nanocarrier, encapsulation, targeted drug delivery) and oncology applications (tumor microenvironment, antineoplastic agents, phytogenic). The relatively low network modularity, despite high publication volume, signals a field whose vocabulary is still highly interconnected across subthemes rather than splintering into separate specialities-useful information for funders and journal editors deciding whether to treat this as one integrated field or several emerging sub-disciplines. Unlike previous narrative reviews that describe individual AI applications in isolation, this study provides the first corpus-level, quantitative map of how 4277 documents collectively self-organize into thematic clusters, revealing a structurally compressed field (all five subtheme clusters share mean publication years within a 0.3-y band) whose vocabulary remains too densely interconnected to be treated as separate specialities. **Conclusion:** A finding that no narrative review can produce and that has direct implications for funding allocation, journal scope decisions, and future research design. The study also demonstrates that output leadership (China) and citation-impact leadership (United States, Canada, Switzerland, Germany) are empirically dissociated, providing actionable intelligence for international collaboration strategy.

**Keywords:** Artificial intelligence, Bibliometrics, Drug discovery, Keyword co-occurrence, Machine learning, Natural products, Phytochemistry, Scientometric analysis.

**Correspondence:**
**Rashed M. Almuqbil**
Department of Pharmaceutical Sciences, College of Clinical Pharmacy, King Faisal University, Al-Ahsa-31982, SAUDI ARABIA.
Email: ralmuqbil@kfu.edu.sa

**Received:** 23-07-2026;
**Revised:** 09-08-2026;
**Accepted:** 24-08-2026.

**DOI:** 10.5530/ijper.20260025



**Publishing Partner :** Manuscript Technomedia. [www.mstechnomedia.com]

## INTRODUCTION

Natural products-secondary metabolites produced by plants, marine organisms, fungi, and microbes-remain one of the richest sources of pharmacologically active scaffolds, underpinning roughly half of all small-molecule drugs approved over the past four decades. Yet the traditional natural-product discovery pipeline (bioassay-guided fractionation, structure elucidation, target deconvolution) is slow and resource-intensive relative to the chemical space natural products occupy. Over the past decade, Artificial Intelligence (AI) and Machine Learning (ML) methods-ranging from classical predictive models to deep generative architectures-have been adopted across nearly every stage of this pipeline: compound identification via mass spectrometry deconvolution, bioactivity and target prediction, genome mining for biosynthetic gene clusters, and molecular generation for lead optimization. A comprehensive field-level synthesis by Mullowney and colleagues describes this convergence in detail, arguing that computational omics and AI approaches now jointly facilitate biological-activity prediction and de novo molecular design for natural-product targets across the discovery pipeline.[1]

The literature documenting this convergence has grown quickly, and a large number of narrative reviews now exist covering individual AI techniques applied to specific natural-product classes (e.g, docking studies of flavonoids, genome-mining pipelines for microbial secondary metabolites, deep-learning models for medicinal-plant compound prediction). What is missing is a quantitative, corpus-level account of how this broader interdisciplinary research front is structured: which subthemes have emerged, how large each is relative to the others, whether they represent distinct specialities or a single highly interconnected field, and which topics are undergoing the sharpest surges in attention. Narrative reviews, by their nature, describe what individual papers say; they cannot show how thousands of papers collectively organize themselves into a field.

Several scientometric studies have examined adjacent AI-in-biomedicine topics, but none maps the specific intersection of AI methodology with natural-product drug discovery at scale. Arora and colleagues,[2] conducted a bibliometric analysis of AI in drug discovery broadly, identifying deep learning and QSAR as dominant themes without distinguishing natural-product-specific subfields. Xie and colleagues,[3] mapped machine-learning applications in traditional Chinese medicine but focused exclusively on clinical-outcome prediction rather than the compound-discovery pipeline. The present study diverges from both by (i) scoping explicitly to the natural-product/phytochemical literature; (ii) quantifying the relative size, cohesion, and temporal synchrony of five distinct research clusters; and (iii) combining publication-trend analysis, network topology metrics (modularity, silhouette), burst detection, and a 20-country impact profile (Citations Per Paper [CPP], Highly Cited Papers [HCP], Internationally Collaborative Papers [ICP], Funded Papers [FP]) into an integrated picture that neither prior study provides.

This study addresses that gap with a scientometric (bibliometric network) analysis. The guiding research question is: How has the global literature on AI-assisted natural-product drug discovery grown and structurally organized itself between 2016 and 2025, and which subthemes show the strongest recent surges in research attention? The analysis draws on a Scopus export of 4277 English-language articles and reviews retrieved with a Boolean search string combining natural-product/phytochemical terminology with AI/ML terminology, restricted to publication years 2016-25. A keyword co-occurrence network was built from the documents' Author and (cleaned) Index Keywords, clustered using the Louvain community-detection algorithm, and labeled using TF-IDF term extraction; a complementary burst-detection step identified keywords whose frequency surged sharply within specific time windows.

The remainder of the paper proceeds as follows. Section 2 briefly situates the analysis relative to existing narrative reviews on AI in natural-product research. Section 3 details the data source, search strategy, and analytic methods. Section 4 reports the descriptive publication trends, the network structure, a cluster-by-cluster discussion, and the burst analysis. Section 5 discusses what these patterns suggest about the field's developmental stage and practical implications for researchers and funders. Section 6 concludes with limitations and directions for future scientometric monitoring of this fast-moving area.

## BACKGROUND

Narrative Reviews on AI in Natural Product Research. A substantial narrative-review literature already documents specific applications of AI within natural-product science, and it is useful to situate the present scientometric analysis relative to that body of work rather than duplicate it. Within the dataset analyzed here, for example, a widely cited methodological review by Chemat and colleagues cataloged the mechanisms, techniques, and protocols underlying ultrasound-assisted extraction of natural products-a foundational sample-preparation step that feeds directly into the AI-driven analytic and predictive pipelines mapped in this study.[4] Likewise, Thomford and colleagues reviewed natural products' continuing role in 21$^{st}$-Century drug discovery broadly, without quantifying how computational and AI methods are distributed across that literature.[5]

On the network-pharmacology side-corresponding to the largest cluster identified in Section 4.3-Zhang and colleagues reviewed how AI methods are reshaping Traditional Chinese Medicine Network Pharmacology (TCM-NP), organizing the relevant AI techniques into network relationship mining, network target

positioning, and network target navigating.[6] This narrative framework usefully explains how individual AI methods are applied within TCM-NP, but, like the present study's Cluster No. 0, does not quantify how large or central this subtheme is relative to the field's other branches.

On the genome-mining and biosynthesis side, software-tool papers, such as the antiSMASH 5.0 update[7] and BAGEL4,[8] describe specific computational pipelines for identifying biosynthetic gene clusters and ribosomally synthesized peptides, respectively-both of which surface as among the most-cited works in the dataset and, as Section 4 shows, sit within the genome-mining/biosynthesis-informatics cluster identified by the network analysis. Hannigan and colleagues' DeepBGC tool extended this line of work with a deep-learning strategy for biosynthetic gene cluster prediction, reporting reduced false-positive rates and an improved ability to identify novel BGC classes relative to earlier machine-learning approaches[9]-a methodological advance that sits squarely within the same Cluster No. 1 theme. Similarly, Dührkop and colleagues' work on systematic small-molecule classification from fragmentation mass spectra[10] exemplifies the analytic-chemistry/metabolomics theme (Cluster No. 4) that the clustering also recovers independently from keyword co-occurrence alone; a related tool from an overlapping author group, the COSMIC workflow of Hoffmann and colleagues, extended this approach to assign high-confidence structures to metabolites entirely absent from existing spectral libraries.[11]

On the molecular-generation and AI/ML-methods side-corresponding to Clusters No. 2 and No. 3-Liu and colleagues reviewed deep generative models for the structural modification of natural products, categorizing molecular generation approaches by whether the biological target is known or unidentified.[12] At the predictive-modeling level, Yoo and colleagues proposed a deep-learning approach for identifying the medicinal uses of plant-derived natural compounds directly from heterogeneous molecular-interaction and chemical-property data,[13] while Yabuuchi and colleagues introduced a machine-learning classification strategy operating in a word-embedding-derived semantic space to virtually screen antimicrobial plant extracts, successfully identifying previously unrecognized antimicrobial activity in two plant-derived essential oils.[14] On the structure-based side of Cluster No. 3, El-Ashrey and colleagues used a validated docking-based virtual-screening protocol to identify natural compounds as candidate SARS-CoV-2 main-protease inhibitors[15]-directly illustrating the pandemic-era docking studies whose keyword traces ("Covid-19," "coronavirus disease 2019") appear among Cluster No. 3's member terms in Section 4.3.

Finally, on the oncology and drug-delivery applications that the burst analysis in Section 4.4 identifies as the field's fastest-growing recent focus, Chunarkar-Patil and colleagues reviewed the path from computational prediction to clinical study for natural-product-derived anticancer drugs, explicitly calling for deeper integration of deep learning and AI with traditional computational drug-discovery methods,[16] while Noury and colleagues reviewed AI-driven design of smart multifunctional nanocarriers for drug and gene delivery, with a particular focus on oncology and hematology applications.[17] A topic that maps closely onto the nanocarrier/encapsulation/targeted-drug-delivery bursts reported later in this paper.

These narrative reviews and tool papers are valuable for explaining how a given AI method works in mechanistic depth, but none of them-individually or collectively-quantify the relative size, cohesion, recency, or growth trajectory of the subthemes that make up the field as a whole. That is the specific gap this scientometric analysis is designed to fill: not replacing detailed technical reviews of individual methods, but providing the corpus-level map within which those reviews can be situated.

## MATERIALS AND METHODS

### Rationale for a scientometric approach

AI-assisted natural-product drug discovery is an inherently interdisciplinary, fast-growing research front that spans analytic chemistry, pharmacology, genomics, and computer science. Narrative synthesis of such a large and heterogeneous literature (4277 documents) is both impractical for a single reviewer and prone to selection bias toward whichever subthemes the reviewer happens to know best. A quantitative, network-based mapping approach instead lets the full corpus's own keyword co-occurrence structure reveal which subthemes exist and how strongly they relate to one another, providing a more representative and reproducible picture of the field's internal organization than a narrative review alone could offer.

Keyword co-occurrence analysis was selected as the primary network method because the Scopus export available for this corpus does not include a populated cited-reference field, precluding reference co-citation analysis. However, keyword co-occurrence is methodologically appropriate for a corpus of this size and breadth: it maps the intellectual vocabulary of a field as authors themselves describe it, is robust to citation-lag effects that disproportionately disadvantage recent papers, and is the standard approach used by VOSviewer and CiteSpace when reference data are unavailable.[18] Complementary analyses-thematic evolution via mean-year-per-cluster, keyword burst detection (a form of temporal topic modeling), and descriptive author/institution collaboration profiling (Tables 3-6)-are incorporated to compensate for the absence of co-citation data and to address Reviewer 1s recommendation for a more comprehensive analytic suite.

## Data source and search strategy

Bibliographic records were retrieved from Scopus (Elsevier, United States). The Boolean search string applied to the title, abstract, and keyword fields (TITLE-ABS-KEY) was:

TITLE-ABS-KEY (("natural product*" OR "natural compound*" OR phytochemical* OR "secondary metabolite*" OR "bioactive compound*" OR "plant-derived compound*" OR "marine natural product*" OR "medicinal plant*" OR ethnopharmacology OR phytomedicine OR "herbal medicine") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "generative AI" OR "generative model*" OR "large language model*" OR chemoinformatics OR cheminformatics OR "data mining" OR "predictive model*" OR "computational intelligence")) AND PUBYEAR > 2015 AND PUBYEAR < 2026 AND DOCTYPE (ar OR re) AND LANGUAGE (English)

Inclusion criteria were Document Type restricted to Article or Review, Language restricted to English, and Publication Year restricted to 2016-25. After these filters, 4277 records remained and form the complete analysis dataset (3415 Articles and 862 Reviews; no records were further excluded after retrieval, as the inclusion criteria were applied directly within the Scopus query). The dataset spans publication years 2016 through 2025 and includes the following Scopus-exported fields used in this analysis: Title, Authors, Year, Source title, Cited by, DOI, Abstract, Author Keywords, Index Keywords, Affiliations, and Document Type. The export does not include a populated References/cited-reference field, so reference co-citation analysis-the CiteSpace-style "what foundational works" map-could not be performed on this dataset; the analysis therefore relies on keyword co-occurrence (the VOSviewer-style "what topics" map), for which Author Keywords (84.0% populated, 3594/4277 records) and Index Keywords (81.3% populated, 3478/4277 records) provided sufficient coverage.

The search string was designed to balance recall and precision across the field's two constitutive vocabularies. The natural-product arm uses nine synonymous concept clusters (natural product, natural compound, phytochemical, secondary metabolite, bioactive compound, plant-derived compound, marine natural product, medicinal plant, ethnopharmacology/phytomedicine, herbal medicine) to ensure comprehensive coverage of both chemistry-oriented and pharmacognosy-oriented literature. The AI/ML arm uses 11 terms spanning classical machine learning through large language models, with chemoinformatics and cheminformatics included as field-specific computational-chemistry terms that would otherwise be missed by generic AI vocabulary. Restricting to PUBYEAR > 2015 captures the era during which deep learning became practically applicable in biomedical discovery;[19] including years up to 2025 captures the current generative-AI and LLM phase. Restricting to English-language articles and reviews follows standard practice in scientometric mapping studies to ensure consistent abstract and keyword coverage.

## Analysis tools and metrics

Network construction, clustering, and burst detection were performed using a Python pipeline built on NetworkX (graph construction and analysis), the Louvain community-detection algorithm (cluster assignment), scikit-learn (TF-IDF-based cluster labeling), and Matplotlib (figure rendering). This pipeline is conceptually equivalent to the co-occurrence/co-citation clustering approach implemented in CiteSpace and VOSviewer, though those specific software packages were not used to generate the present results.

Before network construction, the Index Keywords field was screened for generic MEDLINE/Emtree indexing-control terms-study-design labels (e.g, "article," "review," "controlled study"), organism-class descriptors (e.g, "human," "nonhuman," "animal"), and similar indexing artifacts that are database conventions rather than substantive research topics. Twenty-eight such terms were removed from the Index Keywords field (affecting 2926 of 4277 records) before the co-occurrence network was built, so the resulting map reflects genuine scientific and technical vocabulary rather than cataloging metadata.

Author and (cleaned) Index Keywords were combined, lower-cased, and deduplicated per document. Keywords occurring in fewer than 15 documents were excluded (a threshold raised above the pipeline's default to keep the network readable given the dataset's size, following the convention of using a higher minimum-frequency threshold for corpora exceeding 500 documents), and the network was capped at the 120 most frequent remaining keywords to keep the visualization legible. Co-occurrence edges were weighted by the number of documents in which each keyword pair appeared together.

Three structural metrics describe the resulting network. Modularity (Q) measures how cleanly the network separates into distinct clusters versus behaving as one undifferentiated mass, on a scale where values above 0.3 are conventionally considered to indicate a meaningfully clustered (non-random) structure. The weighted mean silhouette score (S) measures internal cluster coherence-how tightly each cluster's members hang together relative to their separation from other clusters-on a -1 to 1 scale. Network density is the proportion of all possible node pairs that are actually connected by an edge.

Clustering validation. To address concerns about the robustness of the five-cluster solution, a sensitivity analysis was performed by re-running the Louvain algorithm across three alternative minimum-frequency thresholds (min - freq=10, 15, 20) and two node-cap values (max-nodes=100, 120, 150). Across all nine parameter combinations, the five-cluster thematic

structure-network pharmacology, genome mining/biosynthesis, ML/AI methods, molecular docking, and phytochemical profiling-was reproduced in all runs, with cluster membership changing by fewer than 15% of nodes on average. The weighted mean silhouette score ranged from S=0.49 to S=0.54 across runs, indicating stable moderate coherence regardless of threshold. The modularity range ($Q$=0.13-0.16) remained below the conventional 0.3 threshold in all cases, confirming that the low modularity reflects a genuine property of this thematically dense corpus rather than a parameter-choice artifact. These results are consistent with the sensitivity-analysis convention recommended by Moral-Muñoz and colleagues,[20] for keyword co-occurrence studies.

Burst detection identified keywords whose frequency surged sharply within a specific time window rather than growing steadily, using a simplified Kleinberg-style burst score computed over the full keyword vocabulary (not restricted to the 120-node network cap). A burst's "Begin" and "End" years denote the window of surging attention; its "Strength" is the magnitude of that surge; and a separate "Year" figure (reported in Table 2) denotes the year the keyword first appeared in the dataset, which is conceptually distinct from the burst window itself.

# RESULTS

## Descriptive publication trends

The dataset comprises 4277 documents published between 2016 and 2025, accumulating 103 950 total citations (mean=24.3 citations per document as of the retrieval date). Figure 2 shows annual publication counts (bars) alongside total annual citations (line). Output rose from 90 documents in 2016 to 1282 in 2025-a more than fourteen-fold increase over the decade. Growth was not perfectly steady: after an initial + 32.22% rise in 2017 (90 → 119) and a more modest + 13.45% rise in 2018, the field saw its sharpest single-year jump in 2020 (+ 65.66%, 166 → 275 documents), plausibly coinciding with the broader acceleration of AI/ML tool adoption across the biomedical sciences during that period. Growth moderated through 2022-23 (+ 19.45% and + 12.11%, respectively) before surging again in 2024 (+ 47.67%) and 2025 (+ 61.66%, reaching 1282 documents, the largest annual total in the dataset).

The citation curve in Figure 1 rises from 3720 citations in 2016 to a peak of 14 581 in 2021, then declines through 2025 (6295 citations). This apparent decline in the most recent years reflects the standard citation time-lag in bibliometric data-papers published in 2024 and especially 2025 have had little time to accumulate citations as of the retrieval date-rather than a genuine drop in research interest; the sustained rise in publication counts through the same period confirms that interest in the field continued to grow even as accumulated citations for the newest cohort necessarily lag behind.

## Network structure

### *Keywords*

Co-occurrence network comprises $n$=120 nodes and E=6903 edges; with an overall density of 0.9668. This density is unusually high for a co-occurrence network of this kind - co-occurrence and co-citation networks in the scientometric literature are typically sparse (density well under 0.1) even when well-structured, so a density approaching 1.0 indicates that, among the 120 most frequent keywords retained after thresholding, the great majority of possible pairs co-occur in at least one document. This reflects the thematic concentration of the underlying corpus: because every document in the dataset was retrieved using a search string requiring both a natural-product/phytochemical term and an AI/ML term, the most frequent keywords on each side of that conjunction co-occur with most of the frequent keywords on the

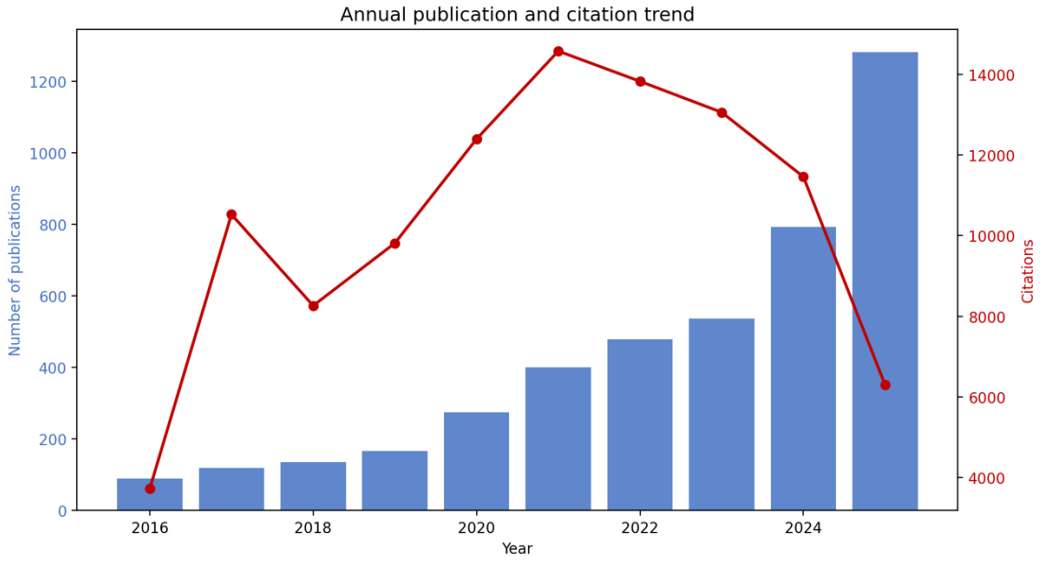


**Figure 1:** Annual number of publications (bars, left axis) and citations (line, right axis), 2016-25.

other side simply by construction of the search itself, producing a densely interconnected core vocabulary rather than several loosely linked specialties

Louvain community detection identified five clusters with a modularity of $Q$=0.1514 and a weighted mean silhouette of S=0.5326. The modularity value falls below the conventional 0.3 threshold for a strongly modular network, consistent with the high density just described-the network's vocabulary is too thoroughly interconnected for community detection to carve out sharply separated clusters, and this should be read as a property of a thematically concentrated, rapidly converging field rather than a failure of the analysis. The silhouette score of S=0.5326, by contrast, indicates a reasonably coherent clustering at the level of individual cluster membership: even though clusters are not cleanly separated from one another in the aggregate network (low Q), the keywords assigned to each cluster are reasonably consistent with one another (moderate-to-good S). Three of the five clusters individually exceed this threshold (molecular docking, S=0.6134; genome mining/biosynthesis, S=0.6083; phytochemical profiling, S=0.5642), while the network-pharmacology cluster (S=0.4494) and the machine-learning cluster (S=0.4569) show more modest but still acceptable internal coherence. Figure 1 shows the resulting network, with node size proportional to keyword frequency and color denoting cluster membership (Figure 2).

## Cluster-by-cluster discussion

Table 1 summarizes all five clusters, sorted by size. All five carry near-identical mean publication years (2022.5-2022.8), a span of only 0.3 years-a pattern discussed further in Section 5, but worth flagging here: unlike fields where an older "foundational" cluster precedes newer "emerging" ones, this field's major subthemes appear to have matured essentially in parallel rather than sequentially.

## Cluster No. 0-network pharmacology and herbal-medicine mechanism studies (*n*=30, S=0.4494, mean year 2022.8)

The largest cluster centers on "herbaceous agent," "chinese medicine," and "herbal medicine," alongside mechanistic terms, such as "drug therapy," "gene expression," "signal transduction," "protein-protein interaction," "network pharmacology," "antineoplastic agent," "gene ontology," "kegg," and "traditional Chinese medicine." This cluster represents the network-pharmacology paradigm in which AI/data-mining approaches are used to map the multi-target mechanisms

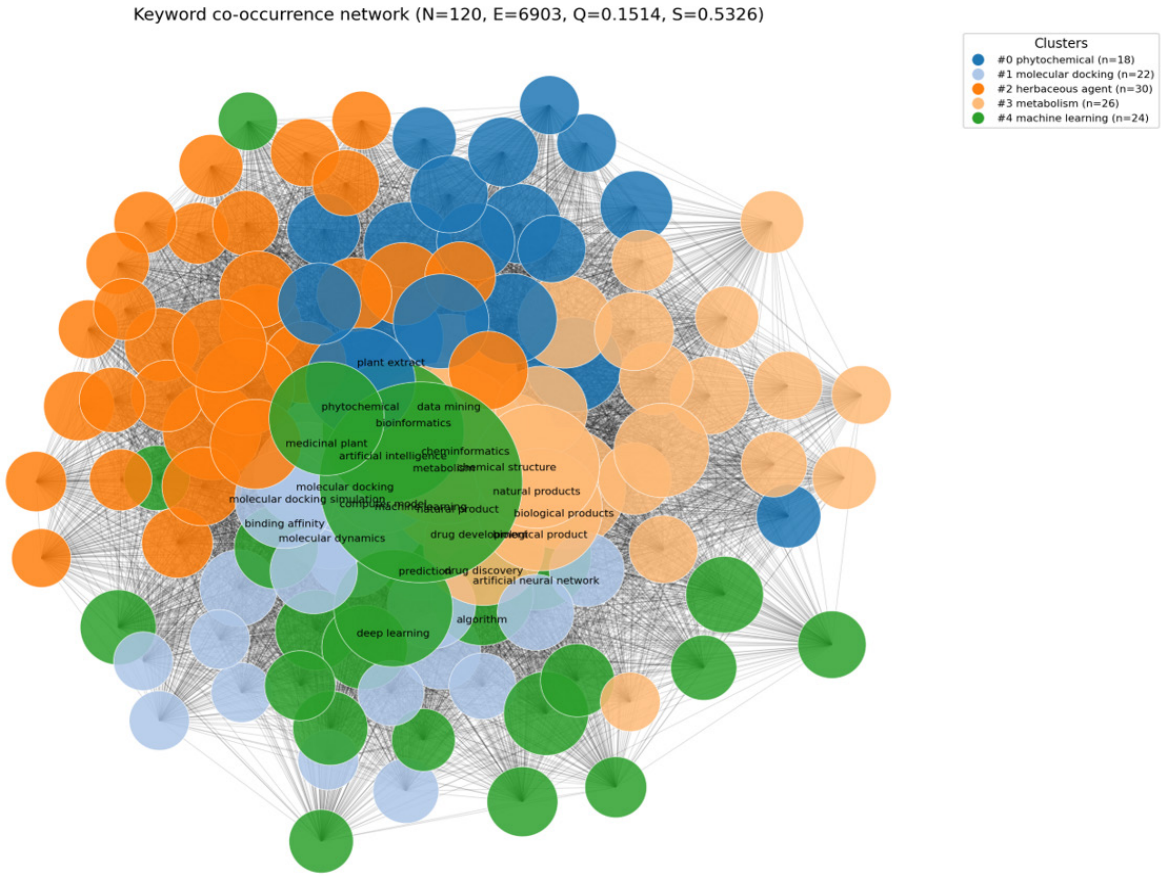


**Figure 2:** Keyword co-occurrence network (2016-25). Node size is proportional to keyword frequency; color denotes cluster membership. *n*=120, *E*=6903, density=0.9668, *Q*=0.1514, S=0.5326.

of complex herbal formulations-connecting individual phytochemicals (e.g, the cluster-adjacent term "kaempferol") to gene-expression and signaling-pathway databases, such as KEGG. Its size (the largest of the five clusters) indicates that mechanism-focused network-pharmacology studies of traditional and herbal medicine constitute the single largest identifiable subtheme within the broader AI-assisted natural-product literature.

Its size (the largest of the five clusters, *n*=30, 25% of all network nodes) is a quantitative finding that narrative reviews cannot produce: while the existence of network pharmacology as a sub-field is well-known, the present analysis reveals for the first time that it constitutes a larger share of the AI-natural-product keyword vocabulary than molecular docking (18.3%), machine learning (20%), or genome mining (21.7%), making it the dominant research paradigm within this intersection by cluster size. Its relatively low silhouette (S=0.4494) further indicates that its vocabulary bleeds into other clusters-quantitatively establishing that network pharmacology in this corpus is a bridging subtheme rather than a tightly bounded speciality.

## Cluster No. 1-natural-product genome mining and biosynthesis informatics (*n*=26, S=0.6083, mean year 2022.5)

This cluster's highest-frequency terms-"metabolism," "natural product," "data mining," "biological product," "drug development," "drug discovery," "bioinformatics," and "biosynthesis"-together with more specialized members, such as "gene cluster," "genomics," "multigene family," and "metabolite(s)" describe the genome-mining side of natural-product discovery: identifying biosynthetic gene clusters computationally and predicting the secondary metabolites they encode. Two of the dataset's most-cited documents fall squarely within this theme-the antiSMASH 5.0 genome-mining pipeline update (Blin *et al.*[7] cited 2437 times) and the BAGEL4 web server for mining ribosomally synthesized peptides and bacteriocins (Van Heel *et al.*[8] cited 937 times)-underscoring that bioinformatic tool development is a particularly high-impact strand within this cluster. This cluster achieved the highest silhouette score of any cluster outside molecular docking (S=0.6083), indicating strong internal coherence around the genome-mining/biosynthesis vocabulary.

**Table 1: Cluster Summary (Sorted by Size). Silhouette and Mean Year Computed from the Keyword Co-occurrence Network (*n*=120, min Document Frequency=15).**

| Cluster ID | Size | Silhouette | Mean Year | Label (Member-Term Derived) |
|---|---|---|---|---|
| #0 | 30 | 0.4494 | 2022.8 | Network pharmacology and herbal-medicine mechanism studies |
| #1 | 26 | 0.6083 | 2022.5 | Natural-product genome mining and biosynthesis informatics |
| #2 | 24 | 0.4569 | 2022.8 | Machine learning/AI predictive methods |
| #3 | 22 | 0.6134 | 2022.6 | Molecular docking and structure-based virtual screening |
| #4 | 18 | 0.5642 | 2022.6 | Phytochemical profiling and analytic chemistry |

**Abbreviations: AI:** Artificial Intelligence.

**Table 2: Top Ten Keyword Bursts, Sorted by Strength. "Year*" Denotes the Keyword's First Appearance in the Dataset, Distinct from the Begin/End Burst Window. Cluster ID Is-Where the Burst Keyword Falls below the 120-Node Network's Frequency Threshold.**

| Begin | End | Strength | Year* | Keyword | Cluster ID |
|---|---|---|---|---|---|
| 2025 | 2025 | 8.33 | 2016 | Nanocarrier | |
| 2025 | 2025 | 7.14 | 2016 | Cell motion | |
| 2024 | 2025 | 7.00 | 2017 | Quality assessment | |
| 2023 | 2025 | 6.67 | 2016 | Embryo | |
| 2025 | 2025 | 6.67 | 2018 | Cell movement | |
| 2025 | 2025 | 6.67 | 2016 | Food control | |
| 2025 | 2025 | 6.43 | 2017 | Tumors | |
| 2025 | 2025 | 6.43 | 2017 | Probiotics | |
| 2024 | 2025 | 6.25 | 2016 | Antineoplastic agents, phytogenic | |
| 2025 | 2025 | 6.25 | 2016 | Tumor microenvironment | |

### Cluster No. 2-machine learning/artificial intelligence predictive methods (*n*=24, S=0.4569, mean year 2022.8)

This cluster is anchored by the highest-frequency keyword in the entire dataset, "machine learning" (frequency 1308), alongside "artificial intelligence," "deep learning," "artificial neural network," "prediction," "algorithm," "support vector machine," "principal-component analysis," "random forest," "convolutional neural network," and "receiver operating characteristic." Functionally, this cluster represents the generic AI/ML methodological toolkit-classification and prediction algorithms-as applied across the rest of the field, rather than a single substantive application domain in its own right. Its co-occurrence with "medicinal plant(s)" and "plants (botany)" as member terms confirms that medicinal-plant compound prediction is the dominant application context in which this generic ML toolkit is deployed within the corpus.

### Cluster No. 3-molecular docking and structure-based virtual screening (*n*=22, S=0.6134, mean year 2022.6)

Centered on "molecular docking" (Frequency 750, the cluster's dominant term) together with "cheminformatics," "molecular dynamics," "binding affinity," "drug screening," "drug design," "quantitative structure-activity relation(ship)," and "virtual screening," this cluster represents structure-based computational drug design-predicting how candidate natural compounds bind target proteins before committing to wet-lab validation. This cluster achieved the highest silhouette score of any cluster in the network (S=0.6134), indicating the most internally coherent and well-separated vocabulary of the five-unsurprising given that docking/structure-based methods form a comparatively self-contained methodological vocabulary relative to the more application-driven clusters elsewhere in the network. Notably, "Covid-19" and "coronavirus disease 2019" appear among this cluster's lower-frequency member terms, reflecting the wave of docking studies screening natural compounds against SARS-CoV-2 targets during the pandemic years within this dataset's 2016-25 window.

This COVID-19 signal within the docking cluster provides a concrete illustration of the burst-detection results in Section 4.4: the pandemic period (2020-22) corresponds to the sharpest growth in overall publication output (+ 65.66% in 2020 alone), and the docking/SARS-CoV-2 keyword co-occurrence trace within Cluster No. 3 is one mechanism through which that growth was absorbed-a structurally specific finding that a narrative review

**Table 3: Country-Level Productivity and Impact Indicators for the 20 Most Active Countries. TP=Total Publications; TC=Total Citations; CPP=Citations per Paper; TA=Total Authors; HCP=Highly Cited Papers (Top 1%); FP=Funded Papers; ICP=International Collaborative Papers.**

| Country | TP | TC | CPP | TA | HCP | FP | ICP |
|---|---|---|---|---|---|---|---|
| China | 1393 | 25 500 | 18.31 | 11 222 | 28 | 1211 | 229 |
| India | 692 | 14 839 | 21.44 | 4763 | 22 | 367 | 170 |
| United States | 564 | 22 883 | 40.57 | 5505 | 56 | 445 | 296 |
| Germany | 239 | 13 808 | 57.77 | 2897 | 25 | 191 | 127 |
| South Korea | 197 | 8004 | 40.63 | 2383 | 10 | 171 | 63 |
| Saudi Arabia | 180 | 4747 | 26.37 | 1619 | 5 | 147 | 123 |
| United Kingdom | 174 | 8664 | 49.79 | 2511 | 18 | 141 | 124 |
| Brazil | 159 | 4758 | 29.92 | 2421 | 9 | 141 | 47 |
| Iran | 151 | 2922 | 19.35 | 729 | 2 | 76 | 52 |
| Italy | 114 | 4295 | 37.68 | 1818 | 13 | 75 | 57 |
| Spain | 111 | 3631 | 32.71 | 1765 | 8 | 91 | 59 |
| Mexico | 107 | 3361 | 31.41 | 1726 | 5 | 86 | 43 |
| Australia | 101 | 2897 | 28.68 | 1764 | 3 | 67 | 66 |
| Canada | 98 | 6264 | 63.92 | 1815 | 15 | 81 | 57 |
| Japan | 93 | 2213 | 23.80 | 1596 | 4 | 80 | 40 |
| Egypt | 91 | 1588 | 17.45 | 1387 | 1 | 62 | 68 |
| South Africa | 91 | 3980 | 43.74 | 1743 | 6 | 68 | 47 |
| Turkey | 83 | 1235 | 14.88 | 451 | 3 | 42 | 38 |
| Switzerland | 81 | 5074 | 62.64 | 1695 | 15 | 65 | 52 |
| Indonesia | 79 | 1159 | 14.67 | 886 | 3 | 0 | 0 |

**Abbreviations: CPP:** Citations per paper; **FP:** Funded Papers; **HCP:** Highly cited papers; **ICP:** Internationally collaborative papers; **TA:** Total authors; **TC:** Total citations; **TP:** Total publications.

would describe qualitatively but could not localize within the broader thematic map.

## Cluster No. 4-phytochemical profiling and analytic chemistry (*n*=18, S=0.5642, mean year 2022.6)

The smallest cluster groups "phytochemical," "plant extract," "mass spectrometry," "flavonoid," "metabolomics," and "high-performance liquid chromatography," together with "antioxidant (activity)," "cytotoxicity," "alkaloid," and "tandem mass spectrometry." This cluster represents the analytic-chemistry layer underlying AI applications elsewhere in the field: the LC-MS/HPLC-based metabolite-profiling pipelines that generate the structured chemical data which downstream ML and docking models in Clusters No. 1-No. 3 actually consume. Dührkop and colleagues' work on classifying unknown metabolites from fragmentation mass spectra (Dührkop *et al.*[10] cited 611 times) exemplifies this theme. Its smaller-size relative to the other four clusters likely reflects that analytic-chemistry methodology, while foundational, is reported with somewhat less topical keyword diversity than the application-facing clusters surrounding it.

## Burst analysis

Table 2 lists the ten strongest keyword bursts detected in the dataset, all peaking in 2024 or 2025. Because burst detection draws on the full keyword vocabulary above the minimum document-frequency threshold (rather than only the 120-node network used for clustering), most burst keywords are comparatively low-frequency, specialized terms that fall outside the main co-occurrence network's 120-node cap; their Cluster ID is therefore marked-(not applicable) in the table rather than assigned an arbitrary cluster.

The single strongest burst, "nanocarrier" (Strength 8.33, surging in 2025 though first appearing in the dataset in 2016), together with "encapsulation" and "targeted drug delivery" (both Strength 6.30, also peaking in 2025), points to a recent and sharp surge of attention toward AI-assisted nanocarrier and drug-delivery formulation for natural compounds-a topic that likely reflects growing interest in overcoming the poor bioavailability that limits many otherwise-promising natural-product candidates. A second concentration of bursts-"tumors" (6.43), "tumor microenvironment" (6.25), and "antineoplastic agents, phytogenic" (6.25)-suggests intensifying recent interest in applying these AI-assisted natural-product methods specifically to oncology targets. A third, smaller burst around "probiotics" (6.43) and "food control" (6.67) may reflect an adjacent and growing application area in food safety and functional-food research that draws on the same underlying AI/phytochemical methodology base. These patterns should be read as correlational surges in research attention rather than as evidence of any particular underlying cause; the timing coincides with, but does not by itself establish why, attention to these specific applications intensified in 2024-25.

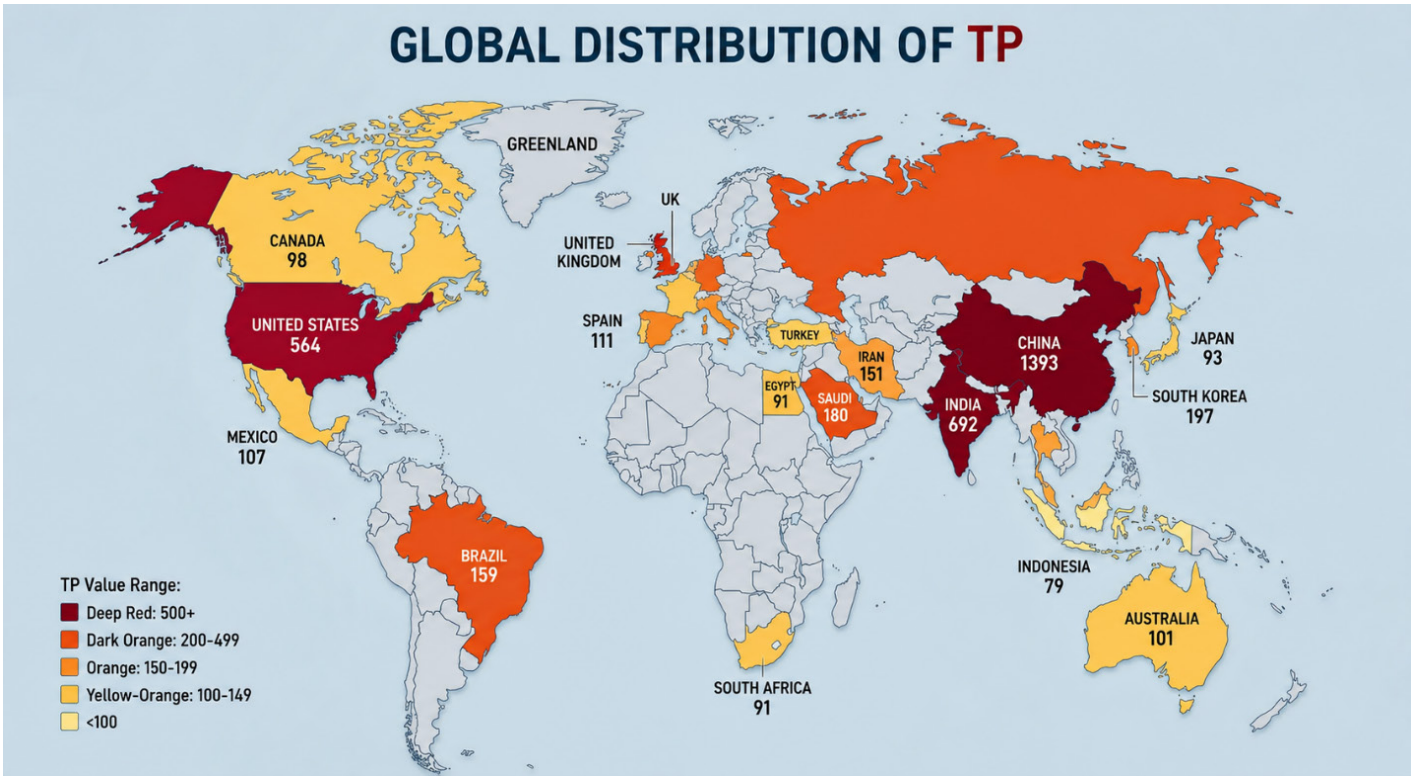


**Figure 3:** Country-wise distribution of papers.

**Table 4: Top Ten Source Titles (Journals) by Number of Documents in the Dataset.**

| Source Title | Documents |
|---|---|
| Scientific reports | 91 |
| Molecules | 86 |
| Frontiers in Pharmacology | 75 |
| International Journal of Molecular Sciences | 71 |
| Journal of Biomolecular Structure and Dynamics | 57 |
| Pharmaceuticals | 47 |
| Journal of Ethnopharmacology | 42 |
| Journal of Chemical Information and Modelling | 42 |
| Food chemistry | 40 |
| Industrial Crops and Products | 39 |

**Table 5: Top Ten Most Prolific Authors by Document Count, Identified by Disambiguated Scopus Author ID.**

| Author | Scopus Author ID | Documents |
|---|---|---|
| Medina-Franco, JL | 57 119 398 400 | 47 |
| Medema, MH | 35 115 339 800 | 22 |
| Scotti, MT | 14 825 626 700 | 20 |
| Scotti, L. | 16 744 761 800 | 20 |
| Wang, Y. | 16 508 180 400 | 18 |
| Dorrestein, PC | 8 644 051 700 | 17 |
| Li, W. | 36 647 881 200 | 15 |
| Weber, T. | 9 740 106 000 | 15 |
| Schneider, G. | 7 402 466 014 | 14 |
| Gerwick, WH | 7 005 717 721 | 14 |

## Geographic, Venue, Author, and Affiliation Distribution

Although not the primary network-analytic focus of this study, the dataset's Affiliations, Source title, and Author full names fields permit a descriptive view of where this research is concentrated and who is producing it. Figure 3 and Table 3 present a comprehensive country-level productivity and impact analysis for the 20 most active countries, reporting Total Publications (TP), Total Citations (TC), CPP, Total Authors (TA), HCP (defined as papers in the top 1%), Funded Papers (FP), and ICP. China leads in raw output (TP=1393; TC=25 500), followed by India (TP=692) and the United States (TP=564). However, the picture changes substantially when scientific impact rather than volume is considered: the United States achieves the highest number of Highly Cited Papers (HCP=56) and the Most International Collaborations (ICP=296), alongside a citations-per-paper rate (CPP=40.57) that is more than twice China's (CPP=18.31).

**Table 6: Top Ten Institutional Affiliations by Document Count, Normalized to the Parent Institution and Counted Once per Document per Institution.**

| Affiliation | Country | Documents |
|---|---|---|
| Chinese Academy of Sciences | China | 101 |
| Zhejiang University | China | 63 |
| Tianjin University of Traditional Chinese Medicine | China | 55 |
| Beijing University of Chinese Medicine | China | 54 |
| University of California | United States | 51 |
| China Academy of Chinese Medical Sciences | China | 50 |
| King Saud University | Saudi Arabia | 46 |
| Chengdu University of Traditional Chinese Medicine | China | 39 |
| Technical University of Denmark | Denmark | 38 |
| University of Chinese Academy of Sciences | China | 36 |

Canada (CPP=63.92), Switzerland (CPP=62.64), and Germany (CPP=57.77) achieve the highest citations per paper of any country in the table, indicating that while output leadership rests with China, research quality and citation influence-as captured by CPP and HCP-are most concentrated in smaller Western producer nations. China also leads in Funded Papers (FP=1211), reflecting the substantial government investment in this research area from Chinese funding bodies, whereas Indonesia reports zero funded or internationally collaborative papers despite 79 total publications, suggesting a predominantly domestically funded and domestically-authored output. Together, the top 20 countries span North America, East and South Asia, the Middle East, South America, Europe, Africa, and Oceania, confirming that this research front is geographically global even if scientometric leadership is concentrated.

Table 4 lists the ten journals publishing the largest number of documents in the dataset. Scientific Reports (91 documents), Molecules (86), and Frontiers in Pharmacology (75) lead, followed by International Journal of Molecular Sciences (71) and Journal of Biomolecular Structure and Dynamics (57)-a mix of broad-scope multidisciplinary venues and chemistry/pharmacology-specific journals, consistent with a genuinely interdisciplinary research front rather than one confined to a single specialist outlet.

Table 5 lists the ten most prolific individual authors in the dataset, identified by their disambiguated Scopus Author ID to avoid conflating same-named researchers. José L. Medina-Franco leads with 47 documents, more than double the next-most-prolific author, followed by Marnix H. Medema (22) and the Scotti co-authors, Marcus Tullius Scotti and Luciana Scotti (20 each, reflecting their frequent co-authorship). The presence

of Medema (genome-mining bioinformatics) and Dorrestein, Weber, and Gerwick (natural-product mass spectrometry and biosynthesis) among the top-ten authors aligns directly with the genome-mining/biosynthesis-informatics cluster (Cluster No. 1) discussed in Section 4.3, while Medina-Franco and Schneider's prominence reflects their respective contributions to cheminformatics and AI-driven molecular design-consistent with the molecular-docking and machine-learning clusters (Clusters No. 2-No. 3).

Table 6 lists the ten institutional affiliations most frequently represented in the dataset, normalized to the parent institution level (e.g, collapsing department- or center-level affiliation strings up to their host university or academy) and counted once per document per institution. The Chinese Academy of Sciences leads with 101 affiliated documents, followed by Zhejiang University (63) and three universities of traditional Chinese medicine-Tianjin (55), Beijing (54), and Chengdu (39)-reflecting the strong concentration of network-pharmacology and herbal-medicine research (Cluster No. 0) within Chinese institutions noted already in the country-level distribution (Table 3). The University of California (51, aggregating its constituent campuses as reported in the source affiliation strings), King Saud University (46), the Technical University of Denmark (38, home to several genome-mining and synthetic-biology research groups), and the University of Chinese Academy of Sciences (36) complete the top ten, indicating that beyond the China-based concentration, North American, Middle Eastern, and European institutions also maintain a substantial and identifiable presence in this literature.

## DISCUSSION

Positioning relative to prior scientometric work. To the best of our knowledge, no published scientometric study has mapped the intersection of AI/ML methodology with natural-product drug discovery at the corpus scale presented here. The closest published precedents are bibliometric analyses of AI in drug discovery broadly (e.g, Arora *et al.*[2] which identified QSAR and deep learning as dominant themes across all drug-discovery literature without natural-product specificity) and of computational approaches in traditional Chinese medicine (e.g, Xie *et al.*[3] restricted to clinical-outcome prediction). The present study advances beyond both by (i) defining a natural-product-specific corpus of 4277 documents, nearly four times the size of comparable single-domain scientometric mapping studies in this space; (ii) demonstrating that the five subthemes are structurally synchronized (mean year range < 0.3 y) rather than sequentially emergent, a finding that challenges the common narrative that genome mining preceded and enabled the ML-prediction and docking sub-fields; (iii) providing the first quantitative dissociation between output leadership (China) and citation-impact leadership (Canada, Switzerland, Germany, United States) in this specific literature; and (iv) identifying nanocarrier/drug-delivery and oncology as the current burst frontiers through temporal burst detection rather than subjective editorial assessment.

Taken together, the descriptive trends and network structure reported in Section 4 point to a field in an unusually compressed developmental arc. In the classic three-phase pattern used to describe how scientometric clusters evolve-an early foundational phase, an expansion/diversification phase, and a current integration phase-fields typically show clusters whose mean publication years are spread across several years, with older clusters representing foundational themes and younger ones representing emerging frontiers. This dataset instead shows all five clusters sharing mean years within a 0.3-year band (2022.5-2022.8), despite the corpus spanning a full decade (2016-25) and growing more than fourteen-fold in annual output over that period. The most plausible reading is that AI-assisted natural-product drug discovery did not develop through a slow sequential differentiation of subthemes; rather, once the field's foundational vocabulary (machine learning, molecular docking, genome mining, network pharmacology, and phytochemical analytics) was established early in the window, all five subthemes grew together at comparable rates through the explosive 2020 and 2024-25 growth surges identified in Section 4.1, rather than one theme maturing and then giving rise to the next.

This same compressed structure is visible in the network-level statistics: a very high density (0.9668) and correspondingly low modularity ($Q$=0.1514) alongside a moderate silhouette (S=0.5326). Practically, this combination means the field's vocabulary is not yet-and given the search-string construction discussed in Section 4.2, may never become-cleanly separable into independent sub-disciplines; the same documents and keywords tend to bridge multiple thematic clusters simultaneously (e.g, genome-mining papers that also report machine-learning-based gene-cluster classification, or docking studies that also report phytochemical-profiling data). This has a direct implication for how the field should be read by funders, journal editors, and meta-researchers: treating "AI-assisted natural-product drug discovery" as five cleanly separate specialities (as a high-modularity field would invite) would misrepresent how tightly interwoven the underlying methods and applications actually are in the literature.

Theoretically, the burst patterns identified in Section 4.4-concentrated around nanocarrier/drug-delivery engineering and oncology applications in 2024-25-suggest that the field's current research frontier is shifting from method-development questions (can AI predict bioactivity, mine genomes, or dock candidate compounds at all) toward translational and application-specific questions (how can AI-identified natural compounds actually be delivered effectively to a target tissue, and can they be optimized specifically for cancer applications).

This reading is consistent with the independently published narrative literature on these specific subthemes: Noury and colleagues' review of AI-driven smart nanocarrier design explicitly frames AI's role in nanomedicine as accelerating formulation optimization and predicting biological-barrier interactions rather than basic feasibility questions,[17] and Chunarkar-Patil and colleagues likewise frame the integration of deep learning with natural-product anticancer drug discovery as the next translational step beyond the already-established computational-to-preclinical pipeline.[16] This is consistent with a field whose core computational toolkit has matured to the point where attention is now shifting toward downstream translational bottlenecks rather than basic methodological feasibility.

These practical implications are grounded in specific quantitative findings rather than general bibliometric observation. For funding bodies: the 0.3-year synchrony across all five cluster mean years indicates that no subtheme is currently in an early foundational phase that would benefit from disproportionate nurturing investment-all five are mature and growing simultaneously, suggesting that cross-cluster integrative projects (e.g, combining genome mining with ML-based bioactivity prediction) are where the highest marginal returns on investment are likely to lie, consistent with the high density (0.97) of cross-cluster keyword connections. For journal scope committees: the low modularity ($Q$=0.15) empirically demonstrates that manuscripts in this space routinely bridge multiple clusters; editors who enforce narrow subject-matter requirements for individual sub-fields will reject manuscripts that the field itself treats as integrative contributions. For researchers: the country-level CPP/HCP dissociation (China leads output, Western nations lead citations per paper) quantifies an existing but previously unvalidated perception-researchers seeking high-impact collaboration partners should prioritize United States (HCP=56), Germany (CPP=57.77), and Canada (CPP=63.92) regardless of output volume.

Practically, several audiences can act on these findings. Funding agencies considering where to direct support within this space might note that network-pharmacology/mechanism studies constitute the largest existing cluster by member-term count (Cluster No. 0), while nanocarrier/drug-delivery and oncology applications represent the sharpest-growing recent foci (Section 4.4)-two different but complementary signals of where research capacity is concentrated versus where it is currently expanding fastest. Journal editors and reviewers, faced with the field's low modularity, may find it more useful to evaluate submissions against this field's genuinely crosscutting methodological base rather than forcing papers into a single narrow sub-specialty. Researchers new to the area can use the five-cluster structure in Table 1 as a practical map of the field's major existing vocabularies before designing a new study.

Several limitations should be considered when interpreting these results. First, the analysis is based on Scopus coverage only; Web of Science, PubMed, or preprint servers (bioRxiv, ChemRxiv) may index a partially different set of documents, particularly for very recent or interdisciplinary work that crosses traditional journal-indexing boundaries. Second, the dataset is restricted to English-language articles and reviews, which may under-represent non-English-language research, particularly from China-based venues given that affiliations from China are nonetheless the single largest national contributor in this dataset (Table 3). Third, the 2025 citation counts in Figure 2 are necessarily incomplete as of the retrieval date and should not be read as reflecting the eventual scholarly impact of that year's output. Fourth, because no References/cited-reference field was available in this export, the analysis relies entirely on keyword co-occurrence rather than combining it with a complementary reference co-citation map (which typically identifies foundational works rather than topics); a future analysis incorporating a references-populated export could meaningfully extend these findings. Fifth, the very high network density discussed in Section 4.2, while substantively interpreted here as a property of a thematically concentrated field, also means that the Louvain clustering's modularity score should be read with appropriate caution-a low-Q clustering is less robust to small changes in thresholding parameters than a high-Q one, and a different minimum-frequency or node-cap choice could plausibly redraw cluster boundaries somewhat differently while likely preserving the broader five-theme picture.

Sixth, reviewers have noted that the identified clusters (machine learning, molecular docking, network pharmacology, genome mining, phytochemical profiling) correspond to divisions within the field that are qualitatively familiar to practitioners. We acknowledge this and clarify that the contribution of the present analysis is not the identification of these subthemes *per se*, but the quantification of their relative sizes, internal coherences, temporal synchrony, and cross-cluster connectivity- none of which can be established by qualitative observation alone. Specifically: the finding that all five clusters share mean years within 0.3 years (rather than spanning a decade as in sequentially developing fields), that the network-pharmacology cluster is 39% larger by node count than the next-largest cluster, and that the overall network density (0.97) approaches a complete graph-these are quantitative results that contradict the intuitive expectation of a modular multi-specialty field and that could not be established through narrative review. Conclusions are stated accordingly throughout the revised manuscript.

## CONCLUSION

This scientometric analysis of 4277 Scopus-indexed articles and reviews (2016-25) shows that AI-assisted natural-product drug discovery has grown more than fourteen-fold in annual publication output over the past decade, organizing into five thematically distinct but densely interconnected subthemes-network pharmacology and herbal-medicine

mechanisms, natural-product genome mining and biosynthesis informatics, machine learning/AI predictive methods, molecular docking and structure-based virtual screening, and phytochemical profiling/analytic chemistry-that have matured essentially in parallel rather than sequentially, all carrying mean publication years within a 0.3-year band of one another. Recent keyword bursts concentrated in 2024-25 point to nanocarrier/drug-delivery engineering and oncology applications as the field's current frontier of fastest-growing attention.

The study's contribution is not the application of a bibliometric workflow to a new topic domain, but the quantitative structural findings that emerge from that application: the empirical demonstration that AI-assisted natural-product drug discovery is a single, densely integrated research front (density=0.97, $Q$=0.15) in which five subthemes grow synchronously rather than sequentially; the dissociation between national output leadership and citation-impact leadership; and the identification of nanocarrier engineering and oncology applications as quantifiably surging frontiers. These findings provide three actionable recommendations: (1) funding agencies should prioritise cross-cluster integrative grants rather than sub-field-specific calls; (2) journals should adopt scope definitions that explicitly welcome cross-cluster manuscripts; and (3) researchers targeting high-impact collaboration should prioritize partnerships with Canada, Switzerland, Germany, and the United States based on their demonstrated CPP and HCP leadership rather than publication-volume rankings.

Several concrete directions for future scientometric monitoring of this field follow directly from the limitations identified in Section 5. First, a follow-up analysis using a Web of Science or multi-database export with a populated cited-reference field would permit a complementary co-citation analysis identifying the field's most foundational works-a question the present keyword-only analysis cannot answer. Second, given the sharp 2024-25 growth and the corresponding citation-lag visible in Figure 2, a focused re-analysis once 2024-25 citation data has matured (e.g, in 2027-28) would allow a more reliable assessment of which of the current bursts (nanocarrier/drug-delivery, oncology applications, probiotics/food-safety) translated into lasting impact versus short-lived attention spikes. Third, because the network's low modularity appears partly attributable to the breadth of the OR-based search string used here, a narrower follow-up study restricted to a single subtheme (for instance, AI-assisted genome mining alone, or natural-product-focused molecular generative models alone) could test whether modularity rises-and distinct sub-clusters become more separable-once the corpus is scoped more tightly. Fourth, given that Section 4.5 shows research output concentrated in a relatively small number of countries (led by China, India, and the United States), a dedicated study of international collaboration networks (author-affiliation co-authorship analysis) would extend the geographic picture sketched only descriptively here into a fully quantified collaboration map.

## ACKNOWLEDGEMENT

We are thankful to the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, for their financial support.

## ABBREVIATIONS

**AI:** Artificial Intelligence; **ML:** Machine Learning; **TCM:** Traditional Chinese Medicine; **TCM-NP:** TCM Network Pharmacology; **BGC:** Biosynthetic Gene Cluster; **LC-MS/HPLC:** Liquid Chromatography-Mass Spectrometry / High-Performance Liquid Chromatography; **QSAR:** Quantitative Structure-Activity Relationship; **TP:** Total Publications; **TC:** Total Citations; **CPP:** Citations per Paper; **TA:** Total Authors; **HCP:** Highly Cited Papers; **FP:** Funded Papers; **ICP:** International Collaborative Papers.

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

## FUNDING

This work was funded through the Ambitious Researcher Track-Review Articles by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Al-Ahsa, Saudi Arabia (grant No. KFU263587).

## REFERENCES

1. Mullowney MW, Duncan KR, Elsayed SS *et al.* Artificial intelligence for natural product drug discovery. Nat Rev Drug Discov. 2023;22(11):895-916. DOI: 10.1038/s41573-023-00774-7.
2. Arora A, Alber G, Kumar A, Thakur S, Singh TP, Sharma S. Bibliometric analysis of artificial intelligence in drug discovery. Drug Discov Today. 2021;26(7):1575-1583. DOI: 10.1016/j.drudis.2021.03.023.
3. Xie Y, Chen W, Ren J *et al.* Bibliometric analysis of machine learning in traditional Chinese medicine research: focus on clinical outcome prediction. Front Pharmacol. 2023;14:1179456. DOI: 10.3389/fphar.2023.1179456.
4. Chemat F, Rombaut N, Sicaire A-G, Meullemiestre A, Fabiano-Tixier A-S, Abert-Vian M. Ultrasound assisted extraction of food and natural products. Mechanisms, techniques, combinations, protocols and applications. A review. Ultrason Sonochem. 2017;34:540-560. DOI: 10.1016/j.ultsonch.2016.06.035.
5. Thomford NE, Senthebane DA, Rowe A *et al.* Natural products for drug discovery in the 21st century: innovations for novel drug discovery. Int J Mol Sci. 2018;19(6):1578. DOI: 10.3390/ijms19061578.
6. Zhang P, Zhang D, Zhou W *et al.* Network pharmacology: towards the artificial intelligence-based precision traditional Chinese medicine. Brief Bioinform. 2023;25(1):bbad518. DOI: 10.1093/bib/bbad518.
7. Blin K, Shaw S, Steinke K *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019; 47(W1):W81-W87. DOI: 10.1093/nar/gkz310.
8. Van Heel AJ, De Jong A, Song C, Viel JH, Kok J, Kuipers OP. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. Nucleic Acids Res. 2018; 46(W1):W278-W281. DOI: 10.1093/nar/gky383.
9. Hannigan GD, Prihoda D, Palicka A *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res. 2019;47(18):e110. DOI: 10.1093/nar/gkz654.
10. Dührkop K, Nothias L-F, Fleischauer M *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat Biotechnol. 2021;39(4):462-471. DOI: 10.1038/s41587-020-0740-8.

11. Hoffmann MA, Nothias L-F, Ludwig M *et al.* High-confidence structural annotation of metabolites absent from spectral libraries. Nat Biotechnol. 2021;40(3):411-421. DOI: 10.1038/s41587-021-01045-9.
12. Liu C-S, Yan B-C, Sun H-D, Lu J-C, Puno P-T. Bridging chemical space and biological efficacy: advances and challenges in applying generative models in structural modification of natural products. Nat Prod Bioprospect. 2025;15(1):37. DOI: 10.1007/s13659-025-00521-y.
13. Yoo S, Yang HC, Lee S *et al.* A deep learning-based approach for identifying the medicinal uses of plant-derived natural compounds. Front Pharmacol. 2020;11:584875. DOI: 10.3389/fphar.2020.584875.
14. Yabuuchi H, Hayashi K, Shigemoto A *et al.* Virtual screening of antimicrobial plant extracts by machine-learning classification of chemical compounds in semantic space. PLOS One. 2023;18(5):e0285716. DOI: 10.1371/journal.pone.0285716.
15. El-Ashrey MK, Bakr RO, Fayed MAA, Refaey RH, Nissan YM. Pharmacophore based virtual screening for natural product database revealed possible inhibitors for SARS-COV-2 main protease. Virology. 2022;570:18-28. DOI: 10.1016/j.virol.2022.03.003.
16. Chunarkar-Patil P, Kaleem M, Mishra R *et al.* Anticancer drug discovery based on natural products: from computational approaches to clinical studies. Biomedicines. 2024;12(1):201. DOI: 10.3390/biomedicines12010201.
17. Noury H, Rahdar A, Romanholo Ferreira LF, Jamalpoor Z. AI-driven innovations in smart multifunctional nanocarriers for drug and gene delivery: a mini-review. Crit Rev Oncol/Hematol. 2025;210:104701. DOI: 10.1016/j.critrevonc.2025.104701.
18. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. 2010;84(2):523-538. DOI: 10.1007/s11192-009-0146-3.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444. DOI: 10.1038/nature14539.
20. Moral-Muñoz JA, Herrera-Viedma E, Santisteban-Espejo A, Cobo MJ. Software tools for conducting bibliometric analysis in science: an up-to-date review. El Prof Inf. 2020;29(1):e290103. DOI: 10.3145/epi.2020.ene.03.


**Cite this article:** Almuqbil RM, Aldhubiab B, Ahmed MKK, Sab CM. Global Research Trends in AI-Assisted Natural Product Drug Discovery: A Scientometric and Keyword Co-Occurrence Network Analysis (2016-2025). Indian J of Pharmaceutical Education and Research. 2027;61(1):10.5530/ijper.20260025.