ELSEVIER

KEYNOTE

From chemical space to synthetic reality: A framework for makeable-by-design generative AI in drug discovery

Sebastián O. Simonetti [1], Aldana B. Moroni [2], Maria M. Zanardi [2], Frank Kozielski [3], Geoffrey Wells [3,*], Exequiel O.J. Porta [3,*]

[1] IQUIR-CONICET, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531, S2002LRK Rosario, Argentina
[2] Instituto de Investigaciones en Ingeniería Ambiental, Química y Biotecnología Aplicada (INGEBIO), Facultad de Química e Ingeniería del Rosario, Pontificia Universidad Católica Argentina, S2002QEO Rosario, Argentina
[3] UCL School of Pharmacy, University College London, 29/39 Brunswick Square, London WC1N 1AX, UK

Generative artificial intelligence (AI) is reshaping molecular design, but many outputs are still judged by proxy metrics that do not establish whether proposed compounds can be synthesized, tested and advanced. This review reframes the field around 'synthetic reality': the extent to which AI-generated molecules survive route planning, precursor availability, experimental execution and medicinal chemistry decision-making. We examine the limits of heuristic synthesizability metrics, assess route-aware and execution-facing systems and propose maturity levels, reporting standards, actionability metrics and decision gates for auditable real-world evaluation. Future progress will depend on shifting from molecule generation alone to experimentally actionable design.

Keywords: generative AI; drug discovery; molecular design; synthetic accessibility; retrosynthesis; self-driving laboratories; synthetic reality



Maria M. Zanardi holds a PhD in Chemical Sciences and is an Associate Researcher at CONICET, based at INGEBIO (Pontificia Universidad Católica Argentina), where she serves as Director. She is also Professor of Organic and Biological Chemistry in the Schools of Engineering and Medicine at Universidad Católica Argentina. Her research integrates quantum chemistry, DFT calculations, machine learning and chemoinformatics to support structural elucidation, conformational analysis and structure-activity relationships. She has particular expertise in computer-assisted structure elucidation and molecular prediction, applying computational and AI-enabled approaches to chemical, biological and pharmaceutical problems.



Geoffrey Wells is an Associate Professor in Medicinal Chemistry at UCL School of Pharmacy, University College London, where he directs the MSc programs in Drug Discovery and Development and Drug Discovery and Pharma Management. He is a medicinal chemist whose research focuses on the design and synthesis of agents for cancer, neurodegenerative diseases and other serious conditions. His work integrates synthetic chemistry, structure-guided drug design, molecular modeling, docking and biological evaluation to support the rational design and optimization of small-molecule therapeutics. He previously worked as a Drug Discovery Project Leader at Pharminox and has extensive experience in translational medicinal chemistry.



Exequiel O.J. Porta is a Research Fellow in Chemistry at UCL School of Pharmacy, University College London, and a Lecturer (Assistant Professor) in Natural Sciences (Biology and Chemistry) at The Open University. He holds a PhD in Chemical Sciences and a degree in Biotechnology. His research integrates medicinal chemistry, chemical biology, fragment-based drug discovery, organic synthesis, activity-based protein profiling, chemoproteomics and AI-enabled molecular design. His current work focuses on next-generation pan-coronavirus therapeutics, chemical probes, target validation and anti-infective drug discovery, with emphasis on synthesis-aware strategies that connect digital design with experimentally actionable chemistry.

* *Corresponding authors at: University College London, London, UK. Wells, G. (g.wells@ucl.ac.uk), Porta, E.O.J. (e.porta@ucl.ac.uk).*

1359-6446/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
https://doi.org/10.1016/j.drudis.2026.104794

Introduction: Generative AI and the promise (and pitfalls) of 'infinite' chemical space

Modern generative models promise to move molecular design beyond library-centric search, including enumerated, virtual and evolutionary libraries, toward the exploration of effectively unbounded chemical space. The number of potentially drug-like molecules is commonly estimated to be on the order of 10^{60}, far beyond exhaustive enumeration.[p1] Deep-learning architectures, including variational autoencoders, transformers, generative adversarial networks, diffusion models and reinforcement-learning frameworks, can now propose *de novo* structures for multiparameter objectives without relying solely on brute-force enumeration and screening.[p2],[p3],[p4],[p5] These advances expand computational molecular exploration toward underexplored chemotypes and, increasingly, 3D, geometry-aware or property-guided molecular generation.[p5],[p6],[p7],[p8] However, these advances do not resolve the distinct constraint that determines whether digital designs have practical value: can the proposed molecules be made, tested and reproduced under real experimental conditions?

The promise is not hypothetical. Artificial intelligence (AI)-enabled pipelines have already delivered striking examples of accelerated discovery, including the generative design and experimental validation of inhibitors of discoidin domain receptor 1, a receptor tyrosine kinase, within 46 days,[p9] a generative-model workflow for antimicrobial peptide discovery that identified, synthesized and experimentally tested 20 candidate antimicrobial peptides within 48 days,[p10] and a generative pipeline that progressed from target nomination to preclinical candidate nomination for Rentosertib in approximately 18 months, followed by Phase I and Phase IIa evaluation in idiopathic pulmonary fibrosis.[p11],[p12] At the same time, setbacks in AI-enabled clinical programs, including Exscientia's EXS-21546 and BenevolentAI's BEN-2293, illustrate that accelerated design and prioritization do not guarantee translational success.[p13],[p14] In fact, an even earlier bottleneck often intervenes first: many digitally attractive molecules cannot be synthesized efficiently, reproducibly or at all.[p2] This distinction becomes sharper as upstream models improve. Target-aware generators, such as TamGen, condition molecular design on protein-pocket information, while structure- and binding-affinity prediction systems, including Boltz-2 and proprietary reports on Isomorphic Labs Drug Design Engine (IsoDDE), suggest that upstream predictive tools might continue to improve, although not all such systems are equally auditable.[p15],[p16],[p17] Yet better target awareness does not remove the need for executable chemistry; it magnifies it, because a more plausible binder is still a translational dead end if it cannot be made, procured and reproduced.

Synthesizability is not a single concept. In molecular generation, it is commonly operationalized through synthetic accessibility, synthetic complexity, retrosynthetic accessibility or feasibility scores, including synthetic accessibility score (SAScore), synthetic complexity score (SCScore) and retrosynthetic accessibility score (RAScore), and more recent human-in-the-loop approaches such as focused synthesizability score (FSscore) (discussed below). These metrics are useful for ranking, prioritizing or flagging candidates, but they do not by themselves establish whether a proposed molecule can be planned, procured, synthesized, tested and iterated upon in a defined experimental setting. We therefore use 'synthetic reality' to describe a broader design-route-execution property: whether a candidate is actionable under practical constraints, including reaction feasibility, precursor accessibility, route robustness, selectivity risk, procurement constraints and platform capabilities. It is, consequently, broader than synthesizability scores or computational route existence alone, but remains a discovery-stage execution concept rather than a substitute for full manufacturability or Chemistry, Manufacturing and Controls assessment. In practical terms, synthetic reality asks which AI-generated proposals warrant laboratory commitment. Without that alignment, unconstrained generators can produce synthesis-orphan structures: molecules that score highly *in silico* yet lack a credible route from accessible precursors. The bottleneck then shifts downstream, where chemists must manually triage designs that appear computationally compelling but fail during route planning, procurement, synthesis or the first experimental reality check.

This tension has a clear historical analog. The combinatorial chemistry era greatly expanded access to large compound libraries, but library scale alone did not ensure that libraries occupied chemical space closely aligned with drugs, natural products or other biologically relevant collections. The broader lesson was that increasing library size did not, by itself, ensure greater coverage of drug-relevant chemical space.[p18] Generative AI now faces the same inflection point. Without native integration of synthesis, procurement and experimental robustness, generative AI risks shifting feasibility triage downstream rather than eliminating it, echoing well-documented failure modes of unconstrained molecular generation.[p19],[p20],[p21] Conversely, when makeability is treated as a design objective rather than a downstream filter, the field moves toward a more consequential standard: not infinite chemical imagination, but executable chemistry.

Herein, we examine why the gap between generative imagination and experimental feasibility persists, how fast proxy scores address it only partially and how reaction-aware models, negative data, self-driving laboratories and explicit governance can help narrow it. Our central argument is simple: synthetic reality is increasingly a necessary condition for generative AI to function as a reliable engine for makeable-by-design drug discovery (Figure 1).

The synthesizability gap: When generators propose unmakeable molecules

Early *de novo* generators made a fundamental limitation visible: formally valid molecules are often not practically makeable. Gao and Coley used an automated system for knowledge-based continuous organic synthesis (ASKCOS) to evaluate outputs from then-state-of-the-art generators and found that many high-scoring molecules lacked readily identifiable routes to purchasable precursors.[p2] The failure modes were recognizable. Models driven hard against surrogate objectives can produce molecular structures that score well computationally but fall outside reliable applicability domains or fail route-based feasibility checks.[p22],[p23],[p24] A particularly clear example is docking-driven reward hacking, in which generators append flexible

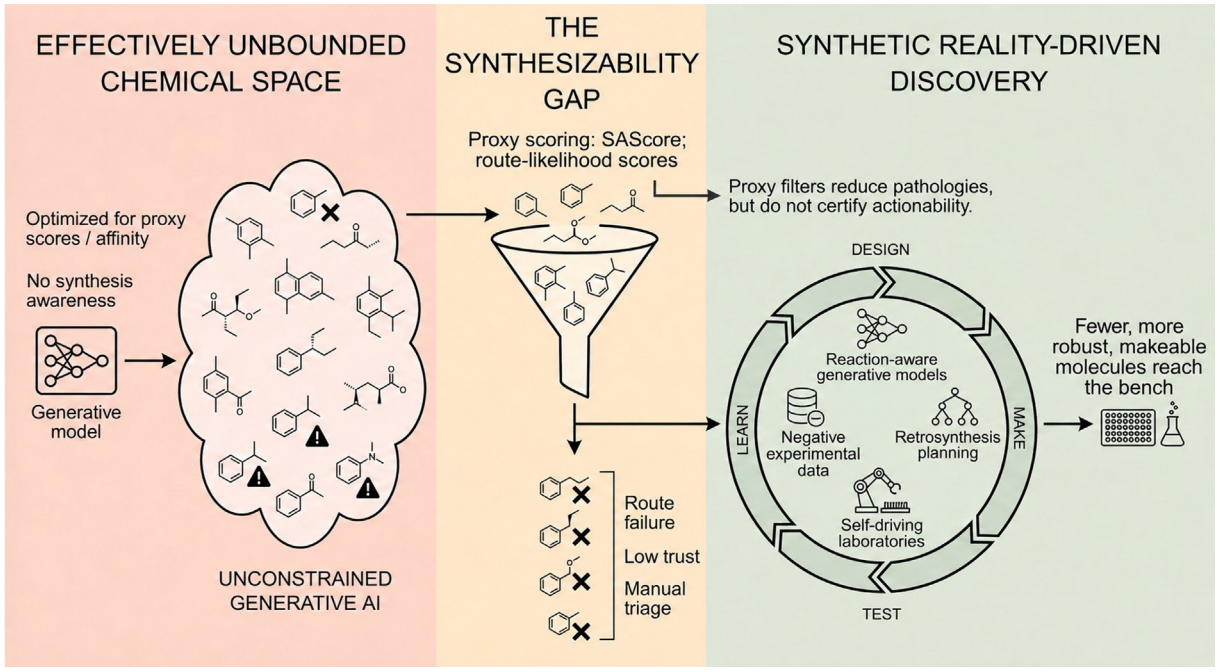


FIGURE 1

From unconstrained molecular generation to synthetic reality-driven drug discovery. Unconstrained generative AI models explore vast regions of theoretical chemical space; however, they often propose molecules that are not connected to practical synthesis. Heuristic filters partially mitigate this issue but fail to guarantee executability. By contrast, synthetic reality-driven workflows can integrate reaction-aware generation, retrosynthesis constraints, negative experimental data and closed-loop self-driving laboratories to produce molecules that are makeable-by-design and suitable for experimental validation. This figure was created using BioRender (http://biorender.com/). AI: artificial intelligence; SAScore: synthetic accessibility score.

'stringy tails' that improve predicted contact scores while increasing flexibility, reducing synthetic feasibility and risking ADMET (absorption, distribution, metabolism, excretion and toxicity) liabilities.[p21] Recent studies underscore that synthesizability remains a central unresolved barrier rather than a historical curiosity.[p24],[p25],[p26],[p27],[p28]

Mechanistically, this is unsurprising. Many generators learn correlations in molecular space from bioactive datasets, benchmark collections or purchasable catalogs, whereas makeability is determined in synthesis space by reactions, inventories and feasible routes. Under narrow optimization pressure, models can extrapolate beyond the regime in which learned regularities remain trustworthy, producing unusual structural combinations, exploiting weaknesses in scoring functions or drifting into high-reward regions with poor route reachability. This mismatch is also reflected in widely used benchmark suites such as GuacaMol and MOSES, which evaluate goal-directed and/or distribution-learning performance using molecular and statistical metrics but do not themselves establish route-level feasibility.[p29],[p30],[p31] Additionally, high-scoring generative outputs can nevertheless perform poorly when subjected to synthesis-planning checks.[p2]

This does not imply that ambitious chemistry should be excluded, nor that contemporary molecular generation is broadly synthesis-blind. A substantial body of synthesizability-by-design work already introduces chemical constraints upstream. Reaction-rule and virtual-synthesis systems such as Design of Innovative New chemical entities Generated by Optimization Strategies (DINGOS),[p32] RENATE[p33] and Reaction-driven Objective Reinforcement (REACTOR)[p34] construct candidates through learned or curated transformations and available building blocks; pathway-first and reaction-constrained models, including SynFormer[p26] and ClickGen,[p35] explicitly search reaction- or pathway-linked spaces, while Tanimoto Group Overlap (TANGO)[p36] and other retrosynthesis-in-the-loop approaches use synthesis-planning information as optimization feedback.[p27] These methods represent substantive progress beyond unconstrained generation and post-hoc accessibility filtering.

However, 'synthesis-aware' remains an umbrella term spanning distinct evidentiary claims. Scaffold-, fragment-, functionality- or building-block-guided systems, including DeepScaffold,[p37] ScaffoldGVAE,[p38] relevant scaffold- or structure-constrained modes within REINVENT4[p39] and VeGA-RX/VeGA-SCX,[p40] based on the Versatile Generative Architecture (VeGA), primarily provide chemical guidance by directing generation toward specified cores, motifs or building-block-compatible regions; unless explicit reaction logic is retained, such constraints do not, by themselves, establish that an executable route exists. Reaction-based systems such as RENATE,[p33] REACTOR[p34] and LibINVENT,[p41] together with methods linked to synthetic trees or computer-aided synthesis planning (CASP), provide stronger route grounding by connecting

proposals to transformations, reactants or planned routes. Some systems, including DINGOS,[p32] additionally report prospective synthesis, but, under the framework proposed here, full execution-grounded synthetic reality requires transparent evidence of precursor accessibility, procurement, safety and compliance, attempted synthesis, failures, human intervention and denominator-based outcomes. A key remaining gap is therefore not a lack of attention to synthesizability, but inconsistency in what the broad label 'synthesis-aware' denotes and in the strength of evidence supporting such claims.

Auditing synthetic reality: Standards, benchmarks and prospective validation

To make synthetic reality actionable, evaluation must move beyond molecule-level scores toward evidence that proposed compounds can be executed under explicit constraints. We therefore frame assessment as an evidence hierarchy rather than a rigid filtering sequence. Heuristic scores provide rapid, low-cost triage, but they should not be used as hard-exclusion filters that discard chemically novel yet route-accessible structures before higher-fidelity analysis. Stronger evidence comes from route-level support under declared inventory and feasibility constraints, and the strongest evidence comes from prospective execution with transparent reporting of both successes and failures. Across this hierarchy, the key practical endpoint is the synthesis-actionable rate: the fraction of prioritized or top-ranked designs that are supported by at least one executable route from accessible precursors and pass declared feasibility, safety and compliance checks.

Heuristic and learned proxies: Synthetic accessibility, feasibility and their limits

An early pragmatic response to unrealistic generative outputs was to incorporate fast, feasibility-oriented proxies into optimization and filtering pipelines, including synthetic accessibility, synthetic complexity and retrosynthetic accessibility scores (Table 1). Drug-likeness criteria such as the quantitative estimate of drug-likeness (QED) are frequently optimized alongside these measures to constrain conventional small-molecule property space, but they constitute an independent objective and provide no direct evidence of synthetic accessibility or route feasibility. QED is included in Table 1 only to clarify this complementary role and should be applied cautiously to beyond-Lipinski modalities such as proteolysis-targeting chimeras (PROTACs) or macrocycles. Among early scalar synthesizability constraints, SAScore became especially influential because it is fast, interpretable and readily incorporated into generative objectives.[p42] QED became similarly widespread as a complementary drug-likeness criterion, although neither metric was designed specifically for modern generative optimization.[p43] Later metrics such as SCScore, RAScore and FSscore attempt to capture synthetic complexity, route-finding likelihood or context-dependent feasibility more directly.[p44],[p45],[p46] Additional approaches incorporate building-block availability or reaction awareness, as in building-block and reaction-aware SAScore (BR-SAScore), or embed feasibility terms within broader benchmarking and optimization frameworks such as MolScore.[p25],[p47]

SAScore estimates synthetic difficulty on an approximate 1–10 scale by combining fragment frequencies from large collections of known molecules with penalties for molecular complexity, including molecular size, ring complexity, macrocyclic structures and stereochemical burden.[p25],[p42] Because these fragment frequencies are derived from historical chemical collections, SAScore can conflate novelty with difficulty and penalize rare but accessible chemotypes. QED, by contrast, compresses multiple physicochemical properties into a drug-likeness desirability score. It is not a synthesizability metric and provides no direct information about reaction feasibility, precursor availability or route existence.[p43] Related learned metrics draw on different sources of evidence. SCScore is trained on reaction corpora and estimates synthetic complexity from reactant-product relationships, whereas RAScore estimates whether the retrosynthetic planner represented in its training labels is likely to identify a route to a target molecule.[p44],[p45] Human-in-the-loop approaches, such as FSscore, further emphasize that synthetic feasibility can be context-dependent and can benefit from refinement for specific chemical spaces.[p46]

In practice, these proxies are useful for penalizing overcomplex outputs, triaging large candidate sets and balancing feasibility against potency, physicochemical properties and predicted ADMET behavior in multiparameter optimization workflows.[p2],[p45],[p47] However, they differ substantially in operational burden and interpretation. SAScore remains attractive as a first-pass approximation because it is simple and scalable, whereas SCScore, RAScore and FSscore depend on learned models, training data, retrosynthetic-planning outputs or scope-specific refinement. Consequently, scalar scores are most informative when complemented by route-level evidence, including predicted step count, solve status, inventory-defined precursor availability and route plausibility from CASP tools such as ASKCOS or AiZynthFinder.[p48],[p49]

Their limitations are equally important. These proxies remain historical or model-dependent approximations and do not fully account for reaction-level realities such as chemoselectivity, regioselectivity, stereocontrol, protecting-group requirements, step compatibility, purification burden or precursor availability. A molecule can appear 'easy' by heuristic criteria and still require a brittle or impractical route. Conversely, unfamiliar structural motifs with viable synthesis plans can be discarded because the proxy fails to recognize them. Consistent with this, benchmarking against AiZynthFinder shows that synthetic-accessibility scores differ substantially in how well they predict planner outcomes.[p50] In that study, SAScore and RAScore outperformed SCScore and SYnthetic Bayesian Accessibility (SYBA), although RAScore is inherently planner-dependent because it was trained on AiZynthFinder outcomes.

The broader risk is strategic. If proxy constraints are used as hard exclusion filters, they can bias models toward conservative, familiar chemotypes and reduce the exploratory advantage that makes generative design attractive. Their strongest role is therefore not to define accessible chemical space on their own, but

KEYNOTE

TABLE 1

Comparison of commonly used synthesizability-related metrics and route-level evaluation methods in AI-driven molecular design, highlighting their conceptual basis, typical use, strengths and limitations[a]

Metric	Conceptual basis	Strengths	Limitations
SAScore[p42]	Fragment-frequency statistics from large collections of known molecules, combined with penalties for molecular complexity, including molecular size, ring complexity, macrocycles and stereochemical burden.	Very fast; intuitive; effective for filtering clearly over-complex or 'chemically awkward' structures at scale.	Proxy derived from historical chemistry; can conflate novelty with difficulty; does not reason over reaction steps or guarantee that a viable synthetic route exists; because it is history-dependent, can lag behind advances in synthetic methodology unless periodically updated.
QED[p43]	Weighted desirability function over physicochemical properties associated with approved drugs (e.g. MW, log *P*, HBD/HBA, PSA and aromaticity).	Useful for steering generation toward broadly drug-like physicochemical space; fast, interpretable and widely adopted as an independent objective in multiparameter optimization.	Provides no evidence of synthetic accessibility, reaction feasibility, precursor availability or route existence. Its use should therefore remain analytically separate from synthesizability and route-level evaluation.
SCScore[p44]	Machine-learned estimate of synthetic complexity learned from reaction corpora (captures patterns of how 'advanced' products tend to be relative to precursors in known chemistry).	Captures aspects of perceived synthetic difficulty beyond simple structural heuristics; useful as a learned proxy in optimization loops.	Data- and domain-dependent; can be unreliable outside the training distribution; does not propose or validate an explicit route and therefore cannot certify feasibility.
RAScore[p45]	Machine-learned classifier trained on outcomes of computer-aided retrosynthetic planning, providing a rapid estimate of whether a route is likely to be found.	Fast approximation aligned to route-finding success; practical for multiparameter optimization as an explicit constraint/objective.	Depends on the underlying retrosynthesis engine, settings, inventory model and training data; returns a route-finding likelihood rather than an explicit, experimentally actionable synthetic plan.
FSscore[p46]	Machine-learned focused synthesizability score trained from reaction-derived pairwise comparisons and refined with expert human feedback, database look-ups or custom scores for a defined chemical space.	Context-adaptable; incorporates expert judgement; can be refined for specific chemical scopes; potentially useful as a guidance or ranking term in generative workflows.	Requires scope-specific refinement or labelled preference data; performance depends on the quality and representativeness of expert feedback or database-derived labels; remains a learned feasibility proxy and does not provide an explicit route, procurement assessment or experimental validation.
Retrosynthesis-based evaluation (e.g. ASKCOS and AiZynthFinder[p48],[p49]	Explicit construction of reaction sequences linking a target molecule to available starting materials using reaction templates/learned policies and purchasable inventories.	Highest-fidelity computational proxy for makeability: assesses whether a planner can identify plausible routes, estimated route complexity and precursor availability under defined settings.	Computationally expensive, and integration into large-scale generative optimization loops is nontrivial. Results depend on the planning engine, software version, search settings, reaction-rule coverage, inventory assumptions and database content. Commercial tools can further limit independent verification. Route-finding success does not establish experimental feasibility. Independent expert review can strengthen route plausibility, but prospective execution provides stronger evidence of practical feasibility.

HBD/HBA: hydrogen-bond donor(s)/hydrogen-bond acceptor(s); log *P*: octanol/water partition coefficient; MW: molecular weight; PSA: polar surface area.

[a] Take-home message: Scalar scores and property metrics are useful for rapid screening, while retrosynthesis-based tools provide higher-fidelity route-level evidence. However, none of these approaches alone certifies transformation robustness, starting-material quality, procurement practicality or laboratory execution. Consequently, score-informed and route-informed prioritization should be treated as decision support: useful for scale, but insufficient for audit without explicit constraints and, where claims require it, prospective execution. QED is included because it is frequently optimized alongside synthesizability constraints in generative workflows; it is an independent drug-likeness criterion and should not be interpreted as evidence of synthetic accessibility or route feasibility.

to support early ranking, prioritization and diagnostic review of large candidate sets. A more robust path forward is not to abandon heuristics, but to embed them within workflows where score-informed prioritization is followed by route-grounded evidence and, increasingly, generation methods that operate through chemistry rather than merely being filtered by it.

Operational evaluation and reporting standards for synthetic reality

To move from narrative claims to operational rigor, synthetic reality should be assessed through a minimum reporting framework rather than inferred from heuristic proxies alone. If a workflow claims to incorporate synthetic reality, it should

report whether at least one executable route can be identified and justified under predefined project constraints, including step-count limits, precursor availability, procurement and execution-time limits, approximate cost or resource constraints, inventory assumptions and the CASP environment used. It should also disclose whether starting materials are genuinely accessible rather than only theoretically listed in a catalog, as well as key route fragilities, including high-risk transformations, purification bottlenecks or selectivity concerns, safety and compliance checks and the fraction of prioritized designs that meet the declared synthesis-actionable criteria rather than merely scoring well *in silico*. Together, these elements define the minimum evidence needed to support auditable claims of synthetic reality.

Route-level evidence can be supported by open-source tools, commercial planning platforms or subscription databases. Commercial route-planning systems such as Synthia and databases such as Reaxys and CAS SciFinder-n can provide complementary route proposals, reaction precedents, experimental conditions, supplier records and precursor-availability evidence. Their proprietary algorithms, restricted access, evolving content coverage and sometimes opaque versioning can nevertheless limit independent verification. Restricted access does not invalidate route-level evidence, but it changes its reproducibility status. Authors should therefore report the tool or database name, version and access date; the database, inventory or content snapshot where available; search settings, computational limits and stopping criteria; reaction, building-block and supplier assumptions; and any manual filtering or route modification. Where fixed version identifiers or snapshots are unavailable, that limitation should be stated explicitly. Reports should also distinguish a route proposed by software from one independently reviewed by an experienced chemist, modified after expert assessment or prospectively attempted, because these constitute different levels of evidentiary support.

Benchmarking, prospective validation and reproducibility: How to know a model deserves its scores

Despite rapid progress in molecular generation, it remains difficult to prove that a system genuinely captures synthetic reality. Retrospective benchmarks such as GuacaMol and Molecular Sets (MOSES) are useful for early method development, but they predominantly evaluate molecule-level properties, including validity, novelty, diversity and proxy objectives, rather than synthesis-space executability or experimental robustness.[(p29),(p30),(p51)] Several limitations are especially relevant for synthetic reality.[(p24),(p31),(p52),(p53),(p54)] First, objective-metric misalignment can permit reward hacking, especially when docking or learned predictors dominate optimization. Second, dataset overlap, memorization or validation leakage can inflate apparent generalization. Third, models trained on curated datasets might not transfer cleanly to laboratory execution, where reagent quality, operator practices, analytical workflows and local platform constraints can shape yields and failure modes.

Operational benchmarking should therefore treat the route-planning environment, whether open or proprietary, as a versioned methodological dependency rather than a black box. Authors should report the planning engine and version or access date, search settings and computational limits, inventory assumptions and any reaction-template, building-block or database snapshot used for evaluation.[(p28),(p48),(p52),(p55)] Outcomes should include solve rate, step-count distributions, route diversity or convergence where relevant and explicit denominators. PaRoutes illustrates how route-level benchmarks can move comparison beyond 'solved/not solved' toward auditable route quality and diversity.[(p52)]

Prospective validation provides the strongest evidence because it tests whether a model can survive the transition from digital proposal to laboratory execution.[(p28),(p35),(p56)] In a meaningful study, the generator, scoring functions and selection rules should be frozen before candidate generation. Designs should then proceed through a predefined pipeline covering route support under declared constraints, procurement feasibility and experimental execution, with transparent reporting of successes and failures. Even small batches can be informative when their limited statistical precision is acknowledged and they disclose the progression from design to experiment: how many candidates had viable routes, were procurable and were successfully delivered at the required purity and quantity, why failures occurred, whether yields were acceptable under the declared constraints and how many iterations were needed. Collectively, these outcomes test prospective synthetic reality and allow the synthesis-actionable rate to be evaluated against observed execution, while experimental-delivery and synthesis-success rates remain distinct outcomes. Human intervention should also be reported because a route executed as proposed, a chemist-edited route and an extensively redesigned route are not equivalent evidentiary outcomes, even if all eventually deliver material.

Prospective validation is only as credible as its reproducibility. Provenance-rich reporting should therefore link outcomes to the experimental context, including reagents, equivalents, solvent, temperature, time, atmosphere, analytical confirmation and deviations from the planned route, preferably through an electronic lab notebook, a laboratory information management system or repository-linked metadata.[(p53),(p57)] In parallel, authors should report explicit denominator chains from scored sets to route-certified subsets under declared constraints, procurable, synthesized subsets and, where biological testing is within scope, biologically tested and active subsets, so that small prospective demonstrations do not overstate practical readiness. We summarize the minimum expectations as a Minimal Prospective Validation Checklist (MPVC) to enable comparisons of prospective claims across studies.

Minimal Prospective Validation Checklist for makeable-by-design pipelines

The MPVC (Box 1) is proposed as a minimum evidentiary standard, motivated by the recurring reporting gaps, denominator ambiguities and failure modes identified above in prospective makeable-by-design studies.

Box 1 Minimal Prospective Validation Checklist (MPVC)

1 Frozen model and protocol

The generator, ranking criteria, and selection rules are fixed prior to generation.

2 Explicit feasibility envelope

Prespecified constraints are declared, including step count, precursor or inventory assumptions, procurement/execution-time limits, platform constraints, CASP environment, and inventory snapshot. Cost, sustainability, or process-efficiency criteria are reported when used prospectively as decision thresholds.

3 Route certification

Each selected molecule is accompanied by at least one executable route within the declared feasibility envelope; any material deviation between the planned and executed routes is recorded.

4 Precursor accessibility/availability

Starting materials are either commercially available or accessible through routine in-house synthesis, with lead-time or custom-synthesis requirements explicitly indicated.

5 Safety and compliance screening

Routes and precursors are screened for exceptional or non-routine hazards, restricted substances, controlled precursors, incompatible conditions, and platform- or institution-specific safety constraints before execution.

6 Prospective execution attempt

A defined number of designs proceeds to synthesis under predeclared success criteria, such as identity confirmation, purity, and quantity sufficient for the intended assay; yield may be reported as a route-robustness indicator rather than a universal pass/fail threshold.

7 Outcome reporting

Successes and failures are reported with categorized failure modes, including no-route, procurement, reaction, purification, workup and instability, together with biological inactivity and downstream ADMET or developability attrition where these are within the scope of the study.

8 Reproducibility context

Essential experimental and analytical details are disclosed, including the solvent, temperature, time, concentration, equivalents, atmosphere, and analytical confirmation.

9 Iteration accounting

The number of route-replanning, human-rescue, or Design-Make-Test-Learn (DMTL) iterations is reported, together with their time or resource burden where relevant.

10 Synthesis-actionable rate

The fraction of prioritized or top-ranked designs that satisfy declared route, precursor-accessibility, safety/compliance, and execution-feasibility criteria is disclosed.

Minimum reporting criteria for assessing whether makeable-by-design generative workflows support auditable claims of synthetic reality.

Table 2 provides a worked example using the SyntheMol campaign and extends the MPVC audit to ClickGen, highlighting both genuine progress and remaining gaps in auditability.[(p35),(p56)] The SyntheMol denominator chain distinguishes procurement progression, conditional synthesis success and downstream biological activity, but these values do not constitute a strict synthesis-actionable rate because the complete pre-commitment actionability set was not reported, including formal safety/compliance screening and declared execution-feasibility criteria; route-realization status was also not disclosed. Clearer locked-protocol statements, better procurement disclosure, formal safety screening, explicit failure taxonomies and denominator-based reporting of actionability remain important for fully auditable claims. For future studies, we recommend disclosing a route-realization class (model-native, chemist-edited or expert-redesigned) so that genuine model actionability can be distinguished from substantial expert intervention. Beyond route existence, we also regard deliverability, including time, cost, synthetic effort and operational complexity, as a valuable extension to MPVC rather than a core requirement, consistent with

TABLE 2

MPVC audit of SyntheMol and ClickGen prospective campaigns[a,b]

MPVC criterion (auditable requirement)	SyntheMol[p56]	ClickGen[p35]	Additional reporting needed for strict MPVC compliance
(1) Frozen model and protocol	Partial Workflow follows training, generating, filtering and ordering/synthesizing/testing; however, an explicit 'locked'/pre-registered protocol statement is not emphasized.	Partial Workflow implies a fixed generator and virtual screening before selecting three leads; an explicit 'locked protocol' is not stated.	Explicit statement that the model, scoring functions and selection pipeline were fixed before generation; ideally, include a short locked-protocol description.
(2) Explicit feasibility envelope	Met for chemistry constraints Generation was constrained by REAL Space/building-block chemistry and a defined 13-reaction synthetic toolbox, with single-reaction molecules prioritized for rapid and cost-effective synthesis.	Met for chemistry constraints Generation was constrained by a predefined modular toolbox centered on click chemistry and amidation.	In addition to chemistry constraints, report the full project-level feasibility envelope, including step limits, time/resource limits, platform constraints, allowed reagents/conditions and inventory snapshot.
(3) Route certification	Partial/route-supported by construction Generated molecules were assembled from defined building blocks and reaction rules, providing explicit one-step route logic. However, reaction-template compatibility does not by itself certify experimental executability for every selected design.	Partial Planned synthesis routes were evaluated for 30 selected compounds, including raw-material availability and reaction complexity, and three leads spanning different synthetic difficulty were then synthesized. However, a standardized route-certification gate was not reported for the broader candidate set.	Report route support or certification per committed design, ideally with defined constraints such as step cap, inventory model, permitted reactions and/or independent CASP confirmation.
(4) Precursor accessibility/availability	Partial Inventory is catalog-constrained, yet orderability is not universal: 70 out of 150 selected molecules were available to procure for synthesis. Lead-time/purity flags are not standardized per candidate.	Partial Raw-material availability and synthon acquisition are discussed for selected routes, but supplier-level, lead-time and custom-synthesis flags are not reported as a standardized denominator for the committed design set.	Report the fraction with purchasable or readily preparable precursors, supplier or source category where possible and lead-time/custom-synthesis flags for each committed design set.
(5) Safety and compliance screening	Not reported No formal route/precursor safety, restricted-substance, controlled-precursor or platform-compliance screen is reported.	Not reported No formal route/precursor safety, restricted-substance, controlled-precursor or platform-compliance screen is reported.	Add an explicit route/precursor safety and compliance screen, including restricted substances, controlled precursors, exceptional hazards and platform-specific constraints, and report pass/fail rates.
(6) Prospective execution attempt	MetOrdered 70; successfully synthesized 58 out of 70 (83%) in approximately 4 weeks, with reported purity criteria (>90%).	Met Three selected lead compounds were synthesized within 10 days; production and bioactivity testing were reported within 20 days.	No strict deficiency if the intended prospective scope is declared. Larger prospective batches would improve precision for synthesis-actionable rates and failure-mode estimates.
(7) Outcome reporting	Partial Biological outcomes reported for the synthesized set, including 6/58 active; negative biological outcomes are reported at aggregate level, but synthesis, procurement, and route failures are not categorized by MPVC failure mode.	Partial Outcomes reported for the three synthesized leads; structured negative reporting for nonadvanced or failed designs is not provided.	Provide a minimal failure taxonomy with counts, including route failure, procurement failure, reaction failure, purification bottleneck, instability, inactivity and downstream ADMET/developability attrition.
(8) Reproducibility context	Partial Purity/QC reporting is provided, including LC–MS or ^{1}H NMR in specified cases, but full condition-level provenance per synthesis attempt is not the emphasis.	Partial Analytical characterization data, including ^{1}H NMR, ^{13}C NMR and HRMS, are provided for synthesized leads, but full MPVC-style condition-level provenance per attempt is not the main reporting focus.	Attach condition-level metadata, including reagents, equivalents, solvent, temperature, time, atmosphere, analytical confirmation and deviations from the planned route.

TABLE 2 (*CONTINUED*)

MPVC criterion (auditable requirement)	SyntheMol[p56]	ClickGen[p35]	Additional reporting needed for strict MPVC compliance
(9) Iteration accounting	Not reported Computational MCTS iterations are reported, but DMTL iteration counts or route replanning iterations toward executable synthesis are not reported.	Not reported Not reported as DMTL iteration counts for an executable plan.	Report route-replanning, human-rescue or DMTL iteration counts, plus time/resource burden per loop.
(10) Synthesis-actionable rate	Partial Practical rates can be reported separately: 70/150 selected compounds were available in REAL Space and ordered for synthesis; 58/70 were successfully synthesized. As a downstream biological outcome, 6/58 synthesized compounds were active. Route support by construction is clear, but the complete pre-commitment actionability set was not reported.	Partial The paper reports a virtual-screening funnel, including top-1% docked molecules, approximately 700 clustered molecules, 30 selected compounds across synthetic-difficulty classes, and three synthesized leads. However, no MPVC-style denominator is reported for top-N designs passing route, procurement, safety/compliance, and declared execution-feasibility criteria; at the committed-lead stage, 3/3 selected leads were synthesized.	Define the denominator, such as top-N after scoring or the committed design set, and report pass rates for route support, procurement, safety/compliance and declared execution-feasibility criteria; report actual synthesis and biological outcomes separately.

HRMS: high-resolution mass spectrometry; MCTS: Monte Carlo Tree Search; QC: quality control.

[a] ClickGen is treated as a workflow exemplar rather than a robust estimator of synthesis-actionable rates because only three committed leads were experimentally advanced.

[b] Selected worked example: Using the 150 selected SyntheMol candidates as the initial denominator, 70/150 (46.7%) were available in REAL Space and ordered for synthesis; 58/70 (82.9%) were successfully synthesized (58/150, 38.7% of the selected set); and 6/58 (10.3%) were biologically active (6/150, 4.0% of the selected set). These denominator-specific outcomes distinguish procurement progression, synthesis success and downstream biological activity. They do not constitute a strict synthesis-actionable rate because the complete pre-commitment actionability set was not reported, including formal safety/compliance screening and explicitly declared execution-feasibility criteria. Route-realization status was also not disclosed.

route-cost frameworks such as RouteScore and Synthesis Planning And Rewards-based Route Optimization Workflow (SPARROW).[p58],[p59] The implication for method comparison is straightforward: molecule-level scores remain useful, but they should no longer be the primary endpoint once a workflow claims to be synthesis-aware. At that point, route-, experiment- and project-level outcomes become the relevant evidence.

Within this framework, the concepts and tools serve distinct but complementary functions. Synthetic reality is the overarching concept: the practical actionability of a molecule–route–precursor–execution chain under defined project constraints. The MPVC specifies the evidence that should be reported to substantiate prospective makeable-by-design claims; the synthesis-actionable rate quantifies the fraction of prioritized designs that satisfy declared route, precursor-accessibility, safety and compliance and execution-feasibility criteria; the SRMM (introduced below) classifies the intended design level and evidenced maturity of a workflow; and the decision gates translate this evidence into a candidate-level go/no-go decision about laboratory commitment. Outcomes generated after commitment should then be reported in accordance with the MPVC, used to compare predicted actionability with observed procurement and execution outcomes and to update future synthesis-actionable-rate estimates and incorporated into reassessment of the workflow's evidenced SRMM maturity.

Synthesis-aware generative design: From chemical guidance to execution-grounded evidence

Once synthetic reality is treated as a design variable, the question becomes how to embed it upstream rather than applying it after generation. We assess synthesis-aware methods along two related but distinct dimensions: how synthesis information enters the workflow and what strongest evidence is reported. Chemical-guidance methods constrain generation toward preferred scaffolds, fragments, functional groups or building-block-compatible regions without necessarily retaining an explicit route. Route-feasibility methods link candidates to reaction rules, reactants, synthetic trees, retrosynthetic logic or CASP-supported routes. Execution-grounded synthetic reality requires prospective synthesis or attempted execution together with evidence on procurement, safety and compliance, failures, human intervention and explicit denominators. These categories are related but not interchangeable, and 'synthesis-aware' is used below only as an umbrella term.

Chemical guidance

Chemical-guidance methods narrow the molecular space explored without preserving the transformations needed to make each output. Scaffold-, fragment- and functionality-conditioned approaches can direct generation toward medicinally relevant cores, known pharmacophores or structurally defined local neighborhoods. DeepScaffold and ScaffoldGVAE, together with scaffold-, linker- and molecule-transformation modes in REINVENT4 and SMILES Arbitrary Target Specification (SMARTS)-guided VeGA-RX/VeGA-SCX, exemplify this strategy through scaffold generation, scaffold hopping, controlled decoration or structural conditioning.[p37],[p38],[p39],[p40] These constraints can improve chemical plausibility and support analog generation, but they do not establish that an executable route exists. When generation applies specified reaction templates to compatible

KEYNOTE

reactants, the evidence moves toward route feasibility; LibINVENT, for example, links scaffold decoration to reactions and compatible building blocks.[p41]

Route feasibility

Route-grounded methods retain a connection between proposed molecules and transformations, reactants, synthetic trees or computationally supported routes. Reaction-rule and virtual-synthesis approaches, such as RENATE, REACTOR and LibINVENT, construct candidates within reaction-linked spaces, whereas DINGOS combines generative design with rule-driven virtual synthesis and additionally reports the prospective synthesis of selected outputs.[p32],[p33],[p34],[p41]

Pathway-first and synthetic-tree systems make route linkage explicit. SynFormer constructs reaction sequences from curated transformations and commercially available building blocks, while SynLlama generates molecules together with synthetic pathways assembled from accessible building blocks and reaction templates.[p26],[p60] ChemBO searches a synthesis graph defined by available reagents and transformations, whereas ChemProjector projects molecules into pathway-defined, synthesizable space, and SynNet generates synthetic trees from purchasable starting materials and curated reaction templates.[p61],[p62],[p63] These systems provide computational evidence of how a candidate could be made under the assumptions encoded in their reaction sets, building-block inventories and search procedures.

Constrained optimization provides a related form of route grounding. TANGO directly optimizes starting-material, intermediate and divergent-synthesis constraints, while RxnFlow and 3DSynthFlow generate routes through predefined building blocks and reaction templates, with the latter extending this logic to the joint design of synthetic pathways and three-dimensional binding poses.[p36],[p64],[p65] Retrosynthesis-in-the-loop workflows instead filter, penalize, re-rank or optimize generated molecules using CASP tools such as AiZynthFinder, ASKCOS or Synthia under specified inventory, route-length, search-depth and computational constraints.[p48],[p49] Because full route planning is expensive, some pipelines use learned route-solvability surrogates such as RAScore or sample-efficient architectures that reduce the number of route or oracle calls; Saturn and the approach of Guo and Schwaller illustrate this strategy.[p27],[p45],[p64],[p65],[p66]

These approaches remain sensitive to their reaction sets, inventories, search settings and training data. Restricting generation to cataloged building blocks or established reaction classes can improve immediate actionability but biases exploration toward commercially accessible chemistry. Artificial synthetic trees can scale route-aware learning, yet omit selectivity, protecting-group, purification and empirical execution constraints.[p28] Route-grounded systems are therefore most useful when they optimize beyond route existence toward deliverability, including robustness, safety, sustainability, cost and resource burden. RouteScore, SPARROW and the IM-204 process-optimization case illustrate this broader objective.[p58],[p59],[p67] Route existence remains a higher-fidelity proxy than heuristic accessibility, but it does not establish that precursors are procurable, transformations will succeed locally or the route can be executed without substantial expert modification.

Execution-grounded evidence

Prospective execution tests whether a design survives procurement, experimental chemistry, purification, characterization and, where included in the campaign, downstream biological evaluation. Successful synthesis of selected compounds, however, does not by itself establish a fully audited synthesis-actionable rate. Evidentiary strength depends on the denominator from which compounds were selected, the completeness of procurement reporting, the treatment of unsuccessful attempts, safety and compliance checks and the degree of human intervention.

Prospective campaigns, including DINGOS, ClickGen, SyntheMol, SyntheMol-RL and LambdaZero, show that chemistry-constrained generation can advance selected designs to synthesis and testing.[p32],[p35],[p56],[p68],[p69] ClickGen produced three poly (ADP-ribose) polymerase 1 (PARP1)-targeted leads, two of which showed nanomolar PARP1 inhibition and antiproliferative activity.[p35] SyntheMol yielded six synthesized compounds with antibacterial activity.[p56] SyntheMol-RL reported 79 synthesized and tested compounds, 13 potent *in vitro* hits and *in vivo* activity for synthecin, while LambdaZero synthesized and evaluated 35 analogs from a larger proposed quinazoline series (24/35 inhibited soluble epoxide hydrolase), while 49 additional proposed analogs were not synthesized because of reported solubility and purification issues.[p68],[p69]

These studies provide different degrees of execution-grounded evidence rather than binary certification of synthetic reality. Evidence is strongest when the complete progression from prioritized candidates to route-supported, procurable, attempted, delivered and biologically tested compounds is reported together with failures, route modifications, safety and compliance outcomes and human intervention. A route executed as proposed, a chemist-edited route and one extensively redesigned by experts should not be treated as equivalent merely because each ultimately delivers material.

Procedure-level execution support

At the execution-facing edge of this hierarchy, procedure-generation systems narrow the gap between route planning and laboratory action. Multiple Optimized Specialists for AI-assisted Chemical Prediction (MOSAIC) uses specialized reaction-domain models to generate executable experimental protocols with confidence estimates and reports an overall experimental success rate of 71% and the realization of more than 35 new compounds.[p70] Although not a *de novo* molecular generator, it is relevant because it moves synthetic reality from route-level plausibility toward procedure-level actionability.

Taken together, these approaches show why 'synthesis-aware' should not be treated as a homogeneous evidentiary category. Chemical guidance narrows exploration toward chemically relevant regions; route-feasibility methods connect proposals to transformations, reactants or planned routes; and execution-grounded studies test whether selected proposals can be procured, attempted, delivered and evaluated under declared constraints. Table 3 applies this hierarchy to representative systems and summarizes the strongest publicly reported evidence and the remaining auditability gaps.

TABLE 3

Representative synthesis-aware molecular-design and execution-facing workflows classified according to how synthesis information enters the workflow and the strongest publicly reported evidence (selected examples; not exhaustive)[a]

System or method class (year)	Primary reality hook	Model/search strategy	Primary evidence category and reported validation
Scaffold- and functionality-guided methods: DeepScaffold, ScaffoldGVAE, REINVENT4 scaffold-, linker- and molecule-transformation modes and VeGA-RX/VeGA-SCX[(p37),(p38),(p39),(p40)]	Conditioning on preferred scaffolds, fragments, linkers, functional patterns, SMARTS constraints or related structural features	Scaffold generation and hopping, controlled decoration, linker design, molecule transformation and structure- or functionality-conditioned generation	Chemical guidance: directs generation toward specified cores, motifs or structurally defined regions of chemical space. Validation is primarily computational, and an explicit reaction sequence or executable route is not necessarily retained.
RENATE, REACTOR and LibINVENT[(p33),(p34),(p41)]	Reaction vectors, chemically valid reaction actions, or specified reactions with compatible reactants and building blocks	Pseudo-retrosynthetic generation, reinforcement learning over reaction-defined actions or reaction-based scaffold decoration	Route feasibility: products remain connected to specified transformations and compatible reactants. Evidence is primarily computational or retrospective; prospective execution was not the principal validation.
SynFormer (2025)[(p26)]	Synthetic pathways assembled from reaction transformations and commercially available building blocks	Transformer-based pathway generation with building-block selection and optimization	Route feasibility: each proposal is accompanied by explicit pathway and building-block logic. Prospective laboratory synthesis was not the primary reported evidence.
SynLlama (2025)[(p60)]	Molecules generated together with full synthetic pathways assembled from accessible building blocks and robust reaction templates	LLM fine-tuned for synthetic-pathway generation, analog generation and hit expansion	Route feasibility: computational pathway generation, synthesis-linked analog generation and hit expansion were demonstrated. Prospective execution was not the primary reported evidence.
ChemBO (2020)[(p63)]	Search constrained to a synthesis graph containing available reagents and permitted transformations	Gaussian-process Bayesian optimization over synthesis graphs	Route feasibility: produces synthesis-linked molecular recommendations and corresponding recipes in computational optimization benchmarks. Prospective synthesis was not reported as the principal validation.
ChemProjector (2024)[(p61)]	Projection of molecular structures into synthesizable space through synthetic-pathway postfix representations	Transformer encoder–decoder translating molecular graphs into synthesis-pathway representations	Route feasibility: demonstrates computational projection and planning within pathway-defined synthesizable space. Prospective laboratory execution was not the primary evidence.
SynNet (2021)[(p62)]	Synthetic trees constructed from purchasable starting materials and expert-curated reaction templates	Amortized synthetic-tree generation formulated as a Markov decision process conditioned on molecular embeddings	Route feasibility: demonstrates retrospective synthetic-tree recovery, generation of synthesizable analogs and *in silico* molecular optimization. Prospective execution was not the primary reported evidence.
RxnFlow (2024)[(p64)]	Predefined building blocks and reaction templates constrain sequential molecular assembly	GFlowNet-based sampling over synthetic pathways with action-space subsampling	Route feasibility: generation remains grounded in specified reactions and building blocks, including pocket-conditioned optimization. Validation was computational; prospective synthesis was not the primary evidence.
TANGO (2026)[(p36)]	Direct optimization against starting-material, intermediate and divergent-synthesis constraints	Reinforcement learning using a constraint-aware reward applied to general-purpose molecular generators	Route feasibility: demonstrates design-time constrained generation linked to specified synthesis relationships. Prospective laboratory execution was not reported as the primary validation.

(continued on next page)

TABLE 3 (*CONTINUED*)

System or method class (year)	Primary reality hook	Model/search strategy	Primary evidence category and reported validation
DINGOS (2019)[p32]	Rule-driven virtual synthesis using accessible starting materials and permitted transformations	Hybrid machine intelligence combining molecular design with rule-based synthetic assembly	Execution-grounded evidence (partial): four selected generated compounds were synthesized according to the proposed routes. Complete MPVC-style procurement, safety and compliance, failure, intervention and denominator reporting was not provided.
ClickGen (2024)[p35]	Hard-constrained modular chemistry centered on click and amide-forming reactions	Reinforcement learning and molecular inpainting within a restricted reaction space	Execution-grounded evidence (partial): three selected PARP1-targeted leads were synthesized within 10 days; two showed nanomolar PARP1 inhibition and antiproliferative activity. Complete procurement, safety and compliance, failure-mode, human-intervention and denominator reporting was not available.
SyntheMol (2024)[p56]	Enamine REAL Space building blocks and a defined reaction toolbox, with single-reaction products prioritized for rapid synthesis	Monte Carlo tree search guided by antibacterial property predictors within a predefined combinatorial space	Execution-grounded evidence (partial): of 150 selected candidates, 70 were available in REAL Space and ordered for synthesis; 58/70 were successfully synthesized and 6/58 showed antibacterial activity. The denominator chain is comparatively informative, but formal safety and compliance screening, structured synthesis-failure reporting and route-realization accounting were incomplete.
SyntheMol-RL (2026)[p69]	Enamine REAL and WuXi GalaXi building-block and reaction spaces supporting multi-parameter optimization within a 46-billion-compound synthesizable space	Reinforcement learning over reaction-defined combinatorial spaces guided by property-prediction models	Execution-grounded evidence (partial), with comparatively strong campaign-level denominators: 79 SyntheMol-RL compounds were synthesized and tested, 13 showed potent *in vitro* activity, and synthecin was active in a murine MRSA wound-infection model. Availability and synthesis denominators were reported across the comparative campaign, but formal safety and compliance screening, structured failure categorization and route-realization or human-intervention accounting remained incomplete.
LambdaZero (2024)[p68]	Structure-guided generation combined with drug-likeness and RetroGNN synthesizability constraints	Generative active learning using deep reinforcement learning, a surrogate docking oracle and a fragment-based molecular search space	Execution-grounded evidence (partial): 35 quinazoline-based analogs were synthesized and evaluated; 24 exhibited *in vitro* sEH inhibitory activity, and the lead showed submicromolar inhibition. 49 proposed analogs were not synthesized because of solubility and purification issues, providing useful attrition evidence. A complete MPVC-style route, procurement, safety and human-intervention audit, including a structured failure taxonomy, was not provided.

TABLE 3 (*CONTINUED*)

System or method class (year)	Primary reality hook	Model/search strategy	Primary evidence category and reported validation
MOSAIC (2026)[p70]	Generation of executable experimental procedures with confidence estimates	Specialized reaction-domain models for procedure generation with confidence estimation, followed by prospective experimental execution	Procedure-level execution evidence (partial): prospective execution achieved a reported overall experimental success rate of 71% and the realization of more than 35 new compounds. MOSAIC is not a *de novo* molecular generator but provides relevant evidence at the interface between route plausibility and laboratory execution.

CASP: computer-aided synthesis planning; GFlowNet: generative flow network; MRSA: methicillin-resistant *S. aureus*; sEH: soluble epoxide hydrolase.

[a] Evidence categories represent the authors' interpretation of the strongest publicly reported evidence. Chemical guidance denotes structural or building-block-directed generation without necessarily retaining an explicit route. Route feasibility denotes connection to defined reactions, reactants, synthetic trees, retrosynthetic logic or CASP-supported pathways, but does not imply laboratory execution. 'Execution-grounded evidence (partial)' indicates prospective execution with one or more MPVC elements incompletely reported. Categories can overlap and should not be interpreted as rigid rankings.

Learning from 'failure': The value of negative data in AI training

Reaction-aware generators can constrain chemistry, but they still learn from an experimental record heavily biased toward successes. For the purposes of makeable-by-design discovery, we use negative data broadly to include failed or low-yielding reactions under defined conditions, routes that collapse because of selectivity, purification or instability and biological negatives such as inactivity or unacceptable pharmacological or developability liabilities. Models trained only on successful reactions or active compounds learn an overoptimistic picture of chemistry and biology. Absence from the literature does not necessarily mean 'unexplored'; it might reflect attempts that failed, worked unreliably or were deprioritized before publication.[p71],[p72] The experimental record is therefore richer in failed or discarded hypotheses than the published literature suggests, yet much of this evidence remains uncaptured or inaccessible.

This gap matters because experienced medicinal and synthetic chemists routinely internalize failure. They recognize recurrent problems with reaction conditions, protecting groups, purification and selectivity that a model trained only on successful outcomes cannot inherit. Toniato *et al.* showed quantitatively that including unsuccessful reactions improves reaction-outcome prediction and can markedly improve performance when positive data are scarce; strong results were obtained from as few as 20 positive data points in the controlled dataset, supported by a negative dataset at least 40-fold larger.[p73] In other words, failures are not minor annotations; they are boundary conditions.

Negative data are most informative when matched to the modeling task they are intended to support. Reaction-outcome and condition-prediction models benefit from failed or low-yielding transformations recorded with their substrates, reagents, conditions and analytical outcomes. For retrosynthesis, route-planning and execution-risk modeling, useful negative evidence includes unworkable disconnections, unavailable precursors, incompatible step sequences, unstable intermediates and reaction, workup or purification failures; these data can inform disconnection ranking, inventory-aware planning and execution-risk assessment. By contrast, confirmed biological inactivity is primarily relevant to activity prediction and biological-space definition, whereas ADME, toxicity and other developability liabilities inform downstream property models and multi-parameter prioritization. Instability should likewise be labeled by context: instability during synthesis, isolation or storage informs route and execution models, whereas instability in biological media or formulation is principally a developability signal. Negative outcomes should therefore be captured with task-specific labels and provenance rather than pooled into an undifferentiated failure class.

A broader infrastructure is beginning to emerge. Earlier work in materials science showed that unsuccessful experiments can unlock discoveries inaccessible to models trained only on positive outcomes,[p74] while machine-readable reaction formats increasingly make both positive and negative outcomes easier to capture, exchange and reuse.[p75] In molecular discovery, BioReCS (biologically relevant chemical space) emphasizes that negative biological data are needed to define the boundaries of biologically relevant chemical space, while dark chemical matter and InertDB provide inactive-compound resources that help map repeated inactivity and identify regions requiring alternative biological hypotheses.[p76],[p77],[p78] Reaction and assay outcomes are also context-dependent across laboratories and experimental platforms, making provenance-rich metadata essential if models are to learn transferable constraints rather than local artifacts. Structured reaction and protocol data are equally important. Large language model (LLM)-enabled literature-mining frameworks such as ReactionSeek show that relevant chemical information remains locked in unstructured reaction schemes, procedural text and characterization records and can be converted into standardized machine-readable datasets.[p79] Such extraction should preserve source provenance for auditable model training. A practical next step is to standardize negative data across a task-linked attrition ladder: no route found, procurement barrier, reaction failure, purification or workup failure, product instability, biological inactivity and downstream developability or ADMET liabilities, including poor solubility, permeability limitations, metabolic instability and toxicity.

A practical barrier is that many of the richest failure records reside within pharmaceutical companies.[p72] Federated learning could enable collective learning without centralizing raw

structures or assay outcomes, provided appropriate privacy, security and governance constraints are maintained. Although not designed specifically around failed experiments, initiatives such as Machine Learning Ledger Orchestration for Drug Discovery (MELLODDY) demonstrate the feasibility of cross-company training on proprietary quantitative structure–activity relationship (QSAR) data under privacy-preserving conditions.[p80],[p81] Negative data should not be treated as immutable laws: today's failed transformation might become tomorrow's routine chemistry. Rather, they should be treated as probabilistic evidence of risk that helps models avoid rediscovering the same dead ends while preserving the possibility of future methodological advances.

Closing the loop: Self-driving laboratories as a feedback engine for synthetic reality

Route certification addresses only part of synthetic reality. Closed-loop Design–Make–Test–Learn (DMTL) systems provide a higher-order mechanism for testing generated molecules experimentally and feeding the resulting successes and failures back into subsequent design decisions. In the makeable-by-design setting considered here, a self-driving laboratory (SDL) can couple model-driven selection of candidates or experiments to automated synthesis and characterization, assays or analytical measurements and feedback that informs subsequent decisions. Programmable-chemistry frameworks such as Chemputation further illustrate how synthesis procedures can be encoded as executable chemical programs and compiled to modular robotic hardware, linking digital route specification to laboratory execution.[p82] Over the past decade, SDLs and related digital-chemistry paradigms have moved from concept to working reality across chemistry and materials science, combining automation, machine learning and decision-theoretic experiment planning under predefined operating constraints.[p83],[p84],[p85],[p86],[p87]

In small-molecule discovery, an SDL workflow can couple route-aware generation to modular synthesis or flow chemistry, analytical confirmation by liquid chromatography–mass spectrometry (LC–MS) or nuclear magnetic resonance (NMR) and automated biological testing. Its power lies not only in throughput but also in feedback: experiments are chosen because they reduce uncertainty or improve performance, rather than because they are merely easy to run. Several components of this vision already exist. A complementary next extension is from route proposal to reaction development itself. LLM-based synthesis-development platforms are beginning to support literature search, condition screening, optimization, kinetics, purification and scale-up planning, suggesting that, when coupled to upstream molecular-design and route-planning systems, synthetic reality could increasingly be audited beyond route proposal through experimental reaction development.[p88] Autonomous mobile-robot systems have been shown to operate synthesis platforms, shuttle samples to distributed analytics and automatically check the reproducibility of screening hits before further development.[p89] Other platforms have demonstrated autonomous or Bayesian-guided experiment selection in closed-loop workflows,[p90],[p91] while, in materials discovery, A-Lab autonomously realized 36 of 57 targeted inorganic materials over 17 days of operation.[p92],[p93] In drug discovery, early demonstrations of rapid AI-guided design, synthesis and validation already foreshadow this convergence.[p9]

However, SDLs are neither chemically nor operationally universal. Their accessible reaction space is limited by hardware, safety rules and the difficulty of automating heterogeneous mixtures, workups, separations and purification. Deployment also requires substantial capital and recurring investment in robotic hardware, synthesis modules, analytical and assay integration, software and data infrastructure, validation and qualification, maintenance, consumables and personnel with automation expertise.[p83],[p85] System interoperability, platform reliability, workflow integration and sustained technical support can also materially constrain deployment, particularly outside well-resourced industrial or institutional environments. Platform-specific execution constraints should, where possible, be represented upstream in the feasibility envelope. Penalizing transformations that require awkward unit operations or favoring homogeneous, modular chemistry can increase the likelihood that a proposed route is executable on the available platform.[p28] Purification and analytical throughput can nevertheless remain bottlenecks after a plausible route has been identified. Synthetic reality is therefore not an intrinsic property of a molecule alone, but of a molecule–route–platform–resource combination under defined local constraints. Progress depends not only on better planning algorithms but also on broader automation-ready chemistry, more robust purification strategies and richer feedback linking platform failures back into model design.

Beyond the economics and organizational demands of platform implementation, SDL objectives should also extend beyond immediate feasibility to include early indicators relevant to route cost, supply-chain resilience, environmental impact and later manufacturability. A molecule that is accessible at milligram scale can still fail translationally because the route is unsafe, non-scalable or economically fragile. This distinction should remain explicit: discovery-scale makeability asks whether a molecule can be produced reproducibly for learning and decision-making, whereas manufacturing-scale feasibility asks whether it can later be produced safely, robustly and economically at process scale. From an auditing perspective, SDLs are also a natural infrastructure for MPVC-style evaluation because they can prospectively quantify execution rates, purification bottlenecks, assay reproducibility and iteration costs under defined constraints. In this setting, synthetic reality ceases to be a downstream obstacle and becomes the medium through which models learn the coupled constraints of chemistry and biology. As this loop matures, the limiting factor increasingly shifts from speed alone to decision quality, which in turn motivates a clearer maturity language for synthesis-aware systems.

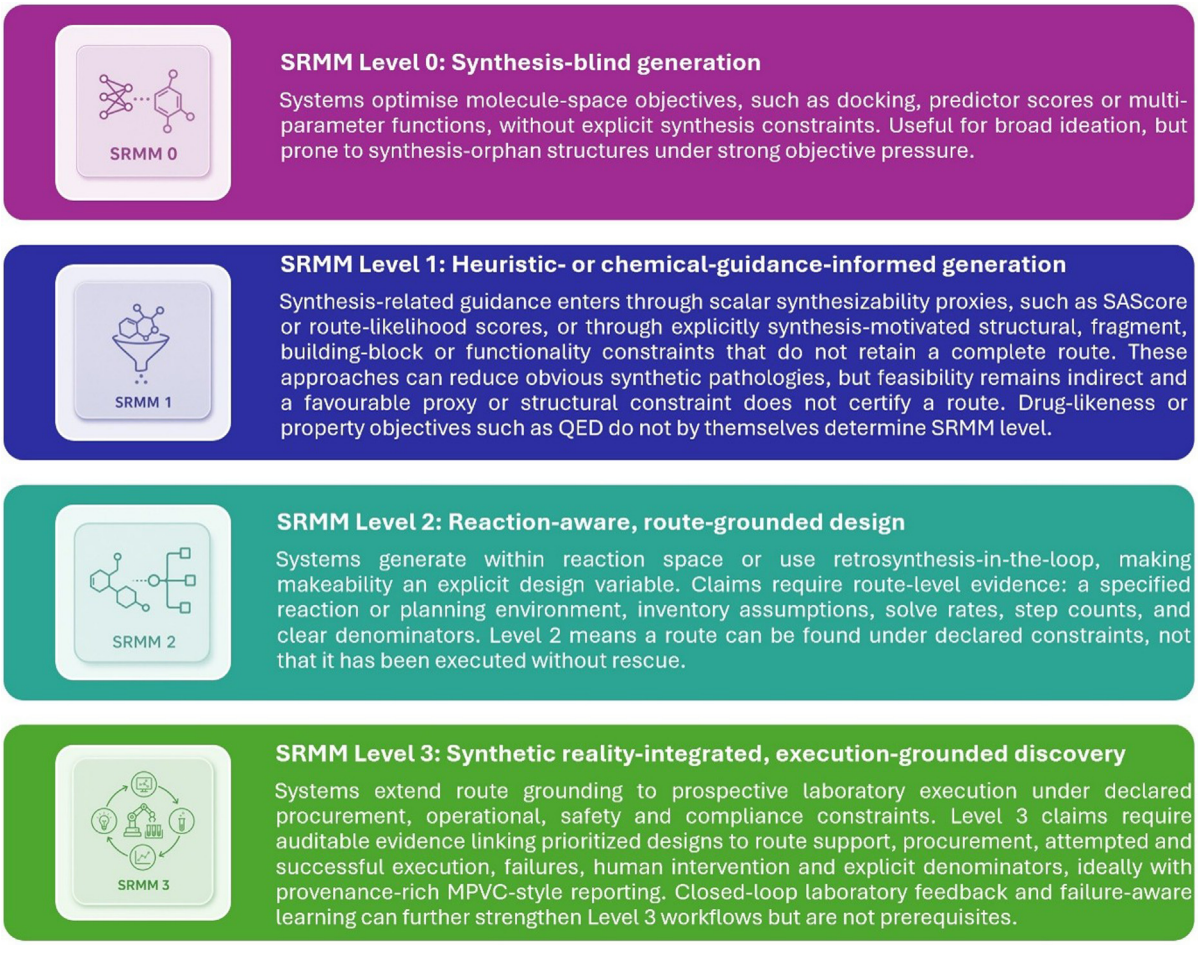
A synthetic reality maturity model for synthesis-aware design

The rapid expansion of generators, retrosynthesis engines and closed-loop platforms has made 'synthesis awareness' an over-

loaded term. Different systems introduce constraints at different stages, use different notions of feasibility and report evidence of varying strength. We therefore propose a practical four-level synthetic reality maturity model (SRMM) as a shared vocabulary rather than a leaderboard (Box 2). The SRMM classifies systems according to where synthesis constraints enter the workflow and what strength of evidence supports their actionability claims, thereby aligning evaluation with the claims being made. Levels 0 and 1 systems can reasonably be judged predominantly by molecule-level and heuristic objectives during early method development. Level 2 claims should be supported primarily by route-level outcomes under declared reaction, inventory and planning constraints, whereas Level 3 claims additionally require execution- and project-level evidence, including prospective outcomes, failures, intervention and iteration burden. For each workflow, we recommend reporting two related but distinct assessments: the intended design level, defined by where and how strongly synthesis constraints are embedded before laboratory commitment, and the evidenced maturity level, defined by the highest SRMM level whose evidentiary requirements are supported by the publicly reported validation. A method can therefore be designed at Level 2 while its evidence remains at Level 1 or can provide partial Level 3 evidence without satisfying all requirements for full execution-grounded maturity. In simple terms, the SRMM classifies how synthesis constraints enter a workflow and the level of evidence supporting its claims, whereas the MPVC specifies the prospective reporting evidence needed to substantiate execution-grounded Level 3 claims; selected MPVC elements can also strengthen the auditability of Level 2 route-feasibility claims.

KEYNOTE

Box 2 Synthetic reality maturity model (SRMM)



A four-level framework for classifying generative workflows according to when synthesis constraints are introduced and what evidence supports their actionability claims.

Applied illustratively to the systems discussed above, scaffold-, fragment- and functionality-guided generators, including DeepScaffold, ScaffoldGVAE and the relevant REINVENT4 and VeGA modes, can generally map to SRMM Level 1 when those constraints explicitly encode synthesis-related chemical guidance; generic scaffold or property conditioning alone does not increase SRMM maturity. SynFormer, TANGO and RxnFlow are intended Level 2 systems, and their published evidence supports route-grounded maturity consistent with Level 2 because reactions, building blocks, synthetic pathways or explicit synthesis constraints are embedded before laboratory commitment, although the completeness of route-level reporting varies. ClickGen, SyntheMol and SyntheMol-RL are likewise designed at Level 2, but their prospective synthesis campaigns provide partial support for Level 3. Full Level 3 maturity would require integrated evidence that committed designs survive route certification, precursor accessibility and procurement, safety and compliance assessment, experimental execution and structured failure reporting and provenance-rich experimental context, with route modifications or human rescue and denominator-based outcomes disclosed. Their present status is therefore best described as intended Level 2 and evidenced at Level 2 with partial Level 3 support. These assignments are illustrative and evidence-dependent rather than fixed rankings and might change as further validation becomes available.

Trust, governance and the human–AI partnership

A proposal accompanied by a transparent route, accessible precursors and clearly identified fragility points is easier for chemists to interrogate, assess and act on than one presented only with a high score. Synthetic reality therefore provides a practical form of route-grounded operational explainability: it does not replace mechanistic interpretation of biological predictions, but answers the operational question of how a molecule could actually be made. Upstream molecular-representation, generation or affinity models can improve the computational signals used to propose and prioritize molecules, but they do not substitute for route-level executability, procurement evidence or prospective synthesis outcomes.[p8],[p16],[p94],[p95]

Governance should focus on how the evidence specified above is traced, reviewed and acted upon. A candidate should be accompanied by a traceable evidence package covering relevant model, data and tool versions; route and inventory assumptions; precursor and procurement status; safety and compliance screening; applicability-domain and uncertainty assessments; and route-realization status (model-native, chemist-edited or expert-redesigned). These records allow chemists to distinguish a genuinely model-supported proposal from one whose apparent feasibility depends on substantial expert repair. Locked selection protocols, explicit denominators, versioned provenance and documented route modifications are also needed to reduce selective reporting and improve auditability.

Human oversight remains essential, including measures to mitigate automation bias. Workflows should define veto criteria for infeasible or hazardous designs, triggers for mandatory expert review and responsibility for route approval, safety assessment, experimental authorization and interpretation of failures. Human experts should retain responsibility for identifying hidden chemoselectivity, purification, stability and platform risks; modifying, rescuing or rejecting proposed routes; and judging whether the available evidence justifies experimental investment.

LLM-based chemistry assistants introduce additional risks because fluent outputs can contain incorrect precedents, conditions, routes or citations. Such outputs should be treated as hypotheses and verified through database retrieval, reaction and retrosynthesis checks, supplier and inventory checks, uncertainty assessment and human review; tool-augmented systems, such as ChemCrow, illustrate the importance of combining generative capability with retrieval and external verification.[p19],[p96],[p97],[p98] Safety and security controls should additionally screen routes and precursors for exceptional hazards, restricted substances, controlled precursors and other dual-use or security-sensitive materials before execution.[p96],[p99] These governance requirements help establish whether a claim is auditable and trustworthy; they do not, by themselves, determine whether a candidate should be synthesized. That candidate-level decision is the role of the gates below.

Decision gates: From algorithmic suggestion to laboratory commitment

The decision gates are the candidate-level component of the framework. They use the pre-commitment evidence corresponding to relevant MPVC criteria, interpret it in light of the workflow's SRMM maturity and the uncertainty associated with the underlying models, and support a go/no-go decision for laboratory commitment. The synthesis-actionable rate remains a portfolio- or workflow-level outcome measure rather than a gate for an individual molecule (Figure 2).

Gate 1: Route, procurement and compliance audit (actionability)

Drawing on the pre-commitment evidence corresponding to the MPVC's route, procurement and safety criteria, the proposal should satisfy predefined feasibility criteria and include at least one route judged executable within the declared feasibility envelope, explicit inventory assumptions, precursor availability, lead-time or custom-synthesis requirements, safety and compliance status and key fragility drivers such as purification bottlenecks, selectivity concerns or high-risk transformations. The route provenance at commitment (model-native, chemist-edited or expert-redesigned) should also be recorded.

Gate 2: Model confidence and applicability (epistemic risk)

Each model used in the workflow should be assessed against its applicability domain. For property, activity or target-aware models, this includes the molecular, target and assay space represented in the training or validation data; for synthesis models, it includes the reaction classes, building blocks, templates, inventories and platform conditions covered by the planning engine. Candidates or routes that fall outside these domains or carry substantial predictive uncertainty should trigger enhanced expert review, which can include similarity-based or unfamiliarity-based out-of-distribution assessment, model-specific stress testing or targeted validation where the additional evidence would change the commitment decision.[p100] Protein conformational

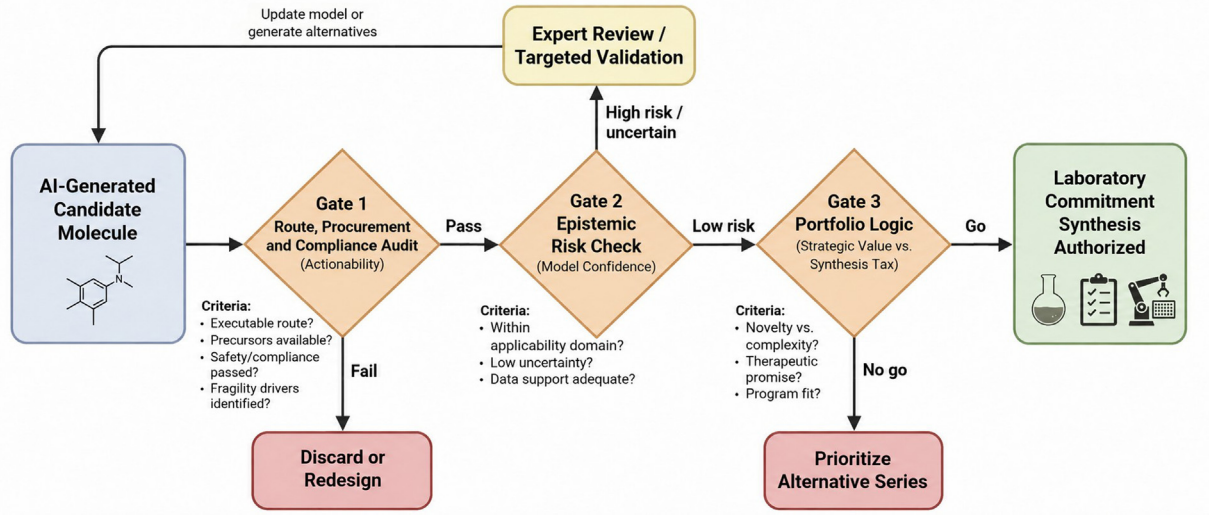


FIGURE 2

Decision gates for laboratory commitment of AI-generated molecules. Practical go/no-go framework for translating synthetic-reality evidence into bench commitment and mitigating automation bias. Gate 1 (route, procurement and compliance audit) assesses route feasibility, precursor availability, procurement constraints, safety and compliance and synthetic fragility. Gate 2 (epistemic risk) assesses model confidence, applicability domain, receptor-state uncertainty and data support. Gate 3 (portfolio logic) balances strategic novelty and therapeutic promise against the synthesis tax, including execution cost, complexity, opportunity cost and iteration burden. Experimental outcomes feed back into MPVC reporting, calibration of future actionability assessments and reassessment of the evidenced SRMM maturity. This figure was created using BioRender (http://biorender.com/). AI: artificial intelligence; MPVC: Minimal Prospective Validation Checklist; SRMM: synthetic reality maturity model.

uncertainty is a specific form of epistemic risk for target-aware generation and docking. Where binding-site flexibility is consequential, candidates should be evaluated across a representative receptor ensemble derived from multiple experimental structures, molecular-dynamics or enhanced-sampling trajectories or ligand-conditioned flexible-structure models. The receptor states used in the assessment should be documented, including how they were generated and selected, how predictions across states were combined and whether candidate ranking is robust across plausible conformations. Strong dependence on a single, poorly supported receptor state should trigger expert review and targeted structural or biophysical validation (where decision-relevant and feasible) before laboratory commitment.[p101],[p102],[p103]

Gate 3: Strategic value versus synthesis tax (portfolio logic and prioritization)

Teams should explicitly balance novelty and biological or therapeutic promise against the projected cost, complexity, opportunity cost and iteration burden of making the molecule. A high synthesis tax does not automatically disqualify a proposal: difficult chemistry can be justified when biological promise or strategic novelty is sufficiently strong, but that trade-off should be made explicit rather than hidden behind molecule-level scores.

In summary, Gate 1 asks whether the proposal can be executed, Gate 2 whether the evidence supporting it can be trusted and Gate 3 whether it merits the required investment. If the proposal proceeds, the resulting procurement, synthesis, failure and testing outcomes should flow back into MPVC reporting, be used to evaluate and recalibrate how predicted actionability translates into observed execution outcomes, and inform reassessment of the workflow's evidenced SRMM maturity. The gates thereby complete an auditable cycle from workflow classification and evidence reporting to laboratory commitment and experimental learning.

Concluding remarks and future perspectives: From generative imagination to synthetic reality in drug discovery

Generative AI has opened access to vast chemical design spaces; however, without synthetic reality, much of this space remains a mirage. The central claim of this review is that progress should be judged less by how many molecules a model can imagine and more by how reliably it can propose molecules that can be made, tested and advanced. In this sense, synthetic reality is not a secondary filter. It is the operational foundation of makeable-by-design discovery.

The framework proposed herein is intentionally practical. Heuristic scores remain useful for triage, but route support or certification, procurement-aware evaluation, safety and compliance screening, provenance-rich reporting and prospective validation provide the evidence needed to audit genuine synthesis awareness. Within this framework, the MPVC defines the evidence that should be reported in prospective validation, the synthesis-actionable rate quantifies the fraction of prioritized designs meeting declared actionability criteria, the SRMM distinguishes intended design from evidenced maturity and the decision gates guide candidate-level laboratory commitment. Synthetic reality is therefore best understood as a discovery-

stage execution standard: broader than classical synthesizability, but not synonymous with full downstream manufacturability.

For AI-generated designs, synthetic reality is a prerequisite for experimental progression, not a guarantee of translational or clinical success. A makeable molecule can still fail because of target biology, clinical pharmacokinetics, efficacy, toxicity or endpoint translation. What synthetic reality does is to reduce an avoidable source of attrition and make DMTL cycles more informative by requiring digital proposals to confront an early experimental reality check. This is precisely why negative outcomes provide learning signals from failure, while SDLs can accelerate closed-loop experimentation and feedback where such integration is appropriate.

Looking forward, the most productive paradigm is a unified design-and-make workflow in which structures and routes are co-designed under explicit constraints, informed by real inventories, experimental outcomes, platform limits and human judgment. The same principles should extend beyond conventional small molecules to modalities such as macrocycles and PROTACs. They are likewise relevant to chemical-probe design, where feasibility, selectivity, target engagement, control compounds and context-of-use constraints can be even sharper.[(p104)] As target-aware and structure-aware models continue to improve, the decisive question will increasingly be not whether AI can rank plausible binders, but whether it can advance them into the laboratory through auditable, reproducible chemistry. Pocket-conditioned and 3D generators should therefore be evaluated for structural and conformational realism and across plausible receptor states where conformational heterogeneity is relevant, without treating binding-aware predictions as substitutes for route- and execution-level validation.[(p101),(p102),(p103),(p105)] Generative AI will become genuinely transformative in drug discovery only when it moves beyond unconstrained ideation to function as a trustworthy, auditable partner in executable chemistry under accountable human oversight.

CRediT authorship contribution statement

Sebastián O. Simonetti: Writing – review & editing, Visualization, Investigation, Conceptualization. **Aldana B. Moroni:** Writing – review & editing, Visualization, Investigation. **Maria M. Zanardi:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Frank Kozielski:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Geoffrey Wells:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Exequiel O.J. Porta:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Funding acquisition, Conceptualization.

Acknowledgments

This work was supported by UK Research and Innovation through the Medical Research Council (MRC; grant number MR/X013995/1) and by the 2025/26 University College London (UCL) Global Engagement Fund Award (project 592980). F.K. and G.W. thank the UCL School of Pharmacy for its support. E.O.J.P. is grateful to the MRC for fellowship support, and A.B.M. thanks Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET; National Scientific and Technical Research Council, Argentina) for fellowship support. S.O.S. is grateful for support from Instituto de Química Rosario (IQUIR; Rosario Institute of Chemistry) and the Universidad Nacional de Rosario, while M.M.Z. thanks Instituto de Investigaciones en Ingeniería Ambiental, Química y Biotecnología Aplicada and the Universidad Católica Argentina for their support. The authors thank Dr Renzo Carlucci (Universidad Nacional de La Plata) for insightful discussions and critical feedback on synthesizability metrics, retrosynthetic assessment and practical aspects of synthesis-aware molecular design.

Declarations of interest

No interests are declared.

Data availability

No new experimental data were generated. This review synthesizes information from published and publicly available sources.

References

1. Reymond J-L, Awale M. Exploring chemical space for drug discovery using the chemical universe database. *ACS Chem Neurosci*. 2012;3:649–657. https://doi.org/10.1021/cn3000422.
2. Gao W, Coley CW. The synthesizability of molecules proposed by generative models. *J Chem Inf Model*. 2020;60:5714–5723. https://doi.org/10.1021/acs.jcim.0c00174.
3. Bhat AR, Ahmed S. Artificial intelligence (AI) in drug design and discovery: a comprehensive review. *Silico Res Biomed*. 2025;1, 100049. https://doi.org/10.1016/j.insi.2025.100049.
4. Das U. Generative AI for drug discovery and protein design: the next frontier in AI-driven molecular science. *Med Drug Discov*. 2025;27, 100213. https://doi.org/10.1016/j.medidd.2025.100213.
5. Özçelik R, Brinkmann H, Criscuolo E, Grisoni F. Generative deep learning for de novo drug design—a Chemical Space Odyssey. *J Chem Inf Model*. 2025;65:7352–7372. https://doi.org/10.1021/acs.jcim.5c00641.
6. Serrano DR et al. Artificial intelligence (AI) applications in drug discovery and drug delivery: revolutionizing personalized medicine. *Pharmaceutics*. 2024;16:1328. https://doi.org/10.3390/pharmaceutics16101328.
7. Wan Z et al. Applications of artificial intelligence in drug repurposing. *Adv Sci*. 2025;12, 2411325. https://doi.org/10.1002/advs.202411325.
8. Zeng C et al. PropMolFlow: property-guided molecule generation with geometry-complete flow matching. *Nat Comput Sci*. 2026;6:233–242. https://doi.org/10.1038/s43588-025-00946-y.
9. Zhavoronkov A et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*. 2019;37:1038–1040. https://doi.org/10.1038/s41587-019-0224-x.
10. Das P et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat Biomed Eng*. 2021;5:613–623. https://doi.org/10.1038/s41551-021-00689-x.
11. Ren F et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nat Biotechnol*. 2025;43:63–75. https://doi.org/10.1038/s41587-024-02143-0.
12. Xu Z et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. *Nat Med*. 2025;31:2602–2610. https://doi.org/10.1038/s41591-025-03743-2.
13. Gangwal A, Lavecchia A. Unleashing the power of generative AI in drug discovery. *Drug Discov Today*. 2024;29, 103992. https://doi.org/10.1016/j.drudis.2024.103992.
14. Dharmasivam M, Kaya B, Akinware A, Gholam Azad M, Richardson DR. Leading artificial intelligence-driven drug discovery platforms: 2025 landscape

and global outlook. *Pharmacol Rev*. 2026;78, 100102. https://doi.org/10.1016/j.pharmr.2025.100102.
15. Wu K et al. TamGen: drug design with target-aware molecule generation through a chemical language model. *Nat Commun*. 2024;15:9360. https://doi.org/10.1038/s41467-024-53632-4.
16. Passaro S et al. Boltz-2: towards accurate and efficient binding affinity prediction. *bioRxiv*. 2025. https://doi.org/10.1101/2025.06.14.659707.
17. Callaway E. 'An AlphaFold 4' scientists marvel at DeepMind drug spin-off's exclusive new AI. *Nature*. 2026;651:18. https://doi.org/10.1038/d41586-026-00365-7.
18. Singh N, Guha R, Giulianotti MA, Pinilla C, Houghten RA, Medina-Franco JL. Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J Chem Inf Model*. 2009;49:1010–1024. https://doi.org/10.1021/ci800426u.
19. Walters WP, Barzilay R. Critical assessment of AI in drug discovery. *Expert Opin Drug Discov*. 2021;16:937–947. https://doi.org/10.1080/17460441.2021.1915982.
20. Renz P, Van Rompaey D, Wegner JK, Hochreiter S, Klambauer G. On failure modes in molecule generation and optimization. *Drug Discov Today Technol*. 2019;32–33:55–63. https://doi.org/10.1016/j.ddtec.2020.09.003.
21. Van den Broek RL, Patel S, Van Westen GJP, Jespers W, Sherman W. In search of beautiful molecules: a perspective on generative modeling for drug design. *J Chem Inf Model*. 2025;65:9383–9397. https://doi.org/10.1021/acs.jcim.5c01203.
22. Hu Q et al. Target-aware 3D molecular generation based on guided equivariant diffusion. *Nat Commun*. 2025;16:7928. https://doi.org/10.1038/s41467-025-63245-0.
23. Cremer J, Le T, Noé F, Clevert DA, Schütt KT. PILOT: equivariant diffusion for pocket-conditioned *de novo* ligand generation with multi-objective guidance *via* importance sampling. *Chem Sci*. 2024;15:14954–14967. https://doi.org/10.1039/D4SC03523B.
24. 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. https://doi.org/10.1038/s41467-025-57582-3.
25. Chen S, Jung Y. Estimating the synthetic accessibility of molecules with building block and reaction-aware SAScore. *J Cheminform*. 2024;16:83. https://doi.org/10.1186/s13321-024-00879-0.
26. Gao W, Luo S, Coley CW. Generative AI for navigating synthesizable chemical space. *Proc Natl Acad Sci USA*. 2025;122, e2415665122. https://doi.org/10.1073/pnas.2415665122.
27. Guo J, Schwaller P. Directly optimizing for synthesizability in generative molecular design using retrosynthesis models. *Chem Sci*. 2025;16:6943–6956. https://doi.org/10.1039/d5sc01476j.
28. Papidocha SM, Burger A, Bernales V, Aspuru-Guzik A. The elephant in the lab: synthesizability in generative small-molecule design. *Curr Opin Chem Eng*. 2026;51, 101217. https://doi.org/10.1016/j.coche.2025.101217.
29. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: BENCHMARKING models for de novo molecular design. *J Chem Inf Model*. 2019;59:1096–1108. https://doi.org/10.1021/acs.jcim.8b00839.
30. Polykovskiy D et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Front Pharmacol*. 2020;11, 565644. https://doi.org/10.3389/fphar.2020.565644.
31. 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. https://doi.org/10.1186/s13321-023-00781-1.
32. Button A, Merk D, Hiss JA, Schneider G. Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. *Nat Mach Intell*. 2019;1:307–315. https://doi.org/10.1038/s42256-019-0067-7.
33. Ghiandoni GM et al. RENATE: a pseudo-retrosynthetic tool for synthetically accessible *de novo* design. *Mol Inform*. 2022;41, e2100207. https://doi.org/10.1002/minf.202100207.
34. Horwood J, Noutahi E. Molecular design in synthetically accessible chemical space via deep reinforcement learning. *ACS Omega*. 2020;5:32984–32994. https://doi.org/10.1021/acsomega.0c04153.
35. Wang M et al. ClickGen: directed exploration of synthesizable chemical space via modular reactions and reinforcement learning. *Nat Commun*. 2024;15:10127. https://doi.org/10.1038/s41467-024-54456-y.
36. Guo J, Schwaller P. TANGO: direct optimization of constrained synthesizability for generative molecular design. *Nat Comput Sci*. 2026;6:260–270. https://doi.org/10.1038/s43588-026-00959-1.
37. Li Y, Hu J, Wang Y, Zhou J, Zhang L, Liu Z. DeepScaffold: a comprehensive tool for scaffold-based de novo drug discovery using deep learning. *J Chem Inf Model*. 2020;60:77–91. https://doi.org/10.1021/acs.jcim.9b00727.
38. Hu C et al. ScaffoldGVAE: scaffold generation and hopping of drug molecules via a variational autoencoder based on multi-view graph neural networks. *J Cheminform*. 2023;15:91. https://doi.org/10.1186/s13321-023-00766-0.
39. Loeffler HH et al. Reinvent 4: modern AI-driven generative molecule design. *J Cheminform*. 2024;16:20. https://doi.org/10.1186/s13321-024-00812-5.
40. Delre P, Bellofatto G, Lavecchia A. VeGA-RX and VeGA-SCX: controllable SMARTS-guided generative transformers for precision-driven *de novo* drug design. *J Chem Inf Model*. 2026;66:5189–5205. https://doi.org/10.1021/acs.jcim.6c00535.
41. Fialková V et al. LibINVENT: reaction-based generative scaffold decoration for *in silico* library design. *J Chem Inf Model*. 2022;62:2046–2063. https://doi.org/10.1021/acs.jcim.1c00469.
42. Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminform*. 2009;1:8. https://doi.org/10.1186/1758-2946-1-8.
43. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. *Nat Chem*. 2012;4:90–98. https://doi.org/10.1038/nchem.1243.
44. Coley CW, Rogers L, Green WH, Jensen KF. SCScore: synthetic complexity learned from a reaction corpus. *J Chem Inf Model*. 2018;58:252–261. https://doi.org/10.1021/acs.jcim.7b00622.
45. Thakkar A, Chadimová V, Bjerrum EJ, Engkvist O, Reymond JL. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem Sci*. 2021;12:3339–3349. https://doi.org/10.1039/d0sc05401a.
46. Neeser RM, Correia B, Schwaller P. FSscore: a personalized machine learning-based synthetic feasibility score. *Chem Methods*. 2024;4, e202400024. https://doi.org/10.1002/cmtd.202400024.
47. Thomas M, O'boyle NM, Bender A, de Graaf C. MolScore: a scoring, evaluation and benchmarking framework for generative models in de novo drug design. *J Cheminform*. 2024;16:64. https://doi.org/10.1186/s13321-024-00861-w.
48. Genheden S, Thakkar A, Chadimová V, Reymond JL, Engkvist O, Bjerrum E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J Cheminform*. 2020;12:70. https://doi.org/10.1186/s13321-020-00472-1.
49. Tu Z et al. ASKCOS: open-source, data-driven synthesis planning. *Acc Chem Res*. 2025;58:1764–1775. https://doi.org/10.1021/acs.accounts.5c00155.
50. Skoraczyński G, Kitlas M, Miasojedow B, Gambin A. Critical assessment of synthetic accessibility scores in computer-assisted synthesis planning. *J Cheminform*. 2023;15:6. https://doi.org/10.1186/s13321-023-00678-z.
51. Cao D et al. Benchmarking real-world applicability of molecular generative models from de novo design to lead optimization with MolGenBench. *bioRxiv*. 2025. https://doi.org/10.1101/2025.11.03.686215.
52. Genheden S, Bjerrum E. PaRoutes: towards a framework for benchmarking retrosynthesis route predictions. *Digit Discov*. 2022;1:527–539. https://doi.org/10.1039/D2DD00015F.
53. Aal E, Ali RS, Meng J, Khan MEI, Jiang X. Machine learning advancements in organic synthesis: a focused exploration of artificial intelligence applications in chemistry. *Artif Intell Chem*. 2024;2, 100049. https://doi.org/10.1016/j.aichem.2024.100049.
54. Hagemann P, Müller S, George J, Benner P. Transport novelty distance: a distributional metric for evaluating material generative models. *J Phys Mater*. 2026;9, 035011. https://doi.org/10.1088/2515-7639/ae869a.
55. Phan T-L, Nguyen Song NN, Stadler PF. SynRXN: an open benchmark and curated dataset for computational reaction modeling. *Sci Data*. 2026;13:625. https://doi.org/10.1038/s41597-026-07260-w.
56. Swanson K, Liu G, Catacutan DB, Arnold A, Zou J, Stokes JM. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. *Nat Mach Intell*. 2024;6:338–353. https://doi.org/10.1038/s42256-024-00809-7.
57. Steinbeck C et al. Proposal NFDI4Chem 2025-2030 in the national research data infrastructure (NFDI)—our vision: all chemists publish FAIR data. *Res Ideas Outcomes*. 2025;11, e177037. https://doi.org/10.3897/rio.11.e177037.
58. Seifrid M et al. Routescore: punching the ticket to more efficient materials development. *ACS Cent Sci*. 2022;8:122–131. https://doi.org/10.1021/acscentsci.1c01002.
59. Fromer JC, Coley CW. An algorithmic framework for synthetic cost-aware decision making in molecular design. *Nat Comput Sci*. 2024;4:440–450. https://doi.org/10.1038/s43588-024-00639-y.

60. Sun K et al. SynLlama: generating synthesizable molecules and their analogs with large language models. *ACS Cent Sci*. 2025;11:2108–2120. https://doi.org/10.1021/acscentsci.5c01285.
61. Luo S, Gao W, Wu Z, Peng J, Coley CW, Ma J. Projecting molecules into synthesizable chemical spaces. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2406.04628.
62. Gao W, Mercado R, Coley CW. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. *arXiv*. 2021. https://doi.org/10.48550/arXiv.2110.06389.
63. Korovina K et al. ChemBO: Bayesian optimization of small organic molecules with synthesizable recommendations. *arXiv*. 2019. https://doi.org/10.48550/arXiv.1908.01425.
64. Seo S et al. Generative flows on synthetic pathway for drug design. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2410.04542.
65. Shen T et al. Compositional flows for 3D molecule and synthesis pathway co-design. *arXiv*. 2025. https://doi.org/10.48550/arXiv.2504.08051.
66. Guo J, Chen J, Gx-Chen A, Schwaller P. Sample-efficient generative molecular design using memory manipulation. *Nat Mach Intell*. 2026;8:449–460. https://doi.org/10.1038/s42256-026-01200-4.
67. Teixeira RI, Andresini M, Luisi R, Benyahia B. Computer-aided retrosynthesis for greener and optimal total synthesis of a helicase-primase inhibitor active pharmaceutical ingredient. *JACS Au*. 2024;4:4263–4272. https://doi.org/10.1021/jacsau.4c00624.
68. Korablyov M et al. Generative active learning for the search of small-molecule protein binders. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2405.01616.
69. Swanson K et al. SyntheMol-RL: a flexible reinforcement learning framework for designing easily synthesizable antibiotics. *Mol Syst Biol*. 2026;22:833–867. https://doi.org/10.1038/s44320-026-00206-9.
70. Li H et al. Collective intelligence for AI-assisted chemical synthesis. *Nature*. 2026;651:107–115. https://doi.org/10.1038/s41586-026-10131-4.
71. Strieth-Kalthoff F, Sandfort F, Kühnemund M, Schäfer FR, Kuchen H, Glorius F. Machine learning for chemical reactivity: the importance of failed experiments. *Angew Chem Int Ed Engl*. 2022;61, e202204647. https://doi.org/10.1002/anie.202204647.
72. Maloney MP et al. Negative data in data sets for machine learning training. *Org Lett*. 2023;25:2945–2947. https://doi.org/10.1021/acs.orglett.3c01282.
73. Toniato A, Vaucher AC, Laino T, Graziani M. Negative chemical data boosts language models in reaction outcome prediction. *Sci Adv*. 2025;11, eadt5578. https://doi.org/10.1126/sciadv.adt5578.
74. Raccuglia P et al. Machine-learning-assisted materials discovery using failed experiments. *Nature*. 2016;533:73–76. https://doi.org/10.1038/nature17439.
75. Nippa DF et al. Simple user-friendly reaction format. *Mol Inform*. 2025;44, e202400361. https://doi.org/10.1002/minf.202400361.
76. Medina-Franco JL, López-López E, Avellaneda-Tamayo JF, Zamora WJ. On the biologically relevant chemical space: BioReCS. *Front Drug Discov*. 2025;5, 1674289. https://doi.org/10.3389/fddsv.2025.1674289.
77. An S et al. InertDB as a generative AI-expanded resource of biologically inactive small molecules from PubChem. *J Cheminform*. 2025;17:49. https://doi.org/10.1186/s13321-025-00999-1.
78. Wassermann AM et al. Dark chemical matter as a promising starting point for drug lead discovery. *Nat Chem Biol*. 2015;11:958–966. https://doi.org/10.1038/nchembio.1936.
79. Li J, Li M, Yang Q, Luo S. ReactionSeek: LLM-powered literature data mining and knowledge discovery in organic synthesis. *Nat Commun*. 2026;17:3356. https://doi.org/10.1038/s41467-026-70180-1.
80. Aledhari M, Razzak R, Parizi RM, Saeed F. Federated learning: a survey on enabling technologies, protocols, and applications. *IEEE Access*. 2020;8:140699–140725. https://doi.org/10.1109/access.2020.3013541.
81. Heyndrickx W et al. MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. *J Chem Inf Model*. 2024;64:2331–2344. https://doi.org/10.1021/acs.jcim.3c00799.
82. Cronin L, Pagel S, Sharma A. Chemputer and chemputation – a universal chemical compound synthesis machine. *Proc Natl Acad Sci USA*. 2026;123, e2511080123. https://doi.org/10.1073/pnas.2511080123.
83. Tom G et al. Self-driving laboratories for chemistry and materials science. *Chem Rev*. 2024;124:9633–9732. https://doi.org/10.1021/acs.chemrev.4c00055.
84. Häse F, Roch LM, Aspuru-Guzik A. Next-generation experimentation with self-driving laboratories. *Trends Chem*. 2019;1:282–291. https://doi.org/10.1016/j.trechm.2019.02.007.
85. Bayley O, Savino E, Slattery A, Noël T. Autonomous chemistry: navigating self-driving labs in chemical and material sciences. *Matter*. 2024;7:2382–2398. https://doi.org/10.1016/j.matt.2024.06.003.
86. Slattery A et al. Automated self-optimization, intensification, and scale-up of photocatalysis in flow. *Science*. 2024;383, eadj1817. https://doi.org/10.1126/science.adj1817.
87. Bonciolini S, Pulcinella A, Noël T. Tech-enhanced synthesis: exploring the synergy between organic chemistry and technology. *J Am Chem Soc*. 2025;147:28523–28545. https://doi.org/10.1021/jacs.5c10303.
88. Ruan Y et al. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nat Commun*. 2024;15:10160. https://doi.org/10.1038/s41467-024-54457-x.
89. Dai T et al. Autonomous mobile robots for exploratory synthetic chemistry. *Nature*. 2024;635:890–897. https://doi.org/10.1038/s41586-024-08173-7.
90. Burger B et al. A mobile robotic chemist. *Nature*. 2020;583:237–241. https://doi.org/10.1038/s41586-020-2442-2.
91. Lohmann AHJ et al. Bayesian optimized electrosynthesis of azobenzenes in a self-optimizing flow set-up: from doe-guidance to gram-scale preparation. *ChemRxiv*. 2026. https://doi.org/10.26434/chemrxiv-2026-8rd5c.
92. MacLeod BP et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci Adv*. 2020;6, eaaz8867. https://doi.org/10.1126/sciadv.aaz8867.
93. Szymanski NJ et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*. 2023;624:86–91. https://doi.org/10.1038/s41586-023-06734-w.
94. 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;4:1256–1264. https://doi.org/10.1038/s42256-022-00580-7.
95. Ross J et al. GP-MoLFormer: a foundation model for molecular generation. *Digit Discov*. 2025;4:2684–2696. https://doi.org/10.1039/d5dd00122f.
96. Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. *Nat Mach Intell*. 2024;6:525–535. https://doi.org/10.1038/s42256-024-00832-8.
97. McGibbon M et al. From intuition to AI: evolution of small molecule representations in drug discovery. *Brief Bioinform*. 2023;25, bbad422. https://doi.org/10.1093/bib/bbad422.
98. Bender A, Schneider N, Segler M, Walters WP, Engkvist O, Rodrigues T. Evaluation guidelines for machine learning tools in the chemical sciences. *Nat Rev Chem*. 2022;6:428–442. https://doi.org/10.1038/s41570-022-00391-9.
99. Urbina F, Lentzos F, Invernizzi C, Ekins S. Dual use of artificial-intelligence-powered drug discovery. *Nat Mach Intell*. 2022;4:189–191. https://doi.org/10.1038/s42256-022-00465-9.
100. Van Tilborg D, Rossen L, Grisoni F. Molecular deep learning at the edge of chemical space. *Nat Mach Intell*. 2026;8:575–587. https://doi.org/10.1038/s42256-026-01216-w.
101. Lu W et al. DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nat Commun*. 2024;15:1071. https://doi.org/10.1038/s41467-024-45461-2.
102. Fischer M, Coleman RG, Fraser JS, Shoichet BK. Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. *Nat Chem*. 2014;6:575–583. https://doi.org/10.1038/nchem.1954.
103. Ricci-Lopez J, Aguila SA, Gilson MK, Brizuela CA. Improving structure-based virtual screening with ensemble docking and machine learning. *J Chem Inf Model*. 2021;61:5362–5376. https://doi.org/10.1021/acs.jcim.1c00511.
104. Carlucci R, Porta EOJ. Artificial intelligence for small-molecule chemical probe design: why drug-centric models may not transfer cleanly. *Front Drug Discov*. 2026;6, 1836983. https://doi.org/10.3389/fddsv.2026.1836983.
105. Sanjrani N, Coupry DE, Pogány P, Palmer DS, Pickett SD. Benchmarking 3D structure-based molecule generators. *J Chem Inf Model*. 2025;65:8006–8021. https://doi.org/10.1021/acs.jcim.5c01020.