Available online **www.ijpras.com**

International Journal of Pharmaceutical Research & Allied Sciences, 2026, 15(1):85-98
https://doi.org/10.51847/95N6WdsK1c

Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Foundation Models for Pharmacoinformatics and QSAR: A Scoping Review of Transfer, Grounding, Calibration, and Chemical Validity

Nguyen Thanh Huy[1]*, Pham Quang Minh[1], Le Thi Bich[2], Tran Van Nam[3]

[1]Department of Foundation Models and Pharmacoinformatics, Faculty of Pharmacy, Hanoi University of Pharmacy, Hanoi, Vietnam.
[2]Department of Transfer and Grounding in QSAR, Faculty of Pharmacy, Can Tho University of Medicine and Pharmacy, Can Tho, Vietnam.
[3]Department of Calibration and Chemical Validity, Faculty of Pharmacy, Hue University, Hue, Vietnam.

***Email:** huy.nguyen@hup.edu.vn*

ABSTRACT

Foundation models are increasingly presented as reusable computational infrastructures for molecular property prediction, compound generation, reaction modelling, biological representation learning, and pharmacoinformatics. Their relevance to quantitative structure–activity relationship modelling, however, cannot be inferred from pretraining scale or benchmark performance alone. This scoping review maps how foundation-model concepts are defined, trained, adapted, transferred, grounded, calibrated, and evaluated across chemically and biologically relevant data modalities. A PRISMA-ScR-compatible and JBI-informed approach was used to define broad mapping questions, eligibility boundaries, information sources, search logic, screening criteria, data-charting fields, and a descriptive thematic synthesis. The evidence was organized by model family, pretraining data, representation type, adaptation mechanism, downstream task, transfer setting, validation stage, uncertainty treatment, chemical-validity assessment, and intended decision context. The literature spans molecular graphs and strings, conformational or multi-view molecular data, reactions, proteins, omics, assay descriptions, and scientific text. Evidence is concentrated in retrospective model development and internal benchmark evaluation, whereas independent external validation, calibration under distribution shift, explicit applicability-domain analysis, prospective experimental confirmation, and decision-context evaluation are less consistently represented. Reported transfer is strongly conditioned by endpoint definition, chemical-space overlap, split strategy, label quality, comparator selection, and access to task-specific data. Scientific grounding remains heterogeneous, ranging from syntactic molecular validity to tool-supported chemical checking, retrosynthetic assessment, and interpretable representation analysis. The review identifies a need for clearer foundation-model definitions, realistic transfer tests, uncertainty evaluation under shift, stronger chemical-validity controls, reproducible workflows, and explicit boundaries between computational performance and pharmaceutical utility. Foundation models may broaden reusable representation learning, but their value for QSAR and pharmacoinformatics remains conditional on evidence appropriate to the intended chemical domain and decision use.

Key words: *Foundation model, QSAR, Pharmacoinformatics, Chemical space, External validation, Applicability domain*

Received: 01 December 2025; **Revised:** 25 February 2026; **Accepted:** 28 February 2026

INTRODUCTION

Machine learning now contributes to multiple stages of pharmaceutical research, including target-related analysis, molecular property prediction, virtual screening, compound design, development modelling, and evidence integration. These activities do not constitute a single validated workflow: their usefulness depends on the scientific question, data-generating process, endpoint definition, evaluation design, and intended downstream decision [1]. Foundation models extend this landscape by pretraining reusable representations or generative capabilities on broad datasets before adaptation to particular tasks. Their scale and reuse potential are important, but neither property establishes reliable transfer to a new endpoint, chemical series, biological system, or pharmaceutical context.

Artificial intelligence in computer-assisted drug discovery includes heterogeneous model classes with different representations, objectives, assumptions, and failure modes. Graph neural networks, sequence models, encoder–decoder systems, multimodal architectures, generative models, and predictive QSAR systems cannot be treated as equivalent merely because they use pretraining or share a foundation-model label [2]. A meaningful assessment must therefore distinguish the information encoded by a representation, the objective used during pretraining, the adaptation mechanism, the comparator, and the validation distribution. Without these distinctions, apparently similar performance claims may refer to substantially different scientific tasks.

The available chemical and biological data introduce additional limitations. Molecular records can contain duplicated structures, inconsistent standardization, assay heterogeneity, uncertain labels, missing experimental context, and non-random coverage of chemical space. Biological observations may similarly reflect measurement conditions, selection effects, proxy endpoints, or incomplete mechanistic annotation. Increasing training scale does not automatically correct these problems, and a model may learn dataset-specific regularities that transfer poorly outside the distributions from which the labels were constructed [3]. Consequently, foundation-model assessment requires attention to data provenance and label meaning as well as architecture and benchmark scores.

A scoping review is appropriate because the relevant literature spans model-development studies, benchmark analyses, representation-learning investigations, methodological comparisons, validation studies, software articles, and field-level syntheses. PRISMA-ScR principles support transparent reporting of review objectives, eligibility criteria, information sources, search logic, selection, charting, and synthesis without converting the review into an effectiveness assessment [4]. JBI guidance further supports a broad population–concept–context framework, iterative refinement of search and charting procedures, and descriptive or thematic mapping when concepts and evidence maturity are heterogeneous [5]. The objective of this review is therefore to map how foundation models are being used in pharmacoinformatics and QSAR, identify the conditions under which transfer claims are evaluated, and clarify the evidence needed to establish grounding, calibration, chemical validity, and externally relevant performance.

Scoping-review objectives and eligibility framework

The review addressed six connected mapping questions: which foundation-model concepts, technologies, data types, and pharmacoinformatics applications have been studied; how the literature can be classified; which pretraining and adaptation stages are represented; how transfer, validation, uncertainty, grounding, and translation are evaluated; which terms or contexts remain inconsistently defined; and which research priorities emerge from the mapped evidence. These questions were framed to characterize the extent, nature, and boundaries of the literature rather than determine a pooled effect or identify a universally superior model. The reporting framework follows the scoping-review elements of explicit objectives, eligibility, information sources, selection, charting, and synthesis [4].

Eligibility was structured around the computational systems under study, the relevant concepts, and the contexts in which claims were evaluated. Eligible evidence directly examined foundation or broadly pretrained models, transferable representations, QSAR or molecular property prediction, chemical or biological data integration, applicability domains, uncertainty, chemical validity, or external validation. Charting treated model architecture, pretraining corpus, representation, learning objective, adaptation mechanism, endpoint, chemical domain, split strategy, validation stage, calibration, and decision use as separate dimensions. This approach is consistent with JBI guidance for conceptually heterogeneous evidence and prevents the foundation-model label from functioning as an assumed indicator of performance or maturity [5].

The resulting framework locates foundation models within the wider pharmaceutical machine-learning landscape while separating model construction from evidence of downstream utility. Broad pharmaceutical applications provide the outer context [1]. Differences in representation and learning objective define distinct methodological families [2]. Data quality, label construction, and chemical-space coverage impose cross-cutting constraints on

every family [3]. **Figure 1** presents the original conceptual synthesis for foundation-model landscape for pharmacoinformatics, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

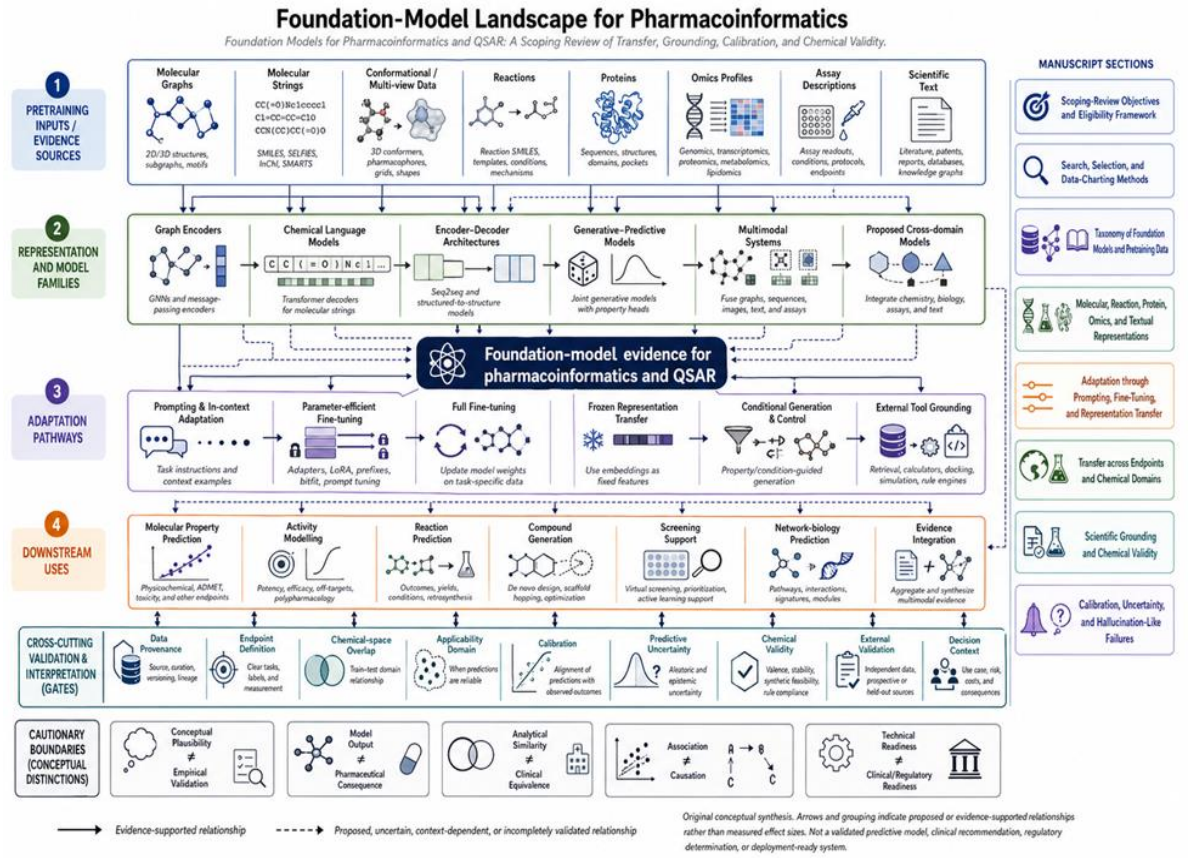

Figure 1. Foundation-Model Landscape for Pharmacoinformatics

Search, selection, and data-charting methods

Information sources comprised biomedical literature indexes, publisher records, DOI-indexed bibliographic metadata, and journal-ranking records used to verify eligibility. The search combined foundation-model terminology with QSAR, pharmacoinformatics, molecular property prediction, transfer learning, pretraining, fine-tuning, prompting, representation learning, molecular graphs, chemical language models, reaction modelling, proteins, omics, external validation, applicability domains, chemical space, calibration, predictive uncertainty, chemical validity, synthesizability, explainability, causal modelling, and reproducibility. Searches were iteratively refined when terminology varied across disciplines, consistent with transparent scoping-review reporting [4]. No quantitative synthesis or search-flow estimate was inferred from the search process.

Titles and abstracts were assessed for direct relevance to the review questions, followed by full-text assessment of scientific scope, publication type, DOI availability, and extractable evidence. Articles were retained only when they could support a defined review construct, paragraph-level claim, table entry, figure component, limitation, or research priority. Version and DOI checks were used to avoid duplicate records, preprint–journal duplication, and double counting of corrected and original articles. Ambiguous decisions were resolved by returning to the operational eligibility definitions rather than broadening inclusion solely because an article used related artificial-intelligence terminology. This staged and documented selection logic follows the iterative approach permitted for scoping reviews [5].

Data charting captured bibliographic identity, study type, model family, pretraining data, representation modality, objective, adaptation mechanism, application context, endpoint, comparator, split strategy, validation stage, reported capability, calibration or uncertainty treatment, chemical-validity assessment, principal limitation, and intended decision context. The synthesis was descriptive and thematic: evidence was classified by model and data taxonomy, then mapped across transfer, grounding, uncertainty, validation, and translation dimensions. Formal risk-of-bias scoring was not used as an exclusion mechanism because the review questions concern conceptual coverage and methodological maturity rather than comparative intervention effects. Instead, limitations in

validation design, data externality, applicability-domain definition, calibration, and reporting were retained as charted evidence attributes [5].

Taxonomy of foundation models and pretraining data

A useful taxonomy begins with the relation between the pretraining object and the information required by downstream QSAR tasks. Graph foundation models treat atoms and bonds as structured entities and can learn reusable molecular representations through self-supervised objectives. MolE illustrates this category through disentangled attention intended to separate molecular factors within graph representations and support transfer across evaluated downstream tasks [6]. This evidence supports graph-based pretraining as a distinct model family, but it does not show that one representation transfers uniformly across all endpoints, scaffold classes, stereochemical contexts, or assay regimes.

A second category integrates prediction and generation within a shared model. Bidirectional molecular modelling can condition on a structure to predict properties or condition on desired properties to generate candidate structures. A single molecular foundation model has demonstrated this bidirectional structure–property capability under its reported benchmark conditions [7]. The architecture therefore expands the range of possible downstream interactions, but bidirectionality should not be interpreted as causal knowledge of why a structural modification changes a biological or physicochemical endpoint.

A third category enlarges the pretraining modality beyond a static molecular graph or one-dimensional string. Molecular-video pretraining represents compounds through sequences of molecular views or conformational frames and may capture information unavailable to a single static encoding. A video-derived molecular foundation model has shown that such multi-view pretraining can support several scientific drug-discovery tasks in the evaluated settings [8]. Nevertheless, learned variation across frames is not automatically equivalent to a physically validated conformational ensemble, thermodynamic distribution, or experimentally observed molecular mechanism.

A fourth category comprises scalable encoder–decoder model families trained on chemical strings or text-like scientific representations. Open chemistry foundation models can support common pretraining infrastructure, multiple model sizes, task adaptation, and more inspectable comparisons across downstream applications. An open-source encoder–decoder family has demonstrated the feasibility of applying a shared chemistry-oriented architecture across multiple tasks [9]. Openness improves access and reproducibility potential, but it does not remove dependence on corpus composition, tokenization, benchmark construction, computational resources, or task-specific validation.

Across these categories, "foundation model" is best treated as a charted combination of pretraining scope, reusable architecture, objective, data modality, adaptation pathway, and downstream evidence rather than as a performance designation. Graph, bidirectional generative–predictive, dynamic multi-view, and encoder–decoder families encode different inductive biases and omit different forms of chemical information. Their categories are therefore non-exclusive, and a model may occupy more than one class. The review taxonomy does not rank these families globally; it identifies what was pretrained, what was transferred, how performance was evaluated, and which chemical or decision boundaries remain unresolved. **Table 1** organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for scoping-review eligibility and data-charting framework for foundation models for pharmacoinformatics and qsar.

Table 1. Scoping-review eligibility and data-charting framework for Foundation Models for Pharmacoinformatics and QSAR.

Evidence domain or review construct	Review question	Eligible evidence type	Key extraction fields	Appraisal or relevance criterion	Synthesis role	Main limitation	Interpretation boundary
Review population and system	Which computational systems fall within pharmacoinformatics or QSAR?	Model studies, validation studies, methodological analyses, software articles, and	Pharmaceutical task, endpoint, biological or chemical system, intended use	Direct relevance to a defined pharmacoinformatics or QSAR question	Establishes the review's population and context	Field boundaries overlap with cheminformatics, bioinformatics, and general drug-discovery AI	Inclusion indicates relevance, not demonstrated effectiveness or clinical utility [1].

		field syntheses					
Foundation-model construct	How is a foundation model distinguished from a task-specific pretrained model?	Studies reporting broad pretraining, reusable representations, generative capabilities, or multi-task adaptation	Architecture, corpus, objective, reuse mechanism, downstream tasks, openness	Model must describe a reusable pretrained capability rather than apply the label without operational detail	Supports model-family classification	Terminology is inconsistent and may reflect scale or branding rather than function	"Foundation model" is a proposed charting class, not a validated performance tier [2].
QSAR and pharmacoinformatics task	What downstream property, activity, interaction, or pharmaceutical question is addressed?	Predictive, generative, representation-learning, screening, or evidence-integration studies	Endpoint definition, label origin, task type, comparator, decision context	Endpoint and intended use must be sufficiently specified for interpretation	Connects pretraining to downstream application	Similar task names may conceal different assays, conditions, or label definitions	Analytical prediction cannot alone establish pharmaceutical consequence [1].
Pretraining data and provenance	What information was available to the model before downstream adaptation?	Molecular graph, string, conformational, reaction, protein, omics, assay-description, or scientific-text studies	Source, modality, curation, standardization, duplication control, coverage, missingness	Data origin and relation to downstream evaluation must be identifiable	Maps information content and possible leakage pathways	Corpus composition and preprocessing are often incompletely reported	Larger corpora do not automatically correct biased, heterogeneous, or weakly defined labels [3].
Representation and objective	What chemical or biological relations can the representation encode or omit?	Graph, sequence, encoder–decoder, multimodal, contrastive, generative, or predictive studies	Tokenization, structural encoding, objective, augmentation, dimensionality, model family	Representation and objective must be connected to a specific downstream capability	Distinguishes model mechanisms and inductive biases	Learned features may be difficult to interpret chemically	Representation similarity does not establish mechanistic or causal grounding [2].
Eligibility and transparent reporting	Are objectives, sources, selection, charting, and synthesis sufficiently reported?	Scoping-review methods and reporting guidance	Review questions, eligibility boundaries, information sources, search concepts, selection, charting, synthesis	Procedures must be explicit enough to understand how evidence entered the map	Supports reproducible review conduct	Reporting completeness does not determine the validity of included models	PRISMA-ScR-compatible reporting is not a risk-of-bias score or evidence grade [4].
Population–concept–context charting	How should heterogeneous evidence be organized without forcing effect comparison?	Methodological guidance and eligible domain studies	Population or system, concept, application context, evidence type,	Charting categories must answer a review question and remain revisable when terminology differs	Supports descriptive and thematic synthesis	Iterative categories may change as evidence is examined	The framework is an original pharmacoinformatics-specific adaptation of JBI-informed charting [5].

			validation stage				
Transfer and validation	Under what conditions is a pretrained capability evaluated beyond its development distribution?	Internal, scaffold, temporal, external, out-of-domain, prospective, or experimental evaluations	Split strategy, overlap, applicability domain, externality, comparator, calibration, failure analysis	Validation design must be distinguished from ordinary internal performance testing	Maps evidence maturity and transport boundaries	"External" datasets may still overlap in chemistry, provenance, or label construction	Transfer is bounded to the evaluated endpoint, chemical space, and data-generating process [3].
Chemical validity and decision use	Does the output satisfy inspectable chemical constraints and support the intended decision?	Validity checks, retrosynthetic or tool-grounded assessments, expert review, experimental follow-up, decision-context studies	Structural validity, synthesizability, mechanistic support, uncertainty, action linked to output	The check must correspond to the claim being made about the output	Separates plausible computation from consequential use	Proxy validity measures may omit feasibility, biology, safety, or implementation	Passing a computational check does not establish experimental success, clinical equivalence, regulatory acceptance, or deployment readiness [1].

Molecular, reaction, protein, omics, and textual representations

Reaction models commonly represent reactants, reagents, and products as token sequences, allowing transformer architectures to learn reaction transformations without manually specified reaction rules. The Molecular Transformer demonstrated that sequence-based reaction prediction can be coupled with confidence estimation under curated evaluation conditions [10]. Such confidence is model- and distribution-dependent and should not be interpreted as universal evidence of reaction feasibility, yield, selectivity, or laboratory reproducibility.

Protein language models expand the foundation-model concept from chemical structures to biological sequences. Large-scale unsupervised learning from protein sequences has produced representations associated with structural and functional characteristics, indicating that biologically meaningful regularities can emerge from sequence pretraining [11]. Nevertheless, representation-level associations do not establish causal biological mechanisms, ligand-specific effects, or transfer to a particular pharmacological endpoint without targeted validation.

Omics foundation models encode cellular states rather than isolated compounds. Transfer learning from transcriptomic data has supported predictions involving network biology and perturbation responses, illustrating how pretrained cellular representations may extend pharmacoinformatics beyond conventional structure-only QSAR [12]. These models address a different inferential unit from molecular property prediction; successful cell-state transfer therefore cannot be assumed to establish compound-level generalization or clinical response.

Multimodal representations attempt to connect chemical structures with assay descriptions, biological annotations, or scientific language. Combining molecular structures with bioassay text has improved property-prediction performance in evaluated settings by supplying contextual information not contained in structure alone [13]. However, textual modalities may also transmit annotation inconsistencies, experimental ambiguity, historical biases, or target leakage. Multimodal performance must therefore be interpreted in relation to the provenance and independence of every input channel.

Adaptation through prompting, fine-tuning, and representation transfer

Fine-tuning adapts pretrained parameters to a defined downstream task and is especially attractive when labelled QSAR data are limited. MolPMoFiT showed that inductive transfer from chemical language modelling can support molecular activity prediction in low-data settings [14]. The evidence supports endpoint-specific utility, but it does not establish that fine-tuning will improve every assay, scaffold family, class distribution, or external chemical domain.

Frozen representation transfer provides another adaptation route. Transformer embeddings derived from screening information have been used as inputs for binding-related prediction, showing that pretrained features can retain

information useful for downstream modelling [15]. Apparent transfer may nevertheless depend on unrecognized similarity between source and target assays, compounds, or labels. Evaluation should therefore document overlap and compare transferred representations with well-tuned task-specific baselines.

Comparative evidence indicates that pretraining benefits are heterogeneous. Controlled evaluations have shown that pretrained molecular representations do not consistently outperform alternative representations or supervised baselines across all tasks and split strategies [16]. Performance gains may arise from architecture, data volume, hyperparameter tuning, or chemical-space overlap rather than pretraining alone. Foundation-model evaluation should isolate these factors instead of treating pretraining as the default explanation for improvement.

Prompting and tool augmentation introduce adaptation without relying exclusively on conventional parameter updating. Chemistry-focused large language models can be connected to external tools for molecular calculations, reaction information, or structured chemical checks, thereby improving the operational grounding of some responses [17]. Tool access does not eliminate hallucination-like failure because the model may select the wrong tool, construct an invalid query, misread the result, or overstate its relevance. **Table 2** organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for concept, method, technology, or application taxonomy for foundation models for pharmacoinformatics and qsar.

Table 2. Concept, method, technology, or application taxonomy for Foundation Models for Pharmacoinformatics and QSAR.

Method, platform, or evidence category	Representative application	Reported strength	Validation context	Generalization concern	Contradictory evidence	Evidence gap	Review interpretation
Chemical-language fine-tuning [14]	Low-data activity prediction	Reuses pretrained molecular syntax	Retrospective endpoint-specific evaluation	Endpoint and scaffold dependence	Gains are not necessarily uniform	Independent low-data comparisons	Evidence-derived; transfer is conditional
Frozen transformer embeddings [15]	Binding prediction from assays	Reduces task-specific representation learning	Screening-derived benchmark data	Hidden assay or chemical overlap	Task-specific baselines may remain competitive	External-series testing	Evidence-derived; overlap must be audited
Molecular pretraining comparison [16]	Property prediction	Tests whether pretraining adds value	Controlled benchmark comparison	Split and tuning sensitivity	Pretraining is not consistently superior	Standardized ablation studies	Evidence-derived critical comparison
Tool-augmented language model [17]	Chemical reasoning and calculation	Connects generation to inspectable tools	Curated chemistry tasks	Tool-selection and interpretation errors	Tool access does not prevent hallucination	Independent adversarial evaluation	Evidence-derived; operational grounding only
Prompt-based adaptation	Task instruction without full retraining	Flexible task formulation	Emerging chemistry applications	Prompt sensitivity and instability	Strong task models may outperform prompting	Reproducible prompt protocols	**Proposed charting category**
Parameter-efficient adaptation	Updating restricted model components	Potentially lower adaptation cost	Limited pharmacoinformatics evidence	Capacity and domain-shift limits	Full fine-tuning may perform differently	Head-to-head external validation	**Proposed charting category**
Multimodal alignment	Linking structures, assays, and text	Adds contextual information	Modality-specific retrospective tests	Annotation bias and leakage	Extra modalities may add noise	Modality-ablation and provenance studies	**Proposed integrative category**
External-tool grounding	Retrosynthesis, property calculation, database checking	Produces inspectable intermediate outputs	Tool-supported workflows	Tool coverage and database dependence	Correct tool output may still be misused	End-to-end error attribution	**Proposed evaluation requirement**

Transfer across endpoints and chemical domains

Transfer claims require benchmarks that make tasks, splits, and metrics sufficiently explicit for comparison. MoleculeNet provided standardized datasets and evaluation conventions for molecular machine learning [18]. Such infrastructure improves comparability, but benchmark performance remains surrogate evidence. It does not establish transfer to an independently generated assay, a new chemical series, a changed experimental protocol, or a pharmaceutical decision context.

Large-scale comparisons further show that performance depends on endpoint characteristics and comparator quality. Evaluation across numerous drug-target prediction tasks found that deep learning could perform strongly while also revealing substantial heterogeneity among assays and the continued competitiveness of other machine-learning approaches [19]. Model-family rankings should therefore remain conditional on label density, imbalance, endpoint complexity, descriptor choice, and tuning effort.

Representation and validation split interact strongly. Learned molecular representations can produce different apparent advantages under random, scaffold-based, or other partitioning strategies, and performance may change when models encounter structurally less familiar compounds [20]. Random splits often preserve close chemical neighbours across training and test sets. They are useful for interpolation assessment but are insufficient as the sole evidence for transfer across chemical domains.

Activity cliffs expose an additional limitation of aggregate metrics. Small structural changes can correspond to large activity differences, creating local prediction failures even when global performance appears acceptable. Systematic activity-cliff analysis has shown that molecular machine-learning models may struggle in precisely these chemically informative regions [21]. Transfer evaluation should therefore combine global metrics with local error analysis rather than infer chemical reliability from average performance alone.

Scientific grounding and chemical validity

Chemical grounding begins by separating formal molecular validity from practical feasibility. Generative models may produce structures that satisfy representation syntax and valence rules while remaining difficult to synthesize, unstable, implausible, or poorly matched to the intended biological context. Assessment of generated molecules has shown that formal validity alone does not establish synthesizability [22]. Chemical validity is therefore multidimensional rather than a single pass–fail property.

Retrosynthetic planning offers a more inspectable form of grounding by proposing routes from candidate products to available starting materials. AiZynthFinder demonstrates how automated retrosynthesis can be incorporated into computational molecular-design workflows [23]. A predicted route provides stronger evidence than syntactic validity, but it remains a computational proposal. Route success, selectivity, purification, safety, scale, and material availability require separate assessment.

Rapid accessibility scores can approximate route-based evidence when full retrosynthetic searches are too costly. RAscore was trained using outputs from automated retrosynthetic planning to classify molecular accessibility more efficiently [24]. Such a score may support triage, but it is neither an executable synthesis protocol nor proof that a compound can be prepared under relevant conditions. Thresholds should remain context-specific and independently checked.

Interpretability may help investigators inspect whether predictions rely on chemically plausible features, but explanations must not be confused with causal mechanisms. Explainable artificial intelligence can support hypothesis generation, error analysis, and communication when its methods and stability are reported [25]. Post hoc feature attribution alone cannot demonstrate that highlighted substructures cause an endpoint or that modifying them will produce the predicted effect. **Table 3** organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for evidence distribution, validation maturity, and translational coverage in foundation models for pharmacoinformatics and qsar.

Table 3. Evidence distribution, validation maturity, and translational coverage in Foundation Models for Pharmacoinformatics and QSAR.

Cross-study pattern or configuration	Context or boundary condition	Mechanism or explanatory factor	Observed or reported outcome	Rival explanation	Certainty or rigour concern	Implication	Research priority
Syntactically valid generation [22]	Representation-level evaluation	Learned molecular grammar	Valid molecular strings or graphs	Memorized structural patterns	Validity omits feasibility	Formal correctness is insufficient	Add synthesis and stability checks

Retrosynthetic route proposal [23]	Tool-supported chemical assessment	Search through reaction templates	Candidate synthetic pathways	Template or database coverage	Routes remain unexecuted	Improves inspectability	Validate route success experimentally
Accessibility proxy [24]	High-throughput candidate triage	Learned approximation to route planning	Rapid synthesizability classification	Training-set dependence	Proxy may fail outside its domain	Useful for prioritization only	Report domain and failure cases
Explainable prediction [25]	Model inspection	Feature attribution or interpretable structure	Chemically inspectable rationale	Correlation or explanation instability	Explanations may be non-causal	Supports review, not proof	Test stability and intervention relevance
High benchmark accuracy	Similar train–test chemistry	Interpolation from neighbouring examples	Strong aggregate metrics	Chemical-space redundancy	Externality may be limited	Does not establish broad transfer	Use temporal and external splits
Multimodal grounding	Structure combined with text or assays	Contextual representation alignment	Improved task performance	Annotation leakage or bias	Input independence unclear	Added modalities require provenance audit	Perform modality ablations
Tool-grounded language modelling	External calculators or chemistry software	Model delegates inspectable operations	More verifiable intermediate outputs	Tool misuse or query errors	End-to-end error remains uncertain	Grounding is operational, not causal	Attribute errors across model and tool
Pharmaceutical translation	Decision-specific downstream use	Combination of model and external evidence	Potential prioritization support	Workflow and selection effects	Prospective evidence is sparse	Utility is context-bound	Validate consequences prospectively

Calibration, uncertainty, and hallucination-like failures

Uncertainty estimation methods differ in computational cost, assumptions, and sensitivity to distribution shift. Comparative evaluation of scalable approaches for molecular property prediction has shown that ensembles, stochastic methods, and related techniques can behave differently across datasets and tasks [26]. No method can be assumed universally optimal; selection should reflect the decision context, available computation, expected shift, and consequences of error.

Predictive accuracy and uncertainty quality are distinct. A model may produce accurate mean predictions while assigning poorly calibrated confidence, or it may appear calibrated internally while failing on unfamiliar chemistry. Neural-network uncertainty studies in molecular property prediction demonstrate the need to evaluate error ranking, calibration, and out-of-domain behaviour separately [27]. Calibration observed on an internal test set cannot establish calibration after endpoint, assay, or chemical-space shift.

Evidential approaches aim to learn predictions and uncertainty in a single model and can guide selection toward candidates that combine predicted desirability with informative uncertainty. Evidential deep learning has been applied to molecular property prediction and discovery-oriented prioritization [28]. This capability may improve candidate selection under specified conditions, but decision benefit depends on uncertainty validity, acquisition policy, laboratory follow-up, and the costs of false confidence.

Chemistry-related uncertainty includes measurement noise, model uncertainty, data sparsity, distributional novelty, label-definition ambiguity, and task mismatch. These components have different causes and require different diagnostics [29]. Hallucination-like failures in foundation models may arise when fluent or structurally plausible outputs conceal unsupported chemistry, misunderstood context, or extrapolation. A single confidence score should not be treated as a complete representation of these failure sources.

Evaluation benchmarks and external validation

Cross-task benchmark infrastructures can improve accessibility, documentation, and comparison across therapeutic machine-learning problems. The Therapeutics Data Commons was developed to organize datasets, tasks, and evaluation resources across therapeutic science [30]. Such infrastructure facilitates reproducible experimentation but cannot guarantee label quality, adequate baselines, absence of overlap, or relevance to a specific downstream pharmaceutical decision.

Benchmark construction may reward memorization rather than generalization when related ligands occur across training and test partitions. Analysis of ligand-based classification benchmarks showed that structural redundancy can inflate apparent predictive performance [31]. External validation therefore requires more than assigning a new dataset name: chemical novelty, provenance independence, endpoint compatibility, and temporal separation must be evaluated explicitly.

Apparent performance is also shaped by choices throughout the modelling pipeline. Systematic assessment of molecular property prediction has demonstrated the influence of representation, model architecture, split strategy, dataset size, and other experimental elements [32]. Because these factors interact, isolated claims that one architecture is superior are difficult to interpret without harmonized tuning, matched data, transparent preprocessing, and sensitivity analyses.

Industry-scale evaluation provides valuable evidence about performance under larger and more operationally relevant datasets. Deep learning has been evaluated for drug-target prediction at industrial scale, showing both practical potential and dependence on task and data conditions [33]. Industrial use does not automatically constitute independent external validation because development and evaluation may still occur within related data systems. **Table 4** organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for research gaps, underrepresented domains, and priority directions for foundation models for pharmacoinformatics and qsar.

Table 4. Research gaps, underrepresented domains, and priority directions for Foundation Models for Pharmacoinformatics and QSAR.

Evidence-maturity domain	Current evidence position	Methodological weakness	Translation barrier	Needed validation	Reporting priority	Decision implication	No-overclaiming boundary
Benchmark infrastructure [30]	Shared datasets and tasks are available	Dataset quality varies	Benchmark relevance may be indirect	Independent task-specific evaluation	Provenance, labels, splits, comparators	Supports comparison, not readiness	A benchmark is not external validation
Train–test independence [31]	Redundancy is a recognized concern	Similar chemistry can cross splits	Memorization may appear as transfer	Scaffold, temporal, and provenance-separated tests	Duplicate and similarity audits	Limits generalization claims	New dataset labels do not prove independence
Pipeline sensitivity [32]	Multiple design choices affect outcomes	Inconsistent tuning and preprocessing	Results may not reproduce	Harmonized ablation and sensitivity studies	Full pipeline specification	Architecture claims remain conditional	Retrospective superiority is not prospective utility
Industry-scale modelling [33]	Large operational datasets have been examined	Externality may remain limited	Organizational data differ across settings	Independent cross-organization testing	Data lineage and deployment context	Supports feasibility assessment	Scale is not clinical or regulatory validation
Classical baseline comparison [34]	Strong non-foundation models remain competitive	Baselines may be weakly tuned	Added complexity may lack benefit	Matched optimization and compute accounting	Hyperparameters and feature stability	Use the simplest adequate model	Novelty does not establish superiority
Reproducible QSAR workflow [35]	Modular open tools are emerging	Applicability domains are inconsistently reported	Workflows are hard to transfer	Re-execution on independent endpoints	Data, code, versions, domains	Improves auditability	Software availability is not endpoint validation
Causal interpretation [36]	Most evidence remains predictive	Association is often overinterpreted	Predictions may not support intervention	Causal assumptions and intervention tests	Estimand and causal graph	Restrict claims to prediction	Predictive importance is not causal effect

Leakage control [37]	Leakage is recognized across ML science	Controls are often post hoc	Inflated evidence may guide decisions	Predefined leakage audits and replication	Complete data lineage	Confidence should reflect evidence independence	Reproducibility checks do not guarantee utility

Evidence gaps and future priorities

Foundation-model studies should be compared with strong, task-appropriate alternatives rather than untuned or obsolete baselines. Practical evaluation of gradient-boosting models has shown that established approaches can remain highly competitive and that interpretation may vary with modelling choices [34]. Future comparisons should harmonize data, splits, descriptors, hyperparameter optimization, computational budgets, and uncertainty assessment before attributing gains to model scale or pretraining.

Reproducible QSAR infrastructure is also needed. QSPRpred illustrates how modular open-source workflows can support data preparation, model building, applicability-domain analysis, and evaluation [35]. Similar transparency should be required for foundation-model adaptation, including pretrained weights, software versions, tokenization, frozen and updated parameters, prompts, tool calls, split generation, duplicate removal, calibration procedures, and failed runs. Open software improves auditability but does not validate a specific endpoint.

A further priority is to separate predictive association from causal and counterfactual interpretation. Guidance from actionable machine learning shows that prediction, causal inference, and counterfactual reasoning answer different questions and require different assumptions [36]. In pharmacoinformatics, feature importance or predicted response should not be described as a causal molecular mechanism unless intervention-relevant assumptions and evidence are supplied.

Leakage control should be defined before model development rather than added after unexpectedly strong results. Leakage and incomplete reproducibility can inflate machine-learning evidence through overlap, preprocessing across partitions, outcome-derived features, or repeated benchmark optimization [37]. **Figure 2** presents the original conceptual synthesis for evaluation pathway from pretraining to chemically valid downstream use, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

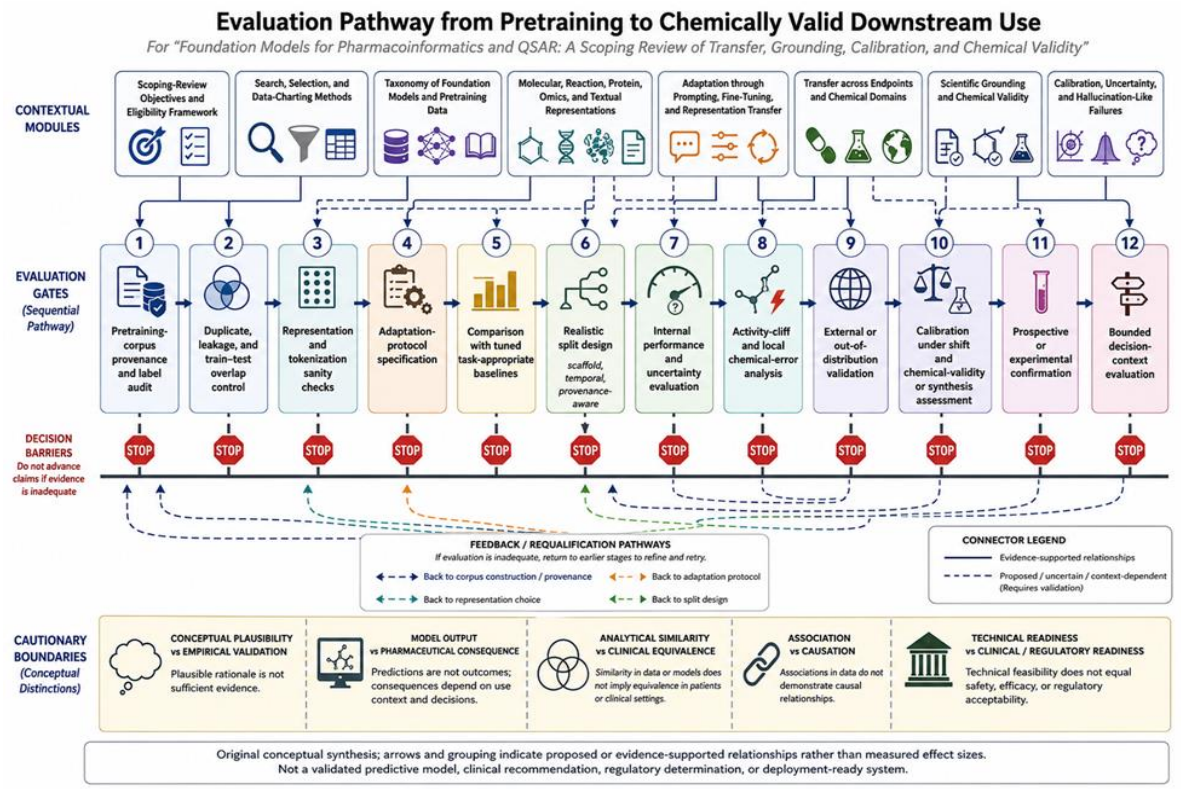

Figure 2. Evaluation Pathway from Pretraining to Chemically Valid Downstream Use

The figure is an original conceptual synthesis developed for "Foundation Models for Pharmacoinformatics and QSAR: A Scoping Review of Transfer, Grounding, Calibration, and Chemical Validity." Arrows and grouping

indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, clinical recommendation, regulatory determination, or deployment-ready system.

Review limitations

The review was bounded by its eligibility definitions, selected information sources, journal criteria, publication formats, and terminology. Relevant studies may describe transferable pretraining without using the term "foundation model," while other articles may use the term without operational detail. PRISMA-ScR supports transparent reporting but cannot guarantee exhaustive capture of an evolving interdisciplinary vocabulary [4]. JBI-informed iteration reduced this problem without eliminating it [5].

Cross-study comparison was constrained by incompatible datasets, endpoint definitions, representations, split strategies, metrics, baseline tuning, and validation stages. Aggregate evidence could not be treated as quantitatively commensurable. Activity-cliff analyses further show that global performance can conceal chemically local failures, complicating comparisons based solely on average metrics [21]. The review therefore maps evidence patterns rather than estimating pooled effects or prevalence.

Several recent foundation models have limited independent replication, prospective testing, or evaluation across truly external chemical and organizational contexts. Publication bias may favour successful architectures, while unsuccessful transfer attempts and negative experimental results may remain underreported. Leakage and incomplete workflow reporting also make some strong performance claims difficult to reproduce or interpret independently [37].

Formal risk-of-bias scoring was not used because the review was designed to map heterogeneous concepts, methods, and validation practices rather than estimate comparative effectiveness. This decision preserves broad coverage but prevents a single summary judgement about evidence quality. The resulting taxonomy and evaluation pathway are interpretive syntheses; they should be tested and refined as definitions, benchmarks, external validations, and prospective evidence mature.

CONCLUSION

Foundation models broaden pharmacoinformatics and QSAR by enabling reusable learning across molecular graphs and strings, reactions, proteins, omics, assays, and scientific text. Their scientific value, however, depends less on scale or nomenclature than on the alignment among pretraining data, representation, adaptation, endpoint, chemical domain, uncertainty, and validation design. The scoped evidence supports conditional transfer and increasingly diverse forms of computational grounding, but it also reveals persistent limitations involving dataset overlap, activity cliffs, calibration under shift, synthesizability, causal interpretation, and independent external validation. Foundation-model outputs should therefore be treated as bounded evidence for defined computational tasks rather than as automatic indicators of pharmaceutical utility. Progress requires transparent data lineage, tuned comparators, applicability-domain analysis, local chemical-error testing, uncertainty evaluation, reproducible workflows, and prospective confirmation where decisions carry material consequences.

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. Yang X, Wang Y, Byrne R, Schneider G, Yang SY. Concepts of artificial intelligence for computer-assisted drug discovery. Chem Rev. 2019;119(18):10520-94. doi:10.1021/acs.chemrev.8b00728

3. Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data used for AI in drug discovery. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
4. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467-73. doi:10.7326/M18-0850
5. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. 2020;18(10):2119-26. doi:10.11124/JBIES-20-00167
6. Mendez-Lucio O, Nicolaou CA, Earnshaw B. MolE: A foundation model for molecular graphs using disentangled attention. Nat Commun. 2024;15(1):9431. doi:10.1038/s41467-024-53751-y
7. Chang J, Ye JC. Bidirectional generation of structure and properties through a single molecular foundation model. Nat Commun. 2024;15(1):2323. doi:10.1038/s41467-024-46440-3
8. Xiang H, Zeng L, Hou L, Li K, Fu Z, Qiu Y, et al. A molecular video-derived foundation model for scientific drug discovery. Nat Commun. 2024;15(1):9696. doi:10.1038/s41467-024-53742-z
9. Soares E, Vital Brazil E, Shirasuna V, Zubarev D, Cerqueira R, Schmidt K. An open-source family of large encoder-decoder foundation models for chemistry. Commun Chem. 2025;8(1):193. doi:10.1038/s42004-025-01585-0
10. Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular Transformer: A model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci. 2019;5(9):1572-83. doi:10.1021/acscentsci.9b00576
11. 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). doi:10.1073/pnas.2016239118
12. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616-24. doi:10.1038/s41586-023-06139-9
13. Schuh MG, Boldini D, Sieber SA. Synergizing chemical structures and bioassay descriptions for enhanced molecular property prediction in drug discovery. J Chem Inf Model. 2024;64(12):4640-50. doi:10.1021/acs.jcim.4c00765
14. Li X, Fourches D. Inductive transfer learning for molecular activity prediction: Next-gen QSAR models with MolPMoFiT. J Cheminform. 2020;12(1):27. doi:10.1186/s13321-020-00430-x
15. Morris P, St Clair R, Hahn WE, Barenholtz E. Predicting binding from screening assays with transformer network embeddings. J Chem Inf Model. 2020;60(9):4191-9. doi:10.1021/acs.jcim.9b01212
16. Zhang Z, Bian Y, Xie A, Han P, Zhou S. Can pretrained models really learn better molecular representations for AI-aided drug discovery? J Chem Inf Model. 2024;64(7):2921-30. doi:10.1021/acs.jcim.3c01707
17. Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6(5):525-35. doi:10.1038/s42256-024-00832-8
18. 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
19. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441-51. doi:10.1039/C8SC00148K
20. Yang K, Swanson K, Jin W, Coley CW, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88. doi:10.1021/acs.jcim.9b00237
21. 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
22. Gao W, Coley CW. The synthesizability of molecules proposed by generative models. J Chem Inf Model. 2020;60(12):5714-23. doi:10.1021/acs.jcim.0c00174
23. Genheden S, Thakkar A, Chadimova V, Reymond JL, Engkvist O, Bjerrum E. AiZynthFinder: A fast, robust and flexible open-source software for retrosynthetic planning. J Cheminform. 2020;12(1):70. doi:10.1186/s13321-020-00472-1
24. Thakkar A, Chadimova V, Bjerrum EJ, Engkvist O, Reymond JL. Retrosynthetic accessibility score (RAscore): Rapid machine learned synthesizability classification from AI driven retrosynthetic planning. Chem Sci. 2021;12(9):3339-49. doi:10.1039/D0SC05401A

25. Jimenez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573-84. doi:10.1038/s42256-020-00236-4
26. Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model. 2020;60(6):2697-717. doi:10.1021/acs.jcim.9b00975
27. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60(8):3770-80. doi:10.1021/acs.jcim.0c00502
28. Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW. Evidential deep learning for guided molecular property prediction and discovery. ACS Cent Sci. 2021;7(8):1356-67. doi:10.1021/acscentsci.1c00546
29. Heid E, McGill CJ, Vermeire FH, Green WH. Characterizing uncertainty in machine learning for chemistry. J Chem Inf Model. 2023;63(13):4012-29. doi:10.1021/acs.jcim.3c00373
30. 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
31. 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
32. Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F. A systematic study of key elements underlying molecular property prediction. Nat Commun. 2023;14(1):6395. doi:10.1038/s41467-023-41948-6
33. Sturm N, Mayr A, Le Van T, Chupakhin V, Ceulemans H, Wegner JK, et al. Industry-scale application and evaluation of deep learning for drug target prediction. J Cheminform. 2020;12(1):26. doi:10.1186/s13321-020-00428-5
34. Boldini D, Grisoni F, Kuhn D, Friedrich L, Sieber SA. Practical guidelines for the use of gradient boosting for molecular property prediction. J Cheminform. 2023;15(1):73. doi:10.1186/s13321-023-00743-7
35. van den Maagdenberg HW, Sicho M, Araripe DA, Luukkonen S, Schoenmaker L, Jespers M, et al. QSPRpred: A flexible open-source quantitative structure-property relationship modelling tool. J Cheminform. 2024;16(1):128. doi:10.1186/s13321-024-00908-y
36. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell. 2020;2(7):369-75. doi:10.1038/s42256-020-0197-y
37. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. doi:10.1016/j.patter.2023.100804