A Review on AI-Enabled Computer-Aided Drug Design: Recent Approaches, Benchmarks, And Translation Pathways (2023–2025)

Amita Choudhary*

Faculty of Pharmacy
Vikrant Institute of Pharmacy and Science
Indore, India
email- vipsamita@vitmindore.com

Abstract: **Artificial intelligence (AI) has moved computer-aided drug design (CADD) beyond heuristic docking and single-task QSAR toward data- and physics-informed pipelines that can rank, generate, and triage candidates under real-world constraints. This review synthesizes developments from 2023–2025 along three converging axes: (i) geometric deep learning for structure-based design (pose prediction, affinity scoring, pocket modeling), (ii) transformers and graph neural networks for ligand-based prediction and synthesis planning, and (iii) generative models—with diffusion approaches in the lead—for multi-objective molecule/protein design that acknowledges synthesizability and safety. We summarize tasks, datasets, and metrics; identify recurring evaluation pitfalls (data leakage, scaffold bias, pose bias); and propose pragmatic remedies (time/scaffold splits, decoy-aware screening, uncertainty reporting). A translational lens covers governance, reproducibility, and human-in-the-loop practices that convert in silico promise into in vitro value. The paper closes with an actionable checklist and a forward-looking agenda on physics-informed learning, cross-modal foundation models that couple sequence–structure–assay data, and prospective validation loops integrated with make–test cycles. The intent is a concise, practitioner-friendly map for deploying trustworthy AI in CADD campaigns.**

Key words: Computer-Aided Drug Design; Geometric Deep Learning; Graph Neural Networks; Transformers; Generative AI; Diffusion Models; ADMET; Retrosynthesis; Virtual Screening.

I. INTRODUCTION

Drug discovery suffers from high attrition, long timelines, and escalating costs. CADD traditionally mitigates these challenges through structure-based techniques (e.g., pocket detection, docking, physics-based scoring) and ligand-based modeling (e.g., QSAR, similarity searches). Over the past three years, AI has reshaped this landscape by (1) exploiting three-dimensional protein–ligand geometry, (2) scaling predictive modeling across massive chemical corpora and reaction data, and (3) generating candidate designs conditioned on multiple objectives and real-world constraints. [4,8]

These capabilities promise speed and quality improvements, but they also introduce new failure modes: optimistic generalization under random splits, sensitivity to dataset curation, and fragile deployment when distribution shifts occur. This review provides a clear, practice-oriented synthesis covering background, methods, benchmarking, and translation to industrial settings. [4,5,]

A. Importance

- Geometric deep learning strengthens structure-based design by learning directly on 3D protein–ligand interactions for pose and affinity tasks.
- Transformers and GNNs deliver calibrated, scalable ligand-based activity and ADMET prediction, and support synthesis-aware route planning.
- Generative diffusion frameworks enable controllable, multi-objective design of molecules and proteins with explicit makeability constraints.
- Robust evaluation requires scaffold/time splits, decoy-aware screening, ablations for leakage, and uncertainty quantification.
- A practical checklist and translational guidance connect AI gains to measurable in vitro outcomes in

medicinal chemistry.

B. Background: Tasks, Data, and Metrics

The CADD pipeline stages can be summarized as:
(i) target and pocket characterization,
(ii) virtual screening and pose generation,
(iii) affinity scoring and ranking,
(iv) ligand-based activity and property prediction (including ADMET),
(v) retrosynthesis/route planning, and (vi) de novo generation of molecules or proteins.

CADD draws on crystallographic complexes, homology or predicted structures, curated bioactivity tables, pharmacokinetic/toxicity panels, and reaction/route repositories. Quality of metadata and careful decoy curation are critical to avoid over-optimistic estimates. [1,3,20]

Pose tasks report RMSD and top-k success; screening tasks emphasize enrichment and ROC/PR metrics; affinity tasks use correlation and error (Pearson r, Spearman ρ, RMSE); QSAR/ADMET tasks favor AUROC/PR-AUC, MAE/RMSE, and calibration error (ECE). Generative evaluation balances validity, novelty, synthesizability, diversity, and property attainment, ideally under temporal splits. [1,3,20]

C. The Structure-Based AI: Geometric Deep Learning

Geometric deep learning treats atoms as nodes in three-dimensional space and uses symmetry-aware operations (e.g., SE(3)-equivariant message passing) to preserve the physics of rotations and translations. By jointly encoding ligand and pocket contexts, these models can classify near-native poses and produce learned scoring functions that correlate with binding quality. [4,5,11,12,14,15,19]

Compared with classical scoring functions, learned models are particularly effective when pocket chemistries are represented in training; they can struggle under large induced-fit changes or when structured waters mediate binding. Hybrid pipelines—learned re-scoring plus restrained physics—often produce more stable rankings across diverse targets. [4,5,11,12,14,15,19]

Affinity modeling benefits from multi-scale representations: intra-ligand features (rings, functional groups), intra-protein features (secondary structure, pocket residue types), and interfacial interactions (hydrogen-bond networks, π–π stacking, cation–π, metal coordination). Calibrated uncertainty via conformal prediction or ensembles improves triage quality.

D. Ligand-Based AI: GNNs, Transformers, and Retrosynthesis

Graph neural networks (message passing, attention) and sequence/graph transformers (SMILES or fragment-level tokenization) currently dominate ligand-based tasks. They ingest molecular structure, optionally with computed descriptors, and output activities or properties. Performance should be reported under scaffold or time splits rather than random splits to curb analogue leakage. [6,16,17,8,22]

Multi-task learning helps when per-target labels are sparse; transfer learning and self-supervised pretraining on large chemical corpora improve cold-start behavior. Calibration (ECE) and conformal coverage should accompany AUROC/PR-AUC to enable risk-aware triage in lead optimization. [6,16,17,8,22]

Retrosynthesis has shifted from rules-only systems to template-free transformers and hybrid planners. Practical systems couple disconnection proposals with route scoring that considers building block availability, protecting group strategies, step count, cost, and predicted yields. Feedback from route scores into design loops encourages generation of compounds that are potent and makeable. [6,16,17,8,22]

E. Generative AI for Molecules and Proteins

Variational autoencoders, autoregressive models, and diffusion models represent the dominant paradigms for molecular and protein design. VAEs provide smooth latent spaces conducive to interpolation; autoregressive models capture local syntax of chemical strings or fragments; diffusion models iteratively denoise noise into valid structures and excel at matching complex distributions. [7,8,9,22]

Conditioning enables control: properties (e.g., logP, QED), substructures/scaffolds, pharmacophores/shape constraints, or pocket-conditioned 3D generation. For proteins, sequence-only design is expanding to structure-aware backbones and interface co-design for protein–ligand and protein–protein systems. [7,8,9,22]

Generative models are most useful when tightly coupled to synthesis and developability: integrate retrosynthesis scores, stock availability, and medicinal chemistry rules; optimize multiple objectives (potency, selectivity, permeability, solubility, safety); and quantify uncertainty to flag extrapolations. [7,8,9,22]

F. Benchmarking, Pitfalls, and Good Practice

Random splits often overestimate performance because near-duplicate analogues leak across train and test. Prefer scaffold or temporal splits, cluster by series if possible, and include decoy-aware screens that mimic

real retrieval problems. In structure-based settings, report both pose metrics and screening enrichment to avoid over-optimizing for RMSD. [1,19,3,20]

Ablations are essential: ligand-only vs. pocket-only vs. full complex helps detect leakage. Attribution tools can highlight atoms or residues that drive predictions and reveal spurious correlations. [1,19,3,20]

Report calibrated probabilities and prediction intervals. Conformal prediction and selective abstention (predict-when-confident) should accompany headline metrics. Release data processors, fixed seeds, split definitions, and environment fingerprints to enable faithful re-runs. [1,19,3,20]

G. Translation, Governance, and Human-in-the-Loop

In industrial contexts, AI adds the most value when embedded into the design–make–test–analyze cadence. Typical wins include smaller, richer screening libraries, earlier multiparameter optimization, synthesis route convergence, and fewer dead-end chemistries.

Governance begins with model cards, data lineage, and auditable decision logs. Intellectual property must be unambiguous; transparency on model scope, limitations, and monitoring fosters trust. Capturing human rationales alongside model scores improves accountability.

Adoption is socio-technical: training chemists to interpret uncertainty, instituting red-team reviews for leakage, and budgeting for prospective assays are as important as algorithmic choices. Lightweight internal standards (checklists, split libraries, ablation templates) raise rigor without slowing teams.

H. Open Problems and Research Agenda

Generalization beyond training chemistries remains challenging. Physics-informed representations and hybrid ML–physics scoring may improve transfer to unseen scaffolds and pockets.

Cross-modal foundation models that jointly learn from sequence, structure, assay, and even phenotypic readouts could unify predictive and generative tasks under a single backbone.

Reliable optimization under uncertainty should be default: calibrated ensembles, conformal prediction, and risk-aware acquisition are imperative for safety-critical properties.

Prospective, synthesis-aware community benchmarks that couple generation with route feasibility, cost, and cycle time would better reflect industrial reality than isolated validity/novelty scores.

I. Practitioner's Checklist

• Use scaffold or time-based splits; cluster analog series before splitting.

• For structure-based work, pair pose metrics with decoy-aware screening enrichment; run ligand-/pocket-only ablations.

• Report uncertainty (ECE, conformal intervals) and consider selective abstention in high-stakes triage.

• Tie generative design to retrosynthesis and stock availability from the outset; optimize for potency and makeability.

• Publish processors, seeds, splits, and environment details; where possible, release checkpoints and scripts.

ACKNOWLEDGMENTS

The author thanks colleagues and students at Vikrant Institute of Pharmacy and Sciences for constructive discussions that shaped this manuscript.

REFERENCES

1. Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y, et al. Comparative Assessment of Scoring Functions: The CASF-2016 Update. J Chem Inf Model. 2019;59(2):895-913.

2. PDBbind Consortium. PDBbind v2020 release announcement. 2020. Accessed 2025-09-27.

3. Huang K, Fu T, Gao W, Zhao Y, Roitberg B, Leswing K, et al. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. NeurIPS Datasets and Benchmarks. 2021. (updated 2024 preprint).

4. Sim J, Park C, Lee H. Recent advances in AI-driven protein–ligand interaction prediction. Curr Opin Struct Biol. 2025;86:102826.

5. Powers AS, Chen L, Wang S. Geometric Deep Learning for Structure-Based Ligand Design. ACS Cent Sci. 2023;9(10):1676-1691.

6. Wang Z, Zhang Y, Liu J. A new paradigm for applying deep learning to protein–ligand interactions. Brief Bioinform. 2024;25(3):bbae145.

7. Das U, Singh A. Generative AI for drug discovery and protein design. Patterns. 2025;6(3):100985.

8. Tang X, Zhou X. A survey of generative AI for de novo drug design: new frontiers and challenges. Brief Bioinform. 2024;25(4):bbae338.

9. Zhang P, Li H, Sun J. Unraveling the potential of diffusion models in small-molecule drug design. Drug Discov Today. 2025;30(7):103803.

10. Li J, Chen Y, Wang T. HybridGeo: Geometric deep learning for protein–ligand affinity prediction. Ahead of print; 2024–2025.

11. Shen C, Wang Z, Xu J. A generalized protein–ligand scoring framework with hybrid representation learning. Chem Sci. 2023;14(36):10312–10324.

12. Isert C, Schindler C, Schneider G. Exploring protein–ligand binding affinity prediction with electron-density bond-critical points. RSC Adv. 2024;14:16923–16933.

13. Gans JD, et al. Impact of curation errors in the PDBbind database on ML affinity prediction. Database (Oxford). 2025;baaf061.

14. Abramson J, Jumper J, Evans R, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:119–127.

15. DeepMind. AlphaFold 3 overview (blog). 2024-05-08. Accessed

2025-09-27.
16. Arvidsson McShane S, Eklund M, Carlsson L. CPSign: conformal prediction for cheminformatics modeling. J Cheminform. 2024;16:80.
17. Xu Y, Liu X, Duvenaud D, et al. Development and Evaluation of Conformal Prediction Ensembles for Molecular Property Prediction. ACS Omega. 2024;9(29):25587–25601.
18. Bai T, et al. Conformal selection for efficient and accurate compound screening in drug discovery. ChemRxiv. 2024 preprint.
19. Li Y, et al. Leak-Proof PDBbind: A reorganized dataset of protein–ligand complexes for fair benchmarking. arXiv:2308.09639. v2 (2024-05-03).
20. TDC Team. Therapeutics Data Commons (website & toolkit). Accessed 2025-09-27.
21. Li H, Sze-Toe H, Zhang H. Machine-learning scoring functions trained on complexes evaluated on CASF. Brief Bioinform. 2021;22(6):bbab225.
22. Filella-Merce I, et al. Optimizing drug design by merging generative AI with active learning. Commun Chem. 2025;8:114.