Foundation Models for Drug Discovery: A Systematic Review of Molecular Language, Representation Learning, Target Prediction, Generative Design, and Translational Risk

Gabriel Costa[1]*, Lucas Ribeiro[1], Ricardo Alves[2]

[1]Department of AI for Anti-Inflammatory Lectins and Glycosides, Faculty of Pharmacy, Federal University of Minas Gerais, Belo Horizonte, Brazil.
[2]Department of Computational Pharmacology for Autoimmune Diseases, Faculty of Pharmacy, University of Coimbra, Coimbra, Portugal.

ABSTRACT

Foundation models are increasingly used in computational drug discovery to learn molecular, protein, graph-based, and multimodal representations, but their translational value remains unevenly supported across tasks and validation settings. This systematic review evaluates how foundation models are applied to molecular language, representation learning, target prediction, compound–target interaction prediction, binding-related modelling, generative molecular design, and translational-risk assessment in drug discovery. PubMed, Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, Wiley Online Library, ACS Publications, Nature Portfolio journals, Oxford Academic journals, and quality-filtered MDPI and Frontiers journals were searched for peer-reviewed journal articles published from 1 January 2017 to 9 July 2026; 580 records were identified, 301 duplicates were removed, 279 records were screened, 105 full-text articles were assessed, 65 full-text articles were excluded with predefined reasons, and 40 studies were included. Evidence was extracted on model category, input modality, pretraining or adaptation strategy, drug-discovery task, benchmark setting, validation type, limitation, and translational risk. The included evidence shows that molecular language models, protein language models, graph-based self-supervised models, and knowledge-guided representation models may improve molecular representation, target-related prediction, compound–target modelling, and generative design support. However, evidence was dominated by benchmark and internal validation, with limited external, prospective, or experimental confirmation across many model categories. Foundation models are reshaping drug-discovery modelling, but benchmark performance should not be equated with chemical validity, biological plausibility, safety, or clinical utility. Stronger dataset provenance, leakage-resistant evaluation, uncertainty reporting, interpretability, prospective testing, and experimental confirmation are required before translational confidence can be justified.

Key Words: *Foundation models, Drug discovery, Molecular language models, Representation learning, Target prediction, Generative design*

eIJPPR 2026; 16(1):35-46

HOW TO CITE THIS ARTICLE: Costa G, Ribeiro L, Alves R. Foundation Models for Drug Discovery: A Systematic Review of Molecular Language, Representation Learning, Target Prediction, Generative Design, and Translational Risk. Int J Pharm Phytopharmacol Res. 2026;16(1):35-46. https://doi.org/10.51847/bWGXsPYdzM

INTRODUCTION

Artificial intelligence has become embedded across drug-discovery workflows, including molecular representation, candidate prioritization, virtual screening support, and automated design logic, but the evidence base has evolved from task-specific modelling toward larger pretrained systems that require more careful interpretation. Earlier work on machine learning and automation in drug

Corresponding author: Gabriel Costa
Address: Department of AI for Anti-Inflammatory Lectins and Glycosides, Faculty of Pharmacy, Federal University of Minas Gerais, Belo Horizonte, Brazil.
E-mail: ✉ gabriel.costa@ufmg.br
Received: 21 October 2025; **Revised:** 21 January 2026; **Accepted:** 27 January 2026

discovery showed that computational prioritization can accelerate parts of discovery, while AI-guided candidate identification still depends on experimental confirmation before biological or therapeutic relevance can be inferred [1-3].

The need for validation is especially clear when computational models are used to select compounds for biological testing rather than merely to rank benchmark examples. A deep learning approach to antibiotic discovery demonstrated how computational screening can identify experimentally testable candidates, but the value of the prediction depended on downstream biological confirmation rather than model output alone [4].

Foundation models now extend this trajectory by learning reusable representations from large molecular, protein, graph, text, or multimodal datasets. In drug discovery, this shift matters because the same pretrained model may be adapted to molecular property prediction, protein representation, target prediction, compound–target interaction prediction, binding affinity estimation, generative molecular design, or safety-related filtering. These capabilities have generated substantial interest, but they also create new risks when benchmark performance is interpreted as evidence of translational readiness.

This systematic review was therefore designed to synthesize peer-reviewed evidence on foundation models for drug discovery, with particular attention to molecular language, representation learning, target-related prediction, generative design, validation practices, and translational-risk boundaries. The objective was not to determine whether foundation models "solve" drug discovery, but to assess what the current evidence supports, where validation remains limited, and which methodological safeguards are needed for responsible translation.

Foundation models in drug discovery

In this review, foundation models are defined as pretrained or large-scale representation-learning systems that can be adapted across multiple downstream drug-discovery tasks rather than trained only for one narrowly specified prediction endpoint. This category includes large language models used in drug discovery, protein language models trained on large protein sequence corpora, and self-supervised biological sequence models that learn transferable representations relevant to downstream protein or target-related tasks [5-7].

Foundation models differ from classical task-specific machine-learning models because their central value lies in reusable representation learning, pretraining, and adaptation. Instead of learning only from a labelled dataset for a single endpoint, a foundation-model workflow may learn from molecular strings, protein sequences, molecular graphs, biomedical text, or multimodal biomedical data and then be fine-tuned, prompted, or otherwise adapted for prediction, prioritization, or design. Therapeutic foundation-model frameworks have therefore been proposed as cross-cutting infrastructure for drug discovery, while also requiring task-specific evidence and governance before practical utility can be inferred [8].

Molecular language models typically encode compounds as chemical strings such as SMILES and apply language-model concepts to molecular representation, property prediction, or generation. Protein language models use amino-acid sequences to learn representations that may support protein function, structure-related inference, or target-related modelling. Graph foundation models and geometry-aware approaches represent molecules as atoms, bonds, and spatial features, while multimodal systems may integrate chemical, biological, textual, assay, or omics-derived information where such data are available.

Fine-tuning, zero-shot, and few-shot settings are important but must be interpreted cautiously. Fine-tuned models may perform well when downstream data resemble pretraining or benchmark distributions, while zero-shot or few-shot claims require especially careful appraisal because apparent generalization can be affected by dataset overlap, task ambiguity, hidden leakage, or benchmark saturation. For this reason, the review treats foundation-model performance as evidence for hypothesis generation or prioritization unless supported by external validation, prospective testing, experimental confirmation, and biologically plausible interpretation.

Review questions and search strategy

This systematic review addressed six questions: how foundation models are applied to molecular language and molecular representation learning; how they are used for target prediction, compound–target interaction prediction, and binding-related tasks; how they support generative molecular design and molecule optimization; what validation strategies are reported; what translational risks are identified; and what research directions are needed for responsible translation. The questions were structured around drug-discovery tasks, input modalities, model categories, validation types, and translational-risk domains.

The database search covered PubMed, Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, Wiley Online Library, ACS Publications, Nature Portfolio journals, Oxford Academic journals, and quality-filtered MDPI and Frontiers journals from 1 January 2017 to 9 July 2026. The search combined terms for foundation models, large language models, molecular language models, protein language models, pretrained models, self-supervised learning, transformers, representation learning,

and generative models with drug-discovery terms including molecular design, target prediction, compound–target interaction, binding affinity, virtual screening, ADMET, toxicity, and molecular property prediction. Benchmark and bioactivity resources were included when they were directly relevant to evaluation, dataset provenance, or extraction logic, including MoleculeNet for molecular machine-learning benchmarks [9]. ChEMBL-related database articles were included because bioactivity data sources are central to model training, benchmarking, target-related prediction, and assay provenance assessment [10]. The updated ChEMBL deposition framework was included to support extraction of provenance-related concerns in bioactivity modelling [11].

Eligible studies were peer-reviewed journal articles published from 2017 to 2026 inclusive, had DOI information, and focused on foundation models, molecular or protein language models, pretrained molecular models, self-supervised representation learning, graph foundation models, multimodal biomedical models, generative AI, target prediction, compound–target interaction, binding affinity prediction, ADMET or toxicity prediction, validation, or translational risk. Articles were excluded if they were conference-only records, preprints only, books, theses, websites, software manuals, GitHub repositories, database webpages, vendor pages, regulatory webpages, policy documents, white papers, outside the search period, missing DOI information, or not aligned with the review questions. Studies focused only on conventional QSAR, docking, molecular dynamics, general clinical language models, or generic AI ethics were excluded unless they had clear foundation-model, representation-learning, drug-discovery, validation, or translational-risk relevance.

The PRISMA-style screening process identified 552 records from databases and 28 records from citation chasing or journal cross-checking, giving 580 total records before deduplication. After removing 301 duplicates, 279 records were screened by title and abstract, of which 174 were excluded. Full texts were assessed for 105 articles, and 65 were excluded for predefined reasons: not a foundation model, pretrained representation model, or relevant large-model method (n=14); not drug-discovery relevant (n=9); conference-only or preprint-only record (n=12); no DOI (n=4); outside 2017–2026 inclusive (n=5); insufficient methodological detail for extraction (n=6); no validation or task-specific evidence (n=5); not a peer-reviewed journal article (n=5); duplicate or substantially overlapping report (n=3); and not aligned with review questions (n=2). The final included-study count was 40.

Figure 1 presents the PRISMA-style flow of record identification, duplicate removal, screening, eligibility assessment, exclusion reasons, and final included studies.

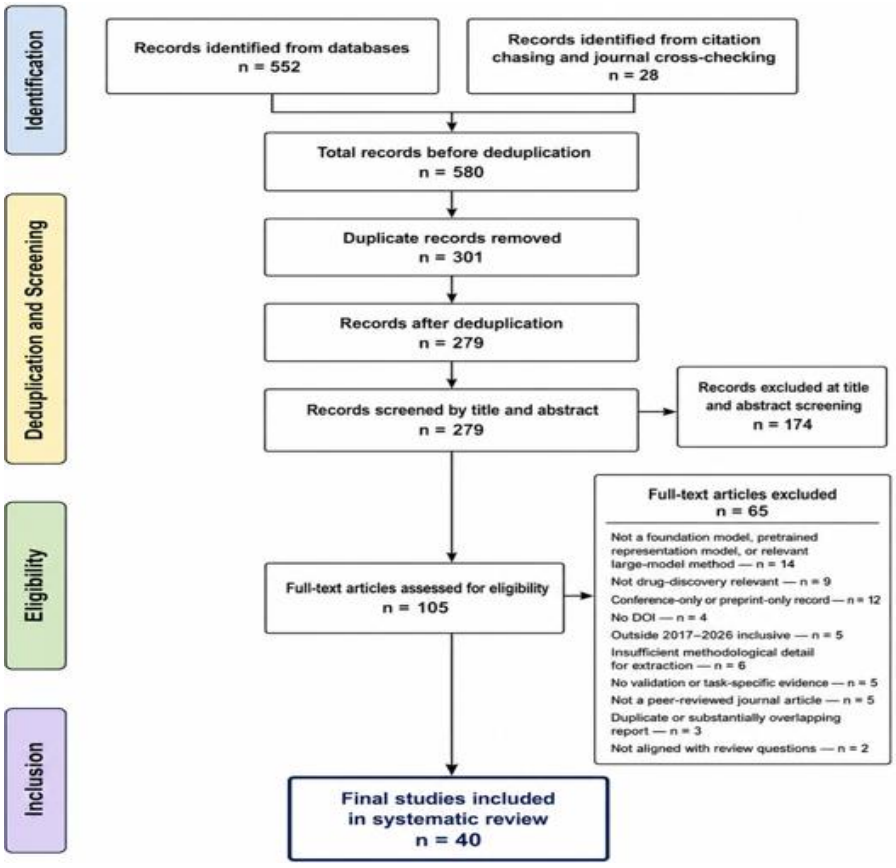

Figure 1. PRISMA-Style Flow Diagram for Study Identification, Screening, Eligibility Assessment, and Inclusion

Alt-text description

A PRISMA-style flow diagram showing 552 database records and 28 additional records, 580 total records before deduplication, 301 duplicates removed, 279 records screened, 174 records excluded, 105 full-text articles assessed, 65 full-text articles excluded with reasons, and 40 final included studies.

Evidence extraction captured foundation model type, molecular or protein input format, pretraining or training data description, fine-tuning or adaptation strategy, downstream task, benchmark or evaluation dataset, reported metric category, validation type, external or prospective validation where reported, experimental validation where reported, main finding, main limitation, and translational risk. Because included studies differed substantially across model architectures, datasets, tasks, evaluation metrics, and validation settings, the synthesis was thematic and narrative rather than meta-analytic; no protocol registration, meta-analysis, inter-rater agreement statistic, or risk-of-bias score is claimed.

Table 1 summarises the review questions, search strategy, eligibility criteria, screening logic, and evidence extraction domains used in this systematic review.

Table 1. Systematic Review Design for Foundation Models in Drug Discovery: Review Questions, Databases, Search Concepts, Eligibility Criteria, Screening Logic, Extraction Domains, Validation Appraisal, and Synthesis Method

Review component	Specification	Rationale	Evidence captured	Risk controlled	Output used in synthesis
Review questions	Six questions covering molecular language, representation learning, target prediction, compound–target interaction, binding-related modelling, generative design, validation, translational risk, and research directions	Ensures the review remains task-specific rather than a generic AI overview	Model category, task type, validation type, limitation, translational-risk issue	Scope drift and unsupported claims	Section-level synthesis and research agenda
Databases searched	PubMed, Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, Wiley Online Library, ACS Publications, Nature Portfolio journals, Oxford Academic journals, and quality-filtered MDPI and Frontiers journals	Captures biomedical, chemical, computational, and pharmacological literature	Peer-reviewed journal records from multiple disciplinary sources	Database selection bias	Search strategy and PRISMA log
Search period	1 January 2017 to 9 July 2026	Covers the modern foundation-model and self-supervised molecular-learning period	Eligible journal articles within fixed date range	Out-of-period evidence	Date-bounded included set
Search concepts	Foundation models, LLMs, molecular language models, protein language models, pretrained models, self-supervised learning, transformers, representation learning, generative models, drug discovery, target prediction, DTI, binding affinity, virtual screening, ADMET, toxicity	Links model families with drug-discovery tasks	Concept-level and task-level evidence	Missing relevant model categories	Thematic synthesis categories
Inclusion criteria	Peer-reviewed journal article, 2017–2026, DOI present, foundation-model or representation-learning relevance, drug-discovery relevance, extractable methodological or reviewable detail	Preserves evidence quality and extractability	Study aim, model type, task, data source, validation, limitation	Non-reviewable or unsupported evidence	Final 40-study evidence base
Exclusion criteria	Non-peer-reviewed records, conference-only or preprint-only records, no DOI, outside period, conventional QSAR without relevant representation learning, clinical LLMs without drug-discovery relevance, docking-only or ethics-only articles without review relevance	Avoids unsupported or off-topic evidence	Exclusion reasons and counts	Inclusion of weak-fit studies	PRISMA-style exclusion log
Screening logic	580 records identified, 301 duplicates removed, 279 screened, 105 full texts assessed, 65 full texts excluded, 40 studies included	Creates transparent eligibility trail	Identification, deduplication, screening, eligibility, inclusion	Invented or opaque study selection	**Figure 1** and systematic-review methods

Extraction domains	Model type, input modality, training or pretraining data, adaptation, task, benchmark, validation, external/prospective/experimental confirmation, finding, limitation, translational risk	Enables cross-study comparison despite heterogeneity	Structured extracted evidence	Random citation placement	Evidence extraction and synthesis tables
Validation appraisal	Dataset provenance, overlap, leakage, benchmark representativeness, split validity, external validation, prospective validation, experimental validation, metric appropriateness, chemical validity, synthetic feasibility, biological plausibility, uncertainty, interpretability, generalizability, reproducibility	Separates model performance from translational confidence	Validation strengths and weaknesses	Benchmark performance overinterpretation	Section 8 and Section 9 synthesis
Synthesis method	Narrative and thematic synthesis, no meta-analysis	Heterogeneous tasks, models, datasets, and metrics prevented quantitative pooling	Thematic patterns across included evidence	False precision from incompatible outcomes	Evidence synthesis and research directions

Molecular language and representation learning

Molecular language modelling in drug discovery began from the premise that chemical structures can be encoded as sequences or substructures and analysed with methods inspired by natural-language processing. Mol2vec provided an early example by learning chemically meaningful substructure embeddings through an unsupervised approach, supporting the idea that molecular fragments can be represented in a latent space with chemical intuition [12].

Transformer-based and pretrained molecular models expanded this logic by learning representations from SMILES strings and adapting them to downstream chemical tasks. Transformer-CNN showed that transformer-derived representations could support QSAR modelling and interpretation, while Chemformer demonstrated that a pretrained transformer can be applied across computational chemistry tasks including prediction and generation [13]. This evidence supports the view that molecular language models may improve transferable chemical representation, although benchmark performance remains task- and dataset-dependent [14].

BERT-style molecular pretraining further illustrates how language-model architectures have been adapted for molecular property prediction and structural representation. Multitask BERT enhanced by SMILES enumeration was reported for molecular property prediction, while a BERT-based SMILES pretraining model was designed to extract molecular structural information from chemical sequences [15]. These studies support the relevance of pretraining and fine-tuning for molecular representation learning, but they do not by themselves establish chemical validity, external generalizability, or experimental relevance outside the evaluated benchmark settings [16].

Graph and knowledge-guided pretraining complement sequence-based molecular language modelling by representing molecular topology, geometry, or prior chemical knowledge. Geometry-enhanced molecular representation learning was reported to improve property prediction, and knowledge-guided pretraining was proposed to strengthen molecular representation by incorporating chemically relevant prior information [17]. These approaches broaden the foundation-model evidence base beyond molecular strings, but their translational interpretation still depends on dataset provenance, benchmark design, external testing, and task-specific validation [18].

Table 2 summarises foundation-model evidence related to molecular language, protein language, and representation learning in drug discovery.

Table 2. Foundation Models for Molecular Language and Representation Learning in Drug Discovery: Model Categories, Input Modalities, Pretraining Strategies, Downstream Tasks, Validation Patterns, Strengths, Limitations, and Translational Concerns

Model or evidence category	Input modality	Pretraining or representation strategy	Drug-discovery task	Reported strength	Validation pattern	Main limitation	Translational concern
Molecular substructure embeddings	Molecular substructures or chemical strings	Unsupervised embedding inspired by	Molecular representation and property-	Captures chemically meaningful	Internal computational validation	Earlier representation approach, not a modern transformer foundation model	Chemical similarity does not guarantee

		natural-language processing	related modelling	substructure relationships			biological relevance
SMILES transformer representations	Molecular strings	Transformer-derived sequence representation	QSAR and molecular property prediction	Supports molecular prediction and interpretability	Benchmark-oriented internal validation	Dataset-dependent performance	Benchmark success may not transfer to new chemistry
Pretrained chemical transformer	Molecular strings	Sequence-to-sequence chemical pretraining	Molecular prediction and generation	Supports multiple computational chemistry tasks	Internal task validation	Limited prospective or experimental validation	Generated or predicted molecules require chemical and biological confirmation
BERT-style molecular model	SMILES strings	BERT-based pretraining and fine-tuning	Molecular property prediction and structural-information extraction	Learns task-adaptable molecular representations	Benchmark validation	External generalization not consistently established	Structural encoding does not equal translational validity
Geometry-enhanced graph representation	Molecular graphs and geometric features	Geometry-aware molecular representation learning	Molecular property prediction	Incorporates spatial or geometric information	Internal benchmark validation	Depends on quality and availability of structural/geometric inputs	Geometry-aware prediction still requires biological validation
Knowledge-guided molecular pretraining	Molecular graphs and chemical knowledge	Pretraining guided by prior chemical knowledge	Molecular representation and property prediction	Integrates prior knowledge with learned representations	Benchmark validation	Knowledge sources may introduce bias or incomplete assumptions	Encoded knowledge must be traceable and biologically relevant
Protein language model evidence	Protein sequences	Self-supervised sequence learning	Protein representation and target-related inference	Learns scalable protein representations	Internal downstream validation	Drug-discovery relevance depends on task adaptation	Protein sequence representation alone does not confirm target mechanism
Cross-modal representation direction	Molecular and biological data where reported	Pretraining or adaptation across multiple data types	Drug-discovery decision support	May connect chemical and biological evidence	Heterogeneous validation	Sparse external and prospective confirmation	Multimodal integration can obscure provenance and uncertainty

Target prediction and generative design

Foundation-model evidence for target-related prediction builds on earlier compound–target interaction and binding-affinity modelling, where drug and protein encoders are combined to estimate interaction probability or affinity. DeepPurpose provided a unified deep-learning framework for drug–target interaction prediction, supporting comparative modelling across drug and protein encoders [19]. DeepDTA showed that sequence-based modelling of SMILES strings and protein sequences could be used for drug–target binding-affinity prediction, establishing an important benchmark-oriented baseline for later language-model and transformer-based approaches [20].

Transformer and pretraining strategies subsequently expanded target-related modelling by representing molecular and protein sequences in more flexible latent spaces. MolTrans used transformer-based interaction modelling for drug–target interaction prediction, supporting the view that learned token-level interaction patterns can assist DTI modelling [21]. GeneralizedDTA combined pretraining and multitask learning for drug–target binding-affinity prediction, particularly in settings framed around unknown-drug discovery, although its evidence remained benchmark-centred rather than prospectively translational [22].

Dual-language and structure-free approaches further illustrate the movement toward molecular and protein

foundation-model encoders in binding-related prediction. DLM-DTI used dual language modelling with hint-based learning for drug–target interaction prediction, showing how molecular and protein language representations may be combined for interaction modelling [23]. A structure-free drug–target affinity framework using protein and molecule language models demonstrated the feasibility of DTA prediction without explicit structural inputs, but the translational meaning of such predictions remains constrained by benchmark composition, domain transfer, interpretability, and biological confirmation [24].

Generative molecular design represents a related but higher-risk use case because model outputs may be syntactically plausible yet chemically impractical, biologically irrelevant, unsafe, or unsupported by experimental evidence. Interpretable bilinear attention with domain adaptation has been used to improve DTI prediction and address aspects of domain transfer, which is important when predictive outputs are used to guide prioritization or design decisions [25]. Pretrained biochemical language models have also been adapted for targeted drug design, supporting hypothesis generation in target-aware design workflows while still requiring chemical-validity, synthetic-feasibility, biological-plausibility, and safety assessment before translational confidence can be inferred [26].

Table 3 maps foundation-model applications in target prediction, compound–target interaction prediction, binding-related tasks, and generative molecular design.

Table 3. Foundation Models for Target Prediction and Generative Molecular Design: Target-Related Tasks, Compound–Target Interaction Prediction, Binding-Related Modeling, Virtual Screening, Molecule Generation, Optimization, Chemical Validity, Safety Relevance, and Validation Boundaries

Application domain	Model function	Data requirement	Reported output type	Evaluation approach	Validation limitation	Risk of overinterpretation	Review synthesis implication
Compound–target interaction prediction	Learns interaction probability or compatibility between compounds and targets	Molecular and protein representations, labelled interaction data	Interaction score or binary prediction	Benchmark comparison using DTI datasets	Often internally validated	Treating predicted interaction as confirmed mechanism	Useful for prioritization, not mechanism confirmation
Binding-affinity prediction	Estimates strength of drug–target binding from sequence or representation inputs	SMILES, protein sequences, affinity labels	Affinity score or regression output	Davis, KIBA, BindingDB-type benchmark evaluation where reported	External and prospective testing often unclear	Treating predicted affinity as experimentally measured binding	Requires independent and experimental confirmation
Transformer-based DTI	Encodes molecular and target tokens with interaction modelling	Tokenized molecular and protein sequences	Interaction probability or ranking	Internal benchmark evaluation	Sensitive to dataset overlap and split design	Assuming transformer attention equals biological explanation	Supports representation learning but not automatic interpretability
Pretraining and multitask DTA	Uses pretrained or multitask representations for affinity prediction	Molecular/protein data and task labels	Binding-related prediction	Benchmark performance comparison	Generalization to unseen targets or chemistry may remain limited	Inferring broad unknown-drug utility from narrow benchmarks	Needs scaffold-, target-, and time-aware validation
Dual molecular/protein language models	Combines molecular and protein language encoders	SMILES or molecular strings plus protein sequences	DTI or DTA prediction	Internal benchmark testing	Domain transfer incompletely established	Assuming sequence representation captures full biological context	Promising for target-related modelling but validation-dependent
Domain-adapted interpretable DTI	Uses attention and domain adaptation to improve transfer	Molecular/protein features and source–target domains	Interaction prediction with interpretable components	Benchmark and domain-adaptation evaluation	Interpretability remains model-dependent	Treating attention weights as definitive mechanistic evidence	Useful for transparency but not sufficient for causal inference

ISSN (Online) 2249-6084 (Print) 2250-1029

Target-aware generative design	Conditions molecule generation on target-related information	Molecular and target representations, training examples	Generated or optimized molecules	Computational design metrics and internal validation	Chemical validity, synthesizability, safety, and biology often undervalidated	Treating generated molecules as drug candidates	Supports hypothesis generation only
Virtual screening support	Ranks candidates using learned representations or target-related scores	Compound libraries, target information, benchmark labels	Ranked compounds or prioritized hits	Enrichment or classification metrics where reported	Benchmark bias and memorization can inflate performance	Treating ranking as experimental activity	Requires bias-aware benchmarks and prospective testing

Translational risk and validation challenges

The strongest cross-cutting limitation in the included evidence was the gap between benchmark performance and translational confidence. Ligand-based classification benchmarks can reward memorization rather than generalization, virtual-screening datasets can contain biases that distort model evaluation, and activity cliffs can expose failures of molecular machine learning in local chemical neighborhoods where structurally similar compounds have sharply different activities [27-29].

Data splitting is therefore not a technical afterthought but a central determinant of whether reported performance reflects transferable learning or information leakage. DataSAIL was included because it directly addresses data splitting to reduce information leakage in biological and chemical machine-learning settings, reinforcing that validation design must account for similarity, overlap, and dependency among molecules, targets, and assays [30].

Chemical validity also requires careful interpretation in molecular language modelling and generative design. Work on invalid SMILES in chemical language models showed that assumptions about invalid generated strings are not straightforward, which means chemical-language behaviour should be evaluated empirically rather than treated as a simple proxy for drug-likeness, synthesizability, biological activity, or safety [31].

Uncertainty reporting remains a major translational need because drug-discovery decisions often involve prioritizing expensive experiments, not merely selecting the top model output in a benchmark table. Uncertainty quantification has been reviewed as an important component of drug-design decision-making, supporting the synthesis conclusion that foundation-model predictions should be accompanied by calibrated confidence, domain-applicability assessment, and clear distinction between computational prediction and experimentally validated evidence [32].

Evidence synthesis

Across the included studies, molecular language and graph-based representation learning formed the largest conceptual foundation for drug-discovery foundation models. Contrastive graph representation learning showed how self-supervised molecular graph methods can learn transferable representations for property prediction, complementing sequence-based molecular language models and geometry-aware approaches [33].

Target-related prediction and generative design were closely connected but should not be interpreted as equivalent evidence streams. DeepDTAGen linked drug–target affinity prediction with target-aware drug generation in a multitask deep-learning framework, illustrating how prediction and generation can be computationally integrated while still requiring validation of chemical feasibility, biological plausibility, and experimental relevance [34].

The broader deep-learning literature in drug discovery and computational chemistry supports the view that representation learning has expanded modelling across activity prediction, design, synthesis planning, and chemical inference. However, broad reviews of deep learning in drug discovery and computational chemistry also show that the evidence base is heterogeneous across task definitions, datasets, validation settings, and claims of practical impact [35].

This heterogeneity means that foundation models should be synthesized by task and validation type rather than treated as a single uniform technology. Deep learning in computational chemistry supports diverse representation strategies, but the translational meaning of any representation depends on its downstream use, data provenance, benchmark representativeness, external validity, and biological interpretation [36].

Taken together, the evidence supports a cautious synthesis: foundation models may improve molecular representation, target-related modelling, and generative design support, but they remain part of a larger evidence chain that includes assay quality, medicinal chemistry constraints, safety interpretation, uncertainty, and experimental confirmation. Reviews of artificial intelligence in drug discovery and development reinforce that computational methods can influence multiple stages of the

pharmaceutical pipeline while requiring careful interpretation across development contexts [37].

Table 4 presents the evidence synthesis across foundation-model applications, validation practices, translational risks, and research directions.

Table 4. Evidence Synthesis of Foundation Models for Drug Discovery: Application Domains, Validation Patterns, Benchmark Limitations, Translational Risks, Evidence Gaps, and Research Directions

Synthesis theme	Included evidence type	Main contribution	Main limitation	Validation gap	Translational risk	Research direction	Confidence boundary
Molecular language and chemical representation learning	Molecular substructure embeddings, SMILES transformers, BERT-style molecular models	Learns reusable chemical representations for property prediction and design support	Often benchmark-centred	Limited external and prospective testing	Benchmark performance may not generalize to novel chemistry	Scaffold-aware, time-aware, and external validation	Supports representation learning, not clinical utility
Protein language models and target-related tasks	Protein sequence foundation models and dual language models	Captures sequence-derived biological information for target-related modelling	Protein context may be incomplete	Limited disease-context and prospective validation	Sequence similarity may be mistaken for target mechanism	Target-context validation and biological confirmation	Supports target hypothesis generation
Compound–target interaction and binding-related prediction	DTI, DTA, affinity-prediction models	Predicts interaction probability or affinity-like outputs	Sensitive to dataset overlap and split design	Sparse independent and prospective validation	Predicted affinity may be treated as measured binding	Leakage-resistant splitting and experimental assays	Supports prioritization, not mechanism confirmation
Generative molecular design and molecule optimization	Chemical transformers, biochemical language models, target-aware generators	Generates or optimizes molecular candidates computationally	Chemical validity and synthetic feasibility may be insufficiently assessed	Limited experimental confirmation	Generated molecules may be impractical or unsafe	Synthesis-aware, safety-aware, experimentally linked design	Supports hypothesis generation only
ADMET, toxicity, and safety-related prediction	Property prediction, uncertainty, safety-related modelling	May assist early filtering and prioritization	Safety evidence is often indirect or benchmark-based	Limited real-world toxicology confirmation	False reassurance about safety	Calibrated uncertainty and safety-relevant external testing	Does not establish compound safety
Multimodal or knowledge-integrated foundation models	Knowledge-guided pretraining, domain adaptation, multimodal guidance	Integrates chemical, biological, and contextual evidence	Data provenance and modality bias may be opaque	Validation across modalities is inconsistent	Multimodal complexity can hide weak evidence	Transparent data governance and modality-specific appraisal	Supports integrative evidence mapping
Validation and benchmark limitations	Benchmark resources, leakage studies, activity-cliff analyses, data-splitting methods	Identifies why internal metrics may overstate utility	Evaluation practices vary widely	External, prospective, and experimental testing remain uneven	Benchmark superiority may be mistaken for translational readiness	Standardized, leakage-resistant reporting	Confidence limited by validation design
Translational risk and clinical-readiness boundaries	Reviews, uncertainty analyses, model-validity studies	Clarifies risks around uncertainty, interpretability, generalizability, safety, and biological plausibility	Evidence often remains computational	Few studies demonstrate prospective experimental translation	Overclaiming clinical or therapeutic readiness	Prospective testing, experimental confirmation, uncertainty calibration	No deployment claim supported

Figure 2 synthesizes how foundation models connect molecular representation, target-related prediction, generative design, validation practices, and translational-risk boundaries in drug discovery.

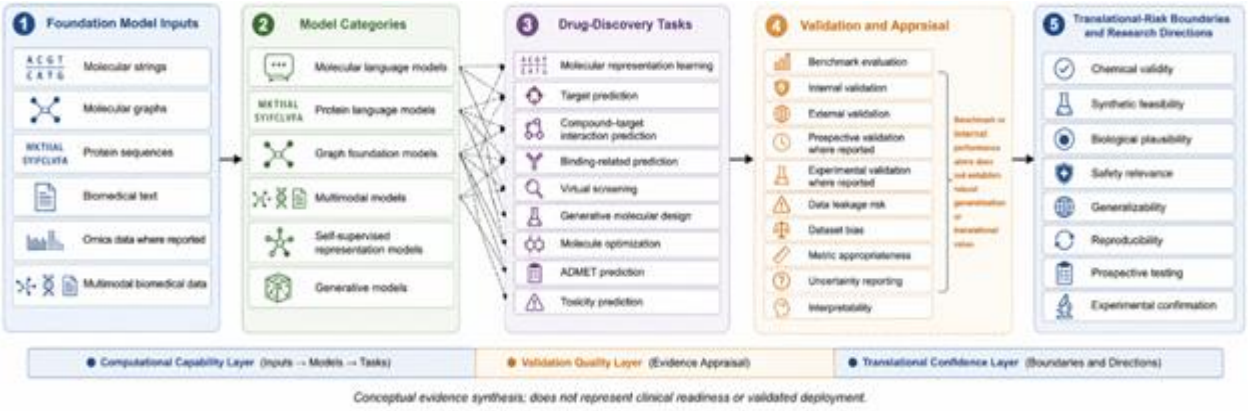

Figure 2. Evidence Synthesis Map of Foundation Models for Drug Discovery

Alt-text description

A conceptual evidence synthesis map showing foundation model categories linked to drug-discovery tasks, validation practices, benchmark limitations, translational risks, and future research directions.

Research directions

Future foundation-model research in drug discovery should move beyond benchmark comparison toward practical evaluation principles that connect pretraining data, task definition, validation design, model uncertainty, and decision context. Recent guidance on foundation and multimodal models in molecular informatics supports the need for clearer evaluation frameworks, practical reporting standards, and cautious interpretation of foundation-model outputs in drug-discovery settings [38].

Systematic evaluation of molecular representation learning foundation models should prioritize leakage-resistant benchmarks, external validation, scaffold- and target-aware splitting, time-split evaluation where feasible, and transparent reporting of pretraining data provenance. A systematic review of molecular representation learning foundation models supports the conclusion that responsible progress requires structured evaluation across model families, data modalities, and downstream tasks rather than isolated benchmark claims [39].

Generative AI and protein-design applications should be linked to medicinal chemistry constraints, target biology, uncertainty calibration, safety-aware filtering, and experimental confirmation. Review evidence on generative AI for drug discovery and protein design supports future work that treats generated molecules and designed sequences as hypotheses requiring synthesis feasibility assessment, biological testing, and translational evidence rather than as validated therapeutic outputs [40].

CONCLUSION

This systematic review shows that foundation models are reshaping molecular representation, protein sequence modelling, target-related prediction, compound–target interaction prediction, binding-related modelling, and generative molecular design in drug discovery. The strongest evidence supports their role as representation-learning and hypothesis-generation tools, while translational confidence remains limited by benchmark dependence, dataset provenance concerns, leakage risk, incomplete external and prospective validation, sparse experimental confirmation, uncertainty gaps, limited interpretability, chemical-feasibility constraints, biological-plausibility uncertainty, and safety-related evidence gaps. Responsible translation will require stronger validation standards, transparent evidence chains, calibrated uncertainty, and experimental confirmation before foundation-model outputs can be treated as reliable drug-discovery decisions.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

[1] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5

[2] Schneider G. Automating drug discovery. Nat Rev Drug Discov. 2018;17(2):97-113. doi:10.1038/nrd.2017.232

[3] 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. doi:10.1038/s41587-019-0224-x

[4] 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.e13. doi:10.1016/j.cell.2020.01.021

[5] Zheng Y, Koh HY, Ju J, Yang M, May LT, Webb GI, et al. Large language models for drug discovery and development. Patterns (N Y). 2025;6(10):101346. doi:10.1016/j.patter.2025.101346

[6] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118(15):e2016239118. doi:10.1073/pnas.2016239118

[7] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Towards cracking the language of Life's code through self-supervised deep learning and high performance computing. IEEE Trans Pattern Anal Mach Intell. 2022;44(10):7112-27. doi:10.1109/TPAMI.2021.3095381

[8] Huang K, Fu T, Gao W, Zhao Y, Roohani Y, Leskovec J, et al. Artificial intelligence foundation for therapeutic science. Nat Chem Biol. 2022;18(10):1033-6. doi:10.1038/s41589-022-01131-2

[9] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A

[10] Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45(D1):D945-54. doi:10.1093/nar/gkw1074

[11] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. 2019;47(D1):D930-40. doi:10.1093/nar/gky1075

[12] Jaeger S, Fulle S, Turk S. Mol2vec: Unsupervised machine learning approach with chemical intuition. J Chem Inf Model. 2018;58(1):27-35. doi:10.1021/acs.jcim.7b00616

[13] Karpov P, Godin G, Tetko IV. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J Cheminform. 2020;12(1):17. doi:10.1186/s13321-020-00423-w

[14] Irwin R, Dimitriadis S, He J, Bjerrum EJ. Chemformer: A pre-trained transformer for computational chemistry. Mach Learn Sci Technol. 2022;3(1):015022. doi:10.1088/2632-2153/ac3ffb

[15] Zhang XC, Wu CK, Yi JC, Zeng XX, Yang CQ, Lu AP, et al. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning BERT enhanced by SMILES enumeration. Research (Wash D C). 2022;2022:0004. doi:10.34133/2022/0004

[16] Zheng X, Tomiura Y. A BERT-based pretraining model for extracting molecular structural information from a SMILES sequence. J Cheminform. 2024;16(1):71. doi:10.1186/s13321-024-00848-7

[17] Fang X, Liu L, Lei J, He D, Zhang S, Zhou J, et al. Geometry-enhanced molecular representation learning for property prediction. Nat Mach Intell. 2022;4(2):127-34. doi:10.1038/s42256-021-00438-4

[18] Li H, Zhang R, Min Y, Ma D, Zhao D, Zeng J, et al. A knowledge-guided pre-training framework for improving molecular representation learning. Nat Commun. 2023;14(1):7568. doi:10.1038/s41467-023-43214-1

[19] 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. doi:10.1093/bioinformatics/btaa1005

[20] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: Deep drug-target binding affinity prediction. Bioinformatics. 2018;34(17):i821-9. doi:10.1093/bioinformatics/bty593

[21] Huang K, Xiao C, Glass LM, Sun J. MolTrans: Molecular Interaction Transformer for drug-target interaction prediction. Bioinformatics. 2021;37(6):830-6. doi:10.1093/bioinformatics/btaa880

[22] Lin S, Shi C, Chen J. GeneralizedDTA: Combining pre-training and multi-task learning to predict drug-target binding affinity for unknown drug discovery. BMC Bioinformatics. 2022;23(1):367. doi:10.1186/s12859-022-04905-6

[23] Lee J, Jun DW, Song I, Kim Y. DLM-DTI: A dual language model for the prediction of drug-target interaction with hint-based learning. J Cheminform. 2024;16(1):14. doi:10.1186/s13321-024-00808-1

[24] Bidgoli AH, Mahdavi M, Malek H. Structure-free drug-target affinity prediction using protein and molecule language models. J Cheminform. 2026;18(1):21. doi:10.1186/s13321-025-01146-6

[25] Bai P, Miljković F, John B, Lu H. Interpretable bilinear attention network with domain adaptation improves drug-target prediction. Nat Mach Intell. 2023;5(2):126-36. doi:10.1038/s42256-022-00605-1

[26] Uludoğan G, Ozkirimli E, Ulgen KO, Karalı N, Özgür A. Exploiting pretrained biochemical language models for targeted drug design. Bioinformatics. 2022;38(Suppl 2):ii155-61. doi:10.1093/bioinformatics/btac482

[27] Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403

[28] Tran-Nguyen VK, Jacquemard C, Rognan D. LIT-PCBA: An unbiased data set for machine learning and virtual screening. J Chem Inf Model. 2020;60(9):4263-73. doi:10.1021/acs.jcim.0c00155

[29] van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model. 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073

[30] Joeres R, Blumenthal DB, Kalinina OV. Data splitting to avoid information leakage with DataSAIL. Nat Commun. 2025;16(1):3337. doi:10.1038/s41467-025-58606-8

[31] Skinnider MA. Invalid SMILES are beneficial rather than detrimental to chemical language models. Nat Mach Intell. 2024;6(4):437-48. doi:10.1038/s42256-024-00821-x

[32] Mervin LH, Johansson S, Semenova E, Giblin KA, Engkvist O. Uncertainty quantification in drug design. Drug Discov Today. 2021;26(2):474-89. doi:10.1016/j.drudis.2020.11.027

[33] Wang Y, Wang J, Cao Z, Barati Farimani A. Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell. 2022;4(3):279-87. doi:10.1038/s42256-022-00447-x

[34] Shah PM, Zhu H, Lu Z, Wang K, Tang J, Li M. DeepDTAGen: A multitask deep learning framework for drug-target affinity prediction and target-aware drugs generation. Nat Commun. 2025;16(1):5021. doi:10.1038/s41467-025-59917-6

[35] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23(6):1241-50. doi:10.1016/j.drudis.2018.01.039

[36] Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry. J Comput Chem. 2017;38(16):1291-307. doi:10.1002/jcc.24764

[37] Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021;26(1):80-93. doi:10.1016/j.drudis.2020.10.010

[38] Pastore EP, De Rango F. Foundation and multimodal models for drug discovery in molecular informatics: Principles, evaluation, and practical guidance. Mol Inform. 2026;45(3):e70027. doi:10.1002/minf.70027

[39] Song B, Zhang J, Liu Y, Liu Y, Jiang J, Yuan S, et al. A systematic review of molecular representation learning foundation models. Brief Bioinform. 2026;27(1):bbaf703. doi:10.1093/bib/bbaf703

[40] Das U. Generative AI for drug discovery and protein design: The next frontier in AI-driven molecular science. Med Drug Discov. 2025;27:100213. doi:10.1016/j.medidd.2025.100213