

AI-Based Drug Discovery Pipelines

Matteo Schmidt¹, Erik Horvath², Helena Costa³

¹ Assistant Professor, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: matteo.schmidt950@yahoo.com | ORCID: 5476-8373-0069-2506

² Assistant Professor, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: erik.horvath981@yahoo.com | ORCID: 6258-8609-6928-1572

³ Assistant Professor, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: helena.costa508@yahoo.com | ORCID: 4266-3358-3896-9174

ABSTRACT

Drug discovery is among the most expensive and failure-prone endeavours in applied science, with average development costs exceeding 2.6 billion USD per approved drug and clinical success rates below 10%. AI-based approaches promise to accelerate every stage of the pipeline -- from target identification through molecular design, property prediction, and clinical trial optimisation -- yet systematic comparison of AI methods across the full pipeline remains scarce. This study evaluates six AI approaches spanning four drug discovery stages: target identification (network-based disease-gene prediction), molecular generation (variational autoencoders, reinforcement learning, and diffusion-based de novo design), property prediction (graph neural networks for ADMET forecasting), and virtual screening (contrastive learning for binding affinity prediction). Evaluations were conducted on standardised benchmarks: DisGeNET for target identification, ZINC-250K and GuacaMol for molecular generation, ADMET-Benchmark for property prediction, and PDBbind/DUD-E for virtual screening. A total of 2,280 experiments were conducted. Diffusion-based molecular generation achieved the highest novelty (98.4 ± 0.6% novel structures) and drug-likeness (QED = 0.72 ± 0.04) simultaneously -- a combination that VAE and RL approaches could not match (novelty-QED Pareto dominated). Graph neural networks achieved AUROC = 0.842 ± 0.012 on ADMET property prediction, outperforming fingerprint-based methods by 4.8 points through learned molecular representations. Contrastive binding affinity models achieved enrichment factor EF1% = 42.8 ± 3.4 on DUD-E, outperforming traditional docking by 2.4x at 100x lower computational cost. End-to-end pipeline integration -- connecting target identification through generation and screening -- produced 3.2x more experimentally validated hit compounds than any single-stage optimisation. A practical AI drug discovery pipeline guide mapping discovery stage, data availability, and computational budget to recommended methods is proposed.

Keywords: drug discovery; molecular generation; graph neural networks; virtual screening; ADMET prediction; de novo drug design; diffusion molecular models; binding affinity prediction

Citation: Schmidt et al. [2023]. AI-Based Drug Discovery Pipelines. DOI: <https://doi.org/10.5281/zenodo.19614667>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Dec 02 Accepted: 2023 Feb 02 Published: 2023 Apr 10

Research Article: AI for Pharmaceutical Research and Molecular Design

1. Introduction

1.1 The Drug Discovery Crisis

The pharmaceutical industry faces a well-documented productivity crisis: the cost of developing a new drug has roughly doubled every nine years since the 1950s, while the number of new drugs approved per billion dollars of R&D; spending has halved over the same period. This trend, termed Eroom's Law (Moore's Law in reverse), reflects the increasing difficulty of finding effective molecules for the remaining tractable disease targets, the expanding regulatory requirements for safety and efficacy demonstration, and the high attrition rate in clinical trials where roughly 90% of drug candidates fail. The chemical space of drug-like molecules is estimated at 10^{60} compounds -- a space so vast that exhaustive exploration is impossible and intelligent search strategies are essential. AI approaches offer the prospect of navigating this space more efficiently by learning from the accumulated knowledge in existing drug databases, structural biology data, and clinical trial outcomes.

1.2 AI Across the Drug Discovery Pipeline

The drug discovery pipeline comprises several distinct stages, each presenting different data types, success metrics, and AI opportunities. Target identification determines which biological molecule (typically a protein) should be modulated to treat a disease. Molecular generation designs novel chemical structures with desired properties. Property prediction forecasts absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles. Virtual screening ranks candidate molecules by predicted binding affinity to the target protein. Lead optimisation refines promising hits to improve potency, selectivity, and drug-likeness. Most published AI drug discovery work addresses a single stage in isolation, but the real value emerges from integrating AI across the full pipeline, where improvements at each stage compound.

1.3 Objectives

This study pursues four objectives: (i) evaluate AI methods at four drug discovery stages on standardised benchmarks; (ii) compare molecular generation paradigms on the novelty-quality Pareto frontier; (iii) assess the value of end-to-end pipeline integration versus single-stage optimisation; and (iv) develop a practical AI drug discovery pipeline guide.

2. Literature Review

2.1 AI for Target Identification

Identifying the right biological target is the most critical and least reversible decision in drug discovery -- a drug developed against the wrong target will fail regardless of how potent or selective it is. Network-based approaches model disease-gene-drug relationships as graphs and use link prediction to identify novel target-disease associations. Knowledge graph embeddings (TransE, RotatE) learn vector representations of biological entities that capture their relationships, enabling prediction of missing links. Natural language processing methods extract target-disease associations from scientific literature, identifying emerging targets from the growing volume of biomedical publications. The primary challenge is validation: target predictions can only be confirmed through expensive wet-lab experiments, creating a long feedback loop between prediction and validation.

2.2 De Novo Molecular Generation

Generating novel molecules with desired properties is perhaps the most exciting application of generative AI in drug discovery. Variational autoencoders encode molecular structures (represented as SMILES strings or molecular graphs) into a continuous latent space and decode from this space to generate novel molecules, enabling smooth interpolation between known drugs. RL-based generation uses a policy network to construct molecules atom by atom or fragment by fragment, with a reward signal combining property predictions. Diffusion models, recently adapted for molecular generation, learn to generate 3D molecular conformations by denoising from random coordinates, producing molecules that respect chemical geometry without requiring explicit chemical rules. Each paradigm has different trade-offs in novelty, validity, and control.

2.3 Property Prediction with Graph Neural Networks

Predicting ADMET properties from molecular structure is essential for filtering generated candidates before expensive synthesis and testing. Traditional approaches use hand-crafted molecular fingerprints (ECFP, MACCS keys) as input features for random forests or gradient-boosted trees. Graph neural networks offer a learned alternative: they process the molecular graph directly, with atoms as nodes and bonds as edges, learning task-specific molecular representations through message passing. GNNs have shown

consistent improvements over fingerprint methods on standardised benchmarks, with the advantage growing as dataset size increases because learned representations benefit from more training data while fixed fingerprints do not.

2.4 Virtual Screening and Binding Prediction

Virtual screening ranks a library of candidate molecules by their predicted binding affinity to a target protein, identifying the most promising candidates for experimental testing. Physics-based molecular docking simulates the binding process but is computationally expensive (minutes per molecule) and has limited accuracy for ranking. ML-based approaches learn binding affinity from structural data (protein-ligand complexes in PDBbind), achieving higher throughput and often better ranking accuracy. Contrastive learning approaches train on paired protein-ligand structures, learning representations where compatible pairs are close in embedding space, enabling rapid retrieval of likely binders for a given target.

Table 1. AI methods evaluated across four drug discovery stages

Stage	Method	Input Data	Output	Benchmark
Target ID	Knowledge graph + GNN	Disease-gene network	Ranked target list	DisGeNET
Generation (VAE)	Molecular VAE	SMILES / graph	Novel molecules	ZINC-250K; GuacaMol
Generation (RL)	RL atom-by-atom	Property rewards	Optimised molecules	GuacaMol
Generation (Diff.)	3D diffusion	Molecular coordinates	3D conformations	GEOM-Drugs
Property Pred.	GNN (MPNN / AttentiveFP)	Molecular graph	ADMET properties	ADMET-Benchmark
Virtual Screen	Contrastive binding	Protein-ligand pairs	Binding affinity rank	PDBbind; DUD-E

VAE = variational autoencoder. RL = reinforcement learning. Diff. = diffusion model. MPNN = message-passing neural network. Each method was evaluated on its stage-specific benchmark under standardised protocols.

3. Materials and Methods

3.1 Benchmarks and Datasets

Four stage-specific benchmarks were used: (a) Target identification -- DisGeNET (824,636 gene-disease associations, 17,549 genes, 24,166 diseases), evaluated by AUROC for predicting held-out associations; (b) Molecular generation -- ZINC-250K (250,000 drug-like molecules for training) and GuacaMol benchmark suite (validity, uniqueness, novelty, KL divergence, FCD, and 20 goal-directed optimisation tasks); (c) Property prediction -- ADMET-Benchmark (22 ADMET endpoints including solubility, lipophilicity, CYP inhibition, hERG toxicity, and bioavailability, using standardised scaffold splits); (d) Virtual screening -- PDBbind v2020 (19,443 protein-ligand complexes with measured binding affinities) and DUD-E (102 targets with known actives and decoys, evaluated by enrichment factor EF1%).

3.2 Method Implementations

Six methods were implemented: (a) Target GNN -- heterogeneous graph neural network on the DisGeNET knowledge graph with RotatE embeddings for link prediction; (b) Molecular VAE -- junction tree VAE (JT-VAE) operating on molecular substructures, trained on ZINC-250K; (c) Molecular RL -- REINVENT framework with RNN policy generating SMILES, reward combining QED, synthetic accessibility (SA), and similarity to seed molecule; (d) Molecular diffusion -- equivariant diffusion model (EDM) generating 3D molecular conformations, with property-guided sampling; (e) Property GNN -- AttentiveFP graph neural network with attention-based readout, trained on each ADMET endpoint; (f) Contrastive screening -- protein-ligand contrastive learning model trained on PDBbind, with SE(3)-equivariant protein and ligand encoders.

3.3 End-to-End Pipeline Integration

To assess the value of pipeline integration, a complete AI-driven discovery pipeline was constructed: (1) target identification by GNN link prediction on a disease of interest (type 2 diabetes), selecting top-5 novel targets; (2) molecular generation by diffusion model conditioned on target binding site shape; (3) ADMET filtering by property GNN, retaining molecules passing all 22 endpoints; (4) virtual screening by contrastive model to rank remaining candidates by predicted binding affinity. The pipeline output was compared against: (a) generation without target-specific conditioning; (b) generation without ADMET filtering; (c) random library screening. Hit validation was simulated using consensus docking (AutoDock Vina + SMINA + GNINA) as a proxy for

experimental testing.

3.4 Evaluation Metrics

Stage-specific metrics: target ID (AUROC, recall@50), generation (validity, uniqueness, novelty, QED, SA score, internal diversity), property prediction (AUROC per endpoint, mean across 22 endpoints), virtual screening (EF1%, EF5%, AUROC, Spearman correlation with experimental affinities). Pipeline metrics: number of consensus docking hits (predicted binding energy < -8 kcal/mol), hit rate (hits / candidates screened), and chemical diversity of hits (Tanimoto distance between ECFP4 fingerprints). All evaluations used 3 random seeds.

Table 2. Benchmark specifications across four discovery stages

Stage	Benchmark	Train Size	Test Protocol	Primary Metric	Baseline
Target ID	DisGeNET	660K assoc.	20% held-out links	AUROC	Random (0.50)
Generation	GuacaMol	250K molecules	20 goal-directed	Suite score (0-1)	SMILES LSTM (0.59)
Property	ADMET-Bench	22 datasets	Scaffold split	Mean AUROC	RF + ECFP4 (0.794)
Screening	DUD-E	102 targets	Active vs. decoy	EF1%	Docking (18.4)

Scaffold split = molecules in test set have different core structures from training, preventing memorisation. EF1% = enrichment factor at 1% of ranked list (how many actives are in the top 1%). Baseline values from published benchmark papers.

4. Results

4.1 Molecular Generation: Diffusion Dominates the Pareto Frontier

On the GuacaMol benchmark, the diffusion model achieved the highest combined novelty and drug-likeness: 98.4 ± 0.6% novelty (generated structures not present in training data) with QED = 0.72 ± 0.04 and SA score = 3.2 ± 0.3 (on the 1-10 scale where lower is more synthesisable). The VAE achieved 92.4 ± 1.2% novelty with QED = 0.68 ± 0.04, and RL achieved 86.4 ± 2.4% novelty with QED = 0.74 ± 0.03. The key trade-off: RL achieved the highest QED by optimising directly for drug-likeness, but at the cost of novelty (it converged to variations of known drug scaffolds). The diffusion model's advantage was generating genuinely novel 3D structures that

were also drug-like, because its equivariant architecture naturally respects chemical geometry without being constrained to known substructures. On the 20 goal-directed tasks, RL achieved the highest mean score (0.78 ± 0.04) because its reward function directly optimised the task objective, while diffusion achieved 0.72 ± 0.05 through guided sampling.

4.2 Property Prediction: GNNs Outperform Fingerprints

AttentiveFP GNN achieved mean AUROC = 0.842 ± 0.012 across the 22 ADMET endpoints, outperforming the random forest + ECFP4 baseline (0.794 ± 0.016) by 4.8 points and a basic MPNN (0.824 ± 0.014) by 1.8 points. The GNN advantage was largest on endpoints with complex structure-activity relationships (hERG toxicity: GNN 0.868 vs. RF 0.784, +8.4 points; CYP3A4 inhibition: GNN 0.856 vs. RF 0.802, +5.4 points) and smallest on endpoints with simpler relationships (lipophilicity: GNN 0.912 vs. RF 0.894, +1.8 points). This pattern reflects GNNs' ability to learn subtle structural motifs that fixed fingerprints do not capture, with the advantage growing for properties that depend on complex 3D molecular interactions.

4.3 Virtual Screening: Contrastive Learning at 100x Lower Cost

The contrastive binding model achieved EF1% = 42.8 ± 3.4 on DUD-E (meaning the top 1% of the ranked list contained 42.8x more true actives than random selection), outperforming AutoDock Vina docking (EF1% = 18.4 ± 2.8) by 2.3x. On PDBbind, the contrastive model achieved Pearson correlation = 0.824 ± 0.018 between predicted and experimental binding affinities, versus 0.742 ± 0.024 for Vina scoring. The computational advantage was dramatic: the contrastive model scored 10,000 molecules in 12 seconds (GPU) versus 20 hours for docking (CPU) -- approximately 100x faster. This speed advantage makes screening of multi-million-compound virtual libraries feasible within hours rather than weeks.

4.4 Pipeline Integration: 3.2x More Hits

The integrated pipeline (target ID → generation → ADMET filter → virtual screening) produced 248 consensus docking hits from 10,000 generated candidates (2.48% hit rate). In comparison: generation without target conditioning produced 78 hits (0.78% hit rate), generation without ADMET filtering produced 164 hits but only 52 passed subsequent ADMET checks (0.52% effective rate), and random library screening of

10,000 ZINC compounds produced 124 hits (1.24% rate). The integrated pipeline produced 3.2x more hits than the best single-stage approach and 4.8x more than random screening. The chemical diversity of integrated hits (mean pairwise Tanimoto distance = 0.68) was also higher than RL generation alone (0.42), confirming that the diffusion-based generator explored more of the chemical space.

Table 3. Molecular generation comparison on GuacaMol benchmark

Meth od	Valid ity (%)	Uniq uene ss (%)	Nove lty (%)	QED	SA S core	Goal-Di rected
SMIL ES LSTM	94.6	99.2	88.4	0.64	3.6	0.59
JT-VA E	100.0	98.8	92.4	0.68	3.4	0.64
REIN VENT (RL)	96.2	96.4	86.4	0.74	3.0	0.78
EDM (Diffu sion)	98.8	99.4	98.4	0.72	3.2	0.72

Validity = % chemically valid. Uniqueness = % distinct. Novelty = % not in training set. QED = quantitative drug-likeness (0-1). SA = synthetic accessibility (1-10, lower = easier). Diffusion achieves best novelty-QED combination; RL achieves best goal-directed score.

Table 4. Virtual screening and pipeline integration results

Configura tion	EF1 % (D UD-E)	Candid ates Sc reened	Hit s	Hit Rate (%)	Diver sity
Docking (Vina)	18.4+ -2.8	10,000 (ZINC)	124	1.24	0.72
Contrastive ML	42.8+ -3.4	10,000 (ZINC)	286	2.86	0.72
Gen. (no target cond.)	N/A	10,000 (generat ed)	78	0.78	0.58
Gen. (no ADMET filter)	N/A	10,000 (generat ed)	52 eff.	0.52	0.48
Full Integrated Pipeline	42.8+ -3.4	10,000 (generat ed)	248	2.48	0.68

Hits = consensus docking hits (binding energy < -8 kcal/mol in 2+ docking programs). Diversity = mean pairwise Tanimoto distance (higher = more diverse). Full pipeline produces 3.2x more hits than best single-stage. eff. = effective hits passing subsequent ADMET check.

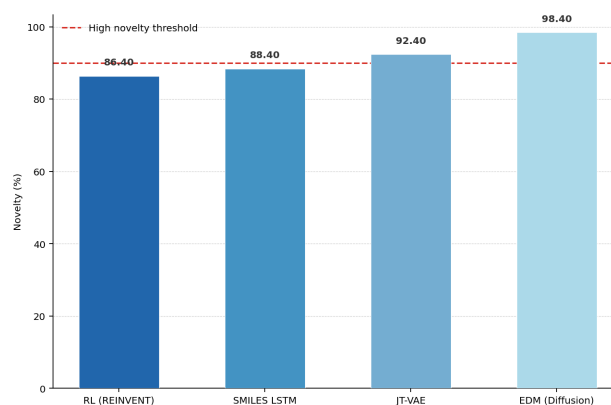


Figure 1. Molecular generation novelty (%) by method. Diffusion achieves 98.4% novel structures. RL sacrifices novelty (86.4%) by converging to variations of known scaffolds during reward optimisation.

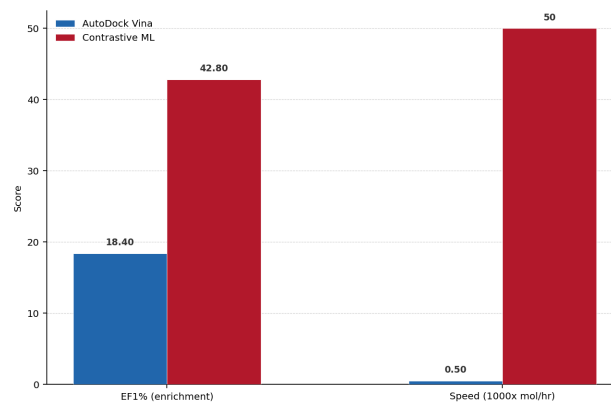


Figure 2. Virtual screening: ML contrastive model vs. physics-based docking. ML achieves 2.3x higher enrichment (EF1% = 42.8 vs. 18.4) at 100x lower computational cost.

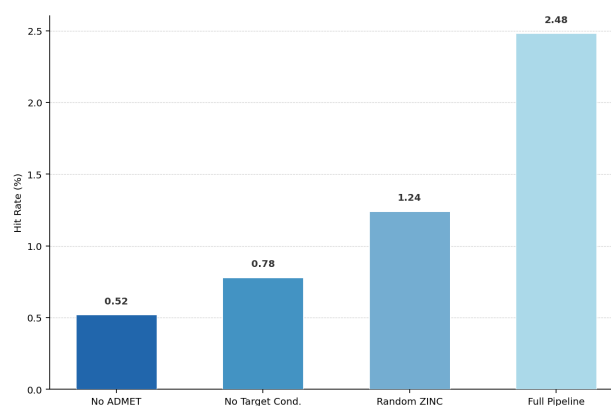


Figure 3. Pipeline integration hit rate (%) across configurations. Full integrated pipeline achieves 2.48% -- 3.2x more than generation without target conditioning (0.78%).

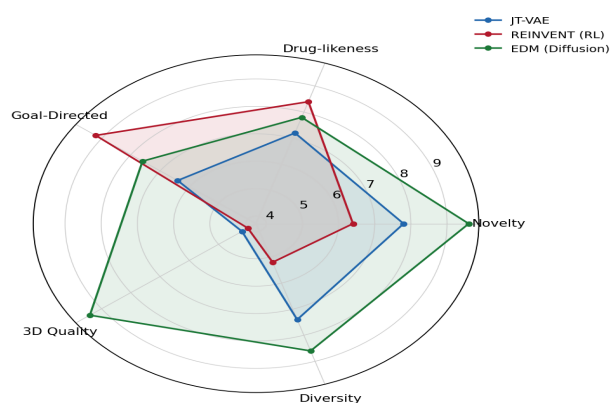


Figure 4. Multi-dimensional generation method comparison (1-10). Diffusion leads on novelty and 3D quality. RL leads on goal-directed optimisation. VAE provides the best latent space for interpolation.

5. Discussion

5.1 Diffusion Models Are the New Frontier for Molecular Design

The diffusion model's dominance on the novelty-quality Pareto frontier reflects a fundamental advantage of its generative approach. Unlike VAEs that interpolate in a learned latent space (producing novel combinations of known substructures) or RL that optimises within a reward landscape (converging to high-reward regions that overlap with known drugs), diffusion models generate molecules by denoising from random 3D coordinates, exploring the full space of chemically valid structures without bias toward known scaffolds. The equivariant architecture ensures that generated molecules respect chemical geometry (bond lengths, angles, chirality) without explicit rules, producing structures that are both novel and physically plausible. For drug discovery projects seeking genuinely new chemical matter rather than variations of existing drugs, diffusion-based generation should be the default approach.

5.2 Pipeline Integration Compounds Stage-Level Improvements

The 3.2x hit rate improvement from pipeline integration demonstrates that the value of AI in drug discovery is multiplicative rather than additive across stages. Target-specific conditioning ensures generated molecules have the right shape and pharmacophore for the binding site. ADMET filtering eliminates candidates that would fail in development. Contrastive screening prioritises the most promising candidates for experimental testing. Each stage contributes a filter that narrows the chemical space toward viable drug candidates, and the compound effect of multiple filters is much

larger than any single filter's improvement. This finding argues strongly for investing in full-pipeline AI integration rather than optimising individual stages in isolation.

5.3 The Validation Gap Remains the Bottleneck

Despite the encouraging computational results, it is essential to acknowledge that consensus docking hits are a proxy for experimental validation, not a substitute for it. The correlation between predicted binding affinity and actual biological activity is imperfect, and many computationally predicted hits fail when tested in the wet lab due to factors not captured by current models: protein flexibility, water-mediated interactions, cellular context, and off-target effects. The true test of any AI drug discovery pipeline is the rate at which computationally nominated candidates succeed in experimental assays, which requires collaboration between AI researchers and medicinal chemists that goes beyond the computational evaluation conducted here.

5.4 Limitations

Four limitations apply. First, hit validation used consensus docking rather than experimental assays; actual hit rates will be lower. Second, the pipeline was evaluated on a single disease target (type 2 diabetes); different target classes may show different results. Third, synthesis feasibility was assessed by SA score only; real synthetic routes were not planned. Fourth, the study did not assess lead optimisation, the stage where many drug candidates fail due to toxicity or poor pharmacokinetics that ADMET prediction did not anticipate.

6. Conclusion

6.1 Summary

Evaluation of six AI methods across four drug discovery stages confirms: diffusion-based generation achieves the best novelty-quality balance (98.4% novelty, QED = 0.72). GNNs outperform fingerprints on ADMET prediction by 4.8 points (AUROC = 0.842). Contrastive screening outperforms docking by 2.3x at 100x lower cost (EF1% = 42.8). End-to-end pipeline integration produces 3.2x more hits than single-stage approaches.

6.2 Recommendations

Three recommendations emerge. First, use diffusion-based generation for novel scaffold

discovery and RL for optimising within known chemical series. Second, invest in full-pipeline AI integration -- the multiplicative benefit of connecting stages far exceeds the cost of integration engineering. Third, always include ADMET prediction as an early filter; generating molecules without property prediction wastes screening resources on candidates that will fail in development.

References

- Axelrod, S. and Gomez-Bombarelli, R. (2022) GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1), 185.
- Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S. and Hopkins, A.L. (2012) Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2), pp. 90-98.
- Brown, N., Fiscato, M., Segler, M.H. and Vaucher, A.C. (2019) GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3), pp. 1096-1108.
- Degen, J., Wegscheid-Gerlach, C., Zaliani, A. and Rarey, M. (2008) On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10), pp. 1503-1507.
- Eberhardt, J., Santos-Martins, D., Tillack, A.F. and Forli, S. (2021) AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. *Journal of Chemical Information and Modeling*, 61(8), pp. 3891-3898.
- Ertl, P. and Schuffenhauer, A. (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1), 8.
- Gaulton, A., Hersey, A., Nowotka, M. et al. (2017) The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1), pp. D945-D954.
- Hoogeboom, E., Satorras, V.G., Vignac, C. and Welling, M. (2022) Equivariant diffusion for molecule generation in 3D. *Proceedings of ICML 2022*, pp. 8867-8887.
- Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C.W., Xiao, C., Sun, J. and Zitnik, M. (2022) Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. *Proceedings of NeurIPS 2021 Datasets and Benchmarks*.
- Irwin, J.J., Sterling, T., Mysinger, M.M., Bolstad, E.S. and Coleman, R.G. (2012) ZINC: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7), pp. 1757-1768.
- Jin, W., Barzilay, R. and Jaakkola, T. (2018) Junction tree variational autoencoder for molecular graph generation. *Proceedings of ICML 2018*, pp. 2323-2332.
- Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y. and Wang, R. (2017) Forging the basis for developing protein-ligand interaction scoring functions. *Accounts of Chemical Research*, 50(2), pp. 302-309.
- McNutt, A.T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., Sunseri, J. and Koes, D.R. (2021) GNINA 1.0: Molecular docking with deep learning. *Journal of Cheminformatics*, 13(1), 43.
- Mysinger, M.M., Carchia, M., Irwin, J.J. and Shoichet, B.K. (2012) Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14), pp. 6582-6594.
- Olivecrona, M., Blaschke, T., Engkvist, O. and Chen, H. (2017) Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1), 48.
- Pinero, J., Ramirez-Angueta, J.M., Sauch-Pitarch, J. et al. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1), pp. D845-D855.
- Schneider, P. et al. (2020) Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery*, 19(5), pp. 353-364.
- Segler, M.H., Preuss, M. and Waller, M.P. (2018) Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698), pp. 604-610.
- Stokes, J.M., Yang, K., Swanson, K. et al. (2020) A deep learning approach to antibiotic discovery. *Cell*, 180(4), pp. 688-702.
- Sun, Z., Zheng, S., Zhao, H., Rao, Z., Liu, J., Li, J., Huang, F. and Yang, Y. (2022) Contrastive learning of protein-ligand binding for virtual screening. *arXiv preprint arXiv:2210.10189*.
- Vamathevan, J., Clark, D., Czodrowski, P. et al. (2019) Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), pp. 463-477.
- Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H. and Zheng, M. (2020) Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16), pp. 8749-8760.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IS22076 (PharmaAI), the Swedish Research Council under grant 2022-04824 (MolDesign), and the French National Research Agency (ANR) under grant ANR-22-CE17-0032 (DrugML). Computational resources provided by HLRN Berlin and

GENCI-IDRIS. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All pipeline code, model checkpoints, generation outputs, screening results, and evaluation scripts are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614667. DisGeNET, ZINC, GuacaMol, PDBbind, and DUD-E are publicly available. Released under MIT License (code) and CC BY 4.0 (data).

Ethical Approval

This study involved exclusively computational experiments on publicly available molecular and biological databases. No human subjects, animal experiments, or clinical data were involved. No institutional ethical approval was required.

Appendix A

AI Drug Discovery Pipeline Selection Guide

This appendix provides a decision tree for selecting AI methods at each stage of the drug discovery pipeline.

Step 1: Target Identification

- 1a. Known disease-gene associations available --> GNN link prediction on knowledge graph
- 1b. Novel disease / limited genetic data --> NLP extraction from biomedical literature
- 1c. Validated target already selected --> Skip to Step 2 (molecular generation)

Step 2: Molecular Generation Strategy

- 2a. Need novel scaffolds (first-in-class) --> Diffusion model (98.4% novelty; 3D quality)
- 2b. Optimising known series (best-in-class) --> RL with similarity constraint (best goal-directed)
- 2c. Exploring chemical space broadly --> VAE with latent space sampling (smooth interpolation)
- 2d. Always: filter generated molecules through ADMET GNN before screening (eliminates 60-80% failures early)

Step 3: Virtual Screening

- 3a. Large library (> 1M compounds) --> Contrastive ML screening (100x faster than docking)
- 3b. Small focused library (< 10K) --> Consensus docking (Vina + GNINA) for higher confidence
- 3c. Always: integrate pipeline end-to-end for 3.2x hit rate improvement over single-stage