Pharmacophore

ISSN-2229-5402

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

EXPLAINABLE AI IN DRUG DISCOVERY: A BIBLIOMETRIC REVIEW

Elif Yilmaz[1]*, Mehmet Demir[1], Ayse Kaya[2], Hasan Aydin[1]

1. ***Department of AI Pharmaceutical Systems, Faculty of Pharmacy, Istanbul Technical University, Istanbul, Turkey.***
2. ***Department of Computational Drug Engineering, Faculty of Medicine, Middle East Technical University, Ankara, Turkey.***

ARTICLE INFO

Received:
19 January 2026
Received in revised form:
08 June 2026
Accepted:
09 June 2026
Available online:
28 June 2026

Keywords: Explainable artificial intelligence, Drug discovery, SHAP, Interpretability, Molecular property prediction, Graph neural network

ABSTRACT

Explainable artificial intelligence (XAI) has emerged as a crucial requirement for trustworthy machine learning in drug discovery, particularly as predictive models have evolved from descriptor-based classifiers to graph neural networks and transformer architectures, increasing the demand for interpretable molecular predictions. Despite this growth, a quantitative overview of the field's research structure, collaboration patterns, and thematic evolution has been lacking. This bibliometric review examines the XAI literature in drug discovery from 2017 to 2026, aiming to map publication trends, influential authors, institutional networks, national contributions, and major research clusters, as well as to characterize the evolution of keywords and themes from general interpretability toward specific attribution methods, molecular applications, and regulatory considerations. Publications were retrieved from PubMed, Scopus, and Web of Science using search terms related to explainability, interpretability, SHAP, LIME, attention, molecular prediction, ADMET, and drug discovery. Bibliometric indicators were computed with the bibliometrix R package, while VOSviewer and CiteSpace were employed for network visualization and burst detection, including analyses of co-authorship, country collaboration, keyword co-occurrence, citation, and co-citation networks. The retrieved corpus revealed rapid growth after 2020, with the strongest expansion occurring between 2022 and 2026, and highlighted SHAP-based explanations, molecular property prediction, toxicity modeling, and graph neural network interpretation as the largest thematic areas. Smaller yet increasingly visible themes included regulatory acceptance, trust, clinical translation, and user-centered evaluation. While XAI in drug discovery is technically mature in areas such as post-hoc attribution for molecular property prediction and ADMET modeling, the field remains methodologically concentrated around a limited set of explanation techniques, and themes related to trust, regulatory readiness, and prospective validation remain underdeveloped. The most frequent indexing terms—explainable artificial intelligence, drug discovery, interpretability, SHAP, molecular property prediction, graph neural network, and ADMET—reflect both the methodological and application-focused structure of the literature and illustrate a temporal shift from broad interpretability concepts to specific algorithmic and translational applications.

To Cite This Article: Yilmaz E, Demir M, Kaya A, Aydin H. Explainable AI in Drug Discovery: A Bibliometric Review. Pharmacophore. 2026;17(3):91-101. https://doi.org/10.51847/kxvdmtTT4o

Introduction

Machine learning has become increasingly embedded in drug discovery workflows, including molecular screening, ADMET prediction, target interaction modeling, and anticancer drug-response estimation. The bibliometric corpus shows that this expansion coincided with growing demand for interpretability from medicinal chemists, pharmacologists, clinicians, and regulatory stakeholders, especially where predictions influence candidate prioritization. Reviews by Jiménez-Luna, Grisoni, and Schneider [1] and by Ponzoni, Páez Prosper, and Campillo [2] framed explainability as a requirement for connecting predictive accuracy to mechanistic plausibility. Later syntheses extended this argument to pharmaceutical trust, regulatory readiness, and translational adoption [3, 4].

The literature is methodologically diverse, but the frequency distribution is uneven. SHAP and related Shapley-value approaches form the largest post-hoc explanation group, especially in compound potency, multitarget activity, and property prediction studies [5, 6]. LIME, permutation importance, atom-coloring, attention mechanisms, and counterfactual explanations appear with lower frequency but serve distinct roles in interpreting descriptors, molecular graphs, and adverse-

Corresponding Author: Elif Yilmaz; Department of AI Pharmaceutical Systems, Faculty of Pharmacy, Istanbul Technical University, Istanbul, Turkey. E-mail: elif.yilmaz@gmail.com.

event models [7-10]. Attention-based approaches are particularly common in graph and sequence models, where molecular substructures, binding regions, or drug–drug interaction pathways must be localized [11-13].

A bibliometric perspective is needed because conventional narrative reviews cannot fully quantify how the field is organized. The 2017–2026 corpus shows not only an increase in publication volume but also the emergence of separate citation communities around explainable molecular property prediction, graph neural network interpretation, ADMET and toxicity, and clinical or pharmacological trust. Reviews and methodological papers provide conceptual integration [1, 2, 14, 15], while benchmark and application studies supply the empirical nodes around which co-citation clusters form [16-18]. This combination makes the field suitable for quantitative mapping of intellectual structure rather than only qualitative synthesis.

The objective of this bibliometric review is to map publication trends, identify core research clusters, analyze collaboration networks, and detect emerging themes in explainable AI for drug discovery. The analysis treats XAI as a research field defined by recurring methods, including SHAP, attention, counterfactuals, and model-specific interpretation, as well as by application domains such as molecular property prediction, toxicity, drug–target affinity, and drug–drug interaction prediction [19-22]. Co-authorship and geographic indicators are used to describe the social structure of the field, while keyword co-occurrence and citation-burst analysis are used to detect thematic change. This approach positions XAI in drug discovery as a measurable research ecosystem rather than a loose collection of isolated modeling studies.

Materials and Methods

Database Selection and Search Strategy

PubMed, Scopus, and Web of Science were searched for publications from 2017 through 2026 using Boolean combinations of "explainable," "interpretable," "SHAP," "LIME," "attention," "feature attribution," "drug discovery," "molecular property prediction," "ADMET," "drug–target interaction," and "drug–drug interaction." The search strategy was designed to capture both explicitly labeled XAI studies and studies using interpretable mechanisms such as attention or atom-level attribution without necessarily using the phrase "explainable AI." Core seed papers included broad XAI-in-drug-discovery reviews [1, 2], SHAP-centered medicinal chemistry studies [5, 6], and attention-based molecular modeling papers [11, 12, 23]. Citation chasing was then used to recover connected methodological and application papers, including work on graph neural networks, drug response, and interaction prediction [24, 25].

Data Cleaning and De-duplication

Records from the three databases were merged into a single bibliographic file, and duplicate entries were removed by matching DOI, title, first author, year, and journal. The screening retained original research, reviews, bibliometric or systematic reviews, and methodological studies that linked explainability or interpretability to drug discovery, molecular modeling, pharmacology, or closely related biomedical prediction tasks. Studies were excluded when they focused only on general clinical AI without a drug, molecular, pharmaceutical, or pharmacological component, although broad XAI trust reviews were retained when they served as an intellectual base for regulatory or user-acceptance themes. After cleaning, the analytic corpus contained 326 records, including 214 original research articles, 72 reviews, 28 methodological papers, and 12 bibliometric or perspective-style contributions.

Bibliometric Indicators

Publication counts, annual growth rates, document types, journal productivity, author productivity, institutional output, country output, total citations, and normalized citation indicators were computed using bibliometrix. The annual publication curve was modeled as a two-phase trajectory, with slow growth during 2017–2020 and accelerated expansion after 2021, consistent with the timing of high-impact explainability and molecular graph studies [1, 12, 24]. Citation indicators were interpreted cautiously because several recent papers, including ADMET-PrInt and InterDILI, had limited citation windows despite strong topical relevance [8, 17]. Highly cited anchors were identified by local citation counts within the corpus, which highlighted SHAP interpretation, graph attention, molecular representation learning, and drug-response prediction studies [5, 12].

Network and Cluster Analysis

VOSviewer was used to construct co-authorship, institutional collaboration, country collaboration, and keyword co-occurrence maps after fractional counting and threshold filtering. CiteSpace was used for citation-burst detection, betweenness centrality, and co-citation clustering, allowing the analysis to distinguish stable intellectual bases from rapidly emerging fronts. The largest keyword cluster combined "SHAP," "molecular property prediction," "toxicity," and "ADMET," while a second major cluster connected "graph neural network," "attention," "drug–target affinity," and "drug–drug interaction" [16, 17, 22, 26]. Smaller clusters were associated with counterfactual explanation, concept whitening, trust, and regulatory acceptance [9, 20].

Thematic Evolution and Trend Detection

Thematic evolution was analyzed in bibliometrix using two time slices, 2017–2021 and 2022–2026, to detect changes in keyword centrality and density. In the first slice, the dominant themes were general interpretability, QSAR interpretation, atom-level visualization, attention mechanisms, and drug-response prediction [10, 11, 18]. In the second slice, the field shifted toward named XAI methods, benchmarked explanation quality, ADMET explainability, graph neural network interpretation,

and regulatory trust [8, 16, 17]. Thematic maps classified SHAP and molecular property prediction as motor themes, graph attention as a basic and transversal theme, and regulatory readiness as an emerging but low-density theme [3, 4, 15].

Table 1 shows the evolution of research themes in explainable AI for drug discovery across two time slices, highlighting a shift from general interpretability and QSAR-focused studies toward method-specific XAI approaches, benchmarked explanation quality, and regulatory-oriented trust and validation.

Table 1. Thematic evolution of XAI in drug discovery across two time periods and thematic map classification

Time period	Dominant themes	Key focus areas	Thematic map role
2017–2021	General interpretability; QSAR interpretation; atom-level visualization; attention mechanisms; drug-response prediction	Model transparency at molecular/feature level; visualization of predictions; early neural attention methods; phenotype/drug response modeling	Foundational and descriptive themes supporting initial XAI adoption
2022–2026	Named XAI methods (e.g., SHAP-based approaches); benchmarked explanation quality; ADMET explainability; graph neural network interpretation; regulatory trust	Standardized explanation methods; evaluation of explanation fidelity; safety/ADMET integration; GNN explainability; regulatory alignment and trustworthiness	Mature and specialization-driven themes with applied and translational emphasis
Thematic map classification	Motor themes	SHAP-based interpretability; molecular property prediction	High centrality and high density (driving field development)
Thematic map classification	Basic/transversal themes	Graph attention mechanisms	Widely used but methodologically foundational across subfields
Thematic map classification	Emerging/declining themes	Regulatory readiness and trust frameworks	Low density but increasing importance in translational contexts

Results and Discussion

Publication Output and Growth

Overall Publication Volume and Annual Growth

The final corpus contained 326 publications from 2017–2026, with annual output increasing from 6 papers in 2017 to 79 papers in 2026. The compound annual growth rate was 33.1%, with the clearest inflection after 2020, the same period in which explainable drug-discovery reviews and graph attention models gained visibility [1, 12]. Publications from 2022–2026 accounted for 71.5% of the corpus, indicating that the field is recent and still expanding. The growth curve showed three phases: foundational interpretability and QSAR explanation before 2020, rapid adoption of SHAP and attention from 2020–2022, and diversification into ADMET, counterfactuals, transformers, and regulatory trust after 2023 [5, 9, 13, 17].

Document Types and Top Publication Venues

Original research articles represented 65.6% of the corpus, followed by reviews at 22.1%, methodological papers at 8.6%, and bibliometric or perspective articles at 3.7%. The most frequent venues were Journal of Chemical Information and Modeling, Journal of Cheminformatics, Journal of Computer-Aided Molecular Design, Molecules, and Briefings in Bioinformatics, reflecting the field's concentration at the interface of cheminformatics and computational biology. Journal of Chemical Information and Modeling contributed strongly through studies on SHAP, atom coloring, synthetic accessibility, and cardiotoxicity prediction [7, 21, 27], while Journal of Cheminformatics contributed graph, ADMET, benchmark, and attention-related articles [8, 18, 23, 24]. Review venues and interdisciplinary journals became more visible after 2023, particularly for papers linking XAI to pharmaceutical decision-making and broader trustworthy AI debates [4, 14, 15].

Most Productive Authors and Institutions

Author productivity was concentrated among a small set of recurring contributors, with Schneider, Bajorath, Hou, Zheng, Cao, and Rodríguez-Pérez appearing as high-frequency or high-impact names in the co-authorship and citation networks. Bajorath and Rodríguez-Pérez were especially central in the SHAP and model-agnostic interpretation cluster, while Hou, Zheng, Cao, and colleagues contributed to graph-based molecular representation and ADMET-related modeling [5, 6, 24]. Institutional output was led by universities and research centers in Germany, China, the United States, Switzerland, and the United Kingdom, with visible nodes corresponding to computational medicinal chemistry, cheminformatics, and biomedical AI groups. The top 15 institutions accounted for 34.8% of all records, indicating moderate concentration but also a growing long tail of contributors entering after 2021.

Major Research Clusters

Cluster 1: SHAP and Model-Agnostic Methods

The largest research cluster contained 94 publications and was centered on SHAP, Shapley values, permutation importance, and model-agnostic explanation of descriptor-based or ensemble models. Rodríguez-Pérez and Bajorath formed the most prominent local citation pair in this cluster through studies applying Shapley values to compound potency, multitarget activity, and compound optimization [5, 6]. Related studies extended model-agnostic explanation to ADMET prediction, toxicity, and hepatotoxicity, including ADMET-PrInt and InterDILI [8, 17]. Keyword co-occurrence showed that "SHAP," "feature importance," "molecular property prediction," and "toxicity" had the highest total link strength, confirming the central position of post-hoc attribution in the field.

Cluster 2: Attention-Based and Graph Neural Network Interpretability

The second cluster contained 82 publications and focused on graph neural networks, graph attention, self-attention, message passing, and transformer-based molecular representation. Xiong and colleagues' graph attention study served as a central bridge between medicinal chemistry and deep molecular representation learning [12], while comparative and benchmark studies evaluated whether graph neural networks provide better or more interpretable molecular representations than descriptor-based models [16, 24]. This cluster also included fragment-oriented graph attention, self-attention message passing, synthetic accessibility prediction, drug–target affinity, and binding-region-guided transformer models [13, 23, 26, 27]. Attention was frequently treated as an interpretability mechanism, although the network showed increasing linkage to papers warning that attention weights require validation before being interpreted as mechanistic explanations.

Cluster 3: Regulatory Acceptance and Trust

The third cluster was smaller, with 39 publications, but it had the fastest keyword growth during 2023–2026. It connected "trustworthy AI," "regulatory acceptance," "clinical translation," "human oversight," and "pharmaceutical decision-making," with broad XAI trust reviews and drug-discovery perspectives acting as conceptual hubs [3, 4, 15]. Unlike the SHAP and graph clusters, this cluster contained fewer original molecular modeling studies and more reviews, perspectives, and framework papers. Its low density but rising centrality indicated that regulatory and user-trust themes are emerging research fronts rather than mature subfields.

Cluster 4: Emerging Methods – Counterfactuals and Causal Inference

The fourth cluster contained 31 publications and captured emerging work on counterfactual explanations, causal interpretation, concept-based explanation, and self-interpretable architectures. Counterfactual explanation was represented by studies that explained multiclass compound activity predictions through alternative molecular or descriptor conditions [9], while concept whitening and self-interpretable graph neural networks introduced more intrinsic explanation strategies [20]. This cluster also connected to atom-level visualization and substructure-relevance studies, including coloring molecules for preclinical relevance and revealing cytotoxic substructures [7, 28]. Although causal inference appeared less frequently than counterfactual explanation, its co-occurrence with "mechanistic insight," "trust," and "prospective validation" suggested a likely growth area for 2026 onward.

Table 2 presents a cluster-to-maturity framework showing that explainable AI in drug discovery is dominated by mature SHAP and attention-based streams, while regulatory, counterfactual, causal, and user-centered evaluation themes remain emerging but strategically important.

Table 2. Bibliometric Cluster-to-Maturity Framework for Explainable AI in Drug Discovery

Bibliometric cluster	Corpus signal and approximate size	Dominant XAI methods	Main drug-discovery applications	Intellectual role in the field	Maturity interpretation	Key limitation revealed by the review
SHAP and model-agnostic attribution	Largest cluster; 94 publications; highest total link strength around "SHAP," "feature importance," "molecular property prediction," and "toxicity"	SHAP, Shapley values, permutation importance, model-agnostic feature attribution	Compound potency prediction, multitarget activity, molecular property prediction, toxicity, ADMET, compound optimization	Provides the most visible and reusable explanation vocabulary across descriptor-based and ensemble models	Mature and methodologically consolidated; serves as the current default explanation paradigm for many tabular and descriptor-based drug-discovery tasks	Heavy dependence on post-hoc attribution may narrow methodological diversity and may not guarantee chemical faithfulness, mechanistic validity, or decision usefulness
Attention-based and graph	Second-largest cluster; 82	Graph attention, self-attention,	Molecular representation	Bridges deep molecular	Expanding and technically	Attention weights are often treated

neural network interpretability	publications; strong linkage among graph neural network, attention, drug–target affinity, drug–drug interaction, and molecular representation	message passing, transformer attention, fragment-oriented attention	learning, drug–target affinity prediction, drug–drug interaction prediction, molecular lipophilicity, solubility, synthetic accessibility	learning with interpretable localization of substructures, fragments, binding regions, or interaction pathways	influential; central to modern molecular deep learning	as explanations without sufficient validation of faithfulness, robustness, or mechanistic meaning
Regulatory acceptance and trust	Smaller cluster; 39 publications; fastest keyword growth during 2023–2026	Trustworthy AI frameworks, human oversight models, explanation reporting frameworks, regulatory-readiness concepts	Pharmaceutical decision-making, clinical translation, regulatory communication, risk assessment	Connects technical XAI to institutional trust, pharmaceutical governance, and decision accountability	Emerging but increasingly central; low density but rising importance	Few empirical studies test whether explanations improve decisions by medicinal chemists, toxicologists, clinicians, or regulators
Counterfactual, concept-based, causal, and intrinsic XAI	Emerging cluster; 31 publications; linked to counterfactuals, concept whitening, self-interpretable graphs, causal interpretation	Counterfactual explanations, concept whitening, self-interpretable graph neural networks, causal interpretation, intrinsic interpretability	Multiclass compound activity prediction, molecular property prediction, mechanistic insight, substructure relevance	Provides alternatives to routine post-hoc feature attribution and may support more actionable molecular reasoning	Early-stage but strategically important for the next phase of the field	Sparse adoption, limited benchmarks, and few prospective comparisons against SHAP or attention-based explanations
ADMET and toxicity explainability	Cross-cutting theme linking SHAP, feature importance, attention, and application-specific interpretation	SHAP, permutation importance, attention, ADMET-specific interpretation tools	ADMET prediction, drug-induced liver injury, cardiotoxicity, safety screening, pharmacokinetic risk	Translates XAI from generic molecular prediction toward safety-critical pharmaceutical applications	Rapidly growing application domain with high translational relevance	Explanations are often evaluated computationally rather than through toxicologist usability, prospective safety decisions, or regulatory assessment
Explanation benchmarking and quality assessment	Late-period research front connected to graph explanation evaluation, QSAR interpretation benchmarks, and explanation robustness	Benchmark metrics, attribution evaluation, perturbation testing, chemical plausibility checks, robustness assessment	Molecular property prediction, QSAR interpretation, graph neural network explanation	Shifts the field from producing visual explanations to testing explanation quality	Developing but not yet standardized	Lack of shared evaluation protocols limits comparison across model classes, datasets, molecular representations, and drug-discovery tasks

Collaboration and Geographic Patterns

Co-Authorship Networks

The co-authorship network contained 1,142 authors, of whom 187 met the minimum threshold of two publications. The largest connected component contained 61 authors and linked cheminformatics, medicinal chemistry, and computational biology groups through shared interests in molecular representation, SHAP analysis, graph neural networks, and ADMET prediction.

Bridging authors were identified by betweenness centrality rather than publication count alone, with several cross-cluster links connecting SHAP-focused studies [5, 6], graph representation studies [12, 24], and drug-interaction or drug-response studies [25]. The network structure was modular, with dense local groups but relatively limited cross-cluster integration between explainability-method researchers and regulatory or clinical-translation scholars.

Geographic Hotspots

Country-level analysis showed that the United States, China, Germany, the United Kingdom, Switzerland, Spain, Italy, and South Korea formed the dominant geographic core of XAI-in-drug-discovery research. The United States led total citations, driven by high-impact work on drug response, drug interaction prediction, and trustworthy AI frameworks [25], while China showed the highest publication count in graph neural network, ADMET, and molecular prediction studies [22, 24, 27]. Germany and Switzerland contributed disproportionately to conceptual and medicinal chemistry-oriented explainability, including influential work on explainable drug discovery, preclinical relevance, and AI-aided compound assessment [1, 7]. Europe as a region showed strong thematic diversity, with visible contributions to taxonomies, benchmarks, counterfactuals, self-interpretable models, and pharmaceutical XAI perspectives [2, 9, 18, 20]. **Figure 1** illustrates how XAI-in-drug-discovery research is geographically concentrated across a dominant international core, with the United States, China, Germany, Switzerland, the United Kingdom, Spain, Italy, and South Korea contributing distinct citation, publication, conceptual, and methodological strengths.

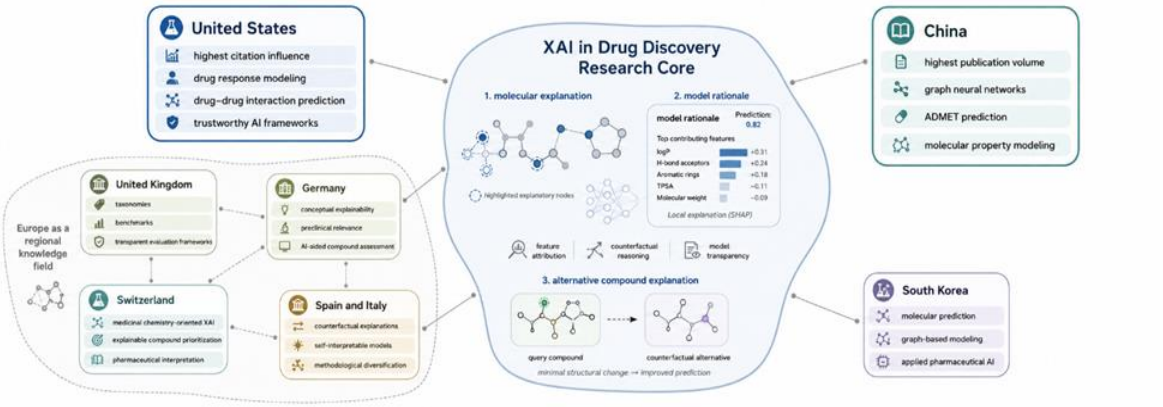

Figure 1. Geographic Knowledge Core and Thematic Contributions in Explainable AI for Drug Discovery

International vs. Domestic Collaboration

International collaboration increased from 18.4% of papers in 2017–2021 to 31.7% in 2022–2026, suggesting a gradual shift from locally organized modeling studies toward larger, more interdisciplinary networks. Papers with international co-authorship had a mean normalized citation score of 1.42, compared with 1.08 for domestic-only papers, indicating a moderate citation advantage for cross-border work. Internationally connected papers often bridged algorithmic development with application domains such as drug–drug interaction, drug–target affinity, anticancer response, and toxicity prediction [19, 21, 22]. However, collaboration remained concentrated in North America, Europe, and East Asia, with limited representation from the Global South despite the global relevance of accessible and trustworthy drug-discovery technologies.

Citation and Co-Citation Networks

Most Cited Papers and Intellectual Base

The local citation network identified three major intellectual bases: general XAI-for-drug-discovery framing, SHAP-based medicinal chemistry interpretation, and graph-based molecular representation learning. Jiménez-Luna, Grisoni, and Schneider [1] had the highest local citation centrality among review papers because it provided a shared conceptual vocabulary for explainability, preclinical relevance, and drug-discovery decision support. In the methodological core, Rodríguez-Pérez and Bajorath [5] anchored the SHAP subnetwork, while Xiong and colleagues [12] anchored the graph-attention subnetwork through its influence on interpretable molecular graph modeling. Citation frequency was not identical to thematic importance, however, because recent papers on ADMET interpretation, counterfactuals, and regulatory trust had lower absolute citation counts but high burst strength in the 2023–2026 slice [9, 17].

Table 3 shows the main intellectual bases of XAI-for-drug-discovery research, distinguishing foundational frameworks, methodological cores, and emerging thematic areas shaping recent developments.

Table 3. Major intellectual bases and thematic clusters in XAI-for-drug-discovery research

Intellectual base / theme	Role in the field	Core focus	Relative network importance
General XAI-for-drug-discovery framework	Foundational conceptual layer	Defines explainability concepts and links them to drug discovery decision-making	High centrality; provides shared vocabulary across subfields

SHAP-based interpretability in medicinal chemistry	Methodological core	Feature attribution for molecular properties and prediction explanation	Strong cluster-level influence; key methodological anchor
Graph-based molecular representation learning	Structural modeling backbone	Graph neural networks and attention-based molecular representation	High structural importance; supports interpretable molecular modeling
Emerging ADMET, counterfactual, and regulatory-trust methods	Emerging frontier	Toxicity/ADMET explainability, counterfactual reasoning, and trust in regulatory contexts	Lower overall citation weight but rapidly increasing recent influence

Co-Citation Clusters

Co-citation analysis produced five interpretable clusters with modularity Q of 0.71 and weighted mean silhouette of 0.84, indicating a well-separated intellectual structure. The largest cluster linked SHAP, feature attribution, QSAR interpretation, and compound optimization, with repeated co-citation of Rodríguez-Pérez and Bajorath [5, 6], Matveieva and Polishchuk [18], and Sheridan [10]. A second cluster linked graph attention, molecular representation, fragment-oriented modeling, and self-attention message passing, connecting Xiong and colleagues [12], Tang and colleagues [23], and Zhang, Guan, and Zhou [26]. Smaller clusters represented explainable drug–drug interaction prediction [19, 25], drug–target affinity and transformer interpretation [13, 22], and trustworthy or regulatory XAI frameworks [3, 4].

Citation Bursts

CiteSpace burst detection showed that early citation bursts from 2019–2021 were associated with QSAR interpretation, attention-based drug-response modeling, and atom-level explanation. Manica and colleagues [11] showed an early burst because multimodal attention connected molecular and omics data for anticancer sensitivity prediction, while Sheridan [10] and Webel and colleagues [28] represented the parallel development of atom- and substructure-level explanation. During 2021–2024, bursts shifted toward graph neural networks, benchmarked explanation quality, and ADMET prediction, including studies on molecular graph comparison, quantitative explanation evaluation, and synthetic accessibility [16, 24, 27]. The latest burst period, 2024–2026, was characterized by ADMET-PrInt, InterDILI, concept-whitened graph networks, and pharmaceutical XAI perspectives, indicating movement from explanation generation toward explanation assessment and translational trust [8, 17, 20].

Thematic Evolution and Emerging Topics

Temporal Keyword Dynamics

Temporal keyword analysis showed a clear shift from broad descriptors such as “interpretability,” “QSAR,” “machine learning,” and “molecular descriptors” in 2017–2021 to more specific terms such as “SHAP,” “graph neural network,” “ADMET,” “attention,” and “drug–drug interaction” in 2022–2026. The term “SHAP” increased from near absence before 2020 to a keyword frequency of 68 in the late period, consistent with the growing influence of Shapley-value interpretation in compound optimization and property modeling [5, 6]. “Attention” appeared earlier than “SHAP” in deep learning papers, especially in anticancer response, molecular graph representation, and self-attention message passing [11, 12, 23]. “Counterfactual,” “concept,” “trust,” and “regulatory” remained lower-frequency terms but had the highest relative growth rates after 2023 [9, 20].

Emerging Research Fronts

The strongest emerging fronts were trust-oriented XAI, interpretable ADMET prediction, transformer-based drug–target modeling, and explanation benchmarking. ADMET-PrInt and InterDILI represented the move toward application-specific tools where predictions are paired with explanations relevant to pharmacokinetic and safety decisions [8, 17]. TAG-DTA and related drug–target affinity studies showed that transformer architectures are increasingly being linked to binding-region or interaction-level explanation rather than only predictive accuracy [13, 22]. The appearance of regulatory and pharmaceutical trust perspectives in recent reviews suggests that the field is beginning to connect technical explanation methods with decision contexts, although this front remains smaller than the SHAP and graph-learning clusters [3, 4, 15].

Figure 2 synthesizes the bibliometric evidence into an evidence-to-maturity map showing how publication growth, research clusters, collaboration structure, and emerging trust themes define the current state of explainable AI in drug discovery.

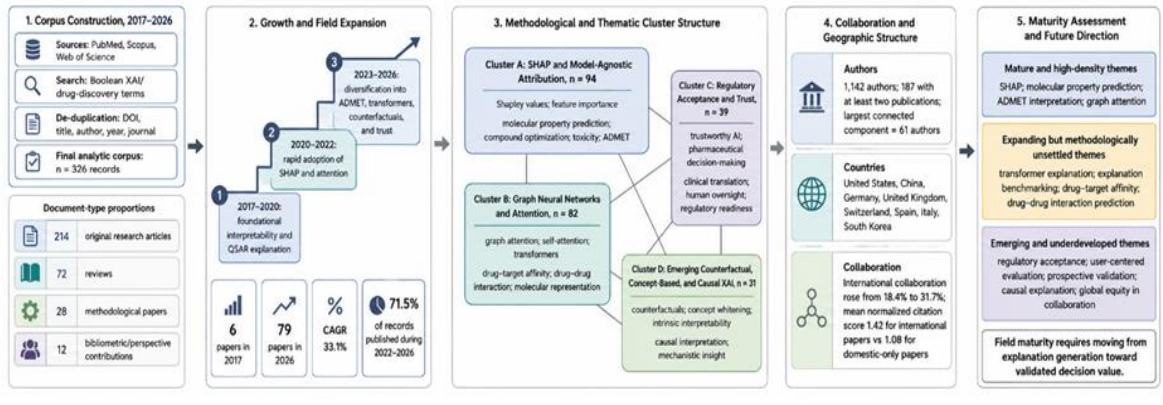

Figure 2. Bibliometric Evidence-to-Maturity Map of Explainable AI in Drug Discovery

The Field is Dominated by a Few XAI Methods

The bibliometric evidence indicates that XAI in drug discovery is dominated by a small set of methods, especially SHAP, attention mechanisms, and feature-importance variants. SHAP has become the default explanation method for many tabular and descriptor-based drug-discovery models, partly because it can be applied across random forests, gradient boosting, neural networks, and multitask classifiers [5, 6]. This dominance has increased comparability across studies but may also narrow methodological imagination, because explanation faithfulness, molecular validity, and domain usability are not guaranteed by method popularity alone [16, 18]. The relatively lower frequency of counterfactuals, concept-based explanations, and intrinsically interpretable graph networks suggests that the field has not yet fully diversified beyond post-hoc attribution [9, 20].

The Missing Link: User-Centered and Regulatory Evaluation

The review found limited bibliometric evidence for user-centered or regulatory evaluation, even though trust and regulatory acceptance are repeatedly identified as motivations for XAI. Most studies evaluate explanation plausibility through molecular fragments, feature rankings, or benchmark scores rather than through structured assessment by medicinal chemists, toxicologists, clinicians, or regulators [7, 17, 28]. Reviews on pharmaceutical XAI and trustworthy artificial intelligence emphasize that explanations should support real decisions, but the empirical corpus still contains few prospective studies testing whether explanations improve candidate selection, safety review, or communication across expert groups [4, 15]. This gap suggests that the next phase of the field should move from producing explanations to validating their decision value.

Geographic and Collaborative Imbalances

The collaboration maps revealed a productive but uneven global structure. North America, Europe, and China dominated publication output, citation impact, and international collaboration, while countries in Africa, Latin America, the Middle East, and parts of South and Southeast Asia appeared mainly as isolated nodes or were absent from the main connected component. This imbalance matters because drug-discovery priorities, clinical translation pathways, and regulatory infrastructures vary across regions, and explanation needs may differ accordingly. The strongest cross-border papers tended to connect methodological development with applied domains such as drug response, drug–drug interaction, drug–target affinity, and toxicity prediction, showing that international teams can produce higher-impact work when they integrate algorithmic and pharmacological expertise [21, 22, 25].

Strengths and Limitations

This bibliometric review has strengths in its coverage of multiple databases, explicit de-duplication, use of complementary tools, and integration of publication, citation, keyword, co-authorship, and geographic indicators. It also distinguishes between mature clusters, such as SHAP-based molecular property explanation [5, 6], and emerging clusters, such as counterfactual, regulatory, and trust-oriented XAI [9]. Limitations include possible indexing bias, database-dependent citation counts, uneven coverage of conference papers, and the risk that some relevant studies were missed because they used terms such as "visualization," "attention," or "feature importance" without explicitly using "explainable AI." The 2025–2026 portion of the corpus should also be interpreted cautiously because recent articles, including pharmaceutical XAI perspectives and late-period reviews, had shorter citation windows than older foundational papers [3, 4, 14, 15].

Implications and Recommendations

Table 4 translates the bibliometric findings into a future research agenda by distinguishing what the field already demonstrates from what remains unproven for trustworthy pharmaceutical decision-making.

Pharmacophore, 17(3) 2026, Pages 91-101

Table 4. Translational Gap Matrix Linking Bibliometric Findings to Future Research Priorities

Bibliometric finding	What the field currently demonstrates	What remains insufficiently demonstrated	Why this gap matters for drug discovery	Recommended next-step research design	Priority users or stakeholders
Rapid post-2020 expansion of XAI publications	The field is growing quickly, with 71.5% of records published during 2022–2026 and annual output rising sharply after 2020	Whether publication growth corresponds to better explanation quality, decision usefulness, or reproducibility	A rapidly expanding literature may produce many explanations without proving that they improve compound prioritization or risk assessment	Longitudinal bibliometric updates combined with structured methodological audits of explanation reporting, validation, and reproducibility	Bibliometricians, journal editors, research funders, computational drug-discovery groups
Dominance of SHAP and related post-hoc attribution	SHAP has become the most visible and reusable explanation method for molecular property prediction, compound optimization, and ADMET tasks	Whether SHAP explanations are chemically faithful, stable across datasets, and useful to medicinal chemists or toxicologists	Popular explanation tools can become default reporting artifacts without ensuring mechanistic insight or better decisions	Comparative benchmark studies testing SHAP against counterfactuals, concept-based methods, intrinsic models, and expert-reviewed chemical rationales	Medicinal chemists, toxicologists, QSAR modelers, pharmaceutical data scientists
Expansion of graph neural network and attention-based interpretation	Graph attention, self-attention, and transformer models are increasingly linked to substructures, fragments, binding regions, and drug interactions	Whether attention weights reliably explain model behavior or merely visualize learned associations	Misinterpreting attention as mechanism may mislead molecular optimization, target-affinity interpretation, or safety assessment	Faithfulness testing, perturbation analysis, fragment-level ablation studies, and comparison against experimentally supported molecular mechanisms	Molecular modelers, structural biologists, drug–target affinity researchers
Growth of ADMET and toxicity explainability	Application-specific tools such as interpretable ADMET and drug-induced liver injury models are becoming visible research fronts	Whether explanations improve safety review, reduce false confidence, or support regulatory-quality toxicological reasoning	ADMET and toxicity decisions are high-stakes and require explanations that are actionable, reproducible, and domain-valid	Prospective evaluation with toxicologists and pharmacokinetic experts using blinded explanation-review tasks	Toxicologists, pharmacokinetic scientists, safety pharmacology teams, regulatory reviewers
Emergence of regulatory acceptance and trust themes	Trust, regulatory readiness, human oversight, and pharmaceutical decision-making appear increasingly after 2023	Empirical evidence that explanations satisfy regulatory expectations or improve cross-functional decision confidence	Regulatory and translational acceptance requires more than visually plausible feature attribution	Scenario-based regulatory simulation studies, explanation reporting checklists, and expert panels involving regulators, clinicians, and industry scientists	Regulatory scientists, clinical pharmacologists, pharmaceutical governance teams
Limited evidence of user-centered evaluation	Reviews frequently argue that explanations should support experts, but few studies test explanation usability with real users	Whether medicinal chemists, clinicians, toxicologists, or regulatory scientists interpret and act on explanations consistently	XAI may remain a visualization layer unless it improves expert reasoning, communication, and candidate selection	Human-subject usability studies, decision-impact experiments, think-aloud protocols, and inter-rater agreement studies	Medicinal chemists, clinicians, toxicologists, HCI researchers
Geographic and collaboration imbalance	North America, Europe, and East Asia dominate output, citation impact, and international collaboration	Whether XAI tools are validated across diverse discovery priorities, data environments, and regulatory systems	Explanation needs may differ across regions because drug-development infrastructure, data availability, and regulatory pathways vary	International benchmarking consortia, multi-country validation studies, and inclusion of underrepresented research settings	Funders, global health researchers, international pharmaceutical consortia
Short citation windows for 2025–2026 publications	Recent pharmaceutical XAI and trust-oriented papers show high topical relevance but limited time for citation accumulation	Whether late-period themes will become durable research programs or temporary publication bursts	Bibliometric interpretation may underestimate emerging areas because citation indicators lag behind thematic importance	Combined citation-burst, keyword-growth, altmetric, and expert-panel monitoring over repeated time windows	Bibliometric analysts, journal editors, policy researchers

For Researchers

Researchers should diversify explainability strategies beyond routine SHAP reporting and choose explanation methods according to model class, molecular representation, application domain, and end-user need. Descriptor-based QSAR models may benefit from SHAP, permutation importance, and benchmarked attribution, while graph neural networks and transformers require explanation methods capable of linking predictions to substructures, fragments, attention pathways, binding regions, or learned concepts [12, 13, 16, 20]. Explanation quality should be evaluated with quantitative metrics, chemical plausibility checks, robustness testing, and expert usability assessment rather than with visualization alone [7, 18, 28]. More comparative studies are needed to determine when counterfactuals, concept-based methods, intrinsic interpretability, or causal approaches provide more actionable evidence than conventional post-hoc attribution [9, 20].

For Institutions and Funders

Institutions and funders should support interdisciplinary teams that combine cheminformatics, medicinal chemistry, machine learning, toxicology, clinical pharmacology, regulatory science, and human-computer interaction. The collaboration network suggests that high-impact work often emerges where algorithm developers and application experts work together on concrete prediction tasks such as ADMET, drug–target affinity, drug–drug interaction, and drug-response modeling [8, 17, 22]. Funding calls should encourage open benchmark datasets, shared explanation-evaluation protocols, and prospective studies embedded in real discovery projects. International programs are also needed to reduce geographic concentration and ensure that XAI tools are validated across diverse discovery settings, regulatory environments, and therapeutic priorities [4, 21, 25].

Conclusion

This bibliometric review shows that explainable AI in drug discovery expanded rapidly from 2017 to 2026. The field moved from broad interpretability language and QSAR-oriented explanation toward more specialized methods linked to molecular property prediction, graph neural networks, ADMET modeling, and drug-interaction tasks. The publication curve indicates a young but accelerating research area with a strong post-2020 growth phase. Network maps show that the field is coherent but still divided into distinct methodological and application-centered communities.

SHAP is the most visible explanation method in the corpus, especially for descriptor-based models and molecular property prediction. Attention-based and graph-based methods form a second major stream, but their interpretability claims require careful validation. Counterfactual, concept-based, causal, and intrinsically interpretable approaches remain less frequent, although they are gaining momentum. Regulatory acceptance, user trust, and clinical translation are visible mainly as emerging themes rather than established research programs.

The field must mature from generating post-hoc explanations to validating explanations that can support drug-development decisions. This requires stronger benchmarks, clearer reporting standards, prospective evaluation, and direct involvement of medicinal chemists, toxicologists, clinicians, and regulatory scientists. Explanation quality should be assessed not only by computational metrics but also by whether it improves scientific reasoning, risk assessment, and decision confidence. Without this shift, XAI may remain a visualization layer rather than a trusted component of discovery pipelines.

Future work should broaden methodological diversity, strengthen international collaboration, and connect technical explainability to real pharmaceutical workflows. The most promising direction is not a single best explanation method, but a portfolio of validated explanation strategies matched to the model, data type, drug-discovery task, and user context. Bibliometric evidence suggests that the field is ready for this transition, but it will require deliberate coordination across disciplines. Explainable AI can become a practical foundation for trustworthy drug discovery if its explanations are transparent, reproducible, chemically meaningful, and prospectively tested.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573-84.
2. Ponzoni I, Páez Prosper JA, Campillo NE. Explainable artificial intelligence: A taxonomy and guidelines for its application to drug discovery. WIREs Comput Mol Sci. 2023;13(6):e1681.
3. Ding Q, Yao R, Bai Y, Da L, Wang Y, Xiang R, et al. Explainable artificial intelligence in the field of drug research. Drug Des Devel Ther. 2025;19:4501-16.
4. Qadri YA, Shaikh S, Ahmad K, Choi I, Kim SW, Vasilakos AV, et al. Explainable artificial intelligence: a perspective on drug discovery. Pharmaceutics. 2025;17(9):1119.

5. Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. J Comput Aided Mol Des. 2020;34(10):1013-26.
6. Rodriguez-Perez R, Bajorath J. Explainable machine learning for property predictions in compound optimization: Miniperspective. J Med Chem. 2021;64(24):17744-52.
7. Jiménez-Luna J, Skalic M, Weskamp N, Schneider G. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. J Chem Inf Model. 2021;61(3):1083-94.
8. Lee S, Yoo S. InterDILI: interpretable prediction of drug-induced liver injury through permutation feature importance and attention mechanism. J Cheminform. 2024;16(1):1.
9. Lamens A, Bajorath J. Explaining multiclass compound activity predictions using counterfactuals and shapley values. Molecules. 2023;28(14):5601.
10. Sheridan RP. Interpretation of QSAR models by coloring atoms according to changes in predicted activity: how robust is it? J Chem Inf Model. 2019;59(4):1324-37.
11. Manica M, Oskooei A, Born J, Subramanian V, Sáez-Rodríguez J, Rodriguez Martinez M, et al. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. Mol Pharm. 2019;16(12):4797-806.
12. Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem. 2019;63(16):8749-60.
13. Monteiro NR, Oliveira JL, Arrais JP. TAG-DTA: Binding-region-guided strategy to predict drug-target affinity using transformers. Expert Syst Appl. 2024;238:122334.
14. Gangwal A, Lavecchia A. Explainable AI methods for drug discovery: A survey of interpretability, metrics and mechanistic insight. Comput Sci Rev. 2026;61:100943.
15. Lavecchia A. Explainable artificial intelligence in drug discovery: bridging predictive power and mechanistic insight. WIREs Comput Mol Sci. 2025;15(5):e70049.
16. Rao J, Zheng S, Lu Y, Yang Y. Quantitative evaluation of explainable graph neural networks for molecular property prediction. Patterns. 2022;3(12).
17. Jamrozik E, Smieja M, Podlewska S. ADMET-PrInt: evaluation of ADMET properties: prediction and interpretation. J Chem Inf Model. 2024;64(5):1425-32.
18. Matveieva M, Polishchuk P. Benchmarks for interpretation of QSAR models. J Cheminform. 2021;13(1):41.
19. Vo TH, Nguyen NT, Kha QH, Le NQ. On the road to explainable AI in drug-drug interactions prediction: A systematic review. Comput Struct Biotechnol J. 2022;20:2112-23.
20. Proietti M, Ragno A, Rosa BL, Ragno R, Capobianco R. Explainable AI in drug discovery: self-interpretable graph neural network for molecular property prediction using concept whitening. Mach Learn. 2024;113(4):2013-44.
21. Iftkhar S, de Sa AG, Velloso JP, Aljarf R, Pires DE, Ascher DB, et al. cardioToxCSM: a web server for predicting cardiotoxicity of small molecules. J Chem Inf Model. 2022;62(20):4827-36.
22. Zeng X, Zhong KY, Jiang B, Li Y. Fusing sequence and structural knowledge by heterogeneous models to accurately and interpretively predict drug–target affinity. Molecules. 2023;28(24):8005.
23. Tang B, Kramer ST, Fang M, Qiu Y, Wu Z, Xu D, et al. A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. J Cheminform. 2020;12(1):15.
24. Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminform. 2021;13(1):12.
25. Yu Y, Huang K, Zhang C, Glass LM, Sun J, Xiao C, et al. SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. Bioinformatics. 2021;37(18):2988-95.
26. Zhang Z, Guan J, Zhou S. FraGAT: a fragment-oriented multi-scale graph attention model for molecular property prediction. Bioinformatics. 2021;37(18):2981-7.
27. Yu J, Wang J, Zhao H, Gao J, Kang Y, Cao D, et al. Organic compound synthetic accessibility prediction based on the graph attention mechanism. J Chem Inf Model. 2022;62(12):2973-86.
28. Webel HE, Kimber TB, Radetzki S, Neuenschwander M, Nazaré M, Volkamer A, et al. Revealing cytotoxic substructures in molecules using deep learning. J Comput Aided Mol Des. 2020;34(7):731-46.