Pharmacophore

ISSN-2229-5402

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

MOLECULAR FOUNDATION MODELS IN PHARMACEUTICAL SCIENCES: A CRITICAL REVIEW

Victor Hugo[1]*, Daniel Cruz[1], Javier Salazar[2]

1. ***Department of Intelligent Pharmaceutical Analytics, Faculty of Pharmacy, National University of Colombia, Bogota, Colombia.***
2. ***Department of Computational Drug Sciences, Faculty of Medicine, University of Antioquia, Medellin, Colombia.***

ARTICLE INFO

Received:
19 February 2026
Received in revised form:
28 May 2026
Accepted:
30 May 2026
Available online:
28 June 2026

Keywords: Molecular foundation models, Pharmaceutical machine learning, Data leakage, Pretraining bias, Transfer learning, ADMET

ABSTRACT

Molecular foundation models, pre-trained on millions of chemical structures, are increasingly promoted as a universal solution for pharmaceutical prediction tasks. Their appeal lies in the possibility that large-scale chemical pretraining can reduce dependence on small, noisy, task-specific datasets. Despite their rapid proliferation, critical examination of their pretraining data, leakage risks, transferability, and validation practices remains limited and fragmented. This is problematic because pharmaceutical machine learning is especially vulnerable to hidden similarities between training and test compounds. This critical review evaluates molecular foundation models in pharmaceutical sciences, focusing on pretraining data quality, data leakage, transferability evidence, and validation rigour. It treats reported benchmark performance as a hypothesis requiring scrutiny rather than as sufficient evidence of utility. The review identifies pervasive data biases, frequent over-optimistic evaluation due to leakage, inconsistent evidence on transferability, and a widespread lack of external or prospective validation. These issues are not incidental limitations but structural weaknesses in how many molecular foundation models are developed and assessed. Uncritical adoption of molecular foundation models risks misleading performance claims and may slow, rather than accelerate, pharmaceutical applications. Greater attention to data provenance, split design, uncertainty, and prospective relevance is necessary before such models can be trusted in drug discovery workflows. A set of recommendations is proposed for more robust pretraining, transparent evaluation, and domain-appropriate validation. Molecular foundation models should be judged not by benchmark novelty alone but by their ability to generalize under conditions that resemble pharmaceutical decision-making.

To Cite This Article: Hugo V, Cruz D, Salazar J. Molecular Foundation Models in Pharmaceutical Sciences: A Critical Review. Pharmacophore. 2026;17(3):72-80. https://doi.org/10.51847/5KyoIKlui1

Introduction

The development of pharmaceutical artificial intelligence has shifted from narrowly trained, task-specific models toward large pre-trained representations intended to serve as reusable foundations across many molecular prediction problems. MoleculeNet helped standardize this transition by collecting diverse molecular benchmarks into a common evaluation setting [1], while directed message-passing architectures showed that learned molecular representations could compete strongly with expert fingerprints in property prediction [2]. Early SMILES-based language models then reframed molecular representation learning as a pretraining problem analogous to natural language processing, suggesting that chemical syntax could be mined at scale before fine-tuning on smaller pharmaceutical datasets [3]. This shift has been intellectually productive, but it has also encouraged a tendency to treat scale and pretraining as intrinsic virtues rather than as design choices that must be validated.

The central promise of molecular foundation models is that a single representation learned from large molecular corpora can be adapted to ADMET prediction, binding affinity estimation, virtual screening, molecular generation, and related drug discovery tasks. Self-supervised graph-transformer models [4], BERT-like molecular encoders [5], chemical language models [6], and transformer-decoder generators such as MolGPT [7] all express this ambition in different architectural forms. In pharmaceutical settings, this promise is attractive because many endpoints are data-poor, expensive to measure, and noisy, making transfer from broader chemical pretraining seem practically valuable. Yet the same features that make transfer appealing also make evaluation fragile, because modest improvements on retrospective datasets may not translate into better experimental decisions.

Corresponding Author: Victor Hugo; Department of Intelligent Pharmaceutical Analytics, Faculty of Pharmacy, National University of Colombia, Bogota, Colombia. E-mail: victor.hugo@gmail.com.

The rapid growth of molecular foundation models has exposed concerns that are insufficiently addressed by headline benchmark comparisons. Studies of activity cliffs show that small structural changes can produce large biological effects that remain difficult for learned molecular representations to capture [8], while distribution-shift evaluations emphasize that models often degrade when test compounds differ meaningfully from training compounds [9]. More broadly, leakage has been identified as a reproducibility threat across machine-learning-based science [10], and molecular data are particularly susceptible because analog series, duplicated measurements, and shared assay origins can silently connect training and test examples. These problems challenge the assumption that pretraining automatically improves generalization.

This review critically evaluates molecular foundation models in pharmaceutical sciences through four linked weaknesses: pretraining data quality, data leakage, transferability, and validation. It draws on foundational benchmarks and architectures [1, 2], language and graph pretraining studies [3-7], contrastive and multimodal representation learning [11-13], and recent critiques of benchmarking and splitting [9, 14-16]. The aim is not to dismiss molecular foundation models, which remain scientifically important, but to distinguish robust evidence from optimistic claims. Sections 2-5 examine the emergence of the field, the composition and bias of pretraining data, leakage and over-optimistic evaluation, and the contested evidence for transferability.

The Rise of Molecular Foundation Models

Architecture, Scale, and Pre-Training Objectives

Molecular foundation models include SMILES language models, graph neural networks, graph transformers, contrastive models, generative decoders, and 3D equivariant systems, each encoding different assumptions about what molecular information matters. SMILES Transformer treated strings as chemical sentences [3], Mol-BERT adapted masked language modeling to molecular tokens [5], MolGPT used autoregressive generation for chemical design [7], and self-supervised graph transformers learned from molecular graph perturbations and contextual prediction [4]. Later work expanded the design space through contrastive graph learning [11], knowledge-enhanced molecular pretraining [12], functional-group-aware encoders [17], and 3D conformation-aware representation learning [18]. The diversity of objectives is a strength, but it complicates comparison because improvements may reflect architecture, data curation, task overlap, or split design rather than a general foundation-model effect.

From Benchmark to Pharma: The Claims of Universality

The universality narrative rests on the idea that molecular pretraining can produce representations useful across therapeutic discovery, from early screening to safety assessment and design. Reviews and perspective articles have framed foundation models as part of a broader artificial intelligence platform for therapeutic science [19], and antimicrobial discovery work has suggested that pre-trained molecular representations can support practical discovery tasks [20]. However, systematic studies caution that apparent gains are not uniform and may depend strongly on the downstream endpoint, fine-tuning regime, and similarity between pretraining and evaluation data [14, 21]. The phrase “foundation model” can therefore obscure a critical distinction between reusable representation learning and genuine cross-domain pharmaceutical utility.

The Industrial and Academic Adoption Trend

Academic adoption has accelerated through open benchmarks, shared software, and increasingly sophisticated molecular pretraining studies, while industrial interest is driven by the prospect of reusing large internal chemical archives across many programs. Chemical language models demonstrated that large corpora could organize chemical space in useful ways [6], and graph and multimodal approaches have extended this logic toward richer molecular structure and biological context [12, 13]. Yet the adoption trend is ahead of the evidentiary base: many models are evaluated on familiar retrospective benchmarks rather than on temporally separated industrial project data or prospective assays. This mismatch means that pharmaceutical integration is often justified by plausibility and benchmark success rather than by rigorous evidence of decision impact.

Pretraining Data: Sources, Biases, and Challenges

Composition of Popular Pretraining Corpora

Popular molecular pretraining corpora commonly draw from PubChem, ChEMBL, ZINC, and related collections, which are large but not chemically neutral. MoleculeNet exemplified the centrality of benchmark datasets derived from public molecular and bioactivity sources [1], while later pretraining studies scaled this approach by learning from broad structure collections before fine-tuning on pharmaceutical endpoints [3-5]. These sources overrepresent synthesizable, drug-like, organic small molecules and underrepresent natural products, organometallics, highly flexible macrocycles, inorganic drugs, biologics, and poorly annotated failed compounds. Consequently, a model can appear broadly chemical while being trained on a biased subset of the chemical universe most convenient for database assembly.

Consequences of Data Bias

Data bias matters because representation learning can amplify the assumptions embedded in pretraining corpora rather than correct them. A model trained mostly on drug-like organic molecules may perform well on benchmark ADMET tasks yet struggle with natural products, peptides, metal-containing therapeutics, covalent fragments, or modalities outside conventional small-molecule space. Activity-cliff analyses show that models can fail even within familiar chemical neighborhoods when

minor structural changes cause major potency shifts [8], and distribution-shift studies indicate that generalization weakens when test compounds depart from training distributions [9]. These findings suggest that pretraining breadth should be assessed by chemical and biological coverage, not merely by the number of molecules.

Data Quality and Artefacts

Pretraining scale does not eliminate the consequences of noisy structures, inconsistent standardization, assay artefacts, duplicated compounds, or conflicting labels. Directed message-passing models and systematic property-prediction studies demonstrate that representation quality is closely tied to the reliability of the molecular and endpoint data used for learning [2, 21]. When pretraining includes salts, tautomers, stereochemical ambiguities, unfiltered reactive species, or assay-biased compounds, the model may encode artefactual regularities that later appear as predictive signal. This is especially concerning for pharmaceutical applications, where false confidence in a learned embedding can guide resource-intensive experimental decisions.

Data Leakage and Over-Optimistic Evaluation

The Insidious Nature of Leakage in Molecular Benchmarks

Data leakage in molecular machine learning is often subtle because chemical series contain near-neighbor compounds whose shared scaffolds, assay histories, or duplicate records can bridge training and test sets. Random splits in MoleculeNet-style evaluations [1] may test interpolation within familiar analog series rather than extrapolation to new chemistry, while scaffold and temporal splits attempt to impose more realistic separation. Leakage is not merely a technical nuisance; it can make a model appear to have learned transferable pharmacological structure when it has mostly learned local chemical similarity. The broader reproducibility literature warns that such leakage can create inflated claims even when modeling code is technically correct [10].

Case Studies Where Leakage Inflated Performance

Several molecular machine learning studies have shown that re-evaluation under stricter splitting or distribution-shift settings can substantially weaken reported conclusions. Scaffold-split analyses in virtual screening argue that even commonly used scaffold-based protocols may overestimate performance if they fail to remove deeper similarity or target-related leakage [15]. DataSAIL was introduced precisely because conventional splitting can leave information pathways between train and test sets in structured biological and chemical data [16]. Together with studies exposing distribution-shift fragility [9], this evidence indicates that benchmark superiority should be interpreted cautiously unless leakage-resistant evaluation is built into the study design.

The Design of Leakage-Proof Benchmarks

Leakage-proof benchmarking requires explicit control over molecular similarity, biological relatedness, temporal availability, duplicate measurements, and preprocessing history. DataSAIL represents a step toward systematic split design by treating leakage avoidance as an optimization problem rather than a post hoc reporting choice [16], while scaffold-split critiques show that nominally rigorous splits can still be inadequate [15]. For molecular foundation models, the problem is even harder because the pretraining corpus itself may contain downstream test molecules or close analogs. A credible benchmark must therefore disclose pretraining sources, remove overlapping structures where possible, and evaluate temporal or external generalization rather than only random or scaffold holdouts.

Table 1 provides a critical appraisal framework for judging whether molecular foundation model claims are supported by evidence that is robust enough for pharmaceutical interpretation.

Table 1. Critical Appraisal Framework for Molecular Foundation Model Claims in Pharmaceutical Sciences

Foundation-model claim	Evidence needed to support the claim	Main threat to validity	Stronger appraisal question	Implication for interpretation
Large-scale molecular pretraining produces broadly useful chemical representations	Clear documentation of pretraining corpus composition, molecule standardization, duplicate removal, and chemical-space coverage	Pretraining corpora may overrepresent drug-like organic small molecules while underrepresenting natural products, macrocycles, organometallics, biologics, failed compounds, and rare chemotypes	Does the pretraining corpus cover the chemical and biological domain in which the model is later evaluated?	Scale alone should not be interpreted as representational breadth; data provenance is part of model validity
Benchmark superiority demonstrates pharmaceutical utility	Performance gains across leakage-resistant splits, strong baselines, external datasets, and decision-relevant endpoints	Random or weak scaffold splits may reward interpolation within familiar analog series rather than extrapolation to new chemistry [1, 15, 16]	Would the model have made a useful prediction before the relevant experimental or project decision was known?	Retrospective benchmark gains should be treated as preliminary evidence, not proof of deployment readiness
Pretraining reduces dependence on small	Robust improvement in low-data settings after	Apparent improvement may reflect overlap between pretraining data	Is the performance gain preserved when close	Low-data benefit is plausible but must be

labelled pharmaceutical datasets	comparison with well-tuned task-specific models and uncertainty-aware methods	and downstream test data or similarity between analog series	analogs, duplicates, shared assay origins, and temporal overlap are controlled?	separated from hidden leakage and dataset familiarity
Molecular foundation models transfer across ADMET, binding, screening, and generation tasks	Endpoint-specific evidence showing when transfer helps, is neutral, or harms performance	Transfer may fail when downstream outcomes depend on assay-specific biology, target context, stereochemistry, formulation effects, or rare mechanisms [14, 21]	Which molecular, biological, and experimental conditions define successful transfer?	Transferability should be reported as conditional and endpoint-specific rather than universal
Learned molecular embeddings capture meaningful chemical structure	Activity-cliff sensitivity, scaffold extrapolation, conformational awareness, and chemically interpretable representation analysis	Embeddings may encode superficial similarity or dataset artefacts rather than causal pharmacological structure [8, 9]	Can the representation distinguish structurally similar molecules with sharply different biological effects?	Strong average performance may conceal poor local discrimination in medicinal chemistry-relevant regions
3D, graph, contrastive, or multimodal architectures improve generalization	Fair comparison across architectures under identical data splits, preprocessing, baselines, and validation protocols	Architectural novelty may be confounded with data curation, split design, pretraining overlap, or evaluation selection	Does the architectural advantage persist under leakage-resistant and temporally separated evaluation?	Architecture should be judged through validation realism, not methodological sophistication alone
Uncertainty estimation makes model predictions safer	Calibrated uncertainty under distribution shift, novel scaffolds, rare chemotypes, and external datasets	Uncertainty cannot compensate for flawed validation design or hidden overlap in training and test data [9, 22]	Does uncertainty increase meaningfully when the model encounters unfamiliar chemistry or assay context?	Uncertainty is useful only when tested under realistic extrapolative conditions
Open models and shared benchmarks ensure reproducibility	Public model weights, preprocessing scripts, pretraining data lineage, split files, hyperparameters, and negative results	Reproducibility remains limited when data lineage, overlap checks, and split construction are opaque [10, 23, 24]	Could an independent group reconstruct the full evaluation and audit leakage pathways?	Reproducibility requires disclosure of the complete modelling and data pipeline, not only code release

Transferability Across Tasks and Domains

Evidence for Positive Transfer

Positive transfer has been reported when pretraining captures chemical patterns relevant to downstream endpoints and when fine-tuning data are limited but aligned with the pretraining distribution. SMILES Transformer showed that learned molecular fingerprints could support low-data drug discovery tasks [3], and contrastive graph representation learning improved molecular property prediction by encouraging invariant representations across molecular views [11]. Functional-group-aware pretraining [17] and conformation-aware pretraining [18] further suggest that chemically meaningful inductive biases can strengthen downstream performance. Nevertheless, these successes are strongest evidence for conditional usefulness, not for universal transfer across pharmaceutical problems.

Evidence for Negative or Absent Transfer

Evidence against automatic transfer is increasingly important because larger or more sophisticated pretraining does not consistently outperform well-tuned task-specific baselines. A systematic study of key elements in molecular property prediction found that data splitting, architecture, and training protocol can be as decisive as pretraining itself [21], while a direct critique asked whether pretrained models truly learn better molecular representations for AI-aided drug discovery [14]. Negative or absent transfer may occur when downstream endpoints depend on assay-specific biology, rare chemotypes, protein context, stereochemistry, formulation effects, or mechanisms absent from the pretraining objective. In such cases, pretraining can act as a prior that is mismatched to the task rather than as a source of useful generalization.

Understanding When Transfer Works

Transfer appears most plausible when the pretraining corpus, pretraining objective, molecular representation, and downstream endpoint share meaningful structure. Knowledge-graph-enhanced molecular contrastive learning attempted to align molecular representations with functional and biological context [12], while multimodal self-supervised learning pursued complementary molecular views rather than relying on a single encoding [13]. These approaches implicitly recognize that chemical syntax alone may be insufficient for tasks involving target biology, pharmacokinetics, toxicity mechanisms, or experimental conditions. The field therefore needs transfer studies that map when representations help, when they are neutral, and when they harm, rather than reporting only favorable aggregate results.

Validation Practices and Reproducibility

The Retrospective Validation Trap

Molecular foundation models are still evaluated predominantly through retrospective benchmarks, internal cross-validation, and static train-test splits that cannot fully represent pharmaceutical deployment. Evidential deep learning has highlighted the importance of uncertainty estimates for molecular prediction [22], but uncertainty alone cannot rescue a validation design that has already allowed chemical or assay information to leak into evaluation. Distribution-shift work shows that models can perform acceptably under familiar retrospective conditions yet become unreliable when exposed to new chemical regions [9]. The central weakness is therefore not simply insufficient model calibration, but the continued use of validation settings that reward interpolation while implying extrapolative utility.

Prospective and Real-World Validation: The Rare Exception

Prospective validation remains rare, which is troubling because drug discovery decisions are prospective by nature: a model recommends compounds before their properties are measured. Pre-trained molecular representations have supported antimicrobial discovery, offering one of the more compelling examples of moving beyond retrospective benchmark claims [20]. However, such studies are still exceptions rather than the dominant evidence base, and even successful prospective screens may not generalize to ADMET, toxicity, formulation, or target-specific medicinal chemistry campaigns. The field needs more studies that report not only whether a foundation model nominated active compounds, but also whether it improved decision-making relative to simpler baselines, expert rules, and project-specific models.

Table 2 shows the current evidence landscape for prospective validation in machine learning–driven drug discovery and highlights the gap between benchmark performance and real-world decision impact.

Table 2. Evidence landscape and limitations of prospective validation in machine learning–guided drug discovery

Category	Description	Strength of Evidence	Key Limitation	Generalization Risk
Retrospective benchmarking studies	Model performance evaluated on historical datasets (e.g., cross-validation, held-out sets)	High volume, standardized metrics	Does not reflect real experimental decision-making	High—often overestimates real-world performance
Pre-trained molecular representation models in antimicrobial discovery	Use of large-scale pretraining to propose novel antimicrobial candidates validated experimentally	Demonstrated successful wet-lab validation in select cases	Narrow domain focus and selective reporting of successes	Moderate to high—uncertain transfer to other therapeutic areas
Prospective screening campaigns	Models used to prioritize compounds before experimental testing	Direct experimental validation of predictions	Typically small-scale and resource-constrained	Moderate—may not scale across projects or targets
ADMET, toxicity, and formulation contexts	Application of models beyond target binding (safety, pharmacokinetics, formulation)	Increasing interest but limited robust prospective studies	Complex biology and multi-factor dependencies	High—poor transferability from discovery models
Decision-impact evaluation frameworks	Comparison against baselines, expert rules, and project-specific models	Emerging but still rare	Lack of standardized reporting of decision improvement	Unknown—depends on rigorous comparative design

Openness, Reproducibility, and Code/Data Availability

Reproducibility is uneven across molecular foundation model research because model weights, pretraining corpora, preprocessing scripts, and negative results are often incompletely disclosed. The broader critique of leakage and reproducibility in machine-learning-based science is directly relevant here, because a model cannot be independently audited if its data lineage and split construction are opaque [10]. Recent reviews of molecular representation learning emphasize the growing diversity of methods but also reveal how difficult it has become to compare models fairly across datasets, objectives, and reporting standards [23, 24]. Without stronger openness norms, confidence in reported performance will depend too heavily on trust in authors' undocumented choices.

Emerging Solutions and Future Directions

Better Pretraining Data Curation and Augmentation

Better foundation models will require better pretraining data, not merely larger datasets. Knowledge-guided pretraining attempts to inject chemically and biologically meaningful structure into representations [25], while multimodal pretraining aims to combine complementary molecular information rather than relying only on SMILES or graph topology [26]. Self-supervised learning from tandem mass spectra further broadens the concept of molecular representation by connecting structures to experimental analytical data [27]. These directions are promising, but they will remain vulnerable to bias unless data curation explicitly addresses chemical coverage, provenance, standardization, assay artefacts, and the inclusion of underrepresented molecular classes.

Next-Generation Benchmarks and Leakage-Aware Evaluation

Next-generation benchmarks should make leakage resistance a design principle rather than an optional sensitivity analysis. DataSAIL demonstrates that split construction can be treated as a formal method for reducing information leakage across structured datasets [16], and scaffold-split critiques show why traditional molecular benchmarks may still be over-optimistic [15]. Benchmark design should also consider pretraining overlap, temporal separation, external validation cohorts, and endpoint-specific biological context. In this respect, the field should move away from leaderboard-style comparisons and toward evaluation protocols that ask whether a model would have been useful at the time a pharmaceutical decision was actually made.

A Critical Assessment Framework

Proposed Dimensions for Evaluating Molecular Foundation Models

A credible assessment framework for molecular foundation models should examine pretraining data, leakage resistance, transferability evidence, validation rigour, uncertainty handling, and reproducibility together rather than in isolation. MolE and related graph foundation models illustrate how architectural advances can be substantial [28], but architecture should not be allowed to distract from whether the model was evaluated under realistic chemical separation. Equivariant 3D transformers add valuable geometric inductive bias [29], yet even 3D pretraining can be overinterpreted if conformer generation, benchmark overlap, or downstream assay context are not carefully controlled. Evaluation should therefore ask not only what the model encodes, but also what evidence shows that encoding survives contact with new chemistry and new decisions.

Minimum Standards for Trustworthy Deployment in Pharma

Before deployment in pharmaceutical workflows, a molecular foundation model should demonstrate transparent pretraining provenance, explicit duplicate and analog control, leakage-aware splitting, task-relevant uncertainty, and external or temporal validation. Optimal masked language modeling for molecules shows that even seemingly technical details of pretraining objective design can influence representation quality [30], while systematic critiques question whether pretraining gains remain robust after fair comparison with strong baselines [14]. A trustworthy deployment case should include evidence of transfer across chemical series, not only across benchmark tasks, and should report failures as carefully as successes. Without these minimum standards, foundation-model language risks becoming a marketing label rather than a scientific claim.

Table 3 consolidates minimum evidence standards that should be met before molecular foundation models are treated as trustworthy tools for pharmaceutical decision-making.

Table 3. Minimum Evidence Standards for Trustworthy Molecular Foundation Models Before Pharmaceutical Deployment

Assessment domain	Minimum evidence standard	Preferred validation approach	Red flags indicating insufficient evidence	Practical pharmaceutical consequence
Pretraining provenance	Full disclosure of databases, version dates, molecular standardization rules, duplicate handling, and exclusion criteria	Pretraining data audit with chemical-class coverage analysis and overlap checks against downstream tasks	Vague descriptions such as “large public chemical database” without data versioning or curation details	Users cannot determine whether the model has learned general chemistry or merely memorized familiar chemical regions
Chemical-space coverage	Evidence that the model has been assessed across conventional and underrepresented molecular classes	Coverage maps by scaffold, molecular weight, stereochemistry, modality, charge state, and chemical class	Strong claims based only on drug-like small-molecule benchmarks	Deployment may fail for natural products, covalent fragments, macrocycles, peptides, metal-containing drugs, or emerging modalities
Leakage control	Explicit removal or accounting for duplicates, near-neighbors, analog series, assay-origin overlap, and pretraining-test overlap	Leakage-aware splitting using scaffold, temporal, external, and optimization-based split strategies such as DataSAIL-type approaches [15, 16]	Exclusive reliance on random splits or poorly justified scaffold splits	Reported accuracy may reflect chemical familiarity rather than true generalization
Transferability evidence	Endpoint-specific testing showing where pretraining improves, does not change, or worsens performance	Cross-endpoint and cross-domain studies with strong task-specific baselines and negative-transfer reporting	Only favorable aggregate benchmark tables are reported	Teams may overuse foundation models in endpoints where task-specific models or expert rules are more reliable
Baseline comparison	Comparison against strong classical fingerprints, directed message-passing models, task-specific neural networks, and simple expert-informed baselines	Matched training budgets, identical splits, and transparent hyperparameter tuning	Weak or outdated baselines used to exaggerate foundation-model gains	Benchmark novelty may be mistaken for real performance improvement
Distribution-shift robustness	Performance tested on new scaffolds, temporally separated compounds, external institutions	External and temporal validation with activity-cliff and out-of-domain sensitivity analyses [8, 9]	High performance only under familiar retrospective benchmark settings	Model predictions may be unreliable for novel medicinal chemistry programs

	or assays, and chemically distant regions			
Uncertainty and failure analysis	Calibration tested under chemical novelty, assay uncertainty, rare endpoints, and shifted distributions	Reliability diagrams, abstention analysis, conformal prediction, or uncertainty-aware decision thresholds	Uncertainty reported only on internal validation data	Teams may treat confident but unsupported predictions as actionable
Prospective relevance	Evidence that the model improves compound prioritization, experimental efficiency, or decision quality before outcomes are known	Prospective assay validation, simulated temporal deployment, or blinded external challenge	Retrospective success presented as equivalent to drug discovery utility	The model may not reduce experimental burden or improve project decisions
Reproducibility and auditability	Public or auditable model weights, code, split files, preprocessing scripts, and reporting of failed evaluations	Independent replication or controlled benchmark re-analysis	Missing split files, unavailable pretraining data, undisclosed preprocessing, selective reporting	Performance claims remain difficult to verify and unsafe to generalize
Governance for deployment	Clear statement of intended use, decision boundaries, human oversight, update policy, and monitoring plan	Model cards or evidence dossiers tailored to pharmaceutical decision contexts	"Foundation model" used as a broad marketing label without deployment constraints	The model may be applied outside the conditions under which it was evaluated

Results and Discussion

Synthesis of the Main Weaknesses

The main weaknesses of molecular foundation models converge around a single theme: reported generality often rests on evidence that is narrower than the claim being made. Public benchmarks enabled rapid progress [1], but they also encouraged repeated evaluation on familiar datasets whose limitations are now apparent. Pretraining can encode useful chemical regularities [3, 4], yet biased corpora, hidden overlap, and activity cliffs can make those regularities fragile in the settings that matter most [8, 9]. The field's central challenge is therefore to distinguish representation learning that is genuinely transferable from representation learning that is merely well matched to existing benchmarks.

Figure 1 synthesizes how molecular foundation model claims must pass through data provenance, leakage control, transferability testing, validation realism, and reproducibility assessment before they can support pharmaceutical decision-making.

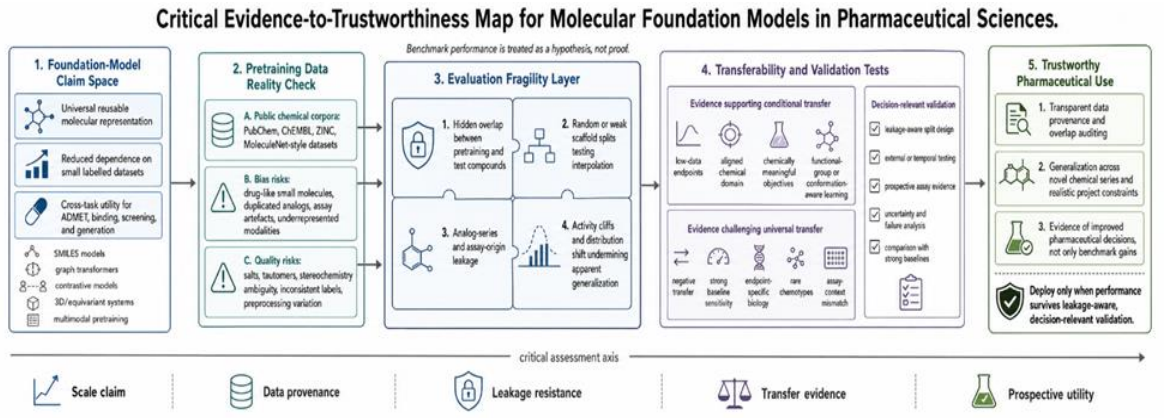

Figure 1. Critical Evidence-to-Trustworthiness Map for Molecular Foundation Models in Pharmaceutical Sciences.

The Gap between Academic and Industrial Realities

Academic evaluations often reward methodological novelty, marginal benchmark gains, and broad task tables, whereas industrial discovery requires reliability under sparse, shifting, confidential, and project-specific conditions. Therapeutic foundation-model perspectives appropriately emphasize the transformative potential of artificial intelligence [19], but pharmaceutical programs must also manage assay drift, synthesis constraints, liability series, intellectual-property boundaries, and decision costs. A model that performs well on curated public endpoints may still be unhelpful when deployed on a chemically novel series with uncertain assay reproducibility. This gap explains why benchmark excellence should be treated as an early screening criterion, not as evidence of operational value.

A Way Forward

Researchers should report pretraining data provenance, overlap checks, split rationale, external validation, uncertainty behaviour, and baseline sensitivity as standard components of molecular foundation model papers. Journal editors and reviewers should require leakage-aware evaluation when claims of generalization or transferability are made, drawing on evidence that conventional random or scaffold splits can mislead [15, 16]. Funders and community benchmark organizers should support prospective and temporally anchored evaluations, especially for ADMET and safety endpoints where retrospective optimism can be costly. The most useful future models will likely be those that combine strong representation learning with transparent data governance, careful validation, and honest reporting of limits.

Strengths and Limitations of this Review

This review's strength is its critical focus on the evidentiary foundations of molecular foundation models rather than on architectural novelty alone. By integrating benchmark papers [1], early language and graph pretraining studies [3, 4], generative and contrastive approaches [7, 11], uncertainty-aware modeling [22], leakage critiques [10, 15, 16], and recent reviews of molecular representation learning [23, 24], it highlights recurring weaknesses across otherwise diverse model families. Its limitation is that the field is moving rapidly, especially in multimodal and 3D molecular pretraining [26, 29], so any critical synthesis may lag behind the newest unpublished or proprietary systems. The review also emphasizes small-molecule pharmaceutical applications and therefore gives less attention to protein language models, biologics, clinical deployment, and fully integrated laboratory automation.

Conclusion

Molecular foundation models hold substantial promise for pharmaceutical sciences because they offer a route to reusable chemical representations in settings where labelled data are often sparse, noisy, or expensive. However, their current evidentiary base is weakened by biased pretraining corpora, hidden data leakage, inconsistent transferability, and validation practices that often remain too retrospective. These weaknesses do not invalidate the field, but they do require a more cautious interpretation of performance claims.

The critical gap between reported benchmark performance and demonstrated pharmaceutical utility must now become a central concern. A model that succeeds on public datasets has not necessarily shown that it can guide medicinal chemistry, reduce experimental burden, or improve decisions under real project constraints. Pharmaceutical usefulness must be demonstrated under conditions that resemble the timing, uncertainty, and novelty of actual drug discovery.

The field should adopt rigorous, leakage-conscious benchmarking, test transferability more systematically, and require external, temporal, or prospective validation whenever strong claims are made. Pretraining data should be documented with the same seriousness as model architecture, and evaluations should include strong baselines, uncertainty assessment, and explicit failure analysis. Negative transfer, poor extrapolation, and irreproducible gains should be reported as scientifically valuable findings rather than treated as inconvenient exceptions.

A community-wide commitment to transparency and realism is essential if molecular foundation models are to mature from impressive computational artefacts into trustworthy pharmaceutical tools. Their future value will depend less on whether they are called foundation models and more on whether they can withstand rigorous, leakage-aware, and decision-relevant validation. The next phase of molecular artificial intelligence should reward models that are not only larger or more elegant, but also more honest, reproducible, and useful.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pande VS. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30.
2. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88.
3. Honda S, Shi S, Ueda HR. SMILES transformer: pre-trained molecular fingerprint for low data drug discovery. arXiv:1911.04738 [Preprint]. 2019.
4. Rong Y, Bian Y, Xu T, Xie W, Wei Y, Huang W, et al. Self-supervised graph transformer on large-scale molecular data. Adv Neural Inf Process Syst. 2020;33:12559-71.
5. Li J, Jiang X. Mol-BERT: an effective molecular representation with BERT for molecular property prediction. Wirel Commun Mob Comput. 2021;2021:7181815.

6. Skinnider MA, Stacey RG, Wishart DS, Foster LJ. Chemical language models enable navigation in sparsely populated chemical space. Nat Mach Intell. 2021;3(9):759-70.
7. Bagal V, Aggarwal R, Vinod PK, Priyakumar UD. MolGPT: molecular generation using a transformer-decoder model. J Chem Inf Model. 2022;62(9):2064-76.
8. 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.
9. Fooladi H, Vu TN, Mathea M, Kirchmair J. Evaluating machine learning models for molecular property prediction: performance and robustness on out-of-distribution data. J Chem Inf Model. 2025;65(19):9871-91.
10. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804.
11. 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.
12. Fang Y, Zhang Q, Zhang N, Chen Z, Zhuang X, Shao X, et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. Nat Mach Intell. 2023;5(5):542-53.
13. Shen A, Yuan M, Ma Y, Du J, Wang M. Complementary multi-modality molecular self-supervised learning via non-overlapping masking for property prediction. Brief Bioinform. 2024;25(4):bbae256.
14. 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.
15. Guo Q, Hernandez-Hernandez S, Ballester PJ. Scaffold splits overestimate virtual screening performance. In: Proceedings of the International Conference on Artificial Neural Networks. Cham: Springer; 2024. p. 58-72.
16. Joeres R, Blumenthal DB, Kalinina OV. Data splitting to avoid information leakage with DataSAIL. Nat Commun. 2025;16(1):3337.
17. Li B, Lin M, Chen T, Wang L. FG-BERT: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. Brief Bioinform. 2023;24(6):bbad398.
18. Qiao J, Jin J, Wang D, Teng S, Zhang J, Yang X, et al. A self-conformation-aware pre-training framework for molecular property prediction with substructure interpretability. Nat Commun. 2025;16(1):4382.
19. 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.
20. Olayo-Alarcon R, Amstalden MK, Zannoni A, Bajramovic M, Sharma CM, Brochado AR, et al. Pre-trained molecular representations enable antimicrobial discovery. Nat Commun. 2025;16(1):3420.
21. 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.
22. 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.
23. Sheshanarayana R, You F. Molecular representation learning: cross-domain foundations and future frontiers. Digit Discov. 2025;4(9):2298-335.
24. 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.
25. Li H, Zhang R, Min Y, Ma D, Zhao D, Zeng J. A knowledge-guided pre-training framework for improving molecular representation learning. Nat Commun. 2023;14(1):7568.
26. Wang X, Wang C, Ji B, Wang J, Zheng M, Song L, et al. Multimodal pre-training models of molecular representation for drug discovery. Natl Sci Rev. 2026;13(1):nwaf495.
27. Bushuiev R, Bushuiev A, Samusevich R, Brungs C, Sivic J, Pluskal T. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. Nat Biotechnol. 2025;43(5):712-25.
28. Méndez-Lucio O, Nicolaou CA, Earnshaw B. MolE: a foundation model for molecular graphs using disentangled attention. Nat Commun. 2024;15(1):9431.
29. Jiao R, Kong X, Zhang L, Yu Z, Ren F, Tan W, et al. An equivariant pretrained transformer for unified 3D molecular representation learning. Nat Commun. 2026;17(2):2456.
30. Krüger FP, Österbacka N, Kabeshov M, Engkvist O, Tetko IV. MolEncoder: towards optimal masked language modeling for molecules. Digit Discov. 2025;4(12):3552-66.