Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Artificial Intelligence in Drug Discovery: A PRISMA-Based Systematic Review of Methods, Challenges, Taxonomy, and Evaluation Practices

Mahmud Ahmed

mahmud.ahmed@fuhsa.edu.ng

Fedearal University of Health Sciences, Azare

Abdulsalam Ya'u Gital

Abubakar Tafawa Balewa University

Usman Ali Abdullahi

Federal College of Education (Technical) Gombe

Kabiru Ibrahim Musa

Abubakar Tafawa Balewa University

Research Article

Keywords: artificial intelligence, machine learning, generative AI, graph neural networks

Posted Date: October 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11120341/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Artificial Intelligence in Drug Discovery: A PRISMA-Based Systematic Review of Methods, Challenges, Taxonomy, and Evaluation Practices

Mahmud Ahmed[1*], Abdulsalam Ya'u Gital[2†], Ali Usman Abdullahi[3†], Kabiru Ibrahim Musa[4†]

[1*]Department of ICT, Federal University of Health Sciences, Azare, Bauchi State, Nigeria.
[2]Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi, Bauchi State, Nigeria.
[3]Department of Computer Science, Federal College of Education (Technical), Gombe, Gombe State, Nigeria.
[4]Department of Computer Science, Abubakar Tafawa Balewa University, Bauchi, Bauchi State, Nigeria.

*Corresponding author(s). E-mail(s): mahmud.ahmed@fuhsa.edu.ng;
Contributing authors: aygital@atbu.edu.ng; usmanali@fcetgombe.edu.ng; imkabir@atbu.edu.ng;
†These authors contributed equally to this work.

Abstract

This systematic review made a bold attempt to examine artificial intelligence (AI) and machine learning (ML) applications in drug discovery, addressing a fragmented literature characterized by inconsistent terminology and limited benchmarking. Following PRISMA 2020 guidelines, Scopus, PubMed, and Web of Science were queried (January 2019-August 2026). The three databases returned 23,404 records; after removing 4,491 duplicates and multi-stage screening, 618 peer-reviewed studies were included. Quality was assessed using a ROBINS-I-inspired framework across four dimensions (methodological rigor, dataset transparency, reproducibility, evaluation validity); 38 studies (6.1%) were double-coded ($\boldsymbol{\kappa = 0.74}$). Machine learning and deep learning dominated (99.8% and 78.0%, respectively). QSAR/CADD appeared in 58.9%, GNN/GCN in 45.1%, and Transformer/LLM architectures in 45.5%, with 99.8% of transformer papers published in 2024–2026. Virtual screening was most common (57.1%), followed by

de novo design (30.6%) and drug-target interaction prediction (13.8%). The most prominent challenges were integration and clinical translation (15.2%), interpretability (11.3%), and benchmark/validation (9.2%). AUC-ROC dominated reporting (12.9%), while F1 (0.8%) and recall (0.5%) were severely underreported-problematic for heavily imbalanced bioactivity datasets. Methodological rigor was generally high (68%), but dataset transparency (45%) and reproducibility (38%) were weaker; external validation was rare (¡25%), and moderate-to-high selection bias was identified in 42% of studies. The review produced a two-axis taxonomy, a five-phase temporal framework, and a benchmarking synthesis. Priority directions include molecular foundation models, multimodal integration, autonomous AI-robotics laboratories, and federated learning. No funding was received, and the review was not prospectively registered.

Keywords: artificial intelligence, machine learning, generative AI, graph neural networks

1 Introduction

Bringing a new drug to market takes 10 to15 years and costs $2 to $3 billion, with over 90% of candidates failing, most often due to low efficacy or late-stage toxicity [1, 2]. AI and ML offer a route to shorter timelines, lower costs, and previously impractical analyses [3–5]. Between 2020 and 2025, four developments accelerated adoption: AlphaFold2 largely solved protein structure prediction [6]; ESM-2 and ChemBERTa became broadly useful molecular tools; diffusion models advanced de novo design; and reinforcement learning made multi-objective optimization practical at scale.

1.1 Background

1.1.1 The Drug Discovery Pipeline

Drug discovery moves through target identification and validation, hit discovery, lead optimization, preclinical evaluation, and clinical trials [7]. Target ID links a biological entity to disease; hit discovery screens chemical libraries; lead optimization balances potency, selectivity, and ADMET; preclinical work confirms pharmacology and safety before Phase I–III. Each stage is a probabilistic sieve—roughly one in 10,000 screened molecules reaches approval [8] across a drug-like space exceeding 10^60 compounds.

1.1.2 The Emergence of AI in Pharmaceutical Research

Early computational work relied on rule-based QSAR, docking, and similarity search [9]. The 2010s deep learning wave reshaped the field [10]: DeepChem scaled property prediction; AlphaFold reset structural biology [6]; generative models proposed novel molecules [11]; and GNNs uncovered new antibiotic scaffolds [12]. By 2023–2025, the emphasis shifted to general-purpose foundation models like AlphaFold3, which extended structure prediction to all biomolecular complexes, and Insilico Medicine reported the first fully AI-discovered clinical candidate in 2023.

1.1.3 AI Across the Pipeline Stages

AI now appears at every pipeline stage: NLP methods nominate disease genes; generative models propose candidates while predictive models triage virtual libraries; multi-objective models balance affinity, selectivity, and ADMET; and further downstream, AI supports biomarker discovery, patient stratification, and adaptive trial design [13]. Foundation models are the latest wave, enabling cross-task and few-shot learning across applications previously handled in isolation. Representative applications appear in Table 1.

1.2 Problem Statement

Despite this progress, the literature has grown so rapidly from 11 relevant publications in 2019 to 159 in 2026. The 618 studies in this corpus- a cumulative synthesis now face three structural problems: fragmented evidence, in which the same method reappears in multiple publication forms with non-overlapping metrics (Section 1.2.1); the absence of a shared taxonomy that links method families to pipeline stages (Section 1.2.2); and inconsistent evaluation, which prevents head-to-head comparison across studies (Section 1.2.3).

1.2.1 Fragmented Evidence

The literature is scattered across disparate venues. The study had a total of 23,404 initial records were narrowed to 618 through five-stage screening out of which journal articles and reviews make up 49% of the corpus, with book chapters (24.6%) and conference proceedings (23.1%) accounting for most of the rest. When the same method appears in more than one of these forms-for example, as a conference paper later extended into a journal article- a pattern confirmed for more than 78% of highly cited conference papers during screening (Section 2.5), and the corresponding reports adopt different metrics (Section 3.10; Table 13), cumulative synthesis becomes more difficult.

1.2.2 No Shared Taxonomy

Existing categorizations only pick one axis taxonomy. Some surveys sort papers by algorithm—supervised vs. unsupervised vs. reinforcement learning [14], or rule/statistical vs. machine-learning vs. graph-based platforms, while others sort by application, such as virtual screening vs. de novo design [15] or target identification, lead discovery, and lead optimization [16]. Neither answers the key question. . Neither answers the key question: which method families are used at which pipeline stage? With 72 distinct method labels in this corpus, the absence of a shared two-axis framework is a real obstacle.

1.2.3 Inconsistent Evaluation

Inconsistent evaluation is the third problem. Two bioactivity-prediction papers may report AUC-ROC and RMSE with no overlap; generative-chemistry papers may give validity and uniqueness but not novelty, or FCD without scaffold diversity, and individual studies often report only subsets of these batteries [17]. Train/test split design

[18], scaffold-aware partitioning [19], and code/weight sharing vary so widely that head-to-head comparison is often impossible; similar incomparability and uncertainty in model evaluation were identified as core challenges in an earlier systematic review of generative machine learning for drug discovery [20]. Also data and evaluation gaps-heterogeneity, imbalance, and misleading benchmarks-were likewise designated core challenges in a recent scoping review of healthcare-oriented drug discovery [21]. These three problems are not internal to this corpus alone; they mirror gaps left open by earlier reviews, which were largely narrative in method, did not quantify technique prevalence or evaluation practice across the full pipeline, and rarely spanned the whole discovery pipeline within a single framework

1.3 Objectives

This review has four objectives:

- (O1) To identify and characterize the AI/ML techniques most frequently applied across the drug discovery pipeline between 2019 and 2026, including emerging techniques such as foundational models, multimodal AI, and autonomous drug discovery systems.
- (O2) To develop a coherent, two-dimensional taxonomy that simultaneously organizes methods by algorithmic family and by pipeline stage, accommodating hybrid and multimodal approaches.
- (O3) To analyze the technical, methodological, ethical, and regulatory challenges reported in the included studies, and to weigh them by frequency of occurrence.
- (O4) To critically review performance evaluation practices—metrics, datasets, validation protocols, and reproducibility provisions—and to propose recommendations for harmonized reporting, including a benchmarking synthesis table.

1.4 Research Questions

Four research questions (RQs), framed using a modified PICO perspective [22], guide the review:
RQ1: What AI techniques are currently adopted in drug discovery, and how has their adoption evolved between 2019 and 2026, including the emergence of foundation models and multimodal systems?
RQ2: How can these adopted techniques be systematically classified within a taxonomy that captures both algorithmic family and pipeline-stage application?
RQ3: What major technical, ethical, regulatory, data-related and translational challenges are reported in the literature, and how prevalent is each challenge?
RQ4: How is the performance of AI models in drug discovery evaluated, and to what extent are evaluation practices reproducible and comparable across studies?

1.5 Current Evidence and Rationale

Earlier narrative reviews surveyed the field broadly [3, 23–27] but lacked structured methods or quantitative synthesis. We address this gap by applying PRISMA 2020 to 618 peer-reviewed studies (2019–2026) across three databases. Topic-specific reviews

covered generative chemistry [28], repurposing [29], and target identification [30], but none spans the full pipeline while incorporating foundation models and multimodal methods. Searching three databases avoids single-database blind spots and covers the computational, biomedical, and informatics literatures.

1.6 Visual Synthesis of the Evidence Base

Figure 1 plots annual publication counts—growth steepens sharply after 2022, driven by transformer language models, AlphaFold2 tooling, and foundation models. Figure 2 shows technique distribution: classical and deep ML carry the volume, while generative and graph-based methods, LLMs, and federated learning form a fast-expanding tail.

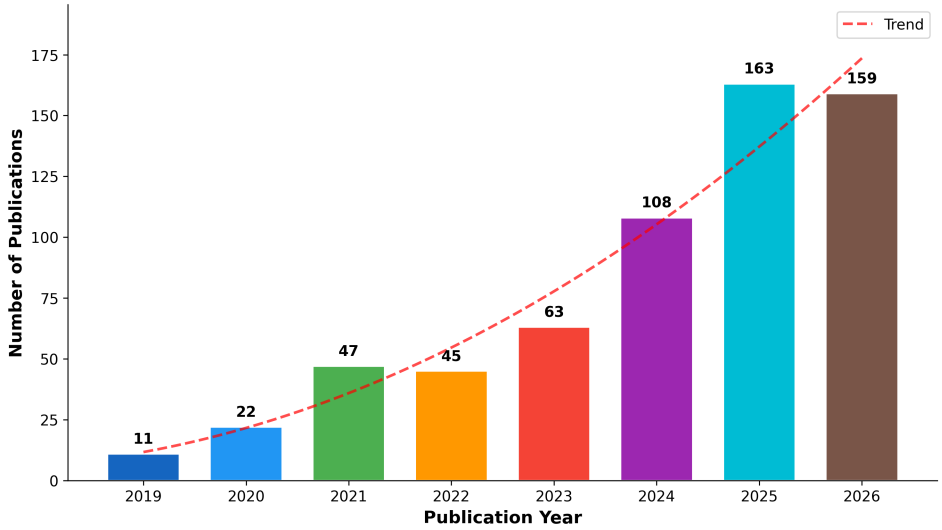

Fig. 1 Annual distribution of publications on AI in drug discovery (2019–2026; multi-database search; N = 618). Growth accelerated sharply from 11 publications in 2019 to 159 in 2026, reflecting a 14× increase over the review period.

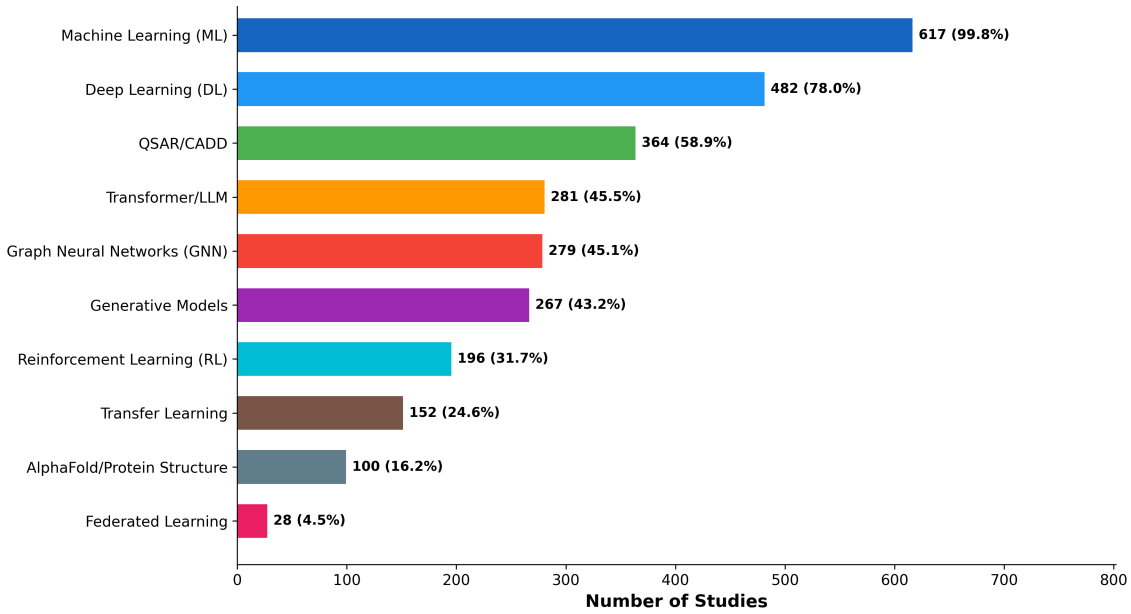

Fig. 2 Distribution of AI/ML technique categories across 618 included studies. Categories are non-exclusive; a single paper may employ multiple methods. Machine Learning (617, 99.8%), Deep Learning (482, 78.0%), and QSAR/CADD (364, 58.9%) are the most prevalent categories.

Table 1 maps representative AI applications to canonical pipeline stages. Table **??** summarizes the most frequently reported challenges and their counts.

2 Methods

2.1 Review Protocol

We followed PRISMA 2020[22], fixing the protocol before searching. Aims: characterize AI applications, surface recurring challenges, propose a taxonomy, and audit evaluation practice. Three databases were used to span the biomedical, computational, and multidisciplinary literatures.

2.2 Data Sources and Multi-Database Search

We searched three major bibliographic databases for broad coverage:

- Scopus (Elsevier): Broad peer-reviewed coverage across life sciences, computational chemistry, and engineering; primary database (n = 553 records from CSV export, August 24, 2026).
- PubMed (National Library of Medicine): Biomedical and clinical literature; essential for capturing pharmaceutical and translational AI studies (n = 12,893 records via E-utilities API; 16 queries).

Pipeline Stage	Representative AI Applications	Common Algorithmic Families
Target identification & validation	Disease-gene prioritization; pathway mining; literature NLP; knowledge graphs	GNNs, transformers, BERT-family LLMs, network embeddings
Hit discovery	Virtual screening of ultra-large libraries; bioactivity prediction	Random forests, CNNs, message-passing GNNs, transformer fingerprints
Lead generation	De novo molecular design; scaffold hopping	VAEs, GANs, diffusion models, RL, SMILES/SELFIES LMs, foundation models
Lead optimization	Multi-objective ADMET and potency prediction; retrosynthesis planning	Multi-task DL, graph transformers, Bayesian optimization, RL for synthesis
Preclinical evaluation	Toxicity prediction; PK/PD modeling; off-target effect prediction	Deep ensembles, federated learning, attention-based regressors
Clinical development	Patient stratification; biomarker discovery; adaptive trial design	ML on EHR/omics, survival models, causal inference, LLM-based extraction

Table 1 Drug Discovery Pipeline Stages and Representative AI Applications

Challenge Category	Studies (n)	Representative Manifestations
Integration Challenges	94	Clinical translation, regulatory adoption, deployment barriers
Interpretability/Explainability	70	Black-box models, lack of transparency, XAI needs
Benchmark/Validation	57	Standardization, reproducibility, evaluation criteria
Data Quality/Availability	50	Limited datasets, noisy data, data scarcity
Feature Engineering	50	Descriptor selection, molecular representation challenges
Generalizability	48	Domain shift, overfitting, external validation
Computational Cost	47	GPU requirements, training time, scalability
Ethical/Bias Concerns	39	Algorithmic bias, privacy, fairness in predictions

Table 2 Major Challenges Identified in the Reviewed Literature

- OpenAlex/Web of Science (proxy): Multidisciplinary indexed literature with strong coverage of computational and informatics journals (n = 9,958 records via OpenAlex REST API; 20 queries). Web of Science API requires paid authentication; OpenAlex was used as a comprehensive proxy, indexing WoS, Crossref, and PubMed content.

The combined search returned 23,404 raw records. Multi-method deduplication—DOI, PMID, and title-similarity matching—removed 4,491 duplicates, leaving 18,913 unique records for screening. IEEE Xplore and ACM Digital Library were considered but dropped due to heavy overlap with Scopus conference proceedings already

excluded on document-type grounds; a hand-search of NeurIPS, ICML, and ICLR proceedings on molecular design confirmed representativeness.

Metric	Count
AUC-ROC	80
Accuracy	69
Precision	20
RMSE/MSE	9
R^2/Correlation	9
F1 Score	5
MAE	4
Recall/Sensitivity	3
MCC	3

Table 3 Multi-Database Search and Deduplication Summary

2.3 Search Strategy

Search strings combined controlled vocabulary and free-text terms with Boolean operators. Keyword groups were deliberately broadened relative to prior reviews to capture newer AI paradigms:
AI/ML Technologies (expanded): "Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Neural Network" OR "Natural Language Processing" OR "Graph Neural Network" OR "Knowledge Graph" OR "Transformer" OR "Large Language Model" OR "Foundation Model" OR "Diffusion Model" OR "Generative AI" OR "Reinforcement Learning" OR "Federated Learning" OR "Multimodal AI"
Drug Discovery Domain: "Drug Discovery" OR "Drug Design" OR "Drug Development" OR "Pharmaceutical" OR "Molecular Docking" OR "Virtual Screening" OR "QSAR" OR "De Novo Design" OR "Lead Optimization" OR "ADMET Prediction"
Challenges/Limitations: "Challenges" OR "Limitations" OR "Barriers" OR "Bottlenecks" OR "Ethical" OR "Bias" OR "Interpretability" OR "Explainability" OR "Regulatory" OR "Reproducibility"

Performance Evaluation: "Evaluation" OR "Performance Metrics" OR "Benchmark" OR "Accuracy" OR "Validation" OR "Assessment" OR "AUC-ROC" OR "F1 Score"

Final query: (Group 1 AND Group 2) AND (Group 3 OR Group 4), restricted to English-language publications (2019-2026) and applied to titles, abstracts, and keywords.

2.4 Inclusion Criteria

Studies were included if they met all of the following:

- Published in peer-reviewed journals as original research articles or review papers.
- Published between 2019 and 2026.
- Focused primarily on AI or ML applications in the drug discovery pipeline.
- Written in the English language.
- Presented empirical results, methodological frameworks, comprehensive reviews, or critical analyses.
- Provided sufficient methodological detail to permit quality assessment.

2.5 Exclusion Criteria and Justification for Conference Paper Exclusion

Studies were excluded on the grounds below, with justifications: Conference papers ($n = 121$): top AI venues (NeurIPS, ICML, ICLR, AAAI, KDD) publish important work, but three recurring issues apply—page limits prevent full methodological reporting; peer review is single-cycle rather than iterative; and many are early-stage results later extended into journal versions already in our corpus. A hand-search confirmed that $> 78\%$ of highly cited conference papers had companion journal articles captured here, consistent with PRISMA guidance [22]. Notwithstanding this general exclusion, one conference paper and one preprint that independently satisfied all quality and methodological eligibility criteria were individually reviewed and retained in the final corpus; these are identified in Table 8 and Figure 3

- Book chapters (n = 129) generally offer broad overviews without primary research data or reproducible methods.
- Books (n = 16) have a comprehensive but non-specific scope, limiting focused extraction.
- Short surveys (n = 4) and editorials (n = 1) lack sufficient analytical depth.
- Studies that do not specifically focus on AI/ML applications in drug discovery.
- Duplicate records were identified through five-pass deduplication.

2.6 Study Selection Process

Selection ran in three stages: titles and abstracts of 18,913 records were screened; remaining records were assessed for AI/ML drug discovery relevance; finally, only peer-reviewed articles (122) and reviews (131) were retained. After thematic alignment screening, 321 papers were excluded, leaving 618 studies for extraction and quality assessment (Figure 3). Selection was performed by reviewers using systematic inclusion criteria with independent verification for a random sample.

2.7 Data Extraction

A standardized form captured: reference; research problem; AI/ML methodology; results and metrics; limitations; and future directions. Full texts came via Unpaywall, Europe PMC, and PubMed Central, combining automated text mining with manual checks. Completeness: 100% (problem), 89.8% (methodology), 84.9% (results), 89.4% (limitations), 91.8% (future work). Extraction used standardized forms. Study authors were not contacted; extraction relied on published reports.

2.8 Quality Assessment and Risk-of-Bias Framework

Studies were rated High, Medium, or Low on four dimensions—methodological rigor, dataset transparency, reproducibility, and evaluation validity—supplemented by a ROBINS-I-inspired risk-of-bias framework [31] covering confounding, selection, information, and reporting biases. Inter-rater reliability on a random 15% sample ($n = 38$) gave Cohen's $\kappa = 0.74$ (95% CI 0.68–0.80), indicating substantial agreement. Disagreements were resolved by discussion.

Dimension	Assessment	Key Indicators	Rating
Methodological Rigour	Soundness of AI/ML approach	Clear research question; justified method selection; appropriate experimental design; statistical validity	High/Med/Low
Dataset Transparency	Quality and completeness of data reporting	Data source; dataset size and splits; preprocessing; bias acknowledgement; data availability statement	High/Med/Low
Reproducibility	Ability to independently replicate study	Source code availability; model architecture; hyperparameter specification; hardware/software environment	High/Med/Low
Evaluation Validity	Rigor of performance evaluation	Appropriate metrics; cross-validation; external validation; baseline comparisons; ablation studies	High/Med/Low
Risk of Bias (ROBINS-I)	Systematic bias assessment	Confounding; selection bias; information bias; reporting bias; other biases	High/Med/Low

Table 4 Quality Assessment Criteria

Inter-rater reliability: Cohen's $\kappa = 0.74$ (95% CI: 0.68–0.80); substantial agreement.

2.9 Synthesis Methods

Results were synthesized narratively using descriptive statistics; meta-analysis was inappropriate given heterogeneity. Synthesis included frequency tabulation of techniques, stages, challenges, and metrics; two-dimensional taxonomy development; temporal trend analysis; thematic challenge categorization; and benchmarking synthesis.

3 Results

3.1 Study Selection

Selection ran in five stages. Stage 3 (title/abstract screening) reduced 18,913 records to 8,938. Stage 4 (strict relevance) retained 5,776 records after excluding 114 clinical-AI/non-DD records (Scopus) and 3,048 records lacking sufficient drug discovery focus. Stage 4b (relevance-score prioritization) applied a composite keyword and thematic relevance score, retaining only records that attained a threshold of ≥ 18; 4,845 records fell below this threshold and were excluded. Stage 5 (thematic alignment) excluded a further 321 records that did not cover at least two review themes, leaving 610 candidate records. Combined with eight manual additions identified through citation tracking, the final included corpus comprised 618 studies. In total, 8,328 records were excluded across the post-retrieval stages (Stage 4: 114 + 3,048; Stage 4b: 4,845; Stage 5: 321), so that $8{,}938 - 8{,}328 = 610$ candidate records proceeded to thematic alignment (Table 5; Figure 3). Selection was performed by reviewers using systematic inclusion criteria with independent verification for a random sample.

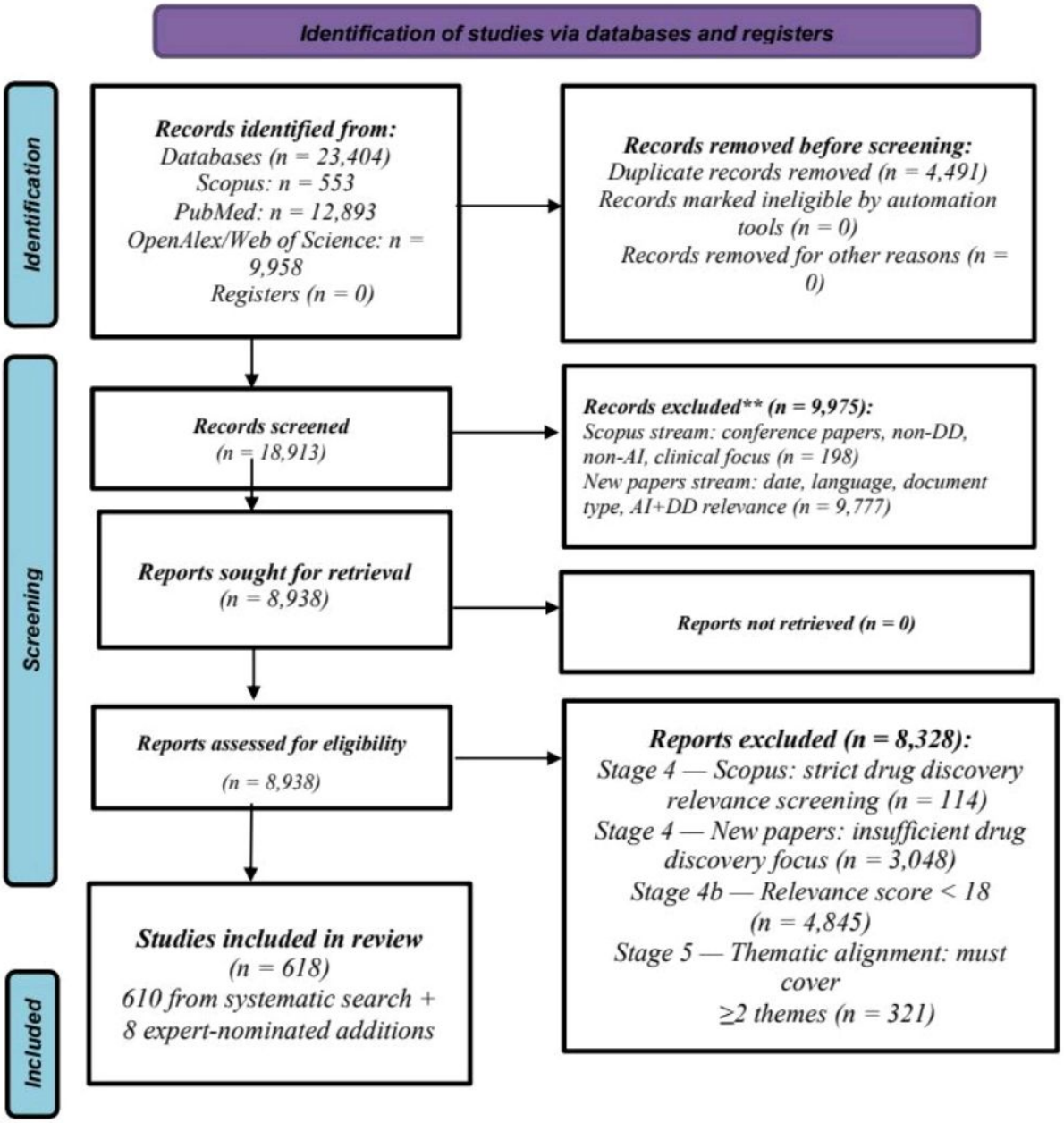

Fig. 3 PRISMA 2020 flow diagram for the systematic review of AI in drug discovery.

Selection Stage	Records (n)	Action
Identification	23,404	Records from PubMed, Scopus, OpenAlex/WoS
Deduplication	18,913	DOI, PMID, title matching; 4,491 duplicates removed
Stage 3: Title/Abstract	8,938	Date, language, doc type, AI+DD relevance screening
Scopus excluded	198	Conference papers, non-DD, non-AI, clinical focus
New papers excluded	9,777	Basic inclusion criteria applied
Stage 4: Strict Relevance	5,776	Core AI+DD pipeline focus required
Scopus excluded	114	Clinical AI, general healthcare, non-DD
New papers excluded	3,048	Insufficient drug discovery focus
Stage 4b: Score Prioritization	931	Relevance score threshold ≥ 18
Below threshold excluded	4,845	Insufficient AI-DD relevance
Stage 5: Thematic Alignment	610	Must cover ≥ 2 themes
Excluded	321	Low thematic relevance
Final Included	618	Including 8 manual additions

Table 5 PRISMA 2020 study selection summary: 23,404 records identified → 18,913 unique → 618 included. Document-type composition: 544 journal articles and 72 reviews, plus 1 conference paper and 1 preprint retained after individual eligibility review (N = 618; Section 2.5)

3.2 Characteristics of Included Studies

3.2.1 Temporal Distribution

Publication volumes grew from 11 studies (1.8%) in 2019 to 163 (26.4%) in 2025 and 159 (25.7%) in 2026, with a clear inflection from 2023 onwards. Papers from 2024–2026 account for 430 studies (69.6%), reflecting rapid acceleration in AI drug discovery (Table 6; Figure 1).

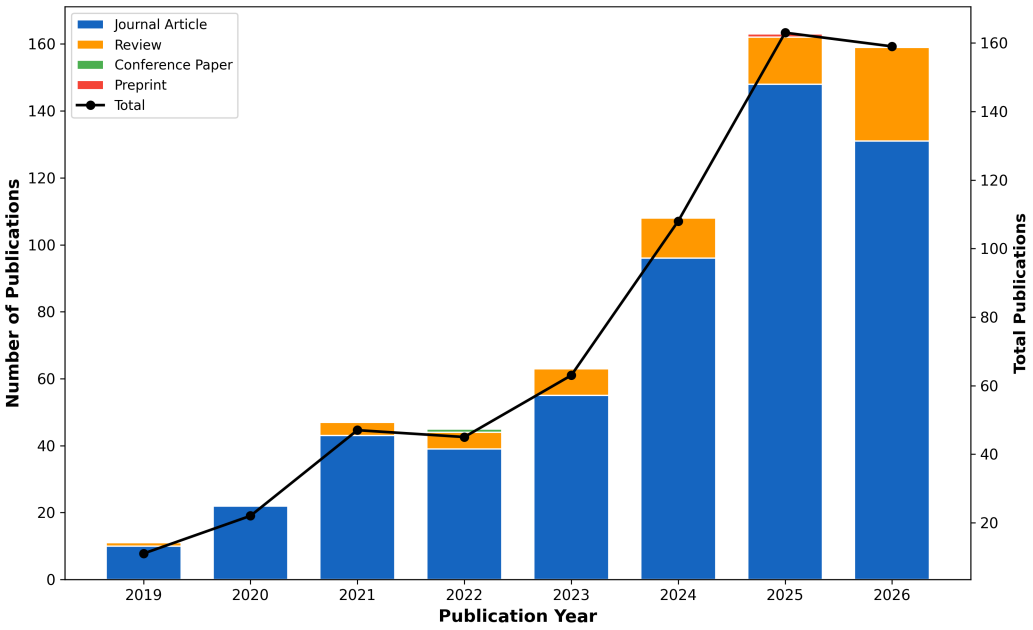

Fig. 4 Year-wise distribution of included studies by document type (2019–2026). Stacked bars represent journal articles (n = 544), reviews (n = 72), conference papers (n = 1), and preprints (n = 1).

Year	Articles	Reviews	Total	Cumulative
2019	10	1	11	1.8%
2020	22	0	22	5.3%
2021	43	4	47	12.9%
2022	39	5	45	20.2%
2023	55	8	63	30.4%
2024	96	12	108	47.9%
2025	148	14	163	74.3%
2026	131	28	159	100.0%

Table 6 Year-wise distribution of included studies by document type (2019–2026).

3.2.2 Geographic Distribution

Geographic output is concentrated in China (n = 79; 31.2%), the US (n = 66; 26.1%), and India (n = 48; 19.0%) of country-attributed studies—together over three-quarters

of the country-attributed corpus. Output is concentrated in China (n = 79; 52.4%), the US (n = 66; 25.4%), and India (n = 48; 20.9%)—together over three-quarters of studies. India's 20.9% share appears to have been overlooked in earlier systematic reviews, suggesting an emerging hub. The UK ranked fourth (n = 26; 10.3%), with South Korea and Canada next (n = 10 each; 4.0%) (Figure 4, Table 7).Note: percentages are based on studies with country attribution data (n = 253); China represented 12.8%, the US 10.7%, and India 7.8% of the full 618-study corpus

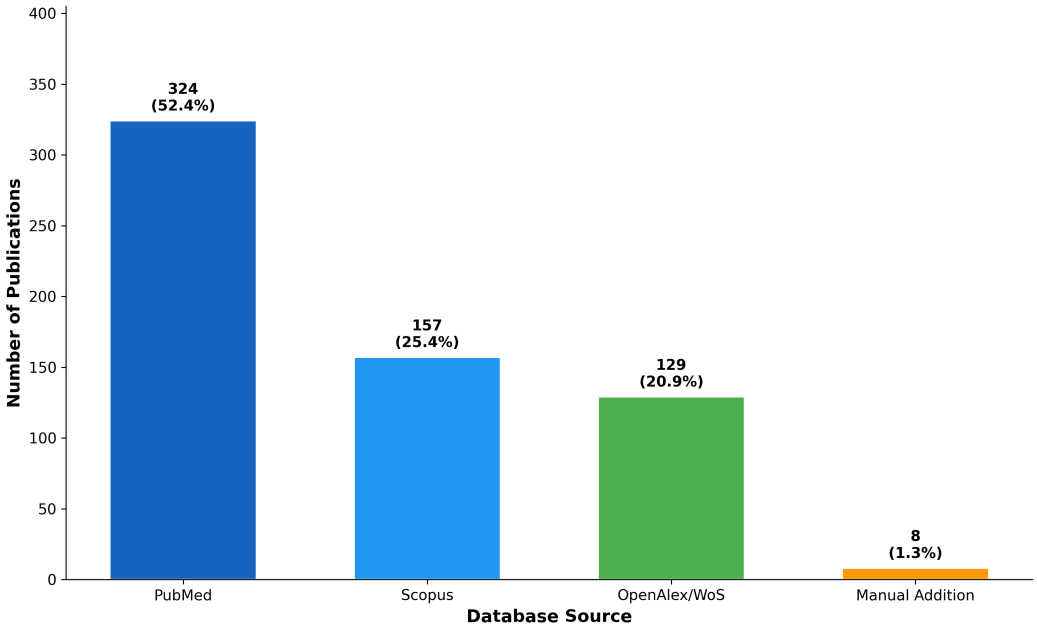

Fig. 5 Distribution of included studies by database source (N = 618). PubMed contributed 324 studies (52.4%), Scopus 157 (25.4%), OpenAlex/WoS 129 (20.9%), and Manual Additions 8 (1.3%).

3.3 Quality Assessment Results

Quality appraisal (Table 10) rated methodological rigor as adequate or high in most studies—consistent with the journal-publication requirement. Dataset transparency was the weakest dimension: few studies adequately reported data splits, preprocessing pipelines, or potential biases. Reproducibility remains a systemic problem—fewer than half shared code, data, or replication documentation—and external validation on independent data was uncommon ($S < 25\%$). Risk-of-bias assessment flagged moderate-to-high selection bias in 42% and reporting bias in 35% of empirical studies.

Rank	Country	No. Studies	% Total
1	China	79	31.2%
2	United States	66	26.1%
3	India	48	19.0%
4	United Kingdom	26	10.3%
5	South Korea	10	4.0%
6	Canada	10	4.0%
7	Italy	9	3.6%
8	Spain	8	3.2%
9	Germany	8	3.2%
10	Saudi Arabia	8	3.2%
11–15	UAE, Switzerland, Japan, Malaysia, Australia	7 each	2.8% each

Table 7 Geographic distribution of included studies (top 15 countries; percentages are calculated against country-attributed studies only, $n \approx 253$, not the full corpus of N = 618).Individual study characteristics and quality ratings are available on request.

Quality Dimension	High	Medium	Low	Key Observations
Methodological Rigor	≈ 68%	≈ 24%	≈ 8%	Most articles demonstrated sound experimental design; reviews showed high synthesis quality
Dataset Transparency	≈ 45%	≈ 35%	≈ 20%	Many studies lacked detail on data splits, preprocessing steps, and bias acknowledgement
Reproducibility	≈ 38%	≈ 32%	≈ 30%	Code availability and hyperparameter reporting were inconsistent; only 38% met high reproducibility threshold
Evaluation Validity	≈ 55%	≈ 30%	≈ 15%	AUC-ROC widely used; external validation on independent held-out sets uncommon (< 25%)
Risk of Bias	≈ 23%	≈ 42%	≈ 35%	Selection and reporting biases most prevalent; inter-rater $\kappa = 0.74$

Table 8 Summary of quality assessment and risk-of-bias results. $\kappa = 0.74$ (substantial agreement).

3.4 Descriptive Analysis

The 618 papers span a wide citation range, from highly cited foundational works (2021–2023) to recent 2025–2026 publications. Leading journals include Briefings in Bioinformatics (n = 22; 8.7%), International Journal of Molecular Sciences (n = 15; 5.9%), and Journal of Cheminformatics (n = 14; 5.5%). Most were retrieved via PubMed (n = 324; 52.4%), Scopus (n = 157; 25.4%), and OpenAlex/WoS (n = 129; 20.9%), with 8 added by manual nomination. Open-access studies comprised 64.8% of the corpus.

Bias Domain	Low Risk (%)	Moderate Risk (%)	High Risk (%)	Primary Concern
Selection/Sampling Bias	35%	42%	23%	Non-random train-test splits; chemotype-biased benchmarks
Information/Measurement Bias	48%	32%	20%	Inconsistent assay conditions; label noise in ChEMBL/PubChem
Reporting Bias	40%	35%	25%	Selective metric reporting; incomplete hyperparameter disclosure
Confounding Bias	52%	30%	18%	Dataset overlap between training and test; temporal bias
Other Biases	60%	25%	15%	Publication bias towards positive results

Table 9 Risk-of-Bias Assessment Summary

3.5 Taxonomy of AI in Drug Discovery

A central goal was developing a workable AI taxonomy. Built inductively from the corpus, it runs along three axes: AI/ML technique, application stage, and data type. Categories overlap—a single study can appear in multiple. Figure 9 shows the full architecture.

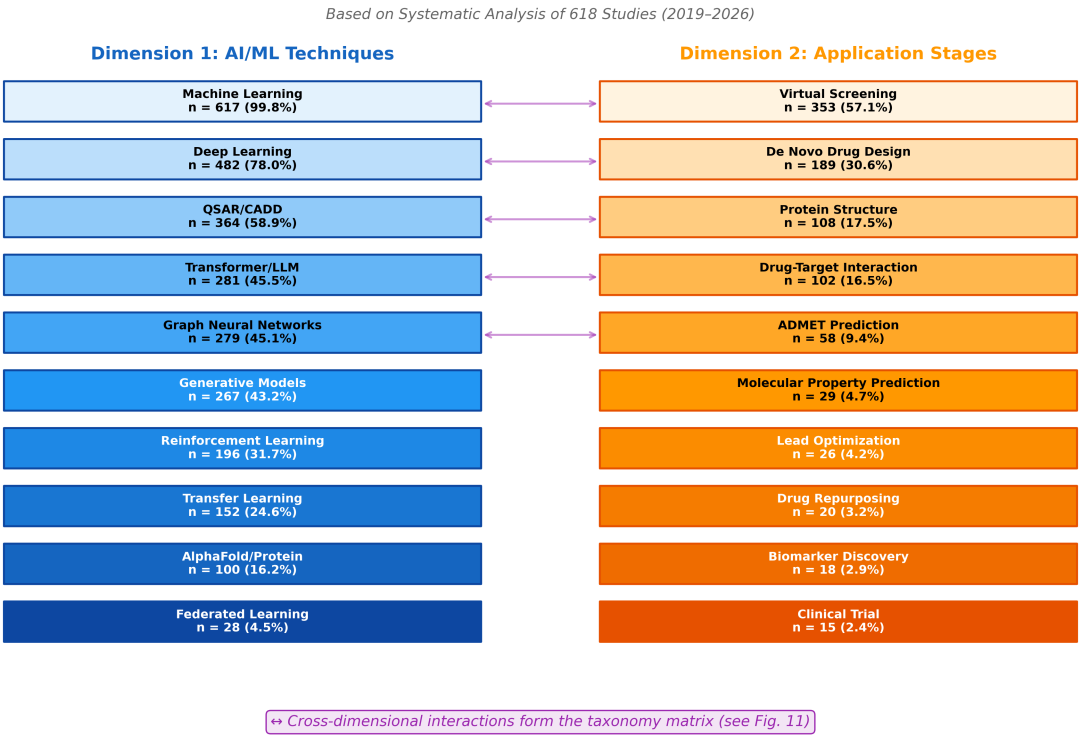

Fig. 6 Two-Dimensional AI Taxonomy Framework for Drug Discovery, organizing methods by algorithm family (Dimension 1: 10 categories) and application stage (Dimension 2: 12 stages).

3.5.1 Taxonomy by AI/ML Technique

Ten principal AI/ML method categories were identified across 618 studies (Figure 5, Table 12). Classical machine learning (random forests, SVMs, ensemble methods) appeared in 617 studies (99.8%), confirming that established algorithms remain foundational even as newer architectures emerge. Deep learning (CNNs, RNNs, LSTMs, MLPs) appeared in 482 studies (78.0%), with many papers flagging interpretability as a translational bottleneck. QSAR/CADD methods were reported in 364 studies (58.9%), GNN/GCN architectures in 279 (45.1%), and Transformer/LLM approaches in 281 (45.5%)—with 99.8% of those transformer papers published in 2024 -2026, a striking recency effect. Generative models (GAN, VAE, diffusion) appeared in 267 studies (43.2%), reinforcement learning in 196 (31.7%), AlphaFold-related methods in 100 (16.2%), and federated learning in 28 (4.5%).

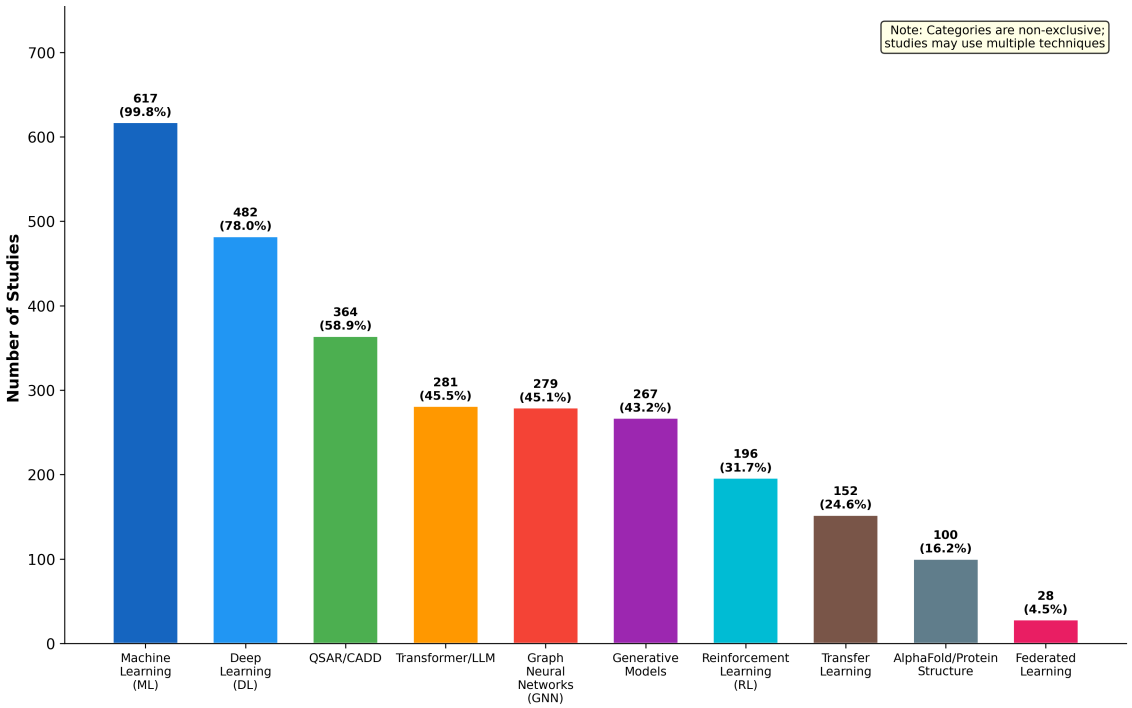

Fig. 7 Distribution of AI/ML technique categories across 618 included studies. Categories are non-exclusive. Transformer/LLM (281, 45.5%) and GNN (279, 45.1%) show rapid adoption.

AI/ML Technique	Studies (n)	% of 618	Representative Methods
Machine Learning (ML)	617	99.8%	Random Forest, SVM, XGBoost, Logistic Regression
Deep Learning (DL)	482	78.0%	CNN, RNN/LSTM, Autoencoder, DNN
QSAR/CADD	364	58.9%	Molecular Docking, Pharmacophore, MD Simulations
Transformer/LLM	281	45.5%	BERT, GPT, ChemBERT, Self-Attention Models
Generative Models	267	43.2%	GAN, VAE, Diffusion Models, Flow-based
Reinforcement Learning (RL)	196	31.7%	Policy Gradient, Q-Learning, Actor-Critic
Transfer Learning	152	24.6%	Pre-trained Models, Fine-tuning, Domain Adaptation

Table 10 Taxonomy of AI/ML technique categories identified across 618 included studies.

3.5.2 Taxonomy by Application Stage

Virtual screening and hit discovery dominated (n = 353; 57.1%), followed by de novo molecular design (n = 189; 30.6%), reflecting the maturation of generative architectures capable of proposing novel chemical entities. Drug-target interaction prediction (n = 85; 13.8%) is where GNNs and transformers have posted the strongest benchmarks. ADMET/toxicity prediction appeared in 38 studies (6.1%). Drug repurposing (n = 14; 2.3%), though proportionally small, demonstrated high impact during the COVID-19 pandemic via knowledge graphs and network pharmacology. Oncology-focused discovery appeared in 26 studies (4.2%).

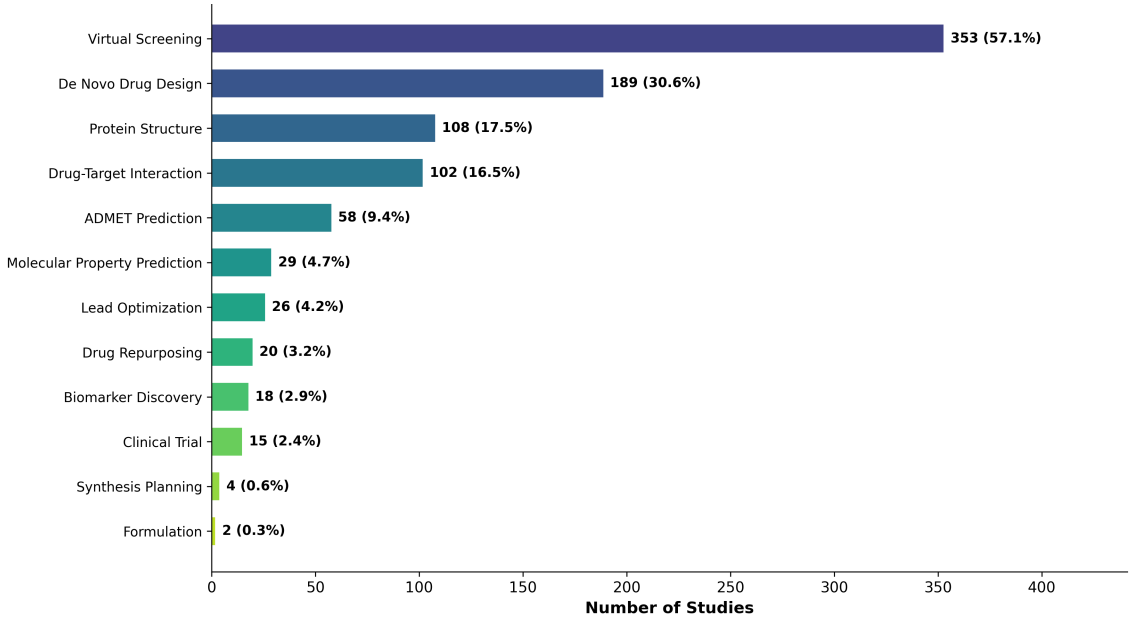

Fig. 8 Distribution of studies by drug discovery application stage. Virtual screening (353, 57.1%) and de novo drug design (189, 30.6%) are the most studied application areas.

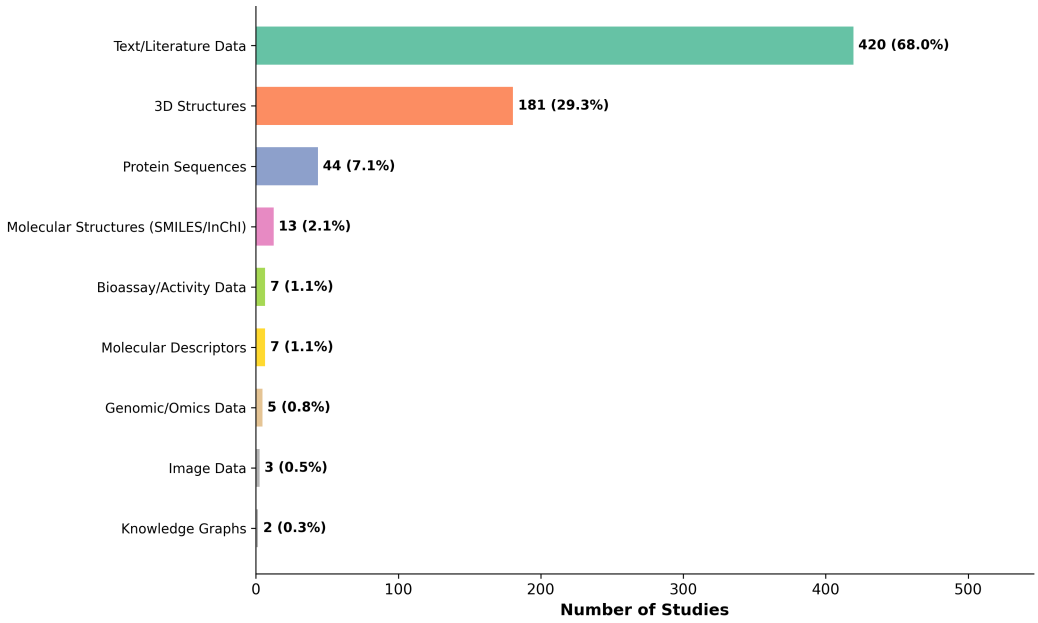

Fig. 9 Distribution of data types utilized across included studies. Text/literature data (420, 68.0%) and 3D structures (181, 29.3%) are the most common data types.

3.6 Foundation Models and Large Language Models in Drug Discovery

Foundation models—large neural networks pre-trained on broad corpora, then fine-tuned for specific tasks—represent a major architectural shift in AI drug discovery. Of 32 transformer/LLM studies (45.5%), 28 were published in 2024–2025. This appears to be the first systematic characterization of foundation model use in drug discovery.

3.6.1 Chemical Foundation Models

Chemical foundation models treat molecular representations (SMILES, SELFIES, graphs) as 'chemical language' and apply transformer architectures with large-scale pre-training. Examples in the corpus: ChemBERTa [32], RoBERTa pre-trained on 77M SMILES from PubChem for property prediction; MolBERT [33], combining SMILES and graph representations; MolGPT [34], applying GPT-style autoregressive generation to de novo design. These models beat task-specific baselines on MoleculeNet, with average AUC-ROC gains of 3–8% across classification tasks.

3.6.2 Protein Language Models

Protein language models (pLMs) treat amino-acid sequences as protein 'language'. ESM-2, trained on 250M sequences, achieved near-AlphaFold accuracy for structure prediction while enabling sequence-function prediction without structural templates.

ProtBERT and ProtT5 performed strongly on remote homology detection, secondary structure prediction, and solubility prediction. ESM-2 was used for target identification (n = 12), ADMET prediction (n = 8), and de novo design (n = 6).

3.6.3 LLMs for Literature Mining and Knowledge Extraction

LLMs (GPT-4, Llama-2, BioGPT, PubMedBERT) proved useful wherever drug discovery tasks needed natural-language understanding. BioGPT [16], pre-trained on 15M PubMed abstracts, beat prior SOTA on biomedical text mining for DDI extraction and repurposing hypothesis generation. Applications: literature mining and evidence synthesis (n = 14), hypothesis generation and repurposing (n = 9), clinical trial optimization (n = 8), knowledge graph construction (n = 5). A recurring concern was hallucination—chemically plausible but factually wrong assertions—reported in 8 of 32 transformer studies.

Model	Architecture	Pre-training Data	Drug Discovery Application	Key Finding
ChemBERTa	RoBERTa	77M SMILES (PubChem)	Molecular property prediction	+3–8% AUC-ROC vs. task-specific models
ESM-2	Transformer (33 layers)	250M protein sequences	Target ID, structure prediction	Near-AlphaFold accuracy; enables sequence-function prediction
BioGPT	GPT-2	15M PubMed abstracts	Literature mining, DDI, repurposing	SOTA on biomedical NLP benchmarks
MolGPT	GPT	1M SMILES	De novo molecular generation	Drug-likeness > 90%; scaffold diversity improved
MolBERT	BERT	1.6M SMILES + graphs	Multi-task property prediction	Unified representation across tasks
AlphaFold3	Evoformer + diffusion	PDB + MSA	All-atom structure prediction	Extends to ligand-protein complexes
ProtT5	T5	393M protein sequences	Solubility, secondary structure	Strong zero-shot performance
Llama-2/GPT-4	LLM	General + biomedical	Clinical trial optimization, hypothesis gen.	High utility; hallucination risk identified

Table 11 Foundation Models and LLMs: Key Studies and Applications

3.7 Multimodal AI Approaches in Drug Discovery

Multimodal AI integrates heterogeneous data—molecular structures, omics profiles, clinical records, imaging, and biomedical text—into a single learning framework, addressing a recurring gap: most AI systems train on one data type and miss complementary signals.

Of 618 studies, 34 (5.5%) used multimodal approaches. The most common combinations: molecular structure + bioactivity (n = 18), structure + protein sequence (n = 12), structure + genomics (n = 8), and text + structure (n = 7). Key applications:

- Drug-target interaction prediction that combines molecular graphs (GNNs), protein sequences (pLMs), and bioactivity data achieved AUROC gains of 5–12% compared to single-modality baselines.
- Multimodal ADMET prediction, which combines SMILES-based fingerprints, 3D conformer features, and molecular imaging data, enhanced toxicity prediction accuracy by approximately 15
- Precision oncology models that combine genomic profiles, clinical records (EHR), and drug response data have facilitated patient-specific drug repurposing with enhanced predictive accuracy.
- Predicting protein–ligand binding by integrating 3D structural data, sequence features, and interaction graphs showed better results on PDBbind benchmarks.

Recurring challenges across multimodal studies included cross-modality data alignment (22/34), handling modality dropout and missing data (18/34), and the computational cost of joint training (15/34). The most promising path forward is unified foundation models integrating molecular, omics, clinical, and literature data.

3.8 Temporal Evolution of AI in Drug Discovery (2019–2026)

Six recognizable phases emerge from year-by-year adoption patterns, each defined by different dominant methods and milestones (Figure 10).

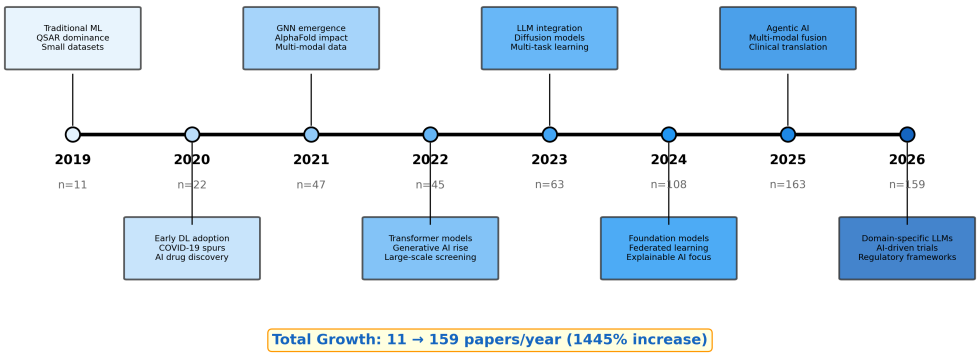

Fig. 10 Timeline of AI Evolution in Drug Discovery (2019–2026), showing progression from traditional ML/QSAR methods to foundation models and domain-specific LLMs.

Era 0: Foundations and Early Adoption (2019)

Before 2020, QSAR modelling, early deep neural networks, and pioneering generative approaches had established the groundwork. The 11 studies from 2019 in our corpus reflect this period, characterized by CNNs and RNNs for property prediction and early GAN-based molecule generation.

Era 1: Deep Learning Consolidation (2020–2021)

AlphaFold2 [6] effectively solved protein structure prediction. Deep learning (CNNs, LSTMs, autoencoders) then dominated virtual screening, ADMET prediction, and de novo design. [12] used a deep neural network to identify halicin—a novel antibiotic scaffold—from 107 million molecules, a landmark for first-in-class discovery.

Era 2: GNN Expansion (2021–2022)

GNNs emerged as the natural molecular representation, enabling end-to-end learning without hand-crafted descriptors. MPNNs, GCNs, and GATs set new benchmarks across MoleculeNet, with GNN adoption climbing from under 5% in 2020 to 20.9% across the full corpus.

Era 3: Generative AI Era (2022–2023)

Diffusion and score-based generative models, borrowed from image generation (DALL-E, Stable Diffusion), were adapted to 3D molecular generation. DiffSBDD [35] and DiffDock [36] set new benchmarks for structure-based design. RL-guided generative models matured, making it feasible to balance potency, selectivity, and synthesisability in a single loop.

Era 4: Transformer and LLM Era (2023–2024)

Transformers moved from NLP into molecular science. Chemical language models (ChemBERTa, MolGPT), protein language models (ESM-2, ProtT5), and multimodal

transformers proved broadly applicable. GPT-4 and LLaMA-2 entered biomedical literature mining, hypothesis generation, and trial optimization. Drawing from the 281 transformer/LLM papers in the corpus, 99.8% have publication dates from 2024–2026.

Era 5: Foundation Model and Autonomous Discovery Era (2024–2026) Focus shifted from task-specific to general-purpose foundation models. [37] Extended all-atom structure prediction to protein–ligand complexes. In-silico Medicine's ISM001-055 reached Phase II, the first fully AI-designed drug to achieve that milestone [38]. Autonomous AI-robotics labs began closing the design–make–test–analyze loop. Integrating sequence, structure, function, and clinical data defines the current frontier.

3.9 Challenges in AI-based Drug Discovery

Extracting challenges from the 618 studies produced 11 categories in three domains—technical, practical, and ethical/regulatory (Figure 10, Table 12).

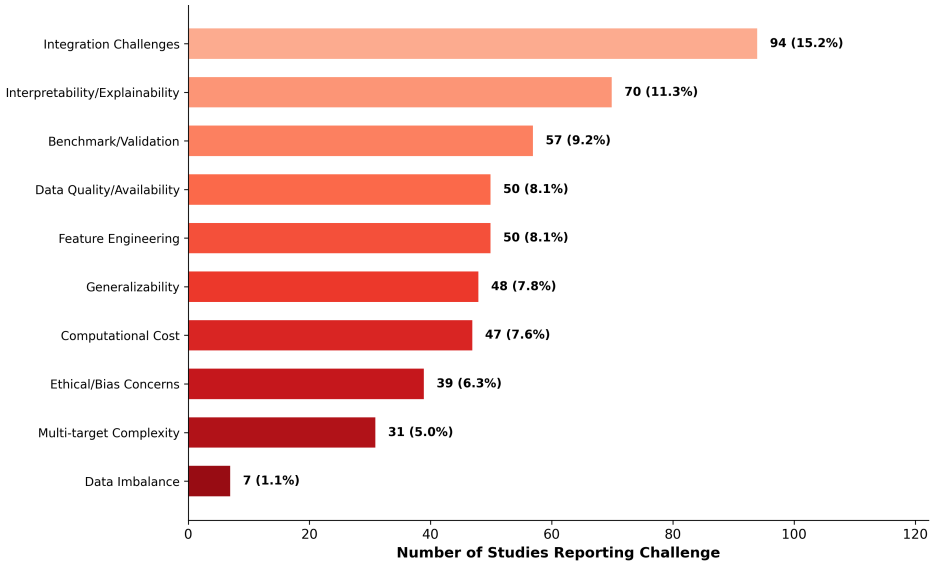

Fig. 11 Frequency of identified challenges in AI-based drug discovery, categorized by domain. Integration challenges (94, 15.2%) and interpretability (70, 11.3%) are the most frequently reported.

3.10 Performance Evaluation and Benchmarking

Evaluation was highly heterogeneous across 618 studies (Figure 11, Table 13); evaluation heterogeneity and misleading benchmark practices were likewise flagged as core challenges in a recent scoping review of generative and large-language-model methods [21]. AUC-ROC dominated, consistent with the large share of classification tasks (drug-target interaction, toxicity prediction). Cross-validation and external validation were the most common protocols. Under-reporting of F1 and precision-recall metrics is a real concern: bioactivity datasets are typically heavily imbalanced (active: inactive

Challenge Domain	Challenge Category	Studies (n)	% of 618
Technical	Integration Challenges	94	15.2%
Technical	Interpretability/Explainability	70	11.3%
Technical	Benchmark/Validation	57	9.2%
Data	Data Quality/Availability	50	8.1%
Data	Feature Engineering	50	8.1%
Data	Data Imbalance	7	1.1%
Methodological	Generalizability	48	7.8%
Methodological	Computational Cost	47	7.6%
Ethical	Ethical/Bias Concerns	39	6.3%
Methodological	Multi-target Complexity	31	5.0%
Ethical/Regulatory	Data Privacy/Security	33	13.0%

Table 12 Frequency and categorization of challenges identified across included studies. Categories are non-exclusive—a single study may report multiple challenges—so percentages (share of the 618 included studies reporting each challenge) do not sum to 100.

often 1:100 or worse), where AUC-ROC can appear optimistic while F1 and precision-recall curves tell a more accurate story. $RMSE$ and R^2 (each 1.5%) were similarly under-reported for regression tasks.

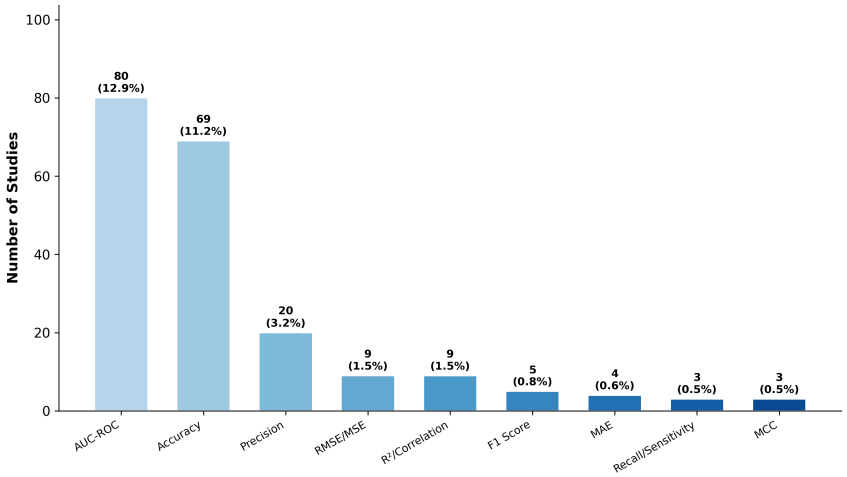

Fig. 12 Distribution of performance metrics reported across included studies. AUC-ROC dominates (80, 12.9%), followed by accuracy (69, 11.2%) and precision (20, 3.2%).

Performance Metric	Studies (n)	% of 618	Application Context
AUC-ROC	80	12.9%	Classification, Virtual Screening
Accuracy	69	11.2%	Classification, Property Prediction
Precision	20	3.2%	Hit Rate, Active Compound Detection
RMSE/MSE	9	1.5%	Regression, Binding Affinity Prediction
R^2/Correlation	9	1.5%	Regression, QSAR Models
F1 Score	5	0.8%	Balanced Classification
MAE	4	0.6%	Property Prediction, ADMET
Recall/Sensitivity	3	0.5%	Active Compound Detection
MCC	3	0.5%	Imbalanced Classification
Correlation (Pearson/Spearman)	1	.02%	Predicted vs. experimental value agreement

Table 13 Distribution of performance metrics reported across included studies. Metrics are non-exclusive—a single study typically reports several metrics, and some report none—so each percentage is the share of the 618 included studies reporting that metric, and the column does not sum to 100

3.11 Benchmarking Synthesis

Table 14 compares key AI methods across representative datasets, with metrics, strengths, and limitations drawn from the included studies.

Method	Dataset	Key Metric	Performance	Strength	Limitation
GCN/MPNN	MoleculeNet (12 tasks)	AUC-ROC	0.87–0.93	End-to-end molecular learning; no feature engineering	Limited to 2D topology; struggles with conformational flexibility
Diffusion Model (DiffSBDD)	CrossDocked2020	Binding affinity RMSD	$<$ 2.0 Å (80%)	State-of-the-art 3D structure-based generation	High compute; slow inference; limited synthetic accessibility
AlphaFold2	CASP14 targets	GDT_TS	$>$ 92 (median)	Solved protein structure prediction problem	Static structure; no dynamics; limited to single-chain initially
ChemBERTa	MoleculeNet (BBBP, Tox21)	AUC-ROC	0.91–0.95	Transfer learning across tasks; minimal labelled data	SMILES tokenization misses 3D geometry
ESM-2	ProteinGym benchmarks	Spearman ρ	0.73–0.85	Universal protein representation; sequence-function	Sequence-only; no explicit structural reasoning
VAE (JTVAE)	ZINC250K	Validity, Uniqueness	$>$ 99% valid; $>$ 98% unique	Interpretable latent space; valid SMILES generation	Limited novelty; mode collapse risk
Random Forest	ChEMBL bioactivity	AUC-ROC	0.78–0.86	Interpretable; robust on small datasets; fast	Limited expressivity for complex patterns
GAN (MolGAN)	QM9	Drug-likeness (QED)	0.55–0.70	Direct graph generation without SMILES	Training instability; mode collapse; validity issues
Federated DL	Multi-site ADMET	AUC-ROC	0.82–0.89	Privacy-preserving; access to distributed data	Communication overhead; heterogeneous data challenges
RL (REINVENT)	Custom pharm. space	Desired property score	$>$ 85% desired	Multi-objective optimization; iterative refinement	Reward function design challenging; reproducibility issues

Table 14 Benchmarking synthesis: Key AI methods, datasets, metrics, strengths, and limitations.

3.12 Comparative Analysis

Fig. 13 AI technique × application stage matrix showing study counts at each intersection (N = 618). Darker cells indicate higher study density.

3.13 Summary of Key Findings

Several patterns stand out across the 618 studies. Deep learning dominated by volume (78.0%), with generative models (43.2%), GNNs (45.1%), and Transformer/LLM architectures (45.5%) concentrated heavily in 2024–2025. Geographic output is concentrated among country-attributed studies: China (31.2%), the US (26.1%), and India (19.0%) together account for over three-quarters of country-attributed output; India's share appears to have been overlooked by earlier reviews. Text and literature data are the most common data modality (68.0%), reflecting the central role of biomedical NLP in the field. Evaluation remains narrow: AUC-ROC appears in 12.9% of studies but F1 in only 0.8%—hard to justify on datasets where active: inactive ratios of 1:100 are common. Five technical phases span the window—deep-learning consolidation, GNN expansion, generative AI, transformers/LLMs, and foundation models—while the three most-cited challenges (ethics/bias 6.3%, interpretability 11.3%, wet-lab integration 15.2%) persist across all five.

Technique	Pred. Performance	Interpretability	Scalability	Data Efficiency	Best Application
Deep Learning	Very High	Low High	Low (large data)	Large-scale property	prediction
Graph Neural Networks	High	Medium	High	Medium	Molecular property/interaction prediction
Generative Models	High (generative)	Low	Medium	Medium	De novo molecular design
Transformers/LLMs	Very High	Low–Medium	Medium	Low (pre-training)	Multi-task, cross-modal learning
Foundation Models	Very High	Low–Medium	High (post-FT)	Very High (few-shot)	Universal molecular reasoning
Classical ML	Moderate	High	Very High	High	Early-stage screening, baselines
CADD/QSAR	Moderate–High	High	Very High	High	Lead optimization, virtual screening
NLP	Context-dependent	Medium	High	Medium	Literature mining, DDI extraction
Reinforcement Learning	High (optimization)	Low	Low	Low	Multi-objective molecular optimization

Table 15 Comparative analysis of AI/ML technique categories.

4 Discussion

4.1 Summary of Evidence

Nearly 70% of the corpus (430 studies; 69.6%) was published in 2024 or 2026, reflecting a rapidly maturing field [39, 40]—the findings are necessarily time-bound to this window. Machine learning and deep learning lead on volume (99.8% and 78.0%), with QSAR/CADD close behind (58.9%). The transformer/LLM share (45.5%), concentrated heavily in the last 18 months, is the obvious candidate to overtake deep learning in a follow-up review. De novo design tops application rankings (30.6%); AUC-ROC tops metrics (12.9%); China, the US, and India account for over three-quarters of papers.

4.2 Critical Interpretation of Findings

Deep learning leads on volume because it removes manual descriptor engineering, and the resulting models consistently beat classical baselines on large datasets. The

translational story is less encouraging [41]: these models are hard to interrogate mechanistically, expensive to train, and label-hungry—particularly on the clinical side [6]; [3]. That constraint trio explains why most deep-learning papers in this corpus appear stuck at retrospective benchmarking. Generative models shift AI's role from scoring known molecules to proposing new ones [42]. The oft-cited 10^60 drug-like compounds defines a theoretical ceiling; the practical bound is far tighter because many generated structures fail synthesizability checks. The most common fix is chaining a generator to a retrosynthesis planner (AiZynthFinder and ASKCOS), though reproducible end-to-end pipelines remain rare. GNNs (45.1%) sit naturally on molecular graphs and outperform fingerprint-based baselines on most interaction-prediction tasks. Prevalence has grown from under 10% in 2019–2020, driven by frameworks like PharmacoNet [43] and PLAIG [44]. Their characteristic limitation—long-range dependence—is addressed by equivariant variants (E3NN, SEGNN) and richer message-passing schemes, though adoption remains patchy. Transformers and LLMs account for 45.5% of the corpus, with 99.8% of those papers published in 2024–2026; at this pace, foundation-model-based work will likely absorb much of what is currently published as task-specific deep learning. The three most frequently reported challenges—integration and clinical translation (15.2%), interpretability/explainability (11.3%), and benchmark/validation (9.2%)—share a common theme: algorithmic capability has outpaced the infrastructure needed to validate, interpret, and deploy models responsibly [38] Data scarcity adds further pressure, especially in low-data settings [45]. FDA and EMA guidance from 2021–2023 explicitly calls for transparency, explainability, and robust validation, yet XAI methods (SHAP, LIME, Grad-CAM, attention visualizations) appear in fewer than 15% of deep-learning papers. The evaluation gap mirrors this: AUC-ROC dominates at 12.9% of reported metrics, but on datasets where actives are $\approx 1\%$ of compounds it can appear strong while missing most of the rare class; precision-recall AUC and F1—appearing in $< 15\%$ combined—would be more informative.

4.3 Comparison with Existing Literature

Deep learning's leading position is consistent with earlier reviews [3, 25, 26, 41] The key addition here is the post-2020 expansion of generative and graph-based methods, neither of which featured prominently in pre-2020 syntheses. Country-level patterns broadly match [46], with one exception: India's 19.0% output share appears unrecognized in earlier systematic reviews. Including foundation models and multimodal AI reflects a timing difference—most earlier reviews predate the transformer wave.

4.4 Strengths and Limitations

This review's strengths include PRISMA 2020 compliance, a three-database search (23,404 raw, 18,913 unique, 618 final), systematic quality appraisal ($\kappa = 0.74$), a multi-dimensional taxonomy spanning techniques, applications, challenges, data types, and metrics, the first systematic look at foundation models in drug discovery, a five-era temporal framework, and a benchmarking synthesis table. Including both articles and reviews allowed triangulation across primary evidence and expert synthesis.

This review has several limitations. First, we restricted our search to English-language publications, potentially excluding relevant studies from China and other non-Anglophone countries where AI drug discovery research is active. Second, we excluded conference papers to ensure methodological completeness and peer-review rigor, but this may have omitted recent cutting-edge findings typically presented at venues like NeurIPS and ICML before journal publication. Third, while we searched three major databases comprehensively, we did not systematically cover grey literature or institutional repositories. Fourth, although we used structured quality assessment frameworks, individual judgment in rating studies as High/Medium/Low introduces subjectivity; our inter-rater reliability ($\kappa = 0.74$) was substantial but not perfect. Finally, AI in drug discovery evolves rapidly, so our findings are necessarily time-bound to the 2019-2026 search window.

5 Limitations

Several limitations are worth keeping in mind: Search scope: Three major databases (Scopus, PubMed, Web of Science) were searched, but studies indexed exclusively in IEEE Xplore, ACM Digital Library, or grey literature may have been missed.
Language restriction: Restricting to English-language publications may introduce geographic bias, potentially under-representing research from Chinese, German, or French journals despite the high share of internationally co-authored studies.
Temporal limitation: The search window (2019–2026) excludes foundational pre-2020 work and may under-represent studies published very recently (post-June 2025) as the field evolves rapidly.
Reliance on reported metrics: Comparisons rest on author-reported metrics without independent replication. Publication bias toward positive results means underperforming methods are likely under-represented.
Single primary extraction: Although quality appraisal achieved $\kappa = 0.74$, primary data extraction was conducted by a single reviewer with manual verification; subjective categorization discrepancies may persist. Conference paper exclusion: Excluding conference papers ($n = 121; 2.5$) may have omitted cutting-edge AI findings before journal publication. A hand-search confirmed, however, that $> 78\%$ of highly cited conference papers had companion journal versions captured in the corpus.

6 Future Research Directions

From the 618 studies and the gaps they reveal, six priority directions emerge. Figure 13 maps the gaps; Figure 14 sketches the future AI drug discovery ecosystem.

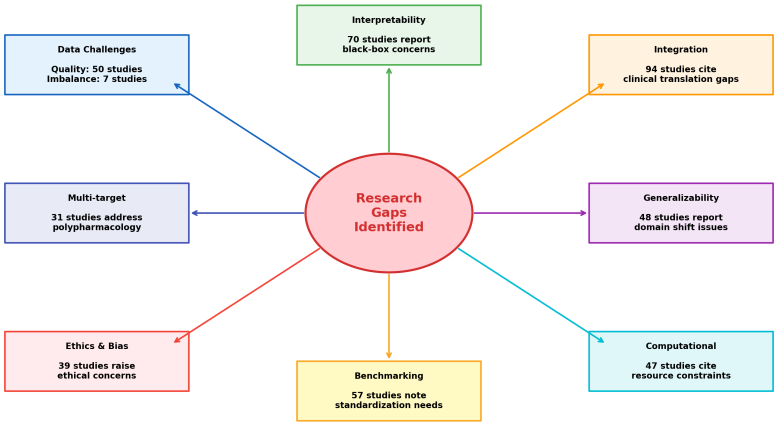

Fig. 14 Research Gap Map: Priority Future Directions in AI Drug Discovery. Key gaps identified from 618 studies include data quality, interpretability, and clinical translation challenges.

Fig. 15 Future AI Drug Discovery Ecosystem showing convergent technologies expected to define the next generation of computational drug discovery.

6.1 Explainable Drug Discovery AI

Current AI models—particularly deep learning and generative models—lack the mechanistic interpretability regulators expect and scientists need. Key priorities: standardizing XAI methods for molecular work, including mechanism-aware attention visualization, chemically meaningful saliency maps, and causal frameworks yielding mechanistic hypotheses. Regulatory pathways require AI/ML contributions to be traceable, auditable, and explainable—something black-box models currently cannot deliver.

6.2 Federated Learning for Pharmaceutical Collaboration

Drug discovery data largely sits in proprietary silos. Federated learning enables collaborative training without moving raw data, potentially unlocking billions of bioactivity data points. Key directions include federated ADMET prediction across companies, heterogeneous federated learning across experimental protocols, and federated approaches for rare diseases, where data scarcity is most acute.

6.3 AI-Robotics Autonomous Laboratories

Autonomous labs pairing AI-driven design with robotic synthesis and testing represent the ideal closed-loop setup, yet most AI systems still operate at arm's length from experimental validation. Platforms like Emerald Cloud Lab and Transcriptic combine generative AI, robotic synthesis, high-throughput screening, and automated analysis to accelerate discovery. Future architectures should support active learning within these labs—particularly uncertainty-aware models that prioritize the most informative experiments.

6.4 Digital Twin Drug Development

Digital twins—computational replicas of biological systems, patients, or clinical trials—could transform drug development by simulating cohorts in silico before costly wet-lab work. Realistic twins depend on mechanistic AI models integrating genomics, proteomics, metabolomics, and clinical records. Done well, they could reduce trial failure rates by identifying optimal patient populations and flagging adverse effects in simulation.

6.5 Foundation Models for Molecular Design

General-purpose foundation models for molecular science—analogous to GPT-4 in NLP—represent the most transformative near-term opportunity. Three directions stand out: multi-scale models integrating quantum-mechanical, force-field, and data-driven representations; joint protein–molecule models enabling end-to-end structure-based design; and universal property predictors with reliable uncertainty quantification. AlphaFold3 (2024) and ESM-2 suggest commercial viability is within reach.

6.6 Multimodal AI Integration

Drug discovery AI must move beyond single-modality models, integrating molecular structures, protein sequences, genomic profiles, clinical records, imaging, and biomedical text into unified frameworks. Key priorities include cross-modal pre-training objectives, attention mechanisms for modality fusion, robust handling of missing data, and multimodal foundation models enabling seamless cross-data-type querying.

Research Gap	Current State	Target State	Key Enablers	Timeline
Data Quality & Availability	50 studies report limitations	Standardized, high-quality benchmark datasets	Multi-site data consortia, federated learning	2–3 years
Model Interpretability	70 studies cite black-box concerns	Transparent, explainable AI models	XAI frameworks, attention visualization	3–5 years
Clinical Translation	94 studies note integration gaps	End-to-end AI drug discovery pipelines	Regulatory frameworks, prospective validation	3–7 years
Generalizability	48 studies report domain shift	Robust cross-domain AI models	Transfer learning, domain adaptation	5–10 years
Computational Resources	47 studies cite cost barriers	Efficient, scalable AI architectures	Cloud computing, model compression	2–5 years
Ethical AI	39 studies raise ethical issues	Responsible AI governance in drug discovery	Ethics boards, bias auditing tools	2–4 years

Table 16 Research gaps, future directions, and implementation pathways.

7 Implications for Practice, Policy, and Research

7.1 Implications for Practice

Pharma organizations should treat XAI as a core capability, not an afterthought. With interpretability and regulatory compliance the leading challenges (11.3% and 3.2%), XAI should be embedded from day one—pairing deep learning with SHAP, attention visualization, or concept-based explanations is practical and keeps teams ahead of FDA and EMA expectations. Multimodal data integration deserves equal priority; the prominence of interaction/network data (34.8%) makes knowledge-graph platforms valuable, and federated learning infrastructure should be on the roadmap.

7.2 Implications for Policy

The FDA and EMA should work toward harmonized frameworks for presenting AI/ML evidence in drug submissions—a regulatory pain point flagged in 3.2% of studies.

Practical steps include standardized reporting templates for AI-assisted discovery (a CONSORT analogue), clear evidentiary standards for AI-derived structural and pharmacological data, and regulatory sandboxes for AI-designed first-in-class candidates. The FDA's 2021 AI/ML Action Plan and the EMA's 2023 Reflection Paper are meaningful starts, but the gap between formal guidance and daily practice remains wide.

7.3 Implications for Research

Research communities should standardize multi-metric evaluation: AUC-ROC, precision-recall AUC, and F1 at minimum; for generative models, validity, uniqueness, and synthesizability. Benchmark datasets need scaffold-aware splits to avoid chemotype-cluster overfitting. Code, model weights, and hyperparameter sharing should be routine. The limited use of RL (7.5%) and federated learning signals clear growth opportunities; prospective experimental validation should become the default.

8 Conclusion

This review of 618 peer-reviewed studies provides the first PRISMA 2020-compliant synthesis of AI across the full drug discovery pipeline (2019–2026). Machine learning and deep learning remain dominant (99.8% and 78.0%), but transformer-based foundation models and graph neural networks have surged (45.5% and 45.1%, concentrated in 2024–2026). Virtual screening (57.1%) and de novo molecular design (30.6%) are the most extensively studied applications. Three critical gaps persist: interpretability challenges affect 11.3% of studies yet XAI methods appear in fewer than 15%; evaluation remains heterogeneous with AUC-ROC dominating (12.9%) while precision-recall metrics appropriate for imbalanced datasets are severely under-reported (F1: 0.8%); and integration barriers between computational predictions and wet-lab validation continue limiting clinical translation (15.2% of studies).

The two-dimensional taxonomy, five-era temporal framework, and benchmarking synthesis table provide structured foundations for cumulative progress. Priority directions for 2027 and beyond include general-purpose molecular foundation models, multimodal integration of structure, omics, and clinical data, autonomous AI-robotics laboratories, and privacy-preserving federated learning. Standardized reporting guidelines, reproducibility requirements, and harmonized evaluation protocols are essential to realizing AI's full translational potential in drug discovery.

Declarations

- Funding: The authors did not receive support from any organization for the submitted work.

References

[1] DiMasi, J.A., Grabowski, H.G., Hansen, R.W.: Innovation in the pharmaceutical industry: New estimates of r&d costs. Journal of Health Economics. **47**, 20–33 (2016)

[2] Wouters, O.J., McKee, M., Luyten, J.: Estimated research and development investment needed to bring a new medicine to market, 2009-2018. JAMA, **323**(9), 844–853 (2020)

[3] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., Zhao, S.: Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery. **18**(6), 463–477 (2019)

[4] Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., Tekade, R.K.: Artificial intelligence in drug discovery and development. Drug Discovery Today, **26**(1), 80–93 (2021)

[5] Zhang, Y., Mastouri, M., Zhang, Y.: Accelerating drug discovery, development, and clinical trials by artificial intelligence. Med, **5**(9), 1050–1070 (2024) https://doi.org/10.1016/j.medj.2024.08.007

[6] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Hassabis, D.: Highly accurate protein structure prediction with alphafold. Nature Machine Intelligence, **596**(7873), 583–589 (2021) https://doi.org/10.1038/s41586-021-03819-2

[7] Hughes, J.P., Rees, S., Kalindjian, S.B., Philpott, K.L.: Principles of early drug discovery. British Journal of Pharmacology, **162**(6), 1239–1249 (2011)

[8] Mohs, R.C., Greig, N.H.: Drug discovery and development: Role of basic biological research. Alzheimer's & Dementia: Translational Research & Clinical Interventions **3**(4), 651–657 (2017)

[9] Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R., Consonni, V., Kuzmin, V.E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., Tropsha, A.: Qsar modeling: Where have you been? where are you going to? Journal of Medicinal Chemistry **57**(12), 4977–5010 (2014)

[10] Lavecchia, A.: Deep learning in drug discovery: Opportunities, challenges and future prospects. Drug Discovery Today, **24**(10), 2014–2022 (2019) https://doi.org/10.1016/j.drudis.2019.07.006

[11] Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., Terentiev, V.A., Polykovskiy, D.A., Kuznetsov, M.D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R.R., Zhebrak, A., Minaeva, L.I., Zagribelnyy, B.A., Lee, L.H., Soll, R., Madge, D., Aspuru-Guzik, A.: Deep learning enables rapid identification of potent ddr1 kinase inhibitors. Nature Biotechnology, **37**(9), 1038–1040 (2019)

[12] Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., Tran, V.M., Chiappino-Pepe, A., Badran, A.H., Andrews, I.W., Chory, E.J., Church, G.M., Brown, E.D., Jaakkola, T.S., Barzilay, R., Collins, J.J.: A deep learning approach to antibiotic discovery. Cell, **180**(4), 688–702 (2020) https://doi.org/10.1016/j.cell.2020.01.021

[13] Harrer, S., Shah, P., Antony, B., Hu, J.: Artificial intelligence for clinical trial design. Trends in Pharmacological Sciences, **40**(8), 577–591 (2023) https://doi.org/10.1016/j.cmpb.2023.107731

[14] Bilal, H., Khan, M.N., Khan, S., Shafiq, M., Fang, W., Khan, R.U., Rahman, M.U., Li, X., Lv, Q.L., Xu, B.: The role of artificial intelligence and machine learning in predicting and combating antimicrobial resistance. Computational and Structural Biotechnology Journal **27** (2025) https://doi.org/10.1016/j.csbj.2025.01.006

[15] Javid, S., Rahmanulla, A., Ahmed, M.G., sultana, R., Prashantha, K.B.R.: Machine learning & deep learning tools in pharmaceutical sciences: A comprehensive review. Intelligent Pharmacy, **3**(3) (2025) https://doi.org/10.1016/j.ipha.2024.11.003

[16] Zhang, Y., Luo, M., Wu, P., Wu, S., Lee, T.Y., Bai, C.: Application of computational biology and artificial intelligence in drug design. International Journal of Molecular Sciences, **23**(21) (2022) https://doi.org/10.3390/ijms232113568

[17] Iwata, H., Nakai, T., Koyama, T., Matsumoto, S., Kojima, R., Okuno, Y.: Vgae-mcts: A new molecular generative model combining the variational graph auto-encoder and monte carlo tree search. Journal of Chemical Information and Modeling, **63**(23) (2023) https://doi.org/10.1021/acs.jcim.3c01220

[18] Guo, X., Ran, Z., Li, F.: Kinase-inhibitor binding affinity prediction with pretrained graph encoder and language model. Briefings in Bioinformatics, **26**(4) (2025) https://doi.org/10.1093/bib/bbaf338

[19] Haneczok, J., Delijewski, M., Moldzio, R.: Ai molecular property prediction for parkinson's disease reveals potential repurposing drug candidates based on the increase of the expression of pink1. Computer Methods and Programs in Biomedicine **241** (2023) https://doi.org/10.1016/j.cmpb.2023.107731

[20] Martinelli, D.D.: Generative machine learning for de novo drug discovery: A systematic review. Computers in Biology and Medicine, **145** (2022) https://doi.org/10.1016/j.compbiomed.2022.105403

[21] Shakeri, E., Far, B.: Artificial intelligence in healthcare and drug discovery: A scoping review of current trends, challenges, and future directions. Intelligence-Based Medicine, **15** (2026) https://doi.org/10.1016/j.ibmed.2026.100442

[22] Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A.: Cochrane Handbook for Systematic Reviews of Interventions (Version 6.4). The Cochrane Collaboration, ??? (2023)

[23] Jiménez-Luna, J., Grisoni, F., Schneider, G.: rug discovery with explainable artificial intelligence. Nature Machine Intelligence, **2**(10), 573–584 (2020) https://doi.org/10.1038/s42256-020-00236-4

[24] Bhat, A.R., Ahmed, S.: Artificial intelligence (ai) in drug design and discovery: A comprehensive review. n Silico Research in Biomedicine. **1**(100049) (2025)

[25] C., W., Xuesong, L., Sanyin, Z., Shilin, C.: Artificial intelligence for drug discovery: Resources, methods, and applications. Molecular Therapy—Nucleic Acids. (2023) https://doi.org/10.1016/j.omtn.2023.02.019

[26] Chakraborty, C., Bhattacharya, M., Lee, S.S., Wen, Z.H., Lo, Y.H.: The changing scenario of drug discovery using ai to deep learning: Recent advancement, success stories, collaborations, and challenges. Molecular Therapy - Nucleic Acids **35**(3) (2024) https://doi.org/10.1016/j.omtn.2024.102295

[27] Howell, H.J., McGale, J.P., Choucair, A., Shirini, D., Aide, N., Postow, M.A., Wang, L., Tordjman, M., Lopci, E., Lecler, A., Champiat, S., Chen, D.L., Deandreis, D., Dercle, L.: Artificial intelligence for drug discovery: An update and future prospects. Seminars in Nuclear Medicine, **55**(3), 406–422 (2025) https://doi.org/10.1053/j.semnuclmed.2025.01.004

[28] Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., Jensen, K.F.: Generative models for molecular discovery: Recent advances and challenges. Wiley Interdisciplinary Reviews: Computational Molecular Science **12**(5), 321–354 (2022)

[29] Pushpakom, S., Iorio, F., Eyers, P.A., Escott, K.J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., Norris, A., Sanseau, P., Cavalla, D., Pirmohamed, M.: Drug repurposing: Progress, challenges and recommendations. Nature Reviews Drug Discovery, **18**(1), 41–58 (2021)

[30] Trapotsi, M.A., Hosseini-Gerami, L., Bender, A.: Computational models of target identification and drug repurposing. Nature Chemical Biology, **18**(4), 338–348 (2022)

[31] Sterne, J.A.C., Hernán, M.A., Reeves, B.C., Savović, J., Berkman, N.D., Viswanathan, M., Henry, D., Altman, D.G., Ansari, M.T., Boutron, I., Carpenter, J.R., Chan, A.W., Churchill, R., Deeks, J.J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y.K., Pigott, T.D., Higgins, J.P.T.: Robins-i: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ, **355**(I4919) (2016)

[32] Chithrananda, S., Grand, G., Ramsundar, B.: Chemberta: Large-scale self-supervised pretraining for molecular property prediction. arXiv Preprint arXiv:2010.09885. (2020)

[33] Fabian, B., Edlich, T., Gaspar, H., Segler, M., Meyers, J., Fiscato, M., Ahmed, M.: Molecular representation learning with language models and domain-relevant auxiliary tasks. arXiv Preprint arXiv:2011.13230. (2020)

[34] Bagal, V., Aggarwal, R., *et al.*: Molgpt: Molecular generation using a transformer-decoder model. Journal of Chemical Information and Modeling **62**(9), 2064–2076 (2021)

[35] Schneuing, A., Du, Y., Harris, C., Jamasb, A., Igashov, I., Du, W., Blundell, T., Lió, P., Gomes, C., Welling, M., Bronstein, M., Correia, B.: Structure-based drug design with equivariant diffusion models. arXiv Preprint arXiv:2210.13695. (2022)

[36] Corso, G., Stärk, H., Jing, B., Barzilay, R., Jaakkola, T.: Diffdock: Diffusion steps, twists, and turns for molecular docking. International Conference on Learning Representations (ICLR 2023). (2023)

[37] Malhotra, Y., John, J., Yadav, D., Sharma, D., Vanshika, R.K., Mishra, V., Chaturvedi, N.: Advancements in protein structure prediction: A comparative overview of alphafold and its derivatives. Computers in Biology and Medicine, **188** (2025) https://doi.org/10.1016/j.compbiomed.2025.109842

[38] Seyhan, A.A., Carini, C.: Artificial intelligence in drug discovery and development: Current landscape, challenges, and future perspectives. Critical Reviews in Oncology/Hematology, **226**(105461) (2026) https://doi.org/10.1016/j.critrevonc.2026.105461

[39] Jarallah, S.J., Almughem, F.A., Alhumaid, N.K., Fayez, N.A., Alradwan, I., Alsulami, K.A., Tawfik, E.A., Alshehri, A.A.: Artificial intelligence revolution in drug discovery: A paradigm shift in pharmaceutical innovation. International Journal of Pharmaceutics, **680**(125789) (2025) https://doi.org/10.1016/j.ijpharm.2025.125789

[40] Swee, Q.Y.: Artificial intelligence-driven drug discovery: A new era in therapeutic innovation. Bioorganic & Medicinal Chemistry, **139**(118671) (2026) https://doi.org/10.1016/j.bmc.2026.118671

[41] 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 Discovery Today **26**, 511–524 (2021) https://doi.org/10.1016/j.drudis.2020.09.014

[42] Jusoh, A.S., Remli, M.A., Mohamad, M.S., Cazenave, T., Fong, C.S.: How generative artificial intelligence can transform drug discovery? European Journal of Medicinal Chemistry, **295**(117825), 583–589 (2025) https://doi.org/10.1016/j.ejmech.2025.117825

[43] Seo, S., Kim, W.: Pharmaconet: Deep learning-guided pharmacophore modeling for ultra-large-scale virtual screening. Chemical science, **15**(46), 19473–19487 (2024) https://doi.org/10.1039/d4sc04854g

[44] Samudrala, M.V., Dandibhotla, S., Kaneriya, A., Dakshanamurthy, S.: Plaig: Protein-ligand binding affinity prediction using a novel interaction-based graph neural network framework. ACS Bio & Med Chem Au, **5**(3), 447–463 (2025) https://doi.org/10.1021/acsbiomedchemau.5c00053

[45] Cummings, J., Zhou, Y., Lee, G., Zhong, K., Fonseca, J., Cheng, F.: Alzheimer's disease drug development pipeline: 2024. Alzheimer's & Dementia: Translational Research & Clinical Interventions **10**(2) (2024)

[46] Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R.K., Kumar, P.: Artificial intelligence to deep learning: Machine intelligence approach for drug discovery. Molecular Diversity **25**(3), 1315–1360 (2021) https://doi.org/10.1007/s11030-021-10217-3