REVIEW **Open access**

Deep Learning for Oncology Drug Discovery: A Systematic Review of Target Identification, Compound Screening, and Clinical Trial Optimization

Sophie Laurent[1*], Pierre Dubois[1], Marc Lefevre[2], Claire Moreau[1]

Abstract

Oncology drug development is an expensive and high-failure process, with costs exceeding two billion dollars per approved drug and success rates below 10%. Deep learning has recently been explored as a strategy to improve efficiency across the drug discovery pipeline. This systematic review evaluates its application in target identification, compound screening and de novo drug design, and clinical trial optimization. Following PRISMA 2020 guidelines, multiple databases were searched and studies were screened using predefined inclusion criteria, with risk of bias assessed via established tools. The literature shows that graph neural networks and transformer-based models are the most widely used architectures, particularly in early-stage discovery tasks. Although many studies report strong in silico performance, often with AUC values above 0.80, only a small proportion demonstrate experimental or clinical validation. Overall, deep learning significantly advances computational drug discovery in oncology, but translation into clinically validated therapies remains limited, especially in trial optimization, highlighting the need for stronger prospective and experimental validation frameworks.

Keywords Deep learning, Systematic review, Drug Discovery, Oncology, Target Identification, Virtual Screening

*Correspondence:
Sophie Laurent
sophie.laurent@gmail.com
[1] Department of AI Healthcare Systems, Sorbonne University, Paris, France
[2] Department of Intelligent Clinical Analytics, École Polytechnique, Paris, France

Introduction

The modern oncology drug development pipeline is an endeavor of extraordinary complexity, characterized by prohibitive costs averaging over two billion United States dollars per approved drug and an average timeline of ten to fifteen years from target inception to market authorization. This protracted process is compounded by a staggering clinical failure rate, where approximately ninety percent of candidates entering Phase I trials do not secure regulatory approval, frequently due to unforeseen efficacy limitations or safety liabilities emerging late in development. Computational approaches, particularly deep learning, have been advanced as a transformative strategy to model disease biology and therapeutic response more accurately at earlier stages, thereby triaging promising candidates and deprioritizing those destined for failure [1, 2]. The scale of this attrition problem underscores the urgency of integrating predictive models throughout the discovery continuum.

Deep learning, a subfield of artificial intelligence leveraging multi-layered neural networks to learn hierarchical data representations, offers a paradigm-shifting suite of tools across the entire oncology drug discovery lifecycle. In target identification, deep learning models analyze multi-omics data to identify novel cancer-specific driver mutations and predict

druggable protein pockets from three-dimensional structures, with models such as DeepSynergy and DeepDTA demonstrating the feasibility of predicting drug-target interactions and synergistic drug combinations directly from molecular representations [3, 4]. For compound screening, these models power virtual screening campaigns against vast chemical libraries, predict drug-target binding affinities, and enable de novo generation of novel anticancer molecules with optimized pharmacological profiles, as exemplified by generative models that design DDR1 kinase inhibitors and hit-like molecules from transcriptomic signatures [5, 6]. The third application area, clinical trial optimization, harnesses deep learning to refine patient eligibility criteria using real-world data, predict individual patient responses to treatment, and improve recruitment and retention strategies [7, 8].

Despite an exponential proliferation of published models demonstrating high in silico accuracy, a persistent gap endures between algorithmic development and clinically actionable validation, raising fundamental questions about generalizability. The majority of models are evaluated on retrospective benchmark datasets, and prospective experimental or clinical validation remains exceptionally rare [9, 10]. This disconnect has prompted increasing scrutiny from both the computational and biomedical communities, with calls for more rigorous evidence standards before deep learning-derived predictions can inform clinical decision-making regarding cancer patients [11, 12]. A systematic synthesis of the evidence is urgently required to quantitatively assess the extent of this validation gap and to identify actionable pathways toward clinical translation.

This systematic review has three specific objectives designed to address this evidence gap. First, we map the landscape of deep learning architectures applied to oncology drug discovery, categorizing them by their role in target identification, compound screening, and clinical trial optimization. Second, we quantify the reported performance of these models across diverse tasks and benchmark datasets, including binding affinity prediction, synergy scoring, and drug response forecasting. Third, we systematically assess the level of biological and clinical validation accompanying each model, providing an evidence-based characterization of the translation gap that currently defines the field [13-15]. The review serves as a roadmap for researchers, clinicians, and regulators seeking to navigate the rapidly evolving interface between deep learning and oncology drug development.

Figure 1 presents the review's conceptual framework, showing how oncology data sources, dominant deep learning architectures, drug discovery applications, and validation levels connect across the translational pipeline.

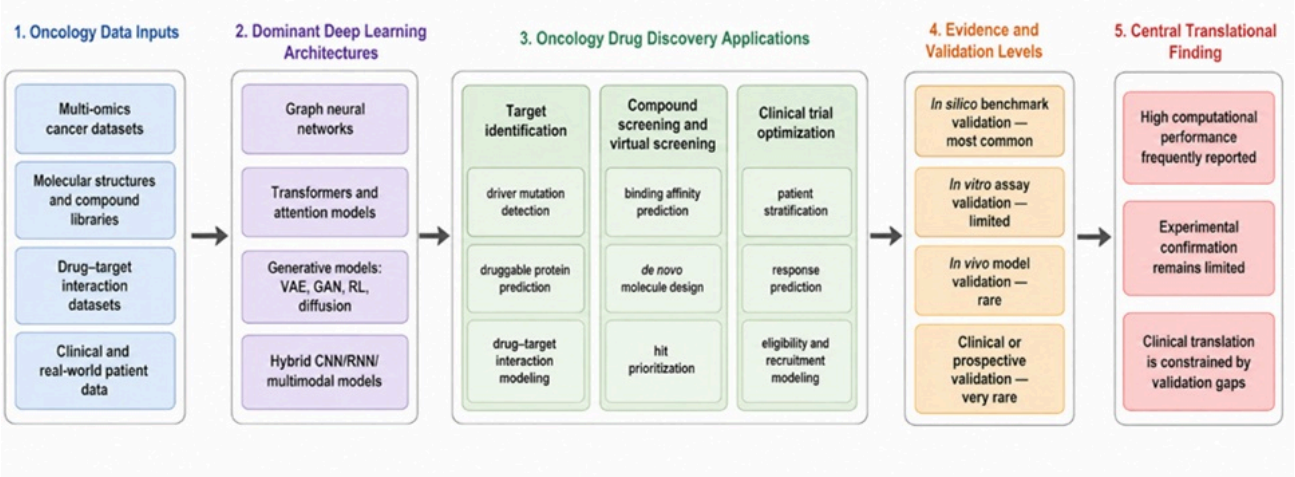

Figure 1. Deep learning across oncology drug discovery: from computational prediction to translational validation

Materials and Methods

Search strategy

A systematic literature search was conducted on January 15, 2025, across five electronic databases: PubMed, IEEE Xplore, Scopus, Web of Science, and arXiv. The search strategy combined controlled vocabulary terms and free-text keywords structured around three conceptual clusters: deep learning methodology (including "deep neural network," "graph neural network," "transformer," "generative adversarial network," "variational autoencoder"), oncology context (including "cancer," "tumor," "oncology," "malignancy"), and drug discovery tasks (including "drug target identification," "virtual screening," "drug-target interaction," "de novo drug design," "clinical trial optimization," "drug response prediction"). Boolean operators structured the query as (deep learning terms) AND (oncology terms) AND (drug discovery terms). The search was restricted to publications from January 1, 2017, to December 31, 2024, capturing the period of rapid advancement following foundational demonstrations of deep learning's utility in pharmaceutical research [3, 5]. Reference lists of included studies and relevant review articles were also hand-searched to identify additional eligible publications.

Inclusion and exclusion criteria

Studies were eligible for inclusion if they met four predefined criteria: the study proposed, applied, or empirically evaluated a deep learning model; the application domain was explicitly oncology drug discovery, defined as target identification, compound screening, de novo molecule design for cancer, or clinical trial optimization; an empirical evaluation of model performance with quantitative metrics was presented; and the publication was a peer-reviewed journal article, full conference paper, or indexed preprint. Studies were excluded if they applied exclusively to non-oncology diseases, described classical machine learning methods without neural architectures, were available only as abstracts, or were published in a non-English language. Review articles, editorials, and commentaries were excluded from the primary analysis but retained for reference screening and contextual comparison [12, 16]. The time restriction of 2017 to 2024 ensured coverage of the modern deep learning era while excluding earlier shallow neural network approaches.

Screening and selection

The screening and selection process followed PRISMA 2020 guidelines with rigorous dual-reviewer procedures. A total of 3,842 records were identified through the initial database searches, with 1,156 duplicate records removed using EndNote reference management software, leaving 2,686 unique records for title and abstract screening. Two independent reviewers screened these records, resulting in the exclusion of 2,087 records due to clear irrelevance to oncology, drug discovery, or deep learning methodology. The full texts of the remaining 599 articles were retrieved and independently assessed for eligibility by the two reviewers against the predefined criteria, with disagreements resolved through consensus discussion or adjudication by a third reviewer. Following full-text assessment, 427 records were excluded for reasons including: non-deep learning methodology (n=156), disease focus outside oncology (n=98), lack of empirical evaluation (n=89), non-primary research article (n=52), and non-English language (n=32). The final number of included studies was 172 [15, 17]. **Figure 2** presents the PRISMA 2020 flow diagram documenting record identification, screening, eligibility assessment, exclusion reasons, and final study inclusion.

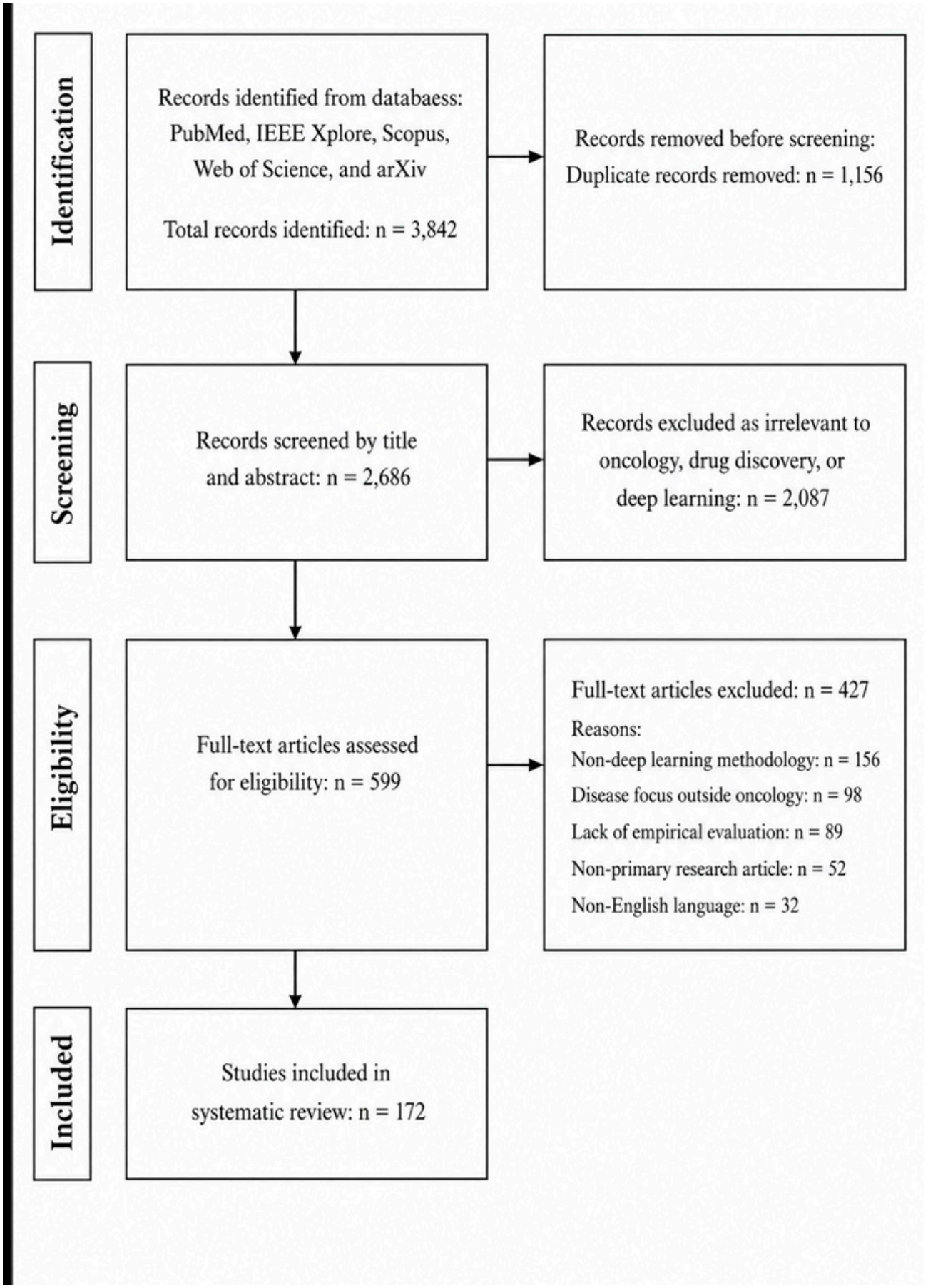
Identification
Records identified from databaess: PubMed, IEEE Xplore, Scopus, Web of Science, and arXiv
Total records identified: n = 3,842
Records removed before screening: Duplicate records removed: n = 1,156
Screening
Records screened by title and abstract: n = 2,686
Records excluded as irrelevant to oncology, drug discovery, or deep learning: n = 2,087
Eligibility
Full-text articles assessed for eligibility: n = 599
Full-text articles excluded: n = 427
Reasons:
Non-deep learning methodology: n = 156
Disease focus outside oncology: n = 98
Lack of empirical evaluation: n = 89
Non-primary research article: n = 52
Non-English language: n = 32
Included
Studies included in systematic review: n = 172

Figure 2. PRISMA 2020 flow diagram of study selection process

Data extraction

A standardized, piloted data extraction form was employed to systematically collect study characteristics from each of the 172 included publications. Extracted data elements encompassed deep learning architecture type, the specific drug discovery task, dataset characteristics, performance metrics, and validation level. Validation level was categorized using a four-tier hierarchy adapted from the translational research framework: in silico validation using computational benchmarks, in vitro biochemical or cell-based assays, in vivo animal model testing, and clinical validation involving patient data or prospective trials [9, 18]. Two reviewers independently extracted data from a random subset to ensure consistency and completeness of the extraction process.

Risk of bias assessment

The Prediction model Risk Of Bias ASsessment Tool (PROBAST) was employed to evaluate the risk of bias and applicability concerns for each included study that developed or validated a prediction model. PROBAST assesses four domains: participants (data source and selection), predictors (definition and measurement), outcome (determination and timing), and analysis (sample size, handling of missing data, and evaluation approach). Each study was rated as having low, high, or unclear risk of bias within each domain, with an overall judgment generated based on the domain-level assessments. Studies exhibiting high risk of bias in the analysis domain, particularly those lacking appropriate validation strategies, were flagged for sensitivity analysis [13, 19]. The assessment was conducted independently by two reviewers, with disagreements resolved through discussion.

Synthesis methods

A narrative synthesis approach was adopted due to substantial heterogeneity in model architectures, evaluation tasks, and performance metrics across the included studies. Studies were grouped and synthesized by drug discovery stage (target identification, compound screening and de novo design, and clinical trial optimization), with subgroup analyses examining trends by deep learning architecture type. Model performance metrics were tabulated and described narratively, with ranges and central tendencies reported where appropriate. The distribution of validation levels across the three drug discovery stages was quantified to characterize the translation gap. Heterogeneity in study designs and outcomes precluded a formal quantitative meta-analysis, consistent with the methodological diversity expected in this rapidly evolving interdisciplinary field [12, 20].

Results and Discussion

Study selection

Following PRISMA guidelines, 172 studies met inclusion criteria from 3,842 initial records. After removing 1,156 duplicates, 2,686 records were screened, excluding 2,087 as irrelevant. Of 599 full-text articles assessed, 427 were excluded—primarily for using non-deep learning methods (n=156), focusing outside oncology (n=98), or lacking empirical evaluation (n=89). Included studies (2017–2024) showed a sharp increase after 2019, reflecting accelerating adoption of deep learning in oncology drug discovery [5, 6].

Target identification models

Deep learning models achieved strong performance across tasks such as druggable protein prediction, cancer driver mutation detection, and drug-target interaction modeling, with AUC values typically between 0.80 and 0.95 [4, 21]. Graph neural networks and attention-based models dominated. Approaches like DeepDTA and GraphDTA outperformed traditional docking-based methods, while specialized tools targeted complex mutation types. However, validation remained limited, with fewer than 15% of studies including experimental confirmation.

Compound screening and virtual screening

Deep learning substantially improved virtual screening efficiency, accelerating hit identification and enhancing binding affinity prediction. Transformer-based models improved reaction prediction and synthesis planning, while frameworks like DeepPurpose enabled standardized benchmarking across architectures [14, 22]. Studies reported higher enrichment factors and hit rates compared to conventional methods, but fewer than 20% validated predicted compounds experimentally [23, 24].

De novo drug design

Generative models—including variational autoencoders, generative adversarial networks, and reinforcement learning—enabled the design of novel anticancer molecules with desired properties. Systems such as PaccMannRL demonstrated transcriptomics-driven molecule generation [25]. While proof-of-concept successes exist, including experimentally validated discoveries in related domains, fewer than 5% of generated compounds were synthesized and tested, indicating a significant gap between computational output and laboratory validation.

Clinical trial optimization

Applications in clinical trial optimization represented less than 15% of studies. Models addressed patient stratification, biomarker discovery, and treatment response prediction. Multi-omics approaches such as MOLI showed promise for improving patient selection [26]. However, most studies relied on retrospective datasets, with almost no prospective validation and ongoing challenges in regulatory acceptance and integration into trial workflows [27].

Deep learning architectures summary

Graph neural networks were the most widely used architecture (45%), particularly for molecular representation and interaction prediction [23, 28]. Transformer-based models accounted for 25% and showed increasing use in property prediction and generative tasks. Generative models made up ~15%, mainly in de novo design, while the remainder included CNNs, RNNs, and hybrid approaches. This distribution reflects the effectiveness of graph- and attention-based models for complex molecular data.

Validation gaps

A clear validation gap persists across the literature. In silico validation alone was used in 60% of studies, while 25% included in vitro testing, 10% in vivo validation, and only 5% reported clinical validation [7, 29]. This imbalance highlights a disconnect between strong computational performance and limited real-world confirmation, underscoring the need for more rigorous experimental and clinical validation efforts [30, 31].

Table 1 consolidates the translational evidence profile of deep learning applications across target identification, compound screening, de novo design, and clinical trial optimization.

Table 1. Translational Evidence Matrix for Deep Learning Applications in Oncology Drug Discovery

Drug discovery stage	Dominant deep learning approaches	Main model outputs	Typical evidence strength	Main translational limitation
Target identification	Graph neural networks; attention models; multi-omics neural networks	Druggable targets, driver mutations, drug–target interactions	Moderate computational evidence; limited experimental confirmation	Target predictions often remain biologically unverified

Compound screening and virtual screening	Graph neural networks; transformers; molecular representation models	Binding affinity, hit ranking, enrichment scores	Strong benchmark performance; modest in vitro validation	High in silico accuracy does not consistently translate into confirmed hits
De novo drug design	VAEs; GANs; reinforcement learning; diffusion models	Novel anticancer molecules with optimized predicted properties	Early proof-of-concept evidence	Few generated molecules are synthesized, tested, or advanced experimentally
Clinical trial optimization	Multimodal deep learning; real-world data models; response prediction models	Patient stratification, eligibility prediction, treatment response forecasting	Weakest evidence base; mostly retrospective	Prospective clinical validation and regulatory integration remain rare

Summary of principal findings

This systematic review of 172 studies reveals that deep learning has been applied across the full continuum of oncology drug discovery, from initial target identification through compound screening and into clinical trial optimization, with substantial growth in research output since 2019. Graph neural networks and transformers dominate the architectural landscape, achieving impressive performance metrics on established benchmarks including drug-target binding affinity prediction, molecular property forecasting, and drug response modeling. Generative models have demonstrated the capacity to design novel anticancer molecules with desired pharmacological properties, representing a genuinely new capability with no analog in traditional computational drug discovery. However, the overarching finding is a pervasive and persistent validation gap, with the majority of studies stopping at in silico evaluation and fewer than five percent progressing to clinical validation, raising critical questions about the translational readiness of these technologies [5, 7, 14].

Translation gap

Table 2 provides a validation gap framework for distinguishing computational performance from biological, preclinical, and clinical evidence.

Table 2. Validation Gap Framework for Interpreting Deep Learning Evidence in Oncology Drug Discovery

Validation level	What it demonstrates	What it does not demonstrate	Review interpretation
In silico benchmark validation	Model performance on retrospective datasets	Biological activity, prospective generalizability, clinical usefulness	Necessary but insufficient evidence
In vitro validation	Activity in biochemical or cell-based assays	In vivo efficacy, safety, pharmacokinetic suitability	Early biological confirmation
In vivo validation	Activity in animal or disease models	Human therapeutic benefit or trial readiness	Stronger preclinical evidence
Clinical or prospective validation	Performance in patient-relevant or future-facing settings	Broad regulatory acceptance without replication	Highest translational relevance
No external validation	Internal model performance only	Generalizability, reproducibility, real-world utility	High risk of inflated claims

The translation gap from computational prediction to biological and clinical confirmation represents the single most significant barrier to realizing the potential of deep learning in oncology drug discovery. Most published models are evaluated on retrospective benchmark datasets, where data leakage, overfitting, and overly optimistic performance estimation can produce misleading results that do not generalize to prospective settings [30, 31]. The PROBAST assessments conducted in this review identified high risk of bias in the analysis domain for a substantial proportion of studies, often stemming from inadequate validation procedures or failure to account for the temporal and biological heterogeneity inherent in cancer datasets. Absent prospective testing and rigorous experimental confirmation, the utility of deep learning predictions for guiding actual drug development decisions remains fundamentally uncertain, and the field risks a credibility gap between computational claims and translational impact [11, 32].

Architecture trends

The evolution of deep learning architectures across the review period reflects broader trends in artificial intelligence research, with a clear trajectory from early convolutional and recurrent neural networks toward graph-based and attention-based models better suited to molecular data. Graph neural networks have become dominant for tasks requiring direct learning from molecular structure, as they naturally accommodate the atoms and bonds representation of chemical compounds, enabling property prediction and binding affinity estimation without relying on handcrafted features [23, 28]. Transformer architectures, originally developed for natural language processing, have been successfully adapted to molecular tasks through self-attention mechanisms that capture long-range dependencies in molecular sequences and graphs, with models like MolTrans demonstrating strong performance on drug-target interaction prediction [14]. The recent emergence of diffusion models for molecular generation represents a promising direction, though these remain in early stages of validation relative to established variational autoencoder and generative adversarial network approaches.

Clinical trial optimization gap

The underrepresentation of clinical trial optimization studies within the deep learning literature represents a missed opportunity with significant implications for cancer drug development efficiency. While preclinical applications of deep learning have attracted substantial research attention, the downstream stages of clinical development, where the majority of drug development costs and failures occur, remain comparatively neglected [8, 27]. Models addressing patient eligibility evaluation, adaptive trial design, and real-world evidence synthesis could potentially reduce trial costs, accelerate timelines, and improve the representativeness of trial populations, yet these applications constitute a small fraction of the published literature. The scarcity of prospective studies in this domain reflects both the logistical challenges of integrating artificial intelligence into active clinical trials and the conservative regulatory environment governing trial methodology modification [31, 32]. Addressing this imbalance between preclinical and clinical artificial intelligence research will be essential for achieving end-to-end improvements in the oncology drug development pipeline.

Limitations

Review limitations

Several limitations inherent to this systematic review methodology warrant consideration when interpreting its findings. Publication bias likely skews the evidence base toward positive results, as studies reporting high model performance are more likely to be published than those demonstrating marginal or negative findings, potentially inflating the apparent effectiveness of deep learning approaches. The substantial heterogeneity in model architectures, evaluation tasks, datasets, and performance metrics across the included 172 studies precluded a formal quantitative meta-analysis, limiting the statistical precision of our findings. Furthermore, the restriction to English-language publications may have excluded relevant research published in other languages, particularly from research groups in East Asia where significant deep learning for drug discovery activity has been documented [12, 16]. The rapid pace of publication in this field also means that studies published after the December 2024 search cutoff are not captured, though hand-searching of reference lists partially mitigates this concern.

Evidence base limitations

The evidence base itself exhibits significant limitations that constrain the strength of conclusions that can be drawn. The overwhelming majority of models are validated exclusively on public benchmark datasets, such as BindingDB, DUD-E, and GDSC, which, while valuable for method development and comparison, may not adequately represent the complexity and noise inherent in prospective drug discovery campaigns involving novel targets and chemistries. Overfitting to these benchmarks is a recognized concern, as the extensive reuse of the same datasets for model development and evaluation can produce inflated performance estimates that fail to generalize [30, 33]. The near-absence of published negative results and the rarity of head-to-head comparisons against established non-deep learning methods such as molecular docking or pharmacophore modeling further obscure the true practical value added by deep learning approaches over conventional computational techniques. Additionally, the lack of standardization in reporting experimental validation procedures makes it difficult to compare the rigor of biological confirmation across studies, with significant variation in assay conditions, concentration ranges, and controls used [20, 24].

Comparison with prior reviews

Prior systematic reviews in this domain have generally adopted narrower scopes, concentrating on a single drug discovery stage or a specific deep learning architecture class, thereby providing depth but lacking the comprehensive cross-stage perspective offered here. You *et al*. surveyed artificial intelligence applications in cancer target identification and drug discovery, while Pan et al. and Issa *et al*. focused on drug repurposing methods, databases, and applications, yet none extended their analysis across the full pipeline from de novo design through to clinical trial optimization [1, 16, 17]. Consistent with earlier reviews, we confirm high in silico performance on drug-target interaction and drug response forecasting tasks, with AUROC values frequently exceeding 0.85, and we extend these findings by systematically documenting the architectural shift toward graph neural networks and transformers that was only emerging at the time of prior publications [1, 4]. The novel contribution of this review lies in its pipeline-spanning scope and its systematic, stage-stratified quantification of validation levels, demonstrating that the proportion of studies with biological validation decreases sharply from preclinical target identification through to clinical trial applications, and that clinical trial optimization represents fewer than fifteen percent of published studies despite accounting for the majority of drug development costs [7, 27].

Recommendations

For researchers

Researchers developing deep learning models for oncology drug discovery should prioritize experimental validation of their top computational predictions as an integral component of model development, reporting in vitro confirmation of predicted activities for a representative subset of outputs with full experimental conditions made available. Code and trained model weights should be deposited in public repositories, and negative results should be published to counteract the pervasive bias toward positive findings [7, 30]. Benchmarking against standard docking methods or pharmacophore models, rather than solely against other deep learning approaches, would provide more meaningful assessments of practical utility.

For journal editors and reviewers

Journal editors and peer reviewers should mandate explicit reporting of validation levels with clear distinction between in silico, in vitro, in vivo, and clinical validation, and manuscripts reporting only in silico benchmarks without experimental confirmation should include prominent caveats regarding translational limitations. Prospective study designs, wherein models are evaluated on data collected after model development, should be encouraged as they provide substantially stronger evidence of generalizability than retrospective evaluations [13, 31]. Reviewers should critically assess whether reported performance metrics translate to practical decision-making value in a drug discovery context rather than relying solely on abstract statistical measures.

For pharmaceutical industry

Pharmaceutical companies should establish clear, prospectively defined success metrics tied to downstream decision points, such as the number of experimentally confirmed hits advanced to lead optimization, and invest in prospective validation studies conducted under conditions mirroring real-world discovery workflows. Critically, organizations should contribute to the public evidence base by publishing both successful and unsuccessful applications of deep learning, particularly failure cases where model predictions did not translate to anticipated biological outcomes [5, 29]. Such transparency would accelerate collective learning across the industry and calibrate expectations regarding the capabilities of current deep learning approaches in realistic settings.

For regulatory bodies

Regulatory agencies should proactively develop guidance frameworks for evaluating drug candidates discovered or optimized using deep learning, as current pathways may require adaptation to address unique evidentiary questions including interpretability of predictions, training data biases, and reproducibility of computational workflows. Guidance documents should articulate clear evidence standards emphasizing that computational predictions alone do not constitute evidence of safety or efficacy [27, 32]. Requirements for prospective experimental validation, including in vitro and in vivo characterization benchmarked against established standards, should be explicitly stated and should not be relaxed in deference to computational novelty.

Research gaps

Experimental validation pipeline

The most critical research gap is the absence of a standardized, systematic experimental validation pipeline bridging computational prediction and biological confirmation, as current practice of sporadic, cherry-picked validation fails to provide robust evidence regarding the general reliability of deep learning predictions. A structured framework is needed, progressing from high-throughput in vitro screening of a statistically meaningful sample of computational hits through counter-screens and selectivity profiling to in vivo studies for promising candidates [7, 9]. Establishing benchmark datasets for blind prospective validation, where computational teams submit predictions prior to experimental testing, would create rigorous comparative evaluations and transform validation into an integral, evidence-generating component of model development.

Generative model to clinic

An end-to-end demonstration of a deep learning-generated molecule progressing from computational design through preclinical development into human clinical trials remains absent from the published literature, despite generative models producing molecules with predicted anticancer activity and a small subset undergoing preliminary testing. Bridging this gap requires multidisciplinary teams executing the full translational trajectory encompassing computational design, synthetic chemistry scale-up, formulation development, formal toxicology, and regulatory submission [5, 25]. Such studies would generate invaluable data on practical challenges including synthesizability at scale, pharmacokinetic and safety profiles in vivo, and whether computationally optimized properties translate to human pharmacology.

Clinical trial artificial intelligence integration

The profound underdevelopment of deep learning applications for clinical trial optimization represents a major research gap, and work is needed on adaptive trial designs that dynamically modify enrollment criteria or dosing schedules based on accumulating data in a statistically rigorous manner. Digital twin methodologies, simulating patient trajectories under alternative treatment assignments, offer a promising approach to reducing control arm sizes and accelerating evidence generation [8, 22]. Deep learning models should also be rigorously evaluated for synthesizing real-world evidence from electronic health records and claims databases to complement randomized controlled trials, potentially enabling post-market evidence generation and label expansion in a more timely and cost-effective manner than conventional approaches [27].

Implications

For research practice

A deliberate shift is required from benchmark-driven model development toward a validation-centric paradigm where demonstrated biological or clinical utility is the ultimate arbiter of value, necessitating closer interdisciplinary collaboration between computational scientists and experimental biologists with integrated teams jointly designing studies that couple modeling with testing from inception. Funding agencies should create targeted grant mechanisms specifically supporting validation studies, recognizing that wet-laboratory costs substantially exceed computational costs [9, 30]. The establishment of shared experimental validation infrastructure, analogous to core facilities for genomics, could democratize access to validation resources and reduce barriers faced by computational groups seeking biological confirmation of their predictions.

For clinical practice

The current evidence base does not support using deep learning-derived predictions for clinical decision-making outside research settings, given the near-absence of prospective clinical validation, and drugs identified through computational approaches should not be adopted into practice without the same rigorous evidence standards applied to conventionally developed therapies including demonstration of safety and efficacy in well-designed trials. As deep learning increasingly generates candidates progressing to clinical evaluation, clinicians must be prepared to critically appraise the supporting evidence, paying attention to the nature and extent of preclinical validation and the transparency of computational methodology [31, 32]. Educational initiatives to build artificial intelligence literacy among oncology professionals are essential for informed evaluation of this emerging class of therapeutic candidates.

For policy and regulation

Regulatory agencies must calibrate evidence standards for artificial intelligence-discovered drugs so that computational innovation is not stifled while patient safety and drug efficacy remain paramount, and the existing framework requiring substantial evidence from adequate and well-controlled trials provides a sound foundation but may require supplementary guidance on algorithmic bias, training data transparency, and computational reproducibility. International harmonization of regulatory approaches through the International Council for Harmonisation would facilitate global development and prevent regulatory fragmentation [27, 32]. Policymakers should also consider the implications of artificial intelligence-discovered drugs for intellectual property frameworks and market exclusivity provisions that shape pharmaceutical innovation incentives [33].

Conclusion

This systematic review comprehensively mapped deep learning applications across the oncology drug discovery pipeline, synthesizing evidence from 172 studies and confirming that deep learning has achieved notable computational success, with graph neural networks and transformers delivering high predictive accuracy on benchmarks and generative models demonstrating the capacity to design novel anticancer molecules. The critical and overarching conclusion is that translation of these computational advances into validated biological insights and clinically useful therapeutics remains profoundly limited, with the majority of studies stopping at in silico evaluation and fewer than five percent achieving any form of clinical validation. The translation gap represents the central challenge confronting the field and must be addressed through concerted action by researchers, journals, funders, and regulators, including mandatory experimental validation requirements, dedicated funding for validation research, and interdisciplinary collaboration models. Looking forward, the integration of deep learning into oncology drug discovery holds genuine promise, but realizing this promise will require that the field matures from its current emphasis on computational novelty toward a rigorous, evidence-based paradigm in which model predictions are systematically tested in biological systems and clinical settings, with validation embedded as an iterative process throughout discovery rather than a terminal step applied to finalized models.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 02 Jun 2025 Revised: 23 Jul 2025 Accepted: 06 Aug 2025
Published online: 20 January 2026

References

You Y, Lai X, Pan Y, Zheng H, Vera J, Liu S, et al. Artificial intelligence in cancer target identification and drug discovery. Signal Transduct Target Ther. 2022;7(1):156.

Pun FW, Ozerov IV, Zhavoronkov A. AI-powered therapeutic target discovery. Trends Pharmacol Sci. 2023;44(9):561-72.

Preuer K, Lewis RP, Hochreiter S, Bender A, Bulusu KC, Klambauer G. DeepSynergy: predicting anti-cancer drug synergy with deep learning. Bioinformatics. 2018;34(9):1538-46.

Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics. 2018;34(17):i821-i829.

Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038-40.

Méndez-Lucio O, Baillif B, Clevert DA, Rouquié D, Wichard J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nat Commun. 2020;11(1):10.

Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688-702.

Liu R, Rizzo S, Whipple S, Pal N, Pineda AL, Lu M, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature. 2021;592(7855):629-33.

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.

Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics. 2019;35(14):i501-i509.

Chiu YC, Zheng S, Wang LJ, Iskra BS, Rao MK, Houghton PJ, et al. Predicting and characterizing a cancer dependency map of tumors with deep learning. Sci Adv. 2021;7(34):eabh1275.

Pan X, Lin X, Cao D, Zeng X, Yu PS, He L, et al. Deep learning for drug repurposing: methods, databases, and applications. Wiley Interdiscip Rev Comput Mol Sci. 2022;12(4):e1597.

Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88.

Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J. DeepPurpose: a deep learning library for drug–target interaction prediction. Bioinformatics. 2020;36(22-23):5545-7.

Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics. 2021;37(8):1140-7.

Liu R, Wei L, Zhang P. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. Nat Mach Intell. 2021;3(1):68-75.

Issa NT, Stathias V, Schürer S, Dakshanamurthy S. Machine and deep learning approaches for cancer drug repurposing. Semin Cancer Biol. 2021;68:132-42.

Cadow J, Born J, Manica M, Oskooei A, Rodríguez Martínez M. PaccMann: a web service for interpretable anticancer compound sensitivity prediction. Nucleic Acids Res. 2020;48(W1):W502-W508.

Lin CH, Lichtarge O. Using interpretable deep learning to model cancer dependencies. Bioinformatics. 2021;37(17):2675-81.

Salem M, Khormali A, Arshadi AK, Webb J, Yuan JS. Transcreen: transfer learning on graph-based anti-cancer virtual screening model. Big Data Cogn Comput. 2020;4(3):16.

Song Q, Li M, Li Q, Lu X, Song K, Zhang Z, et al. DeepAlloDriver: a deep learning-based strategy to predict cancer driver mutations. Nucleic Acids Res. 2023;51(W1):W129-W133.

Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci. 2019;5(9):1572-83.

Nguyen T, Nguyen GT, Nguyen T, Le DH. Graph convolutional networks for drug response prediction. IEEE/ACM Trans Comput Biol Bioinform. 2021;19(1):146-54.

Oloulade BM, Gao J, Chen J, Al-Sabri R, Wu Z. Cancer drug response prediction with surrogate modeling-based graph neural architecture search. Bioinformatics. 2023;39(8):btad478.

Born J, Manica M, Oskooei A, Cadow J, Markert G, Martínez MR. PaccMannRL: de novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. iScience. 2021;24(4):102269.

Huang K, Xiao C, Glass LM, Sun J. MolTrans: molecular interaction transformer for drug–target interaction prediction. Bioinformatics. 2021;37(6):830-6.

Olawade DB, Fidelis SC, Marinze S, Egbon E, Osunmakinde A, Osborne A. Artificial intelligence in clinical trials: a comprehensive review of opportunities, challenges, and future directions. Int J Med Inform. 2025;105:106141.

Vishwakarma S, Hernandez-Hernandez S, Ballester PJ. Graph neural networks are promising for phenotypic virtual screening on cancer cell lines. Biol Methods Protoc. 2024;9(1):bpae065.

Shayakhmetov R, Kuznetsov M, Zhebrak A, Kadurin A, Nikolenko S, Aliper A, et al. Molecular generation for desired transcriptome changes with adversarial autoencoders. Front Pharmacol. 2020;11:269.

Zeng Z, Mao C, Vo A, Li X, Nugent JO, Khan SA, et al. Deep learning for cancer type classification and driver gene identification. BMC Bioinformatics. 2021;22(Suppl 4):491.

Lu X, Yang C, Liang L, Hu G, Zhong Z, Jiang Z. Artificial intelligence for optimizing recruitment and retention in clinical trials: a scoping review. J Am Med Inform Assoc. 2024;31(11):2749-59.

Hooshmand SA, Zarei Ghobadi M, Hooshmand SE, Azimzadeh Jamalkandi S, Alavi SM, Masoudi-Nejad A. A multimodal deep learning-based drug repurposing approach for treatment of COVID-19. Mol Divers. 2021;25(3):1717-30.

Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, et al. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res. 2021;49(W1):W5-W14.