Available online **www.ijpras.com**

International Journal of Pharmaceutical Research & Allied Sciences, 2025, 14(1):46-57
https://doi.org/10.51847/DOiTQHEsxY

Review Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Pharmacoinformatics after Deep Learning: A State-of-the-Art Review of Molecular Representations, Benchmark Design, and Generalization across Chemical Space

Youssef Hariri[1]*, Hassan Nasser[1], Fatima Zahra[2]

[1]Department of Deep Learning Pharmacoinformatics, Faculty of Pharmacy, Mohammed V University, Rabat, Morocco.
[2]Department of Molecular Representations and Generalization, Faculty of Pharmacy, University of Casablanca, Casablanca, Morocco.

***Email:** youssef.hariri@um5.ac.ma*

ABSTRACT

Deep learning has expanded pharmacoinformatics from models built primarily on predefined descriptors and fingerprints toward architectures that learn representations from molecular graphs, strings, three-dimensional structures, assay context, and multiple aligned modalities. This expansion has produced increasingly flexible modelling systems, but it has also encouraged a misleading equation between representational complexity, benchmark performance, and pharmaceutical usefulness. This state-of-the-art review examines the current technical frontier in molecular representation learning, benchmark design, and generalization across chemical space. It organizes descriptor-based, fingerprint-based, graph, sequence, geometric, and multimodal approaches according to the information they encode, the assumptions they impose, and the pharmaceutical questions they can reasonably address. The review also compares supervised, self-supervised, contrastive, transfer-learning, and foundation-model strategies while distinguishing retrospective prediction from externally tested, prospective, experimental, or operational evidence. Particular attention is given to benchmark construction, scaffold and temporal splitting, activity cliffs, out-of-distribution evaluation, predictive calibration, interpretability, assay shift, and reproducibility. The reviewed evidence indicates that learned representations can improve selected molecular-property and bioactivity tasks, yet their relative performance remains dependent on endpoint definition, data curation, chemical-space coverage, split strategy, model tuning, and comparator strength. Richer representations do not automatically yield better generalization, mechanistic understanding, or decision value. Research priorities therefore include assay-aware representation learning, prospective and cross-organizational validation, calibrated uncertainty under distribution shift, chemically meaningful stress testing, transparent reporting, and evaluation frameworks connected to intended pharmaceutical use. The central conclusion is that post-deep-learning pharmacoinformatics should be judged not by representational novelty alone, but by evidence that a model remains informative when chemistry, assays, time, and decision context change.

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

Received: 03 November 2024; **Revised:** 11 February 2025; **Accepted:** 12 February 2025

INTRODUCTION

Machine learning now contributes to molecular-property prediction, virtual screening, target-related modelling, toxicity estimation, pharmacokinetic assessment, and other stages of pharmaceutical research. Its expansion has

© 2025 The International Journal of Pharmaceutical Research and Allied Sciences (IJPRAS). **Open Access** - This article is under the **CC BY NC SA** license (https://creativecommons.org/licenses/by-nc-sa/4.0).

increased the scale and diversity of computational decision support, but the evidentiary maturity of these applications remains uneven across endpoints and development contexts [1].

The dominant narrative often presents deep learning as a transition from manually designed chemical features to autonomous molecular understanding. This framing is insufficient because algorithmic novelty does not itself demonstrate improved scientific decisions, better compounds, or more reliable translation. Pharmaceutical impact depends on data relevance, comparator quality, validation design, uncertainty management, and integration with experimental evidence [2].

Chemical and biological datasets are not neutral collections of molecular facts. They reflect assay protocols, selection practices, publication patterns, organizational priorities, measurement error, inactive-compound definitions, and incomplete coverage of chemical space. Models may therefore learn dataset-specific regularities that resemble general chemical knowledge under familiar benchmarks but weaken when assays, compound series, or decision settings change [3].

This state-of-the-art review evaluates pharmacoinformatics after deep learning through three linked questions: what molecular representations currently define the technical frontier, how benchmark design shapes comparative claims, and what evidence supports generalization beyond familiar chemical space. The review compares major method families without treating one architecture as universally superior and develops an evidence-bounded research agenda rather than a validated readiness framework [4].

Review scope, search approach, and technology taxonomy

The review used a structured contemporary search focused on pharmacoinformatics, QSAR, molecular representation, applicability domains, external validation, dataset shift, calibration, uncertainty, interpretability, foundation models, and chemical-space generalization. Eligible sources were peer-reviewed journal articles with verifiable bibliographic records and direct relevance to a method, benchmark, validation practice, pharmaceutical task, limitation, or emerging research direction. Representation families were charted according to encoded information, dimensionality, invariance, learnability, and task compatibility [5].

The synthesis was designed as a state-of-the-art review rather than a systematic review or meta-analysis. It therefore identifies the technical frontier, compares influential method families, and evaluates the strength and boundaries of their principal claims without asserting exhaustive coverage. Predictive modelling and molecular generation were considered distinct evidence domains because they use different targets, outputs, metrics, and validation expectations [6]. Their results should consequently not be interpreted through a single definition of model success [7].

The technology taxonomy separates predefined descriptors and fingerprints from learned graph, sequence, geometric, and multimodal representations. It also distinguishes representation from learning objective, because the same molecular encoding may support supervised prediction, self-supervised pretraining, contrastive learning, or transfer learning. Graph machine learning was treated as a family spanning molecular topology, molecular interactions, biological networks, and heterogeneous biomedical relations rather than as one uniform architecture [8].

Figure 1 presents the original conceptual synthesis for evolution of molecular representations in pharmacoinformatics, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

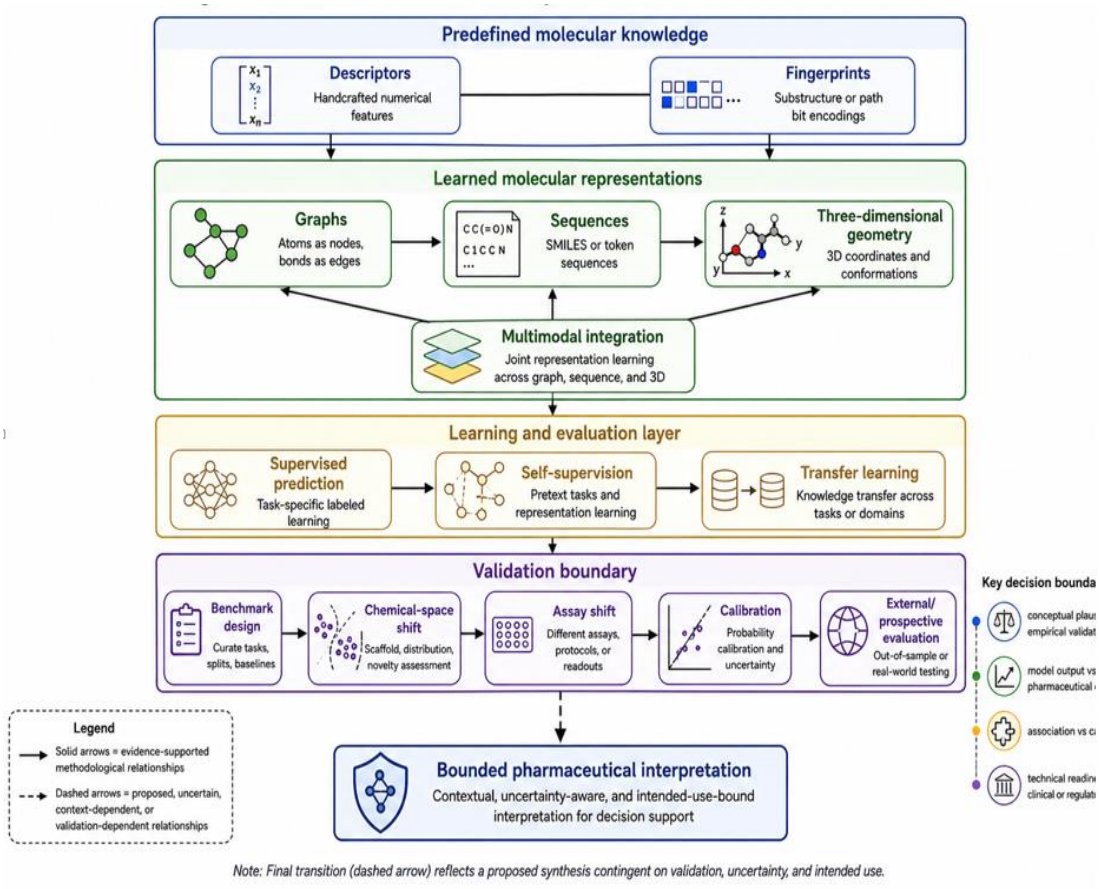

Figure 1. Evolution of Molecular Representations in Pharmacoinformatics

The figure is an original conceptual synthesis developed for "Pharmacoinformatics after Deep Learning: A State-of-the-Art Review of Molecular Representations, Benchmark Design, and Generalization across Chemical Space." 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.

Descriptor-based and fingerprint-based representations

Descriptor-based models encode selected physicochemical, topological, electronic, or structural characteristics as fixed numerical variables. Fingerprints instead represent the presence, absence, or frequency of structural fragments and neighbourhoods. These approaches remain central because they are computationally efficient, compatible with many established algorithms, and often effective when labelled datasets are limited. Deep neural networks have nevertheless outperformed conventional methods in selected large, curated bioactivity benchmarks [9].

Such results are conditional on how representation, algorithm, metric, and dataset interact. Comparative studies across diverse drug-discovery datasets show that the apparent advantage of deep learning changes with endpoint structure, class balance, sample size, metric choice, and baseline implementation. A poorly tuned conventional model is therefore not an adequate reference point for claiming representational progress [10].

Large-scale ChEMBL analyses further indicate that feed-forward neural networks can perform strongly across many target-prediction tasks, while also showing that no model class dominates every endpoint. Averaging performance across heterogeneous assays can conceal substantial differences in task difficulty, label quality, and chemical-series composition [11].

Direct comparisons between graph and descriptor-based systems challenge the assumption that learned molecular representations necessarily replace engineered features. Graph models can reduce manual feature design and learn task-specific encodings, but descriptor-based models remain competitive or superior for some datasets and endpoints [12]. Across major comparative studies, the relative advantage of deep learning remained conditional on dataset construction, representation, endpoint, and evaluation metric [9-11].

Graph, sequence, three-dimensional, and multimodal representations

Graph representations treat atoms as nodes and bonds as edges, enabling message-passing architectures to update atomic and molecular features through local connectivity. Directed message-passing models have achieved strong results across public and industrial property-prediction datasets, particularly when learned graph features are combined with selected molecular descriptors. Their performance nevertheless remains sensitive to dataset size, splitting strategy, and hyperparameter selection [13].

Sequence representations encode molecules as strings, commonly through SMILES or related notations. Recurrent and transformer-based models can learn contextual regularities from these strings, including relationships between substructures and endpoint labels. However, sequence models also inherit representational choices such as atom ordering, tokenization, stereochemical notation, and the existence of multiple valid strings for the same molecular graph [14].

Three-dimensional representations incorporate atomic coordinates, distances, angles, or symmetry-aware geometric relations. SchNet demonstrated how continuous-filter convolutions can learn atomistic environments for molecular and materials properties [15]. Such models offer access to geometry-dependent information, but their pharmaceutical interpretation depends on conformer generation, protonation, tautomeric state, solvent conditions, binding context, and whether the supplied geometry reflects the biological state relevant to the endpoint.

Multimodal approaches seek to align complementary information rather than select one canonical molecular language. MolLM, for example, integrates biomedical text with two-dimensional and three-dimensional molecular representations across several downstream tasks [16]. These systems may support richer transfer and querying, yet they also introduce alignment errors, missing-modality problems, contradictory evidence, and uncertainty about whether shared embeddings preserve causal or assay-relevant information.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for contemporary methods, representations, datasets, and pharmaceutical tasks in pharmacoinformatics after deep learning.

Table 1. Contemporary methods, representations, datasets, and pharmaceutical tasks in Pharmacoinformatics after Deep Learning.

Evidence domain or review construct	Review question	Eligible evidence type	Key extraction fields	Appraisal or relevance criterion	Synthesis role	Main limitation	Interpretation boundary
Fixed descriptors	Which predefined molecular properties remain informative?	Comparative QSAR and property-prediction studies	Descriptor family, endpoint, algorithm, split, comparator	Fair tuning and endpoint relevance	Establishes the conventional baseline	Feature choice may encode prior assumptions	Competitive performance does not establish mechanistic explanation [9]
Structural fingerprints	When do fragment-based encodings remain sufficient?	Multi-dataset model comparisons	Fingerprint type, radius, dimensionality, sparsity, metric	Strong baseline implementation	Tests whether learned representations add value	Limited geometry and global context	Simplicity should not be equated with inferiority [10]
Graph representations	Do learned topological features improve prediction?	Message-passing and descriptor-versus-graph comparisons	Node/edge features, aggregation, depth, split	Comparison against tuned descriptors	Evaluates learned topology	Oversmoothing, locality, dataset dependence	Graph learning does not uniformly dominate engineered features [12]
Molecular sequences	What chemical information can be	Recurrent or transformer sequence studies	Notation, tokenization, augmentation, pretraining objective	Robustness to alternative molecular strings	Assesses sequence-based learning	Representation depends on syntax and enumeration	Language regularity is not identical to chemical causality [14]

	learned from strings?						
Three-dimensional geometry	When does explicit geometry improve relevance?	Atomistic and geometric deep-learning studies	Coordinates, conformers, invariance, target property	State and geometry plausibility	Examines spatial information	Conformer and biological-state uncertainty	Geometric richness does not guarantee biological relevance [15]
Multimodal integration	Can text, 2D, and 3D information be aligned usefully?	Multimodal pretraining and downstream evaluation	Modalities, alignment objective, missing data, transfer task	Ablation and modality-specific validation	Maps the emerging representational frontier	Misalignment and contradictory modalities	Shared embeddings require task-specific validation [16]
Benchmark construction	Does the evaluation test novelty or familiarity?	Benchmark and split-method studies	Dataset provenance, split, metric, leakage controls	Resemblance to intended use	Connects representation to evidence quality	Familiar chemistry may inflate performance	Benchmark success is not deployment evidence [8]
Pharmaceutical decision use	What claim can the model reasonably support?	External, prospective, or experimentally connected studies	Intended use, uncertainty, action, feedback, failure cost	Decision-context alignment	**Original proposed synthesis** linking model evidence to use	No universal readiness criterion	Model output must remain bounded by validation and intended use

Supervised, self-supervised, and transfer-learning approaches

Supervised learning directly optimizes molecular representations against labelled endpoints and can be effective when assay definitions, data volume, and chemical coverage are adequate. Its principal weakness is dependence on labels that may be sparse, heterogeneous, or biased toward previously investigated chemistry.

Self-supervised learning instead derives training signals from molecular structure or alternative encodings. Translation between equivalent molecular strings can generate continuous descriptors that retain useful structural information [17]. Substructure-context learning can likewise create distributed chemical embeddings without endpoint labels [18].

Large-scale chemical-language pretraining extends this logic by learning from extensive molecular corpora before endpoint-specific fine-tuning. MolFormer demonstrated broad retrospective transfer across molecular-property tasks, although such transfer remains conditional on corpus composition, downstream similarity, and evaluation design [19].

Contrastive graph learning uses alternative molecular views and augmentation rules to encourage representation consistency. MolCLR improved several downstream predictions through graph-based pretraining [20]. However, chemically inappropriate augmentations may suppress endpoint-relevant information, so transferability must be demonstrated for each task rather than inferred from pretraining scale alone.

Benchmark datasets and task construction

Benchmark datasets define which molecular question is being asked, which examples are treated as independent, and which errors are rewarded or penalized. MoleculeNet helped standardize molecular-learning datasets, metrics, and split strategies, but it also illustrates that classification, regression, quantum, biophysical, and physiological tasks require different evaluation logic [21].

Generative benchmarks introduce additional distinctions. GuacaMol separates distribution-learning tasks from goal-directed optimization, preventing a model that reproduces a training distribution from being treated as equivalent to one that satisfies a specified objective [22].

MOSES evaluates validity, uniqueness, novelty, and distributional similarity across generated molecular sets [23]. These measures are useful for identifying technical failure, but they cannot independently establish synthetic accessibility, medicinal-chemistry quality, biological activity, developability, or therapeutic relevance.

Benchmark interpretation should therefore begin with the intended claim. A task may assess interpolation, analogue recognition, structural novelty, optimization, or robustness, but no single benchmark captures all these

properties. The benchmark must be treated as an experimental construct rather than a neutral proxy for pharmaceutical usefulness.

Table 2 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for benchmark designs, validation strategies, and comparative evaluation dimensions.

Table 2. Benchmark designs, validation strategies, and comparative evaluation dimensions.

Method, platform, or evidence category	Representative application	Reported strength	Validation context	Generalization concern	Contradictory evidence	Evidence gap	Review interpretation
MoleculeNet	Molecular classification and regression	Standardized tasks, metrics, and splits	Retrospective public datasets	Dataset and split dependence	Strong benchmark results may not survive harder separation	Limited prospective evidence	Useful infrastructure, not deployment evidence [21]
GuacaMol	Molecular generation	Separates distribution learning from goal-directed design	Retrospective generative tasks	Objective gaming and chemical triviality	High scores may reflect metric optimization	Limited experimental confirmation	Capability-specific benchmark [22]
MOSES	Molecular-set generation	Standardized validity, novelty, and diversity measures	Distributional comparison	Similarity does not establish usefulness	Valid molecules may remain impractical	Weak biological grounding	Technical quality assessment only [23]
Random split	Property prediction	Efficient internal comparison	Similar train–test distributions	Analogue leakage	May reward familiarity	Weak evidence for novel chemistry	Supports interpolation claims
Scaffold split	Structural extrapolation	Reduces close scaffold overlap	Retrospective structural separation	Residual analogue similarity	Split severity varies	No universal scaffold definition	Stronger but incomplete generalization test
Activity-cliff test	Local chemical discontinuity	Exposes sensitivity to small structural changes	Matched-pair evaluation	Average metrics hide local failures	Models may rank well overall but fail locally	Limited endpoint coverage	Essential stress test
Temporal split	Future-compound prediction	Approximates chronological use	Earlier training, later testing	Changing assays and project strategy	Time also changes data-generating processes	Often unavailable publicly	Decision-proximal retrospective validation
Prospective evaluation	ADME or screening prediction	Tests genuinely unseen compounds	Prediction before measurement	Organization-specific transportability	Success may not transfer across laboratories	Few replicated studies	Strongest bounded evidence before broader use

Scaffold splitting, temporal splitting, and prospective evaluation

Random train–test splits often place close analogues in both partitions. Under these conditions, a model may achieve favourable performance by recognizing familiar chemical series rather than learning relationships that extend to new regions of chemical space. Some ligand-based benchmarks consequently reward memorization more than target-level generalization [24].

Scaffold splitting attempts to separate core structural frameworks, but its difficulty depends on scaffold definition, dataset diversity, and residual similarity between partitions. Activity-cliff analysis provides a complementary test

because closely related molecules can display sharply different activities, revealing local failures concealed by aggregate metrics [25].

Temporal splitting trains on earlier records and evaluates later compounds, thereby approximating future project use while also incorporating changes in chemistry, assay practice, and organizational priorities. Prospective industrial evaluation of ADME models has shown why this distinction matters: performance on compounds measured after model construction can differ from retrospective expectations [26].

Memorization analysis, activity-cliff testing, and prospective evaluation expose different dimensions of generalization and are not interchangeable validation strategies [24-26]. A defensible validation programme should combine structural separation, chronological evaluation, uncertainty analysis, and endpoint-specific experimental feedback rather than rely on one split as a universal standard.

Generalization under chemical-space and assay shift

Generalization describes performance under a specified change in the data-generating process. Chemical-space shift may involve unfamiliar scaffolds, functional groups, molecular sizes, stereochemistry, or physicochemical regions. Comparative out-of-distribution studies show that model rankings can change substantially with the strategy used to construct the shifted test set [27].

Applicability domains should therefore characterize where model evidence is credible rather than act as binary declarations of validity. Similarity-based boundaries are useful but incomplete because compounds with familiar structures may still be measured under different assays, targets, protocols, or label definitions.

Uncertainty-guided active learning may prioritize measurements that expand informative coverage, but reported gains in out-of-distribution performance are conditional and sometimes modest [28]. Poor uncertainty ranking, biased acquisition pools, or inaccessible regions of chemical space can restrict improvement.

Representation-consistency methods attempt to preserve stable molecular semantics across perturbations or environments. Such approaches have improved selected out-of-distribution benchmarks [29]. Their broader value remains uncertain because benchmark-defined invariance may not match the biological or assay changes encountered in pharmaceutical programmes.

Table 3 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for generalization, uncertainty, interpretability, and reproducibility across the state of the art.

Table 3. Generalization, uncertainty, interpretability, and reproducibility across the state of the art.

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
Strong internal, weaker shifted performance	Scaffold or chemical-space separation	Reduced analogue familiarity	Model ranking changes [27]	Split severity differs	OOD definitions are heterogeneous	Report several shift types	Standardized multi-shift evaluation
Hidden local failure	Activity-cliff regions	Structural similarity with abrupt label change	Average metrics conceal errors [25]	Measurement noise	Cliff definitions vary	Add local stress tests	Assay-aware cliff analysis
Conditional active-learning benefit	Sparse OOD regions	Uncertainty-guided acquisition	Modest or variable improvement [28]	Acquisition bias	UQ may be miscalibrated	Validate ranking quality	Prospective acquisition studies
Apparent invariant representation	Artificial distribution shifts	Consistency objective	Improved selected benchmarks [29]	Benchmark-specific regularization	Limited prospective evidence	Avoid universal robustness claims	Cross-assay replication
Reliable ranking but poor calibration	Retrospective property prediction	Loss optimized for discrimination	Useful ordering with uncertain probabilities	Dataset imbalance	Calibration may drift	Separate ranking from probability claims	Shift-aware calibration

Plausible attribution without mechanism	Graph or descriptor model	Feature-importance method	Chemically interpretable pattern	Correlation or model artifact	Explanations may be unstable	Treat explanations as hypotheses	Experimental falsification
Reproducible score but irreproducible workflow	Complex preprocessing	Unreported data and software decisions	Difficult independent reconstruction	Environment differences	Code alone may be insufficient	Report full provenance	Executable workflows
Proposed integrated configuration	Intended pharmaceutical use	Generalization, calibration, explanation, and reproducibility assessed jointly	More bounded interpretation	Remaining unknown confounding	Not a validated framework	Use as a review synthesis only	Prospective decision-linked testing

Interpretability, uncertainty, and reproducibility

Predictive uncertainty concerns the reliability of a model's confidence or error estimates, not merely the magnitude of average prediction error. Comparative evaluations show that ensembles, dropout-based procedures, and other uncertainty methods differ across datasets and metrics, with no universally superior approach [30].

Interpretability addresses what aspects of the input influence a prediction and how those influences are presented to scientists. Attribution maps, substructure rationales, and counterfactual examples may support model interrogation, but they do not automatically identify biological mechanism or causal molecular determinants [31].

Reproducibility requires reconstructing data selection, preprocessing, molecular standardization, splitting, model configuration, software environment, and analysis. Reporting only an architecture and performance metric is insufficient for independent verification or meaningful comparison [32].

Uncertainty estimation, explanation, and reproducibility answer complementary questions and none can substitute for the others [30-32]. A model may be reproducible yet poorly calibrated, interpretable yet inaccurate, or confident for chemically inappropriate reasons. Trustworthy use therefore requires separate evidence for each property.

Comparative state of the art across pharmaceutical tasks

The strongest evidence for pharmaceutical relevance occurs when computational predictions are followed by experimental testing. Deep learning enabled prioritization of antibacterial candidates with chemical characteristics distinct from familiar training examples, demonstrating bounded capacity to guide experimental selection [33].

Generative modelling has also produced synthesized compounds with measured activity against a defined target. Rapid identification of DDR1 inhibitors illustrates experimental feasibility, but active compounds are not equivalent to optimized leads, development candidates, medicines, or demonstrated clinical benefit [34].

Operational maturity represents another evidence category. An automated machine-learning virtual-screening cascade was implemented within an African drug-discovery centre, showing that integrated computational workflows can be deployed in real research settings [35]. Its significance lies in implementation capability rather than proof of generalized discovery acceleration.

The state of the art is therefore heterogeneous. Descriptor and graph models have mature retrospective evidence; prospective validation remains less common; experimental studies establish selected proof of concept; and operational implementation demonstrates workflow feasibility. None of these evidence levels should be substituted for another.

Table 4 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for maturity assessment, evidence gaps, and emerging research directions for pharmacoinformatics after deep learning.

Table 4. Maturity assessment, evidence gaps, and emerging research directions for Pharmacoinformatics after Deep Learning.

Evidence-maturity domain	Current evidence position	Methodological weakness	Translation barrier	Needed validation	Reporting priority	Decision implication	No-overclaiming boundary

Descriptor-based prediction	Mature retrospective evidence	Endpoint and feature dependence	Limited transfer to novel chemistry	Temporal and external tests	Baseline tuning	Retain as strong comparator	Not obsolete
Graph representation learning	Broad retrospective evidence	Split and hyperparameter sensitivity	Assay and project shift	Multi-site prospective comparison	Graph construction and tuning	Useful when conditional advantage is shown	Not universally superior
Molecular language pretraining	Expanding benchmark evidence	Corpus and tokenization dependence	Weak prospective grounding	Cross-domain transfer tests	Corpus provenance	Promising reusable representation	Not general chemical understanding
Three-dimensional modelling	Strong geometric methodology	Conformer and state uncertainty	Biological geometry may be unknown	State-aware external validation	Coordinate generation	Appropriate for geometry-relevant tasks	Not automatically mechanistic
Multimodal modelling	Emerging evidence	Alignment and missing modalities	Contradictory context	Modality ablation and prospective tests	Pairing and alignment rules	Potential integration layer	Not validated by modality count
OOD generalization	Increasing comparative evidence	No universal shift definition	Future shifts remain unknown	Multiple realistic shift scenarios	Chemical-distance reporting	Supports bounded extrapolation	Not universal robustness
Prospective ADME prediction	Decision-proximal evidence	Organization-specific data	Transportability	Independent replication	Time and project context	Can support specified internal decisions	Not universal external validity
Experimental hit prioritization	Experimental proof of concept	Selection and confirmation bias	Lead optimization and development	Replicated end-to-end programmes	Failed predictions	Supports candidate prioritization	Not medicine discovery proof
Foundation-model scaling	Emerging retrospective evidence	Compute and data confounding	Uncertain decision benefit	Resource-adjusted downstream tests	Data, compute, and model scale	Research-stage capability	Scaling is not readiness
Assay-aware modelling	Emerging contextual evidence	Metadata inconsistency	Cross-laboratory standardization	Prospective assay-transfer tests	Assay provenance	May reduce label-context collapse	Not established generalization

Evidence gaps and emerging research directions

Scaling studies indicate that larger chemical models can improve learning behaviour as data, model capacity, and computation increase [36]. Future evaluation should determine whether these gains survive realistic chemical and assay shifts, justify resource use, and improve decision-relevant outcomes rather than only pretraining loss.

Open encoder–decoder foundation models provide shared architectures for molecular reconstruction, property prediction, and reaction-related tasks [37]. Their value will depend on transparent corpus provenance, contamination analysis, calibration, task adaptation, external validation, and comparison against smaller specialized models.

Assay-aware modelling addresses a deeper representational problem: identical-looking activity labels may arise from different protocols and biological contexts. Encoding assay descriptions alongside compounds and targets may preserve information otherwise collapsed during aggregation [38]. This direction is promising but requires standardized metadata and prospective transfer tests.

Figure 2 presents the original conceptual synthesis for benchmark-to-deployment generalization pathway, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

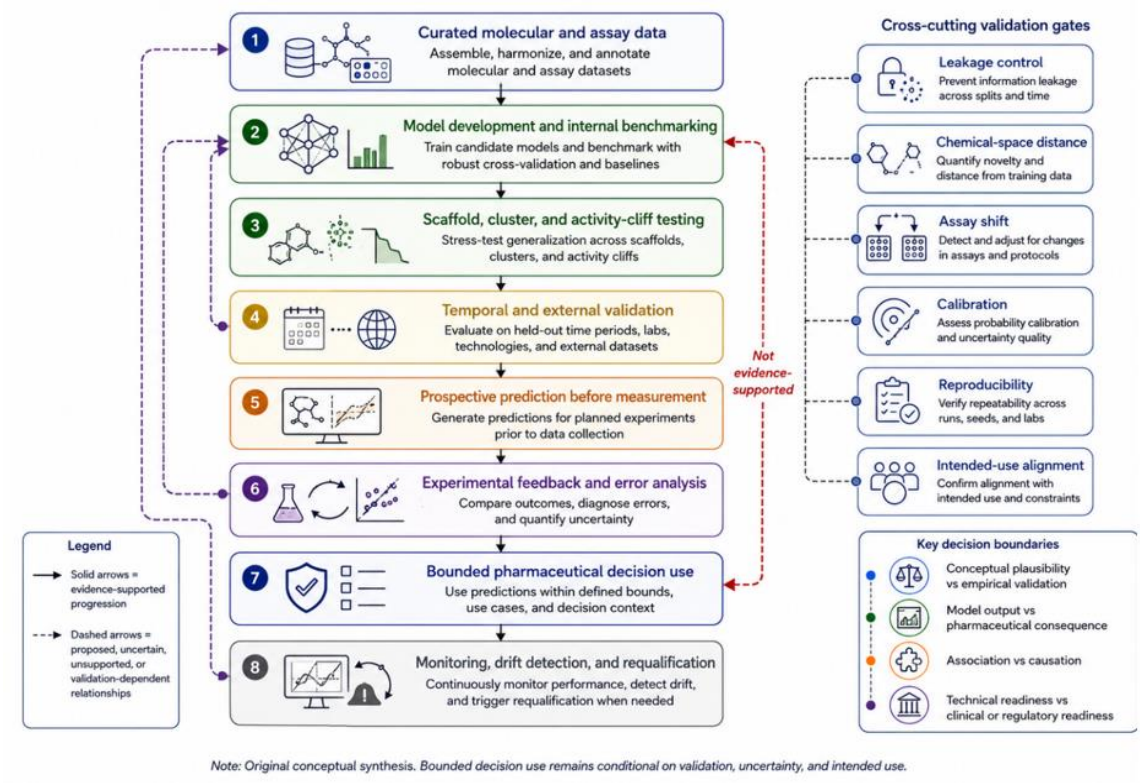

Figure 2. Benchmark-to-Deployment Generalization Pathway.

The figure is an original conceptual synthesis developed for "Pharmacoinformatics after Deep Learning: A State-of-the-Art Review of Molecular Representations, Benchmark Design, and Generalization across Chemical Space." 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.

Limitations of the review

This state-of-the-art review provides a structured comparison of influential technical and validation developments but does not claim exhaustive systematic coverage. Its purpose is to identify the frontier and its boundaries rather than calculate pooled effects or produce a formal evidence grade.

The underlying literature is heterogeneous in datasets, endpoints, molecular standardization, metrics, split strategies, comparator quality, and reporting depth. Direct ranking across studies would therefore risk attributing differences to model architecture when they may arise from task construction or experimental context.

Public studies may underrepresent proprietary failures, abandoned models, negative prospective tests, and organization-specific implementation constraints. Conversely, industrial results may rely on data and workflows that cannot be independently reconstructed.

The review also emphasizes molecular representation and predictive evaluation rather than every aspect of pharmacoinformatics. Conclusions should therefore be applied to specified modelling and decision contexts, not generalized automatically to clinical practice, regulation, manufacturing, or patient outcomes.

CONCLUSION

Pharmacoinformatics after deep learning is defined less by the disappearance of conventional descriptors than by an expanded set of representational and learning choices. Graph, sequence, geometric, multimodal, self-supervised, and foundation-model approaches offer substantial flexibility, but their scientific value remains conditional on data provenance, comparator quality, benchmark construction, chemical-space coverage, assay context, calibration, reproducibility, and prospective evaluation. The field should therefore move from architecture-centred claims toward intended-use-centred evidence. Rich representations are technically important,

yet only validation under relevant chemical, temporal, experimental, and organizational change can establish whether they support a specified pharmaceutical decision.

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. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009
3. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
4. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. Chem Rev. 2019;119(18):10520-94. doi:10.1021/acs.chemrev.8b00728
5. David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: A review and practical guide. J Cheminform. 2020;12(1):56. doi:10.1186/s13321-020-00460-5
6. Elton DC, Boukouvalas Z, Fuge MD, Chung PW. Deep learning for molecular design—a review of the state of the art. Mol Syst Des Eng. 2019;4(4):828-49. doi:10.1039/C9ME00039A
7. Walters WP, Barzilay R. Applications of deep learning in molecule generation and molecular property prediction. Acc Chem Res. 2021;54(2):263-70. doi:10.1021/acs.accounts.0c00699
8. Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, et al. Utilizing graph machine learning within drug discovery and development. Brief Bioinform. 2021;22(6). doi:10.1093/bib/bbab159
9. Lenselink EB, ten Dijke N, Bongers B, Papadatos G, van Vlijmen HWT, Kowalczyk W, et al. Beyond the hype: Deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. J Cheminform. 2017;9(1):45. doi:10.1186/s13321-017-0232-0
10. Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. Mol Pharm. 2017;14(12):4462-75. doi:10.1021/acs.molpharmaceut.7b00578
11. 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
12. Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminform. 2021;13(1):12. doi:10.1186/s13321-020-00479-8
13. 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. doi:10.1021/acs.jcim.9b00237
14. Lin X, Quan Z, Wang ZJ, Huang H, Zeng X. A novel molecular representation with BiGRU neural networks for learning atom. Brief Bioinform. 2020;21(6):2099-111. doi:10.1093/bib/bbz125
15. Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. SchNet—a deep learning architecture for molecules and materials. J Chem Phys. 2018;148(24):241722. doi:10.1063/1.5019779
16. Tang X, Tran A, Tan J, Gerstein MB. MolLM: A unified language model for integrating biomedical text with 2D and 3D molecular representations. Bioinformatics. 2024;40(Suppl 1). doi:10.1093/bioinformatics/btae260

17. Winter R, Montanari F, Noé F, Clevert DA. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci. 2019;10(6):1692-701. doi:10.1039/C8SC04175J
18. Jaeger S, Fulle S, Turk S. Mol2vec: Unsupervised machine learning approach with chemical intuition. J Chem Inf Model. 2018;58(1):27-35. doi:10.1021/acs.jcim.7b00616
19. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell. 2022;4(12):1256-64. doi:10.1038/s42256-022-00580-7
20. Wang Y, Wang J, Cao Z, Barati Farimani A. Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell. 2022;4(3):279-87. doi:10.1038/s42256-022-00447-x
21. 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
22. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: Benchmarking models for de novo molecular design. J Chem Inf Model. 2019;59(3):1096-108. doi:10.1021/acs.jcim.8b00839
23. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, et al. Molecular Sets (MOSES): A benchmarking platform for molecular generation models. Front Pharmacol. 2020;11:565644. doi:10.3389/fphar.2020.565644
24. 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
25. 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
26. Fang C, Wang Y, Grater R, Kapadnis S, Black C, Trapa P, et al. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. J Chem Inf Model. 2023;63(11):3263-74. doi:10.1021/acs.jcim.3c00160
27. Fooladi H, Vu TNL, 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. doi:10.1021/acs.jcim.5c00475
28. Yin T, Panapitiya G, Coda ED, Saldanha EG. Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction. J Cheminform. 2023;15(1):105. doi:10.1186/s13321-023-00753-5
29. Wen X, Liu H, Long W, Wei S, Zhu R. Consistent semantic representation learning for out-of-distribution molecular property prediction. Brief Bioinform. 2025;26(2). doi:10.1093/bib/bbaf147
30. 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
31. Jiménez-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
32. Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18(10):1132-35. doi:10.1038/s41592-021-01256-7
33. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688-702.e13. doi:10.1016/j.cell.2020.01.021
34. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038-40. doi:10.1038/s41587-019-0224-x
35. Turon G, Hlozek J, Woodland JG, Kumar A, Chibale K, Duran-Frigola M. First fully-automated AI/ML virtual screening cascade implemented at a drug discovery centre in Africa. Nat Commun. 2023;14(1):5736. doi:10.1038/s41467-023-41512-2
36. Frey NC, Soklaski R, Axelrod S, Samsi S, Gómez-Bombarelli R, Coley CW, et al. Neural scaling of deep chemical models. Nat Mach Intell. 2023;5(11):1297-305. doi:10.1038/s42256-023-00740-3
37. 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
38. Schoenmaker L, Sastrokarijo EG, Heitman LH, Beltman JB, Jespers W, van Westen GJP. Toward assay-aware bioactivity model(er)s: Getting a grip on biological context. J Chem Inf Model. 2025;65(13):7013-23. doi:10.1021/acs.jcim.5c00603