



Journal of Advanced Pharmaceutical Sciences and Natural Products (JAPSNP)

GENERATIVE ARTIFICIAL INTELLIGENCE IN DE NOVO DRUG DESIGN: VALIDATION FRAMEWORKS, BENCHMARKING FAILURES AND WHAT THE FIRST AI-DISCOVERED CLINICAL CANDIDATES HAVE ACTUALLY TAUGHT US

Dr. Shivaraj D R^{*1}

Ms. Rachana M A²

Ms. T Pushpanjali²

¹Assistant Professor, Department of Pharmacy Practice, Akshaya Institute of Pharmacy, Tumkur 572106, Affiliated to RGUHS, Bengaluru, India. ORCID: 0000-0002-9284-4360

²3rd Pharm D, Department of Pharmacy Practice, Akshaya Institute of Pharmacy, Tumkur 572106, Affiliated to RGUHS, Bengaluru, India.

Corresponding author: Dr. Shivaraj D R, Assistant Professor, Department of Pharmacy Practice, Akshaya Institute of Pharmacy, Tumkur 572106, Affiliated to RGUHS, Bengaluru, India..

Email: Shivaraj.dr27@gmail.com.

ORCID: 0000-0002-9284-4360

ABSTRACT

Generative models now design molecules satisfying almost any computable objective, and the field's benchmarks record that success faithfully. Whether the success means anything is a separate question, answered here from three directions: the internal failure modes of the benchmarks, the prospective exercises that test predictions against experiment, and the clinical record of compounds that reached patients. Goal-directed generators score highly on tasks whose global optimum random sampling can reach, produce molecules that retrosynthesis software

solves only 30.2% of the time when synthesizability is not explicitly rewarded, and exploit scoring functions so reliably that molecules ranked highest by an optimisation model score below 0.2 on an independently trained control. Beneath them lie biased datasets: a network trained on DUD-E reaches an AUC of 0.98 with the protein deleted at test time and falls to 0.53 once the decoy artefact is removed. Deep learning docking methods produce physically valid, accurate poses for 12% of a post-cutoff benchmark on which AutoDock Vina manages 58%, and cofolding accuracy falls to between 10% and 20% on ligands unlike the training data. Prospectively, 23 teams competing against the LRRK2 WD40 domain achieved a hit rate of 3.7% with affinities between 18 and 140 micromolar. The clinical record is more equivocal than either advocates or critics allow: AI-derived molecules completed phase 1 in 21 of 24 cases but only 4 of 10 phase 2 trials, the historical industry rate. Rentosertib raised forced vital capacity by 98.4 mL against a placebo decline of 20.3 mL in 71 patients and dosed its first phase 3 patient in September 2026, and no AI-designed small molecule has yet completed phase 3. Generative chemistry has largely solved the design of molecules satisfying stated objectives and has not solved the statement of objectives worth satisfying, which is why validation must gate on generalisation and prospective experiment rather than on optimisation score.

Keywords: Drug Discovery; Artificial Intelligence; Deep Learning; Molecular Docking Simulation; Clinical Trials; Phase II as Topic; Reproducibility of Results; Pulmonary Fibrosis.



1. Introduction

Pharmaceutical research and development productivity declined for six decades at a rate stable enough to acquire a name, and the diagnosis offered in 2012 attributed most of it to the difficulty of the remaining biology rather than to inefficiency in the laboratory ¹. Generative chemistry entered against that backdrop with a specific promise: if the bottleneck is the cost of proposing and testing candidate molecules, a model that proposes better candidates should move the whole curve. By 2022 the twenty AI-native companies whose pipelines could be reconstructed held roughly 160 disclosed discovery and preclinical programmes and about 15 clinical assets ². Two years later the same analysts counted 75 molecules from AI-native biotechs and their partners entering the clinic since 2015 ³.

The promise is therefore no longer untestable, and three evidentiary streams now exist where once there was only benchmark performance. They disagree, and the disagreement resolves into a single diagnosis. Generative models have become extremely good at optimising whatever function is placed in front of them, and the field's evaluation apparatus measures precisely that capability while failing to measure whether the function was worth optimising. The failure is structural rather than incidental, and it recurs at every level from the scoring function to the training set to the choice of held-out data. Where prospective experiment has been allowed to adjudicate, measured performance has been modest. Where clinical data now exist, the

pattern of success and failure locates the benefit with unexpected precision: AI-derived molecules pass safety-based phase 1 gates at unusually high rates and fail phase 2 efficacy at ordinary ones, which is what one would expect of a technology that optimises molecular properties well and biological hypotheses no better than anyone else.

2. Scope and evidence base

Coverage was built by searching PubMed, Europe PMC, arXiv and bioRxiv for generative molecular design, benchmark critique, structure-based virtual screening bias and AI-discovered clinical candidates from 2018 to September 2026, then snowballing from the reference lists of the retrieved critiques. Clinical outcomes not yet in the peer-reviewed literature were taken from company investor releases and trial registries and are labelled as such, because the alternative is to omit the failures and report only what companies chose to publish. Small molecules are the focus. Claims that could not be verified against a retrieved source were dropped rather than hedged.

3. Benchmarks that cannot fail

GuacaMol and MOSES were built to end an era in which every generative chemistry paper defined its own success criteria ^{4,5}. They succeeded at standardisation and, in doing so, fixed in place a definition of success that has aged badly. GuacaMol's goal-directed tasks are constructed from physicochemical and similarity scoring functions rather than from anything resembling a binding measurement ⁴, while MOSES evaluates a generator against the



distributional properties of its 1.58 million-molecule training partition, which rewards faithful reproduction of ZINC Clean Leads ⁵.

Saturation followed quickly. A re-examination of 25 algorithms across 23 oracle tasks under a budget of 10,000 oracle calls found the drug-likeness objective QED to have a probable global maximum of 0.948 that random sampling reaches, and disabled it because it no longer distinguished between methods ⁶. None of the implemented methods optimised even the toy objectives within hundreds of oracle calls, which matters because a real assay budget is measured in hundreds.

Three failure modes sit underneath the saturation, progressively more serious.

The first is synthesizability. Goal-directed generation on hard tasks, run without any synthesizability term in the objective, produced molecules that a retrosynthetic planner could solve for only 30.2% of ChEMBL-trained output and 32.7% of MOSES-trained output ⁷. Adding a synthetic accessibility term raised those figures to 80.2% and 77.2%, which sounds like a solution until one reads the accompanying finding that no method improved synthesizability relative to its own training set: 89.8% of MOSES molecules and 68.3% of ChEMBL molecules were solvable, against 3.5% for the enumerated GDB-17 ⁷. Generative models inherit tractability rather than creating it (Figure 1a).

The second is that optimisation against a structural objective does not beat looking. Asked to produce 250 unique drug-like molecules

maximising a SMINA docking score across eight targets, the best of three architectures reached roughly minus 9.774 on 5-HT1B, while the top decile of ZINC reached minus 9.894 ⁸. A generator trained on between 1,074 and 10,225 molecules was outperformed by screening a library that already existed. The output is as instructive as the score: one model produced macrocycles of doubtful drug-likeness, another molecules frequently containing fewer than three heavy atoms, and the best-scoring model collapsed to an internal diversity of 0.506 mean Tanimoto distance against 0.787 for its training set ⁸.

The third failure mode is the most consequential, because it cannot be fixed by adding a term to the objective. When a molecule is optimised against a machine-learned activity model and the top-scoring output is then scored by a second model trained independently on the same data, the two disagree systematically at the high-score extreme. Quantified across DRD2, EGFR and JAK2, the mean absolute difference between optimisation and control scores was approximately 0.3, and on EGFR molecules scoring 0.5 to 0.6 on the optimisation model scored below 0.2 on the control ⁹. The proximate cause is unglamorous: datasets of 667 to 842 molecules containing 40 to 140 actives support many statistically indistinguishable models that diverge sharply where the optimiser is pushing ⁹. A generator given such a surrogate does not find active molecules. It finds the surrogate's errors. Multi-objective design on EGFR without reliability constraints produced exactly this



signature: the molecules it favoured had high predicted values but lacked the quinazoline substructure common to validated inhibitors ¹⁰.

4. The bias beneath the benchmark

The surrogate problem has a structural analogue in the datasets on which structure-based methods are trained and tested, and the DUD-E analysis that exposed it remains the most economical demonstration of what a benchmark can measure without anyone noticing. A convolutional network trained on protein-ligand complexes from DUD-E achieved an AUC of 0.98, which would be an excellent result if it reflected learned interaction physics. It did not. A ligand-only network reached the same 0.98, and a network trained with the receptor present and then tested with the receptor deleted also reached 0.98, a mean difference of 0.0006 ¹¹. A nearest-neighbour classifier on ligand fingerprints exceeded an AUC of 0.90 on 97 of 102 targets, and 74 targets could be separated by models trained on functionally unrelated targets. Most damagingly, a model trained on ten targets and tested on the other 92 scored 0.80, falling to 0.53, indistinguishable from chance, once actives from other targets replaced the property-matched decoys ¹¹. Physics-based and differently trained baselines on the same targets scored between 0.632 and 0.725, and the gap between those numbers and 0.98 is the size of the artefact.

The mechanism was understood before it was demonstrated this starkly. Asymmetric validation embedding quantifies train-test redundancy while accounting for similarity among inactives rather

than actives alone, and across seven widely used benchmarks that measure correlated strongly with reported performance regardless of fingerprint, similarity metric or previously applied unbiasing ¹². The conclusion drawn then, that most reported ligand-based performance reflects overfitting rather than prospective accuracy, has not been overturned, and unbiased alternatives built from dose-response assays with confirmed inactives now exist.

None of this is unique to chemistry. A survey of machine learning across the sciences documented leakage in 17 fields, affecting 294 papers, and named the two variants that dominate here: temporal leakage and train-test non-independence, where analogue series straddle the split ¹³. Chemistry's contribution is to have both in its standard benchmarks simultaneously.

5. Structure prediction as a test of generalisation

Accurate structure prediction promised to replace learned surrogates with something closer to physics, and evaluating that promise produced the field's most useful methodological innovation to date. PoseBusters made explicit what root-mean-square deviation had been quietly permitting. A pose can sit within 2 Å of the crystallographic ligand while containing distorted bond lengths, non-planar aromatic rings, inverted stereocentres, internal steric clashes or overlap with the protein itself, and a gate of chemical and physical checks applied alongside RMSD separates the two kinds of success ¹⁴. Applying it was decisive. On the



Astex Diverse Set, DiffDock's 72% success at the 2 Å threshold fell to 47% once physical validity was also required, and Uni-Mol's 45% fell to 12%, while Gold declined from 67% to 64% and AutoDock Vina from 58% to 56% (Figure 1b) ¹⁴. On 308 complexes released after the deep learning methods' training cutoffs the ordering inverted outright: Vina at roughly 58% and Gold at roughly 55% accurate and valid, against 12% for the best deep learning method. Below 30% sequence identity to PDBbind, almost no physically valid poses within 2 Å were generated by any deep learning method ¹⁴.

AlphaFold 3 adopted the PoseBusters set as its ligand evaluation benchmark and reported statistically significant advantages over Vina and over blind cofolding alternatives ¹⁵, and the gate became a de facto standard within months. It did not establish that the underlying generalisation problem had been solved.

Two subsequent lines of work indicate that it has not. A benchmark of 2,600 protein-ligand systems released after all relevant training cutoffs found success rates of roughly 60% to 70% in the highest bin of ligand similarity to training data and roughly 10% to 20% in the lowest, in AlphaFold 3, Boltz-1, Chai-1 and Protenix alike ¹⁶. Ligands with fewer than 100 analogous training structures fell to 10% to 20%, against 40% to 50% for prevalent ligands and cofactors. The same analysis reported that 87% of the PoseBusters benchmark itself carries substantial pocket similarity to the training sets, that model ranking performed not much better than random selection, and that a successor

model trained on more than 25,000 additional PDB entries showed no discernible improvement ¹⁶. These findings come from a preprint and warrant the corresponding caution, but their direction is corroborated by an independent ablation: mutating every binding-site side chain to glycine left cofolding models predicting the same binding mode, with only 42% to 52% of correctly predicted systems changing pose after the interaction surface was destroyed ¹⁷.

For de novo design this is the decisive result rather than a caveat, because a novel chemotype against a target with no close structural precedent is, by construction, the low-similarity case.

6. What prospective experiment reports

Prospective, blinded evaluation with an experimental readout is rare, expensive and by far the most informative evidence available. The CACHE initiative was established to supply it ¹⁸, and its first completed challenge should be read closely by anyone estimating the field's capability. Twenty-three computational teams predicted ligands for the WD40 repeat domain of LRRK2. Of 1,955 nominated molecules, 1,875 were successfully procured, a 93% synthesis rate, and tested by surface plasmon resonance with orthogonal confirmation by differential scanning fluorimetry, fluorine NMR, dynamic light scattering and calorimetry ¹⁹. Seventy-three compounds bound with a dissociation constant below 150 micromolar, a hit rate of 3.7%. Analogue expansion then tested 714 compounds and found 61 with measurable affinity, 31 of them below 150 micromolar, and seven chemical



series survived orthogonal confirmation with affinities between 18 and 140 micromolar¹⁹.

That is the honest measure of the state of the art against a hard, ligand-poor target: a few percent hit rate and high-micromolar starting points. It is a real result and not a trivial one, and it sits far from what benchmark scores imply.

Retrospective evaluation, done honestly, agrees. The most demanding retrospective test yet published asked whether a model trained on a project's early-stage compounds could regenerate its middle and late-stage compounds. On public project data, rediscovery reached 1.60% within the top 100 candidates. On proprietary in-house projects it reached 0.00%, 0.03% and 0.04% at the top 100, 500 and 5,000²⁰. Similarity relationships also inverted between the two settings, consistent with real programmes optimising many parameters at once while public datasets reward potency alone, and the authors concluded that retrospective validation of de novo design may be impossible to do meaningfully. An editorial position taken five years ago, that the value of generative models would be shown through synthesis and biological evaluation rather than documentation-poor reports of in silico performance²¹, reads now less as prophecy than as a description of what the field declined to do.

7. What the clinical candidates have actually taught us

Clinical data cut through the benchmark debate because they are generated by a process nobody can game. Table 1 summarises the candidates

with reported outcomes and Figure 2 the phase-level statistics.

The headline statistics are genuinely favourable and widely misread. Of 24 AI-derived molecules that completed phase 1, 21 succeeded, 87.5% against a historical comparator between roughly 40% and 65%; of ten that completed phase 2, four succeeded, approximately 40%, against a historical 30% to 40% (Figure 2)³. The authors called their work preliminary, and the phase 2 figure rests on ten trials, so a single additional outcome moves it by ten percentage points.

The shape of the result matters more than its precision. A phase 1 gate is predominantly a safety and tolerability gate, and it tests exactly what generative chemistry optimises well: potency, selectivity, absorption, distribution, metabolism, excretion and the avoidance of known structural liabilities. A phase 2 gate tests whether modulating the chosen target changes the chosen disease in humans, which no amount of molecular optimisation addresses. The data are internally consistent with a technology that has improved the molecule and left the hypothesis untouched, and one peer-reviewed assessment states this plainly: AI efforts have not so far produced more effective drugs, and the phase 1 advantage does not carry through²².

The individual histories bear this out, because the failures are failures of biology.

DSP-1181, a 5-HT1A agonist for obsessive-compulsive disorder, was the first AI-designed molecule to enter phase 1, in January 2020²³, and was discontinued after phase 1 despite a



favourable safety profile ²⁴. EXS-21546, an adenosine A2A antagonist, was withdrawn from phase 1/2 roughly ten months after its trial authorisation, the company citing difficulty in reaching a suitable therapeutic index ²⁵. BEN-2293, a topical pan-Trk inhibitor, enrolled 91 patients with mild-to-moderate atopic dermatitis, met its primary safety endpoint and missed its secondary efficacy endpoints, with no effect on pruritus and no EASI benefit outside a post-hoc subgroup with more than 20% affected body surface area ²⁶. REC-994 in cerebral cavernous malformation followed the identical pattern in 62 participants: primary safety endpoint met, patient-reported and physician-reported improvement absent at 12 months, only a trend toward reduced lesion volume on exploratory imaging ²⁷. Its long-term extension then showed the 400 mg arm to be indistinguishable from natural history, and the programme was discontinued alongside REC-2282, which had passed a futility threshold with limited tumour shrinkage across all arms ²⁸.

In each case the molecule behaved as designed and the target did not deliver. That is a coherent finding rather than a list of misfortunes, and it locates the residual attrition.

Rentosertib is the most informative case precisely because both halves came from the platform. A target discovery engine nominated TRAF2- and NCK-interacting kinase as an antifibrotic target *de novo*, a generative chemistry engine produced the inhibitor, and preclinical candidate nomination followed roughly 18 months later ²⁹. The lineage traces to

the 2019 report in which DDR1 inhibitors were generated in 21 days, six compounds synthesised and four found active ³⁰, work that also shows why novelty claims require scrutiny: independent cheminformatic analysis showed the lead closely resembling ponatinib, a marketed drug with published DDR1 activity and published analogue SAR ³¹. That analysis remains grey literature and no formal published rebuttal exists, which is itself a comment on the field's critical apparatus.

The phase 2a trial randomised 71 of 128 screened patients across 21 Chinese sites to placebo or one of three rentosertib regimens for 12 weeks ³². Treatment-emergent adverse events were similar across arms, 70.6% on placebo against 72.2%, 83.3% and 83.3%, but treatment-related events rose monotonically with dose, from 29.4% on placebo to 77.8% at 60 mg once daily, and treatment-related serious events from none to 11.1%. Sixteen patients discontinued, twelve for adverse events, seven of those for liver injury or dysfunction; four of the seven were taking nintedanib concurrently. Forced vital capacity rose by 98.4 mL at 60 mg once daily, confidence interval 10.9 to 185.9, against a decline of 20.3 mL on placebo, and the corresponding 2.82% mean change meets the 2% to 6% minimal clinically important difference for this endpoint ^{32,33}. Patients on 60 mg without concurrent antifibrotic gained 187.8 mL; those taking nintedanib or pirfenidone gained nothing measurable. Serum proteomics showed dose-dependent falls in COL1A1, FAP, FN1 and MMP10, with extracellular matrix organisation



the most downregulated pathway, which is the mechanistic corroboration the FVC signal needs.

Read carefully, this is a positive phase 2a with four material qualifications: arms of 18 patients, a 22.5% discontinuation rate, a dose-dependent hepatic signal concentrated in patients on nintedanib, and three acute exacerbations in the 60 mg arm against one on placebo. The authors name the homogeneity of the Chinese cohort and the 12-week duration as limitations³². A 320-patient, 47-centre phase 3 with a 52-week FVC decline endpoint dosed its first patient in September 2026³⁴, so the earliest possible readout is late 2027. As of September 2026, no AI-designed small molecule has completed a phase 3 trial or received regulatory approval^{24,35}.

The sector's structure meanwhile contracted along with its pipelines. The two most prominent AI-native discovery companies merged in November 2024³⁶ while both were discontinuing clinical assets, and a third delisted from Euronext Amsterdam in March 2025, two years after its phase 2a efficacy miss³⁷. Consolidation accompanied pipeline contraction rather than expansion, and a field whose clinical output was compounding would look different.

8. A validation framework that would bind

The diagnosis dictates the remedy. Four elements are available; only the mechanism to compel them is missing.

Physical and chemical validity belongs in a gate rather than a reported metric. The PoseBusters convention, counting a pose as a success only if

it is both accurate and physically valid, cost nothing to adopt and immediately reordered the published rankings¹⁴. Its equivalent for generative chemistry is a retrosynthesis and validity gate applied before any score is reported, given that unbiased goal-directed output is tractable barely 30% of the time⁷.

Evaluation should be stratified by distance from training data, not averaged over it. A single aggregate number over a benchmark in which 87% of entries resemble the training set¹⁶ reports the memorisation regime and conceals the regime de novo design occupies. Binning performance by similarity to training data converts a misleading average into a usable description.

Surrogate reliability should be established before optimisation rather than inferred afterwards. The control-model comparison is cheap, and the divergence it exposes is large enough to invalidate an entire campaign⁹; reliability thresholds on each objective are one workable implementation¹⁰.

Prospective experimental confirmation should be the minimum evidence standard for any claim that a generative method works. That is the level at which CACHE operates, and the level at which the honest answer turned out to be 3.7%¹⁹. A proposed tiered scheme puts generative molecular design at exactly this tier, below benchmark robustness and above only internal reproducibility³⁸. It is a single-author proposal rather than a community standard, but no competing framework has been offered.



Reporting checklists exist in adjacent fields and work. The four-domain data, optimisation, model and evaluation structure developed for supervised machine learning in biology maps onto generative chemistry with minimal adaptation ³⁹, and clinical prediction modelling, a field with an identical leakage pathology, produced a consensus reporting statement that journals now enforce ⁴⁰. Generative chemistry has no equivalent, and regulation will not supply one. The most developed regulatory credibility framework, a seven-step risk-based assessment anchored to a defined context of use, scopes itself to nonclinical, clinical, postmarketing and manufacturing applications and excludes drug discovery outright ⁴¹. The European reflection paper at least names discovery, setting a conditional trigger under which nonclinical principles apply once results enter the dossier, but imposes nothing on the design step ⁴². Neither agency regulates molecular provenance, and neither should: a compound is assessed on its evidence whatever proposed it. No external body will therefore require generative chemistry to validate itself, so journals and the community must.

9. Conclusion

The evidence assembled here supports a narrower and more useful claim than either the

promotional or the dismissive account of this field. Generative models design molecules that satisfy computable objectives with high reliability, and the clinical record confirms it: an 87.5% phase 1 completion rate is a real achievement in molecular optimisation. The same record shows the benefit stopping where the objective stops, at a phase 2 success rate of roughly 40% that matches the historical industry figure, and at a set of terminated programmes in which the molecule performed and the target did not. Benchmarks obscured this because they measure optimisation against surrogates that are themselves the weakest component; prospective experiment revealed it because it cannot be optimised against. The field's most valuable near-term investment is therefore not a better generator but a validation standard that gates on generalisation and prospective confirmation, adopted by journals because no regulator will impose it. Rentosertib's 52-week phase 3, reading out no earlier than late 2027, will supply the first datum on whether an AI-nominated target survives a properly powered efficacy trial, and that single result will inform the field's prospects more than any benchmark published in the interim.

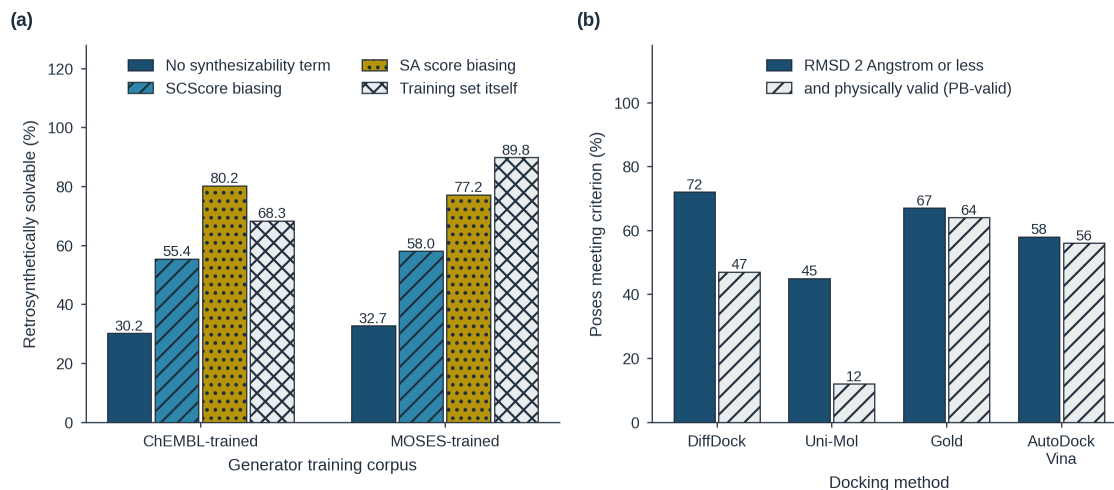


Figure 1. Two gates that reverse published rankings in generative chemistry and structure-based docking. (a) Retrosynthetic solvability of molecules produced by goal-directed generation, as a percentage of generated molecules for which a computer-aided synthesis planner found a route. Bars show generators trained on ChEMBL and on MOSES under four conditions: no synthesizability term in the objective, biasing by SCScore, biasing by synthetic accessibility (SA) score, and the solvability of the generator's own training corpus. No condition exceeds the training corpus by a meaningful margin. Values are percentages of generated molecules and carry no dispersion, because each is a single reported proportion rather than a repeated measurement; data from Gao and Coley⁷. (b) Docking poses meeting the conventional 2 Å root-mean-square deviation criterion alone (solid bars) and meeting it while also passing the PoseBusters physical and chemical validity checks (hatched bars), on the Astex Diverse Set. The two deep learning methods lose 25 and 33 percentage points respectively; the two classical methods lose 3 and 2.

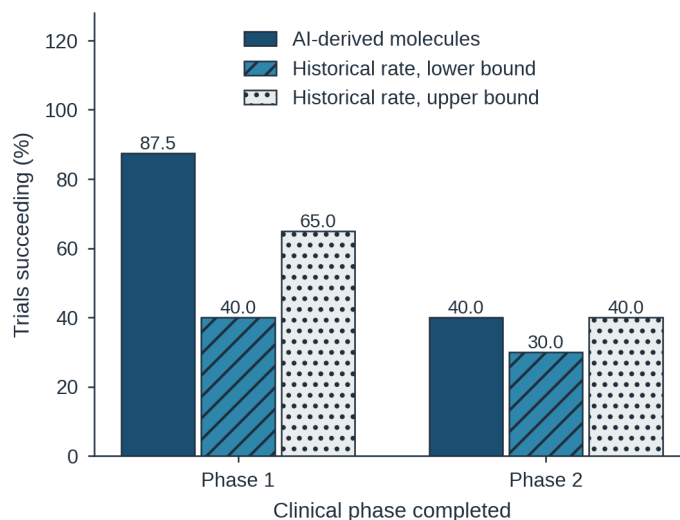


Figure 2. The clinical advantage of AI-derived molecules is confined to the safety-based gate. Bars show the percentage of completed trials judged successful for AI-derived molecules (solid) against the lower and upper bounds of the historical industry comparator reported alongside them (hatched). At phase 1 the AI-derived rate of 87.5%, from 21 successes in 24 completed trials, exceeds the upper bound of the historical range. At phase 2 the AI-derived rate of approximately 40%, from 4 successes in 10 completed trials, sits exactly at the upper bound of the historical range. No confidence intervals are shown because the source reports proportions without them, and the phase 2 denominator of ten trials means a single additional outcome shifts that bar by ten percentage points. Data from Jayatunga et al.³.



Table 1. Small molecules from generative or AI-native discovery platforms with a reported clinical outcome, as of September 2026. Outcomes are given in the sponsor's or authors' own primary and secondary endpoint terms. IPF, idiopathic pulmonary fibrosis; FVC, forced vital capacity; TEAE, treatment-emergent adverse event; OCD, obsessive-compulsive disorder; EASI, Eczema Area and Severity Index.

Candidate	Sponsor	Target and indication	Highest phase reported	Reported outcome	Source class
DSP-1181	Exscientia and Sumitomo Dainippon	5-HT1A agonist, OCD	Phase 1	First AI-designed molecule dosed in humans, January 2020; discontinued after phase 1 despite a favourable safety profile	Company release ²³ ; peer-reviewed ²⁴
EXS-21546	Exscientia	Adenosine A2A antagonist, solid tumours	Phase 1/2	Withdrawn about ten months after trial authorisation; therapeutic index judged unattainable	Trade press ²⁵
BEN-2293	BenevolentAI	Topical pan-Trk inhibitor, atopic dermatitis	Phase 2a	Primary safety endpoint met in 91 patients; secondary efficacy endpoints missed, with no effect on pruritus and no EASI benefit outside a post-hoc subgroup	Company release ²⁶
REC-994	Recursion	Cerebral cavernous malformation	Phase 2	Primary safety endpoint met in 62 participants; no functional improvement at 12 months, only an imaging trend; discontinued after the extension showed results indistinguishable from natural history	Company releases ^{27,28}
REC-2282	Recursion	NF2-mutated meningioma	Phase 2	Futility threshold passed with limited tumour shrinkage across all arms; discontinued	Company release ²⁸
Rentosertib (ISM001-055)	Insilico Medicine	TNIK inhibitor, IPF	Phase 3 ongoing	Phase 2a in 71 patients: TEAE rates comparable to placebo, dose-dependent hepatic signal, FVC +98.4 mL at 60 mg once daily against 20.3 mL decline on placebo; phase 3 first patient dosed September 2026	Peer-reviewed ³² ; company release ³⁴

Declarations

Funding: None.

Conflicts of interest: None declared.



References

1. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov.* 2012;11(3):191-200. doi:10.1038/nrd3681
2. Jayatunga MKP, Xie W, Ruder L, Schulze U, Meier C. AI in small-molecule drug discovery: a coming wave? *Nat Rev Drug Discov.* 2022;21(3):175-176. doi:10.1038/d41573-022-00025-1
3. Jayatunga MKP, Ayers M, Bruens L, Jayanth D, Meier C. How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons. *Drug Discov Today.* 2024;29(6):104009. doi:10.1016/j.drudis.2024.104009
4. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: benchmarking models for de novo molecular design. *J Chem Inf Model.* 2019;59(3):1096-1108. doi:10.1021/acs.jcim.8b00839
5. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, et al. Molecular Sets (MOSES): a benchmarking platform for molecular generation models. *Front Pharmacol.* 2020;11:565644. doi:10.3389/fphar.2020.565644
6. Gao W, Fu T, Sun J, Coley CW. Sample efficiency matters: a benchmark for practical molecular optimization. Preprint at arXiv:2206.12411 (2022).
7. Gao W, Coley CW. The synthesizability of molecules proposed by generative models. *J Chem Inf Model.* 2020;60(12):5714-5723. doi:10.1021/acs.jcim.0c00174
8. Ciepliński T, Danel T, Podlowska S, Jastrzębski S. Generative models should at least be able to design molecules that dock well: a new benchmark. *J Chem Inf Model.* 2023;63(11):3238-3247. doi:10.1021/acs.jcim.2c01355
9. Langevin M, Vuilleumier R, Bianciotto M. Explaining and avoiding failure modes in goal-directed generation of small molecules. *J Cheminform.* 2022;14(1):20. doi:10.1186/s13321-022-00601-y
10. Yoshizawa T, Ishida S, Sato T, Ohta M, Honma T, Terayama K. A data-driven generative strategy to avoid reward hacking in multi-objective molecular design. *Nat Commun.* 2025;16:2409. doi:10.1038/s41467-025-57582-3
11. Chen L, Cruz A, Ramsey S, et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One.* 2019;14(8):e0220113. doi:10.1371/journal.pone.0220113
12. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. *J Chem Inf Model.* 2018;58(5):916-932. doi:10.1021/acs.jcim.7b00403
13. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-



- based science. *Patterns*. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804
14. Buttenschoen M, Morris GM, Deane CM. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem Sci*. 2024;15(9):3130-3139. doi:10.1039/D3SC04185A
15. Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*. 2024;630(8016):493-500. doi:10.1038/s41586-024-07487-w
16. Škrinjar P, Eberhardt J, Tauriello G, Schwede T, Durairaj J. Have protein-ligand cofolding methods moved beyond memorisation? Preprint at bioRxiv. 2025. doi:10.1101/2025.02.03.636309
17. Masters MR, Mahmoud AH, Lill MA. Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions. *Nat Commun*. 2025;16:8854. doi:10.1038/s41467-025-63947-5
18. Ackloo S, Al-awar R, Amaro RE, et al. CACHE (Critical Assessment of Computational Hit-finding Experiments): a public-private partnership benchmarking initiative to enable computational drug discovery. *Nat Rev Chem*. 2022;6(4):287-295. doi:10.1038/s41570-022-00363-z
19. Li F, Ackloo S, Arrowsmith CH, et al. CACHE Challenge #1: targeting the WDR domain of LRRK2, a Parkinson's disease associated protein. *J Chem Inf Model*. 2024;64(22):8521-8536. doi:10.1021/acs.jcim.4c01267
20. Handa K, Thomas MC, Kageyama M, Iijima T, Bender A. On the difficulty of validating molecular generative models realistically: a case study on public and proprietary data. *J Cheminform*. 2023;15:112. doi:10.1186/s13321-023-00781-1
21. Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. *Nat Biotechnol*. 2020;38(2):143-145. doi:10.1038/s41587-020-0418-2
22. Jacobson RD. The AI drug revolution needs a revolution. *npj Drug Discov*. 2025;2:10. doi:10.1038/s44386-025-00013-6
23. Sumitomo Dainippon Pharma and Exscientia. Sumitomo Dainippon Pharma and Exscientia joint development of DSP-1181, a long-acting serotonin 5-HT_{1A} receptor agonist, enters phase 1 clinical study. Press release, 30 January 2020. Available from: <https://www.sumitomo-pharma.com/news/assets/pdf/ene20200130.pdf>
24. Niazi SK. Artificial intelligence in small-molecule drug discovery: a critical review of methods, applications, and real-world outcomes. *Pharmaceuticals*. 2025;18(9):1271. doi:10.3390/ph18091271
25. Taylor NP. Exscientia trims pipeline to focus on top cancer prospects. *pharmaphorum*, 3 October 2023. Available from:



- <https://pharmaphorum.com/news/exscientia-trims-pipeline-focus-top-cancer-prospects>
26. BenevolentAI. BenevolentAI announces results from phase 2a trial of BEN-2293 in atopic dermatitis. Press release, 4 April 2023. Available from: <https://www.businesswire.com/news/home/20230404006090/en/>
27. Recursion Pharmaceuticals. Recursion announces phase 2 data for REC-994 in cerebral cavernous malformation. Investor release, 3 September 2024. Available from: <https://ir.recursion.com/news-releases/news-release-details/recursion-announces-phase-2-data-rec-994-first-disease>
28. Recursion Pharmaceuticals. Recursion reports first quarter 2025 financial results and business updates. Investor release, 5 May 2025. Available from: <https://ir.recursion.com/news-releases/news-release-details/recursion-reports-first-quarter-2025-financial-results-and>
29. Ren F, Aliper A, Chen J, et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nat Biotechnol.* 2025;43(1):63-75. doi:10.1038/s41587-024-02143-0
30. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol.* 2019;37(9):1038-1040. doi:10.1038/s41587-019-0224-x
31. Walters WP. Dissecting the hype with cheminformatics. *Practical Cheminformatics*, 19 September 2019. Available from: <https://practicalcheminformatics.blogspot.com/2019/09/dissecting-hype-with-cheminformatics.html>
32. Xu Z, Ren F, Wang P, et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. *Nat Med.* 2025;31(8):2602-2610. doi:10.1038/s41591-025-03743-2
33. du Bois RM, Weycker D, Albera C, et al. Forced vital capacity in patients with idiopathic pulmonary fibrosis: test properties and minimal clinically important difference. *Am J Respir Crit Care Med.* 2011;184(12):1382-1389. doi:10.1164/rccm.201105-0840OC
34. Insilico Medicine. Insilico initiates phase III clinical trial for rentosertib, its AI-empowered TNIK inhibitor for idiopathic pulmonary fibrosis. Press release, 7 July 2026. Available from: <https://insilico.com/news/xmjnsn4l091-insilico-initiates-phase-iii-clinical-tr>. Trial registration: [ClinicalTrials.gov NCT07687459](https://clinicaltrials.gov/ct2/show/study/NCT07687459)
35. Dharmasivam M, Kaya B, Akinware A, Azad MG, Richardson DR. Leading artificial intelligence-driven drug discovery platforms: 2025 landscape and global outlook. *Pharmacol Rev.* 2026;78(1):100102. doi:10.1016/j.pharmr.2025.100102
36. Recursion Pharmaceuticals and Exscientia plc. Recursion and Exscientia complete merger. Exscientia plc Form 6-K, exhibit



- 99.1, 20 November 2024. Available from: <https://www.fda.gov/media/184830/download>
- https://www.sec.gov/Archives/edgar/data/001865408/000094787124000937/ss4137056_ex9901.htm
37. Euronext. Result of extraordinary general meeting and delisting: BenevolentAI. Company news, 12 March 2025. Available from: <https://live.euronext.com/en/products/equities/company-news/2025-03-12-result-extraordinary-general-meeting-and-delisting>
38. Niazi SK. A risk-tiered validation framework for artificial intelligence in drug discovery: from reproducibility to clinical translation. *Int J Mol Sci.* 2026;27(10):4349. doi:10.3390/ijms27104349
39. Walsh I, Fishman D, Garcia-Gasulla D, et al. DOME: recommendations for supervised machine learning validation in biology. *Nat Methods.* 2021;18(10):1122-1127. doi:10.1038/s41592-021-01205-4
40. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. doi:10.1136/bmj-2023-078378
41. US Food and Drug Administration. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products. Draft guidance for industry and other interested parties, docket FDA-2024-D-4689, 7 January 2025. Available from: <https://www.fda.gov/media/184830/download>
42. European Medicines Agency. Reflection paper on the use of artificial intelligence in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023. Adopted by CHMP 9 September 2024 and by CVMP 11 September 2024.