Pharmacophore

ISSN-2229-5402

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

ARTIFICIAL INTELLIGENCE-ENABLED DRUG DISCOVERY AFTER THE HYPE CYCLE: A BIBLIOMETRIC AND THEMATIC REVIEW OF TWENTY-FIVE YEARS OF RESEARCH

Elena Stankova[1]*, Christopher Brown[2], Siti Aisyah[3], Olivier Durand[4]

1. ***Department of Bibliometric and Thematic Review of AI in Drug Discovery, Faculty of Pharmacy, Karolinska Institute, Stockholm, Sweden.***
2. ***Department of AI-Enabled Drug Discovery and Hype Cycle Analysis, Faculty of Pharmacy, University of Toronto, Toronto, Canada.***
3. ***Department of Twenty-Five Years of AI in Pharma Research, Faculty of Pharmaceutical Sciences, National University of Malaysia, Kuala Lumpur, Malaysia.***
4. ***Department of Post-Hype AI Translation, Faculty of Pharmacy, University of Bordeaux, Bordeaux, France.***

ARTICLE INFO

Received:
15 May 2025
Received in revised form:
28 July 2025
Accepted:
03August 2025
Available online:
28 August 2025

Keywords: Artificial intelligence, Drug discovery, Bibliometric analysis, Science mapping, Thematic analysis, Generative models

ABSTRACT

Artificial intelligence has progressed from a specialized computational aid to a prominent organizing concept across target identification, molecular screening, lead optimization, synthesis planning, drug repurposing, and development strategy. However, the visibility of the field has grown faster than the evidentiary frameworks needed to distinguish methodological novelty from pharmaceutical value. This bibliometric and thematic review examines how the research landscape should be interpreted after the most promotional phase of the AI drug-discovery discourse. The review combines a transparent dual-database corpus-construction protocol with performance analysis, science mapping, and structured thematic coding. Rather than treating publication growth, citation impact, collaboration density, or benchmark performance as direct evidence of maturity, the synthesis evaluates what each indicator can establish and where additional validation is required. The resulting interpretation separates methodological expansion from evidence maturation and distinguishes computational capability, comparative benchmark performance, external generalization, prospective experimental confirmation, developability, clinical evaluation, and routine pharmaceutical readiness. The review further proposes claim-level hype-cycle indicators based on validation realism, reproducibility, data suitability, evidence provenance, prospective confirmation, and correspondence between model endpoints and pharmaceutical decisions. Bibliometric structures are interpreted as properties of the constructed corpus rather than universal representations of the field, while thematic clusters are treated as organizing devices rather than proof of scientific coherence or translational success. Important limitations include database-dependent coverage, incomplete visibility of proprietary industrial research, terminology drift, uneven reporting of negative findings, and the difficulty of attributing program-level outcomes to individual computational components. The central implication is that the next phase of AI-enabled drug discovery should be evaluated through decision-relevant evidence, traceable data practices, realistic validation, and explicit boundaries between prediction and pharmaceutical utility.

To Cite This Article: Stankova E, Brown C, Aisyah S, Durand O. Artificial Intelligence-Enabled Drug Discovery after the Hype Cycle: A Bibliometric and Thematic Review of Twenty-Five Years of Research. Pharmacophore. 2025;16(4):62-71. https://doi.org/10.51847/enwqtllm3h

Introduction

Artificial intelligence-enabled drug discovery now encompasses molecular property prediction, target prioritization, virtual screening, synthesis planning, biological image interpretation, candidate generation, drug repurposing, and selected development activities. This breadth has created a field that is scientifically important but difficult to evaluate as a single technological category. Methods differ in input data, learning objective, pharmaceutical task, validation setting, and proximity to an actionable discovery decision. Machine learning may support multiple stages of discovery and development, but data limitations, repeatability, interpretability, and the connection between model output and experimental decisions remain central

Corresponding Author: Elena Stankova; Department of Bibliometric and Thematic Review of AI in Drug Discovery, Faculty of Pharmacy, Karolinska Institute, Stockholm, Sweden. E-mail: elena.stankova@ki.se.

determinants of impact [1]. Consequently, the field cannot be assessed adequately through a simple contrast between AI-enabled and conventional research. It requires an evidence structure that distinguishes what a model computes from what a pharmaceutical program can conclude or act upon.

The expansion of AI applications has also blurred the distinction between broad participation in the pharmaceutical pipeline and demonstrated improvement in that pipeline. Target identification, compound screening, lead optimization, toxicity prediction, trial design, and development planning involve different evidence standards, yet they are frequently combined under a common narrative of acceleration. Reviews of AI in drug development have documented this expanding application range while also showing that practical prospects remain uneven across stages and use cases [2]. A method that performs well in retrospective molecular prediction may be relevant to early hypothesis generation but remain remote from synthesis, experimental confirmation, developability, or clinical value. The central analytical problem is therefore not whether AI is present in drug discovery, but how its heterogeneous claims should be classified and compared without implying equivalence among fundamentally different evidence levels.

This problem is intensified by the rhetoric of technological cycles. Highly visible benchmark gains, generated molecular structures, or compressed discovery timelines can create an impression of generalized pharmaceutical transformation even when evaluation is limited to narrow datasets, retrospective splits, favourable baselines, or selectively reported examples. Critical examination of the field has shown that discovery impact is often inferred from assumptions that are not established by benchmark performance alone [3]. A post-hype assessment should therefore avoid both indiscriminate enthusiasm and indiscriminate scepticism. The appropriate unit of evaluation is the individual claim: what task was performed, against which comparator, under what data conditions, with what uncertainty, and with what independent experimental or translational confirmation. This claim-level approach allows genuine advances to be recognized without extending their implications beyond the available evidence.

Automation provides a useful illustration of this distinction. Computational design, synthesis planning, robotic experimentation, and iterative learning can potentially form integrated discovery systems, but the presence of each component does not demonstrate that the resulting workflow is reliable, autonomous, or superior to established practice. Reviews of automated drug discovery emphasize that value arises from the integration of design, synthesis, testing, and learning rather than isolated algorithmic performance [4]. Building on that boundary, this article evaluates the intellectual and organizational development of AI-enabled drug discovery through a bibliometric and thematic framework. Its contribution is a proposed synthesis linking publication structures, methodological transitions, collaboration patterns, reproducibility practices, and translational evidence. The framework is interpretive rather than predictive: it does not assign a validated maturity score, forecast clinical success, or establish regulatory or operational readiness.

Bibliometric Questions, Databases, and Corpus-Construction Methods

The bibliometric design is organized around questions that correspond to distinct analytical objects. Performance questions examine how research output, scholarly influence, sources, authors, institutions, and countries are distributed within the constructed literature. Relational questions investigate coauthorship, bibliographic coupling, co-citation, and keyword co-occurrence. Temporal questions examine whether attention moved among classical machine learning, deep learning, generative design, pretrained representations, knowledge graphs, and foundation or tool-augmented models. Translational questions require an additional thematic layer that codes whether publications provide retrospective benchmarks, external validation, prospective experiments, developability evidence, clinical evaluation, or operational implementation. The bibliometrix framework is suitable for integrating performance analysis, network analysis, conceptual-structure mapping, and longitudinal examination within a reproducible workflow [5]. Nevertheless, software output is not self-interpreting: each indicator must be connected to a stated question and bounded claim.

The corpus-construction protocol uses Web of Science Core Collection and Scopus as complementary bibliographic sources. The concept block combines terms for artificial intelligence, machine learning, deep learning, generative models, foundation models, large language models, and graph-based learning. The pharmaceutical block includes drug discovery, drug development, molecular design, virtual screening, lead optimization, target identification, and drug repurposing. Searches are applied to titles, abstracts, and keywords using database-specific field codes and proximity operators, with raw queries and exports retained for reproducibility. Eligible records are peer-reviewed journal articles and reviews that address a substantive computational or pharmaceutical application; conference-only publications, preprints, books, chapters, theses, non-peer-reviewed reports, and records with merely incidental terminology are excluded. Methodological guidance for bibliometric research supports this explicit alignment among research questions, database selection, cleaning rules, analytical procedures, and interpretation [6].

Records are deduplicated first by exact DOI and then by normalized title, publication year, and first-author identity, with manual adjudication of near duplicates, online-first records, corrections, and version-of-record inconsistencies. Author names are normalized using persistent identifiers where available, supplemented by audited handling of initials, diacritics, name changes, and homonyms. Institutions are harmonized through an institution dictionary that distinguishes subsidiaries, historical names, hospitals, universities, biotechnology companies, technology firms, and multinational pharmaceutical organizations. Full and fractional counting must be reported separately where they answer different questions. Database coverage, missing affiliations, incomplete cited-reference fields, indexing differences, and imperfect author disambiguation can materially alter

bibliometric maps and rankings [7]. The review therefore treats every quantitative output as reproducible only within the documented corpus and parameters, rather than as an invariant property of AI-enabled drug discovery.

Table 1 organizes the evidence, constructs, and interpretation boundaries needed to develop bibliometric questions, databases, and corpus-construction methods within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Bibliometric questions, databases, and corpus-construction methods

Evidence domain	Eligible evidence or method	Reported capability or claim	Strength of support	Reproducibility or bias concern	Unresolved gap
Corpus scope and eligibility	Database queries, field restrictions, document-type criteria, manual screening rules	Defines which records can contribute to the bibliometric and thematic synthesis	Strong when complete queries and decisions are archived	Terminology drift, indexing asymmetry, language restrictions, and incidental AI terminology may alter inclusion	No universally accepted operational definition of AI-enabled drug discovery
Database coverage	Complementary exports from Web of Science Core Collection and Scopus	Broadens coverage of journals, affiliations, references, and citation metadata	Moderate to strong when database-specific differences are reported	Coverage varies by journal, year, document type, and cited-reference completeness	Proprietary industrial research and non-indexed outputs remain incompletely visible
Deduplication and version control	DOI matching, normalized-title matching, author-year checks, manual adjudication	Reduces double counting of records and citation relationships	Strong for exact identifiers; moderate for incomplete metadata	Online-first and final records, corrections, and title variants can survive automated rules	Standardized reporting of version-resolution decisions remains uncommon
Author and institution normalization	ORCID, institutional identifiers, curated dictionaries, manual audit	Supports more reliable productivity and collaboration networks	Moderate when high-impact nodes are audited	Homonyms, name changes, subsidiaries, mergers, and multinational affiliations may distort networks	Complete disambiguation is rarely attainable from bibliographic metadata alone
Performance analysis	Publication output, source distribution, age-aware citation indicators, contributor productivity	Describes scholarly activity and influence within the corpus	Strong for documented descriptive patterns	Citation accumulation, field prestige, self-citation, and database coverage influence results	Scholarly influence cannot be equated with pharmaceutical usefulness
Science mapping	Coauthorship, co-citation, bibliographic coupling, and keyword co-occurrence	Reveals observable intellectual, social, and conceptual relations	Moderate when thresholds and counting methods are transparent	Clusters depend on relatedness measures, pruning, resolution, and layout choices	Network proximity does not establish causality, quality, or translational maturity
Thematic and evidence-stage coding	Deductive codebook, inductive refinement, dual-reviewer calibration, adjudication	Connects bibliometric structures to methods, applications, validation settings, and maturity	Moderate to strong when coding rules and disagreements are reported	Abstract-level coding may miss methodological limitations or negative findings	Evidence-stage reporting is inconsistent across computational and pharmaceutical studies
Reproducibility package	Raw exports, cleaning scripts, normalization dictionaries, parameter files, and coding guidance	Permits audit, rerunning, and comparison with future corpora	Strong when materials are complete and accessible	Proprietary database licensing may limit redistribution of raw records	Durable standards for preserving bibliometric workflows remain underdeveloped

Performance Analysis, Science Mapping, and Thematic Coding

Performance analysis describes the observable scholarly structure of the final corpus rather than the performance of AI models. Its outputs may include annual publication activity, document and source distributions, authorship patterns, institutional participation, country affiliations, citation influence, and the relationship between publication age and accumulated citations. These indicators answer questions about visibility, participation, and scholarly concentration, but they do not show whether a model generalizes, whether a generated molecule is synthesizable, or whether an AI-supported decision improved a pharmaceutical program. Science mapping adds a relational layer by examining how publications, authors, organizations, references, and concepts are connected. Citation-based clustering is useful for identifying intellectual groupings, yet the resulting structure depends on choices concerning relatedness measures, clustering resolution, inclusion thresholds, and interpretation [8]. Cluster labels must therefore be supported by representative records and thematic reading rather than generated from network position alone.

Previous bibliometric work in AI-enabled drug discovery illustrates the value and limits of performance indicators. Publication trajectories can identify periods of rising scholarly attention; citation measures can identify influential records; coauthorship can reveal visible collaboration; and keyword analysis can show recurring methodological or application terms. These outputs become misleading, however, when differences in queries, databases, periods, document types, or cleaning procedures are ignored. Domain-specific bibliometric analysis has demonstrated that publication, citation, collaboration, and thematic indicators can organize the development of AI in drug discovery, while its numerical results remain inseparable from the

corpus used to produce them [9]. Accordingly, the present review uses prior bibliometric studies as methodological comparators and contextual evidence, not as sources from which counts, rankings, network centralities, or cluster sizes can be transferred.

Science mapping is paired with thematic coding because automated networks cannot resolve the evidentiary meaning of a publication. A keyword cluster containing terms such as deep learning, virtual screening, molecular generation, or drug repurposing may indicate shared vocabulary while concealing substantial differences in data quality, validation design, pharmaceutical task, and intended decision. Author keywords are normalized for spelling, abbreviations, singular–plural variants, and conceptually equivalent expressions, while generic terms are separated from task-specific or technology-specific language. Temporal slices are interpreted cautiously because terminology changes as methods are renamed, combined, or repositioned. Prior mapping of AI-enabled drug discovery shows that thematic structures are affected by keyword harmonization, clustering decisions, and temporal segmentation [10]. For this reason, cluster naming is treated as a documented interpretive act rather than an objective label automatically produced by the software.

The integrated synthesis uses three mutually constraining layers. Performance analysis identifies who publishes, where the work appears, and how scholarly attention is distributed. Science mapping identifies observable intellectual and collaborative structures. Thematic coding then determines what the records actually claim, which pharmaceutical tasks they address, which data and representations they use, and what level of validation they provide. Long-horizon bibliometric analysis can reveal growth, collaboration, and thematic transition, but findings from one corpus should not be imported as results for another corpus [11]. More importantly, bibliometric research contributes to theory only when indicators are connected to a defensible explanatory interpretation rather than presented as descriptive rankings or visually attractive networks [12]. The proposed interpretation therefore evaluates whether changes in publication structure coincide with changes in methodological focus and evidence maturity, while explicitly withholding causal or translational conclusions that the underlying indicators cannot establish.

Growth Phases in AI-Enabled Drug Discovery

The historical development of AI-enabled drug discovery is better understood as a sequence of overlapping methodological phases than as a series of abrupt technological replacements. Early work was dominated by descriptor-based quantitative structure–activity relationships, supervised classification, regression, docking prioritization, and other task-specific models. The subsequent deep-learning phase expanded the range of learnable representations and applications, including bioactivity prediction, molecular property estimation, synthesis prediction, biological imaging, and molecular generation [13]. This phase did not eliminate classical methods; rather, it increased the number of tasks for which representation learning and larger datasets appeared technically feasible. The bibliometric implication is that publication growth should be interpreted together with changes in task diversity, model architecture, data modality, and validation setting.

The expansion of deep learning also exposed persistent limitations that publication counts and citation growth cannot resolve. Data quality, assay heterogeneity, class imbalance, interpretability, limited applicability domains, and distribution shift continued to constrain generalization [14]. In some areas, architectural complexity increased faster than the quality of evaluation. Random data partitions, retrospective benchmarks, and comparisons against weak or inconsistently tuned baselines could make methodological progress appear more decisive than it was. Accordingly, the deep-learning phase should not be classified as mature merely because its terminology became dominant. Maturity requires evidence that model improvements survive realistic validation, support experimentally meaningful decisions, and remain reproducible across datasets, laboratories, or organizations.

A further growth phase was characterized by the diffusion of AI across a wider pharmaceutical task portfolio. Applications extended from molecular screening and lead optimization to repurposing, toxicity assessment, formulation-related questions, clinical-development support, and workflow coordination [15]. This diversification generated an increasingly heterogeneous literature in which papers labelled as AI-enabled drug discovery could address fundamentally different endpoints. Bibliometric growth therefore reflects both genuine methodological expansion and category broadening. A defensible thematic analysis must distinguish whether a publication predicts an experimentally observable property, proposes a candidate, prioritizes an existing asset, interprets biomedical evidence, or assists an organizational decision. Without this separation, the apparent expansion of the field can obscure large differences in validation burden and pharmaceutical relevance.

The more recent phase includes message-passing networks, symmetry-aware learning, generative molecular design, reaction prediction, large-scale pretraining, knowledge graphs, and tool-augmented language models. These developments indicate increasing specialization of architectures around chemical and biological structure [16]. They also reflect convergence among expanding data resources, computational capacity, experimental automation, and pharmaceutical use cases rather than a single universally transformative algorithm [17]. The proposed growth-phase interpretation therefore separates methodological adoption from evidence maturation. A field may move rapidly toward new representations and model classes while remaining uneven in prospective validation, developability assessment, clinical evidence, and operational integration. Growth is consequently treated as multidimensional: more publications, broader tasks, richer representations, larger collaborative networks, and stronger evidence are related but non-equivalent forms of development.

Shifts from Classical Machine Learning to Generative and Foundation Models

The transition from classical machine learning to newer model families begins with a change in molecular representation. Classical approaches often rely on predefined descriptors, fingerprints, physicochemical variables, or expert-designed features. Learned representations instead derive task-relevant patterns from molecular strings, graphs, three-dimensional structures, reactions, images, sequences, or heterogeneous biomedical relationships. Representation choice determines which information is retained, which invariances are encoded, and which types of extrapolation are plausible [18]. The shift is therefore not simply from "small" to "large" models. It is a transition from manually specified abstractions toward representations learned from data, often with less transparent relationships between the encoded features and the pharmaceutical decision.

Generative modeling introduced a further conceptual change by moving from prediction over predefined candidates to the proposal of new molecular structures. Continuous latent representations enabled molecular optimization within learned chemical spaces and demonstrated that generation could be coupled to property-oriented objectives [19]. Nevertheless, computational validity, novelty, or movement toward a predicted objective does not establish that a generated molecule is synthesizable, stable, selective, developable, safe, or therapeutically valuable. Generative outputs are hypotheses whose usefulness depends on constraints, comparator quality, medicinal-chemistry assessment, and experimental confirmation. Bibliometric growth in generative-design publications must therefore be interpreted alongside the proportion of studies that progress beyond retrospective generation metrics.

Inverse-design perspectives further reframed molecular discovery as a conditional search problem in which desired properties guide the construction of candidate structures [20]. This framing is attractive because it appears to reverse the conventional sequence from molecule to measured property. However, pharmaceutical objectives are incomplete, uncertain, and often conflicting. Potency, selectivity, solubility, permeability, metabolic stability, toxicity, formulation feasibility, intellectual-property position, and commercial context cannot be reduced reliably to a single optimized score. Thematic coding should consequently distinguish unconstrained generation, single-objective optimization, multi-objective design, synthesis-aware generation, experimentally tested generation, and program-level candidate progression. These categories represent materially different evidence states.

Transformer architectures brought language-based modeling into reaction prediction and molecular learning. Reaction prediction could be formulated as translation between chemical expressions while incorporating estimates of predictive uncertainty [21]. Large-scale self-supervised pretraining subsequently showed that molecular language representations can transfer across multiple property-prediction tasks [22]. Yet transfer across benchmarks is not equivalent to transfer across chemical series, organizations, biological mechanisms, or development decisions. **Figure 1** presents the AI drug-discovery hype-cycle evidence timeline, showing how the article's key components and boundaries are connected within shifts from classical machine learning to generative and foundation models.

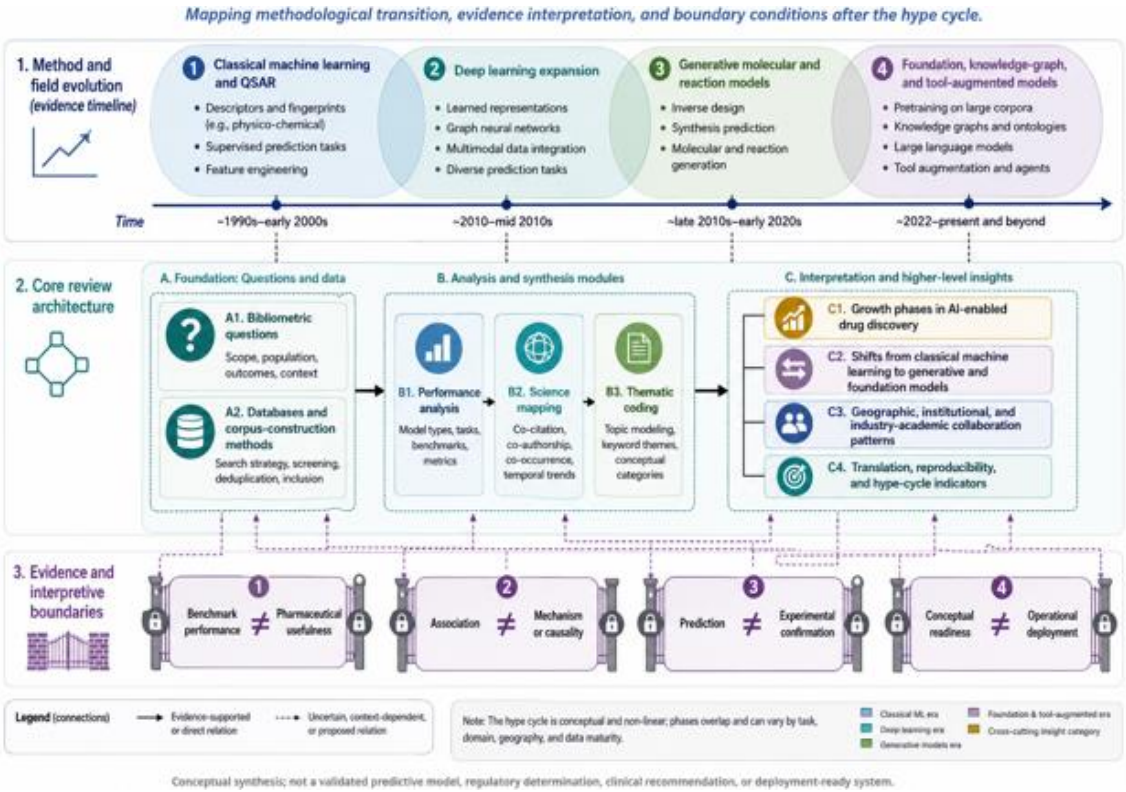

Figure 1. AI Drug-Discovery Hype-Cycle Evidence Timeline.

The figure is an original conceptual synthesis that organizes the article's central contribution across bibliometric questions, databases, and corpus-construction methods, bibliometric questions, corpus-construction methods, performance analysis,

science mapping, and thematic coding, performance analysis, science mapping. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

Geographic, Institutional, and Industry-Academic Collaboration Patterns

Geographic and institutional patterns reveal where visible scholarship is produced and how contributors are connected, but they do not directly measure technological capability or pharmaceutical readiness. Previous bibliometric analyses have shown that apparent country leadership, institutional concentration, and collaboration structure depend on the selected period, database, document types, and operational definition of AI-enabled drug development [23]. Full counting can amplify the apparent contribution of highly collaborative records, whereas fractional counting distributes credit across participating addresses. Multinational affiliations further complicate national attribution. Geographic outputs should therefore be reported with explicit counting rules and interpreted as properties of the documented literature rather than rankings of national scientific superiority.

Institutional analysis requires similar caution. Universities, hospitals, biotechnology firms, technology companies, contract research organizations, and multinational pharmaceutical companies participate through different publication and disclosure practices. Bibliometric mapping can organize institutional, country, funding, and thematic relationships, but each relationship is conditional on the records indexed in the corpus [24]. Proprietary projects, licensing agreements, private datasets, failed programs, internal validation, and commercial partnerships may remain invisible. An institution with fewer publications may nevertheless possess substantial operational capability, while a highly publishing institution may focus on foundational or benchmark research. Institutional productivity and translation must therefore be coded separately.

Industry–academic collaboration is particularly relevant because AI-enabled drug discovery requires expertise spanning chemistry, biology, pharmacology, data engineering, statistics, experimental platforms, clinical development, and regulatory science. Clinical-development applications illustrate that computational models must interact with clinical operations, biostatistics, data governance, and decision processes [25]. Coauthorship networks can identify visible collaboration but cannot establish whether responsibilities were integrated effectively or whether the collaboration improved scientific decisions. Thematic coding should distinguish coauthorship, data-sharing partnerships, method development, experimental validation, asset development, and clinical-development participation where these relationships are reported clearly.

End-to-end drug discovery depends on linked data, experimental evidence, expert judgment, and handoffs across multiple organizational functions [26]. Pharmaceutical productivity also reflects decision quality, target confidence, safety strategy, developability, and organizational discipline rather than computational performance in isolation [27]. The proposed collaboration interpretation therefore combines network position with evidence-stage coding. A central institution in a coauthorship network may be influential in method development without demonstrating prospective pharmaceutical translation. Conversely, translational work may appear peripheral because commercially sensitive activities are not fully published. Collaboration maps should be used to identify observable knowledge exchange and structural concentration, not to infer causal productivity gains or hidden industrial readiness.

Translation, Reproducibility, and Hype-Cycle Indicators

Translation begins with the suitability of the underlying chemical and biological evidence. Experimental measurements may vary across assays, protocols, laboratories, endpoints, and curation practices. Biological labels can be incomplete or context-dependent, while chemical datasets may contain analog bias, duplicate structures, inconsistent stereochemistry, or information leakage. These limitations can create inflated impressions of generalization when models are evaluated on records closely related to their training data [28]. A translational assessment must therefore ask whether the data represent the intended pharmaceutical decision, whether the validation design approximates prospective use, and whether uncertainty or applicability boundaries are reported. Data availability alone does not establish data suitability.

Generative-model evaluation illustrates the need for claim-specific scrutiny. The ability to produce novel or formally valid molecular structures is a computational capability, not a demonstration of medicinal-chemistry value. Proposed compounds must be assessed against realistic comparators, synthesis constraints, structural liabilities, selectivity requirements, and experimentally relevant properties [29]. Reproducibility also requires transparent model configurations, training data, baselines, optimization objectives, random seeds, filtering rules, and failure reporting. Hype-cycle risk increases when attractive examples are emphasized while unsuccessful generations, infeasible structures, or negative experimental results remain unreported.

Benchmark frameworks improve comparability by specifying datasets, split procedures, metrics, and baseline models [30]. Their interpretation remains conditional on whether the split reflects the intended use. Random partitions may test interpolation among related compounds, whereas scaffold, temporal, or external splits may better approximate chemical novelty or prospective deployment. Large-scale model comparisons can reveal relative performance under a defined design, but their conclusions remain specific to the selected targets, assays, preprocessing, metrics, and validation scheme [31]. Reproducibility should therefore be treated as a layered property: computational rerun reproducibility, result reproducibility under independent implementation, and scientific reproducibility across new data or experiments.

Clinical progression represents a distinct evidentiary layer because an AI-supported asset entering human evaluation has passed through many chemical, biological, safety, manufacturing, organizational, and regulatory decisions not captured by a model

benchmark [32]. Even then, attribution is difficult: an asset may be called AI-discovered although computational methods supported only one component of a broader program. Field-wide opportunity, early realism concerns, and limitations in chemical and biological data jointly explain why the same literature can contain genuine advances and exaggerated expectations [1, 3, 28]. **Figure 2** presents the evidence, validation, and translation boundary observatory for artificial intelligence-enabled drug discovery after the hype, showing how the article's key components and boundaries are connected within translation, reproducibility, and hype-cycle indicators.

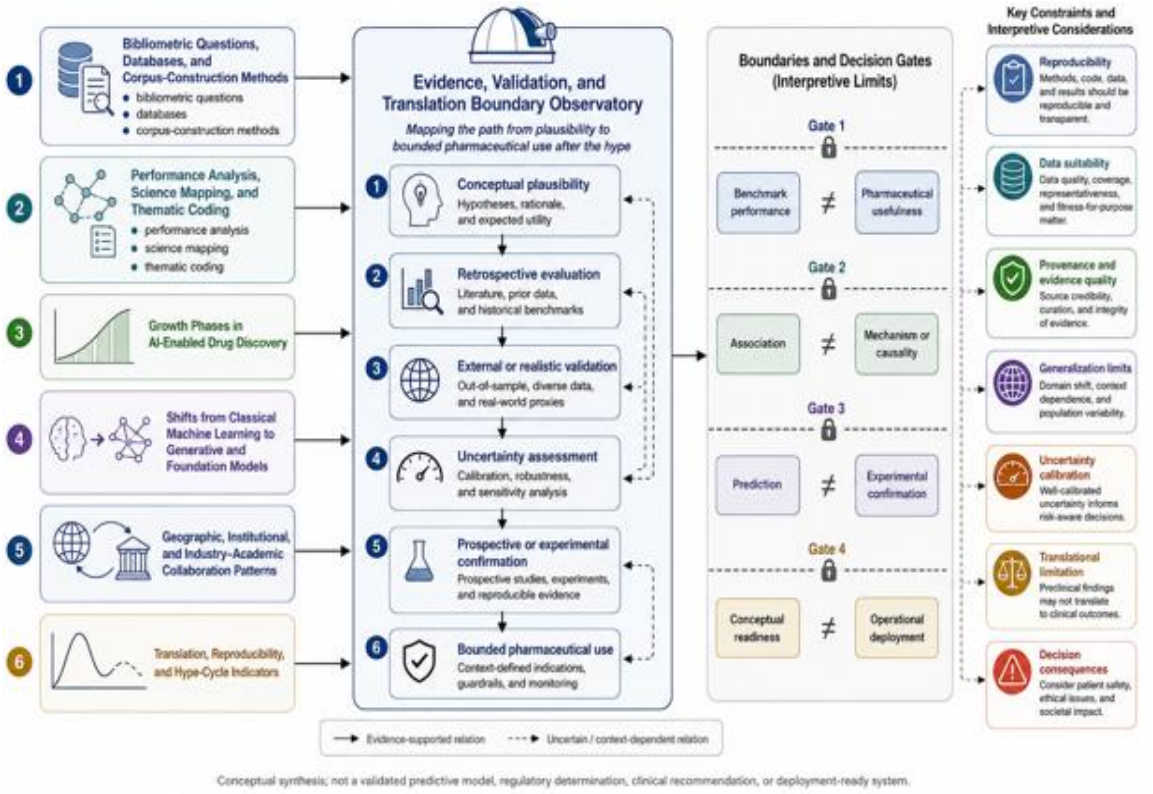

Figure 2. Evidence, Validation, and Translation Boundaries.

The figure is an original conceptual synthesis that shows how claims move from conceptual plausibility through evaluation, uncertainty assessment, and bounded pharmaceutical use. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

Emerging Themes and Priorities for the Next Research Phase

The first priority is evidence infrastructure. Knowledge graphs and integrated biomedical datasets can connect genes, proteins, pathways, diseases, compounds, assays, publications, and clinical observations, but the reliability of those connections depends on provenance, schema design, entity resolution, licensing, and evidence quality [33]. Larger or more connected datasets are not automatically more suitable for pharmaceutical inference. Future research should preserve source attribution, distinguish experimentally supported relations from computational associations, encode uncertainty, and make transformations auditable. Bibliometric analysis can help identify growth in knowledge-graph terminology, but thematic appraisal must determine whether publications address provenance and validation or merely use graph-based representations.

A second priority is the alignment of molecular representation with decision context. The evolution from handcrafted descriptors to learned string, graph, three-dimensional, and multimodal representations has expanded technical options without producing a universally superior representation [34]. Future comparisons should ask which representation supports the intended endpoint, whether structural or biological invariances are appropriate, how uncertainty is calibrated, and whether the model transfers to genuinely new chemical or biological contexts. Representation scale should not substitute for endpoint relevance. The next research phase requires evaluation designs that connect representation choice to the scientific decision being supported.

Foundation-model approaches create opportunities for transfer across tasks and heterogeneous biomedical evidence. A foundation model for clinician-centred drug repurposing illustrates how large-scale relational learning may organize broad repurposing hypotheses across diseases and therapeutic contexts [35]. Such systems may support prioritization, but they do not replace disease-specific mechanistic evidence, pharmacological assessment, experimental confirmation, or clinical evaluation. Their maturity should be judged by external transfer, robustness to missing or conflicting evidence, calibration,

provenance, and the usefulness of outputs to domain experts. Claims of generality require validation across tasks that were not implicitly represented in pretraining.

Tool-augmented language models may extend chemistry-oriented reasoning by connecting natural-language interfaces with calculation, retrieval, structure handling, or specialist software [36]. Reliability, traceability, tool-selection errors, hidden assumptions, and expert oversight remain central evaluation concerns. More broadly, the field's next phase depends on stronger ground truth, realistic validation, reproducibility, and appropriate human intervention [37]. Representation evolution and large-scale pretraining should therefore be assessed together rather than treated as independent evidence of readiness [18, 22, 34]. Credible progress likewise requires medicinal-chemistry scrutiny, clinical evidence, and field-level reproducibility rather than computational novelty alone [29, 32, 37]. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop emerging themes and priorities for the next research phase within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Artificial Intelligence-Enabled Drug Discovery after the Hype Cycle a Bibliometric and Thematic

Approach or application	Data and task context	Evaluation practice	Evidence maturity	Translational limitation	Review interpretation
Descriptor-based and classical machine learning	Curated molecular descriptors, fingerprints, assays, and task-specific endpoints	Cross-validation, external testing, applicability-domain analysis	Established computational capability for bounded tasks	Performance depends strongly on feature design, assay quality, and chemical-domain similarity	Remains useful where data and decisions are well specified; newer models do not automatically supersede it
Deep neural property prediction	Molecular strings, graphs, structures, images, and biological sequences	Realistic splits, baseline parity, calibration, independent datasets	Moderate and task-dependent	High benchmark performance may not transfer to novel chemistry or experimental settings	Evaluate against strong baselines and decision-relevant external data
Generative molecular design	Learned chemical representations and property-conditioned objectives	Validity, novelty, diversity, synthesis feasibility, medicinal-chemistry review, prospective testing	Demonstrated generation capability; uneven experimental maturity	Optimized scores do not establish synthesizability, developability, safety, or therapeutic value	Treat generated structures as hypotheses requiring multi-parameter and experimental evaluation
Reaction and synthesis prediction	Reaction corpora, precursor–product mappings, and chemical language	Temporal testing, uncertainty analysis, route plausibility, expert and laboratory confirmation	Strong benchmark capability in selected settings	Training records may not represent rare chemistry, process constraints, or laboratory feasibility	Useful for prioritization and planning when uncertainty and route constraints are explicit
Molecular foundation models	Large unlabeled or weakly labelled molecular corpora with downstream tasks	External transfer, leakage audit, ablation, calibration, cross-domain testing	Emerging transfer capability	Pretraining overlap and benchmark reuse can exaggerate apparent generality	Generality must be established across genuinely independent tasks and chemical domains
Knowledge graphs and relational learning	Heterogeneous biomedical entities, relations, publications, and experimental evidence	Link-prediction evaluation, provenance audit, expert review, disease-specific validation	Emerging to moderate	Integrated data may preserve hidden bias, uncertain relations, and inconsistent evidence strength	Provenance and evidence weighting are as important as graph scale
Tool-augmented language models	Natural-language requests linked to retrieval, calculation, chemistry tools, or software	End-to-end task testing, tool-call traceability, failure analysis, human comparison	Early demonstrational maturity	Hallucination, incorrect tool selection, hidden assumptions, and unverifiable intermediate steps	Evaluate complete workflows rather than fluency or isolated answers
AI-supported repurposing	Biomedical knowledge, clinical data, molecular evidence, and disease relationships	Retrospective recovery, external disease evaluation, prospective experimental or clinical follow-up	Variable; some externally evaluated systems	Association may be mistaken for mechanism, efficacy, or clinical suitability	Use for prioritization while preserving disease-specific evidentiary requirements
Prospective design–make–test systems	Integrated models, synthesis planning, automation, assays, and iterative learning	Pre-specified experimental protocols, negative-result reporting, cycle-level comparators	Emerging prospective maturity	Attribution, assay relevance, cost, and generalizability across programs remain uncertain	Stronger than retrospective benchmarks but not equivalent to clinical or routine readiness
Clinical and operational adoption	Trial design, candidate progression, portfolio decisions, and recurring pharmaceutical workflows	Prospective endpoints, multi-site validation, workflow comparison, governance, monitoring	Limited and context-specific	Asset outcomes depend on many non-AI factors and long development timelines	Routine readiness must be demonstrated separately from computational and preclinical success

Conclusion

This bibliometric and thematic review proposes a post-hype interpretation of AI-enabled drug discovery in which publication growth, network prominence, methodological novelty, and benchmark performance are treated as informative but insufficient indicators of maturity. Its central contribution is the integration of bibliometric structure with thematic evidence-stage coding, allowing changes in methods, representations, collaborations, and claims to be examined without collapsing them into a single narrative of technological progress. The review distinguishes computational capability from realistic generalization, prospective experimental value, developability, clinical utility, and routine pharmaceutical readiness. It also identifies provenance, reproducibility, realistic validation, negative-result reporting, decision-context alignment, and human–AI workflow evaluation as priorities for the next research phase. The proposed timelines, evidence boundaries, and maturity distinctions are explanatory syntheses rather than validated scoring instruments, regulatory frameworks, or deployment pathways. Their purpose is to make claims more comparable, expose where evidence remains incomplete, and direct attention from the visibility of AI methods toward the quality of the scientific and pharmaceutical decisions they may support.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5
2. Mak KK, Pichika MR. Artificial intelligence in drug development: Present status and future prospects. Drug Discov Today. 2019;24(3):773-80. doi:10.1016/j.drudis.2018.11.014
3. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009
4. Schneider G. Automating drug discovery. Nat Rev Drug Discov. 2018;17(2):97-113. doi:10.1038/nrd.2017.232
5. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. J Informetr. 2017;11(4):959-75. doi:10.1016/j.joi.2017.08.007
6. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: An overview and guidelines. J Bus Res. 2021;133:285-96. doi:10.1016/j.jbusres.2021.04.070
7. Öztürk O, Kocaman R, Kanbach DK. How to design bibliometric research: An overview and a framework proposal. Rev Manag Sci. 2024;18(11):3333-61. doi:10.1007/s11846-024-00738-0
8. van Eck NJ, Waltman L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. Scientometrics. 2017;111(2):1053-70. doi:10.1007/s11192-017-2300-7
9. Tran BX, Nghiem S, Sahin O, Vu TM, Ha GH, Vu GT, et al. Artificial intelligence in drug discovery: Bibliometric analysis of its research development. J Informetr. 2021;15(3):101566. doi:10.1016/j.joi.2021.101566
10. Zhang L, Zhang Y, Chen X, Zhang H. A bibliometric analysis of artificial intelligence in drug discovery and development from 2000 to 2021. Expert Opin Drug Discov. 2022;17(8):813-24. doi:10.1080/17460441.2022.2096016
11. Koçak M, Akçalı Z. The published role of artificial intelligence in drug discovery and development: A bibliometric and social network analysis from 1990 to 2023. J Cheminform. 2025;17(1):71. doi:10.1186/s13321-025-00988-4
12. Mukherjee D, Lim WM, Kumar S, Donthu N. Guidelines for advancing theory and practice through bibliometric research. J Bus Res. 2022;148:101-15. doi:10.1016/j.jbusres.2022.04.042
13. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23(6):1241-50. doi:10.1016/j.drudis.2018.01.039
14. Lavecchia A. Deep learning in drug discovery: Opportunities, challenges and future prospects. Drug Discov Today. 2019;24(10):2017-32. doi:10.1016/j.drudis.2019.07.006
15. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021;26(1):80-93. doi:10.1016/j.drudis.2020.10.010
16. Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G. Artificial intelligence in drug discovery: Recent advances and future perspectives. Expert Opin Drug Discov. 2021;16(9):949-59. doi:10.1080/17460441.2021.1909567
17. Chan HS, Shan H, Dahoun T, Vogel H, Yuan S. Advancing drug discovery via artificial intelligence. Trends Pharmacol Sci. 2019;40(8):592-604. doi:10.1016/j.tips.2019.06.004
18. David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: A review and practical guide. J Cheminform. 2020;12(1):56. doi:10.1186/s13321-020-00460-5

19. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268-76. doi:10.1021/acscentsci.7b00572
20. Sánchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science. 2018;361(6400):360-5. doi:10.1126/science.aat2663
21. Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular Transformer: A model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci. 2019;5(9):1572-83. doi:10.1021/acscentsci.9b00576
22. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell. 2022;4(12):1256-64. doi:10.1038/s42256-022-00580-7
23. Yang X, Wang Y, Byrne R, Schneider G, Yang S. The landscape of artificial intelligence in drug development: Bibliometric analysis of trends in 2014-2019. Front Pharmacol. 2020;11:1551. doi:10.3389/fphar.2020.01551
24. Karger E, Kureljusic M. Using artificial intelligence for drug discovery: A bibliometric study and future research agenda. Pharmaceuticals (Basel). 2022;15(12):1492. doi:10.3390/ph15121492
25. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci. 2019;40(8):577-91. doi:10.1016/j.tips.2019.05.005
26. Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, et al. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater. 2019;18(5):435-41. doi:10.1038/s41563-019-0338-z
27. Morgan P, Brown DG, Lennard S, Anderton MJ, Barrett JC, Eriksson U, et al. Impact of a five-dimensional framework on R&D productivity at AstraZeneca. Nat Rev Drug Discov. 2018;17(3):167-81. doi:10.1038/nrd.2017.244
28. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
29. Walters WP, Murcko MA. Assessing the impact of generative AI on medicinal chemistry. Nat Biotechnol. 2020;38(2):143-5. doi:10.1038/s41587-020-0418-2
30. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A
31. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441-51. doi:10.1039/C8SC00148K
32. Jayatunga MKP, Ayers M, Bruens L, Jayanth D, Meier C. How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons. Drug Discov Today. 2024;29(6):104009. doi:10.1016/j.drudis.2024.104009
33. Bonner S, Barrett IP, Ye C, Swiers R, Engkvist O, Bender A, et al. A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. Brief Bioinform. 2022;23(6). doi:10.1093/bib/bbac404
34. McGibbon M, Shave S, Dong J, Gao Y, Houston DR, Xie J, et al. From intuition to AI: Evolution of small molecule representations in drug discovery. Brief Bioinform. 2024;25(1). doi:10.1093/bib/bbad422
35. Huang K, Chandak P, Wang Q, Havaldar S, Vaid A, Leskovec J, et al. A foundation model for clinician-centered drug repurposing. Nat Med. 2024;30(12):3601-13. doi:10.1038/s41591-024-03233-x
36. Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6(5):525-35. doi:10.1038/s42256-024-00832-8
37. Hasselgren C, Oprea TI. Artificial intelligence for drug discovery: Are we there yet? Annu Rev Pharmacol Toxicol. 2024;64:527-50. doi:10.1146/annurev-pharmtox-040323-040828