Benchmarking AI for Natural Product Drug Discovery: Data Integrity, Model Robustness, Mechanistic Validity, and Translational Utility

Ethan Wright[1]*, Chloe Bennett[2], Jack Turner[1]

[1]Department of Artificial Intelligence for Anti-obesity Phytochemicals, Faculty of Pharmacy, University of Leeds, Leeds, United Kingdom.
[2]Department of Machine Learning for Lipase Inhibition, Faculty of Pharmacy, University of Sheffield, Sheffield, United Kingdom.

ABSTRACT

Artificial intelligence is increasingly used to support natural product identification, bioactivity prediction, target inference, molecular property estimation, and candidate prioritization, yet model evaluation frequently remains dominated by aggregate predictive performance. Such evaluation can obscure structural errors, unreliable annotations, data leakage, narrow chemical-space coverage, distribution shifts, mechanistic uncertainty, and limited relevance to experimental decision-making. This article proposes a conceptual benchmarking framework for AI in natural product drug discovery that organizes evaluation around four interdependent dimensions: data integrity, model robustness, mechanistic validity, and translational utility. The framework begins with natural product identity, chemical structure, stereochemistry, provenance, bioactivity, assay, target, and safety-data audits. It then requires task-aligned split strategies, transparent baseline comparisons, external and out-of-distribution testing, perturbation analysis, uncertainty estimation, bias assessment, and reproducibility documentation. Mechanistic validity is treated as an evidence-integration problem involving compound–target support, biological context, pathway plausibility, omics consistency, assay relevance, exposure considerations, toxicity, and experimental perturbation. Translational utility is evaluated through experimental actionability, safety awareness, developability, source authenticity, external generalizability, implementation feasibility, and expert review. The proposed architecture establishes decision boundaries that prevent benchmark success from being interpreted as proof of mechanism, therapeutic efficacy, clinical utility, safety, deployment readiness, or regulatory acceptance. Its principal contribution is a structured evaluation logic intended to strengthen model comparison, expose limitations, guide research prioritization, and improve the credibility and reproducibility of AI-assisted natural product discovery.

Key Words: *AI benchmarking, Natural product drug discovery, Data integrity, Model robustness, Mechanistic validity, Translational utility*

eIJPPR 2026; 16(2):1-18

HOW TO CITE THIS ARTICLE: Wright E, Bennett C, Turner J. Benchmarking AI for Natural Product Drug Discovery: Data Integrity, Model Robustness, Mechanistic Validity, and Translational Utility. Int J Pharm Phytopharmacol Res. 2026;16(2):1-18. https://doi.org/10.51847/iPQZ02GzuS

INTRODUCTION

Artificial intelligence has become relevant across molecular representation, property prediction, target identification, screening, and development-oriented decision support. In natural product drug discovery, however, the credibility of an AI output depends not only on the selected algorithm but also on the quality of chemical structures, the traceability of compound origins, the biological meaning of labels, and the relationship between the evaluation task and the intended research use. Consequently, evaluation should encompass data quality, natural-product-specific informatics, validation discipline, and realistic evidentiary boundaries rather than predictive

Corresponding author: Ethan Wright
Address:Department of Artificial Intelligence for Anti-obesity Phytochemicals, Faculty of Pharmacy, University of Leeds, Leeds, United Kingdom.
E-mail: ✉ ethan.wright@leeds.ac.uk
Received: 04 December 2025; **Revised:** 10 March 2026; **Accepted:** 11 March 2026

This is an **open access** journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

performance alone [1–3]. This distinction is essential because an apparently accurate model may still be unsuitable for chemically novel natural products, biologically heterogeneous assays, mechanistic interpretation, or experimental prioritization.

Benchmark design provides the formal structure through which models, molecular representations, datasets, split strategies, and metrics can be compared. A benchmark is a defined evaluation arrangement comprising one or more tasks, curated datasets, specified partitions, evaluation measures, and reference baselines, whereas a benchmarking framework additionally establishes the principles by which those elements are selected, audited, interpreted, and connected to decision-making. Standardized molecular machine-learning resources have demonstrated the value of consistent tasks and evaluation procedures for comparative analysis [4]. Nevertheless, benchmark comparability does not ensure that the data are chemically reliable, that test cases are independent of training data, or that performance transfers to an intended natural product discovery setting.

Natural product AI presents an especially demanding benchmarking context. Natural compounds may contain complex ring systems, multiple stereocentres, unusual functional groups, glycosylation patterns, mixtures, salts, uncertain tautomeric forms, or incomplete source information. The same compound may appear under different identifiers, structural representations, organism names, extraction contexts, or activity annotations. Bioactivity records may combine binding, inhibition, phenotypic, cytotoxicity, or organism-level endpoints that are not biologically interchangeable. These characteristics make dataset integrity and chemical-space definition central components of benchmark validity rather than preliminary technical details.

This article therefore proposes an original conceptual benchmarking framework for AI in natural product drug discovery. The framework integrates dataset integrity auditing, chemical-space characterization, leakage control, task-aligned split selection, baseline comparison, robustness testing, external validation, mechanistic evidence appraisal, translational utility assessment, uncertainty and bias analysis, reproducibility documentation, and multidisciplinary expert review. Its purpose is not to establish a universal benchmark standard or to claim empirical validation. Instead, it provides a structured architecture through which benchmark developers, model investigators, pharmacologists, natural product researchers, and experimental teams may compare AI systems more cautiously, expose failure conditions, and distinguish research-prioritization value from therapeutic or clinical claims.

Benchmarking Challenges in Natural Product AI

Natural product datasets are commonly assembled from heterogeneous databases, publications, screening collections, curated repositories, and secondary aggregations. These resources can differ substantially in chemical coverage, identifier practices, structural standardization, biological annotation, provenance detail, and classification logic. Database fragmentation, source aggregation, and learned structural taxonomies therefore create complementary benchmarking challenges involving traceability, comparability, and documentation [5–7]. A benchmark that combines such records without retaining their source history may reproduce circular imports, merge chemically distinct stereoisomers, treat inferred classifications as verified labels, or allow the same underlying evidence to occur repeatedly across training and test partitions.

Data curation can reduce these risks, but it must be treated as a versioned and auditable component of benchmark construction. Recent natural product database curation has highlighted the importance of detecting structural inconsistencies, duplicate records, uncertain classifications, problematic source relationships, and inherited errors [8]. These concerns are directly relevant to benchmarking because apparently minor record-level defects can alter molecular descriptors, similarity relationships, scaffold assignments, class distributions, and test-set independence. Provenance should therefore distinguish primary experimental reports from secondary database transfers, while every structural correction, exclusion, or reconciliation should remain traceable.

Biological labels introduce a second layer of difficulty. Natural product bioactivity may be reported using different units, qualifiers, assay technologies, species, cell lines, target constructs, exposure periods, or endpoint definitions. Negative findings are often less available than positive findings, and inactive records may represent untested compounds, failed assays, insufficient exposure, or true biological inactivity. Pooling these categories can generate label noise and class imbalance while concealing assay-specific uncertainty. Small datasets intensify the problem because individual series, sources, targets, or laboratories may dominate the apparent signal, leaving models vulnerable to spurious correlations and unstable estimates.

Benchmark overfitting can arise when models are repeatedly optimized against familiar datasets, partitions, or public evaluation conventions. Random splits may leave closely related analogues on both sides of the evaluation boundary, and external validation may be absent, too similar to the development data, or indirectly used during model refinement. Reproducibility may also be limited by unavailable preprocessing code, undocumented

ISSN (Online) 2249-6084 (Print) 2250-1029

exclusions, variable molecular standardization, unreported random seeds, or incomplete model configurations. These weaknesses make high benchmark performance difficult to interpret translationally: it may reflect interpolation within a narrow data regime rather than reliable prioritization of new natural products, targets, biological contexts, or experimental hypotheses.

Data integrity and model robustness

Data integrity determines whether benchmark inputs represent coherent chemical and biological entities. Molecular records should be checked for valence, aromaticity, charge, salts, mixtures, tautomers, stereochemistry, and identifier consistency, while natural product provenance should link compounds to organisms, extracts, publications, and database sources where available. Duplicate detection should extend beyond identical text strings to stereochemical variants, salt forms, repeated assay records, close analogues, and shared scaffolds. This is particularly important because benchmark redundancy can reward memorization of dataset-specific chemical patterns rather than generalization to meaningfully new compounds [9]. Data-completeness profiles, label-noise assessments, class distributions, assay metadata, and target mappings should therefore be examined before any model comparison is interpreted.

Bioactivity benchmarking requires careful separation of chemical signal from annotation and assay artifacts. Large-scale comparative work on drug–target prediction illustrates the importance of controlled model evaluation across heterogeneous and imbalanced bioactivity tasks [10]. Molecular representation and partition strategy also materially influence apparent property-prediction performance, meaning that conclusions about algorithms cannot be separated from decisions about input encoding, dataset size, and test-set construction [11]. Benchmark developers should consequently document unit conversions, qualifier handling, replicate aggregation, endpoint definitions, negative-label construction, weighting procedures, and the treatment of missing or censored observations. When assay contexts are not equivalent, stratified evaluation or explicit exclusion may be more defensible than indiscriminate pooling.

Model robustness concerns the stability of conclusions under plausible analytical and chemical changes. Rigorous evaluation should combine activity-cliff-sensitive stress tests, repeated assessment of interactions among datasets, representations, models, and split strategies, and reproducible reporting of analytical decisions and computational artifacts [12–14]. Relevant tests may include repeated random seeds, resampling, alternative molecular representations, controlled label perturbation, removal of dominant series, scaffold-based partitioning, chemically dissimilar test sets, target-held-out evaluation, source-separated testing, and out-of-distribution assessment. Uncertainty estimates should be examined for calibration and error awareness rather than accepted as self-validating confidence measures. Similarly, robustness under one perturbation does not demonstrate general biological validity; each test characterizes only a defined failure mode.

Within the proposed framework, data integrity and robustness form sequential validity gates. A benchmark should first establish chemically valid records, traceable provenance, interpretable labels, sufficiently described assays, duplicate control, and transparent missing-data rules. It should then characterize chemical-space coverage and applicability limits before selecting a split strategy that corresponds to the intended generalization claim. Internal comparisons should include transparent baselines under equivalent tuning conditions, while external datasets should remain independent of model development. Robustness, uncertainty, and reproducibility findings should be reported as distributions, limitations, or use restrictions rather than converted into unsupported claims of mechanistic correctness or translational readiness.

Table 1 summarises data integrity and model robustness requirements for benchmarking AI in natural product drug discovery.

Table 1. Data Integrity and Model Robustness Requirements for Benchmarking AI in Natural Product Drug Discovery: Chemical Structure Quality, Stereochemistry, Compound Provenance, Bioactivity Annotation, Assay Metadata, Data Leakage, Chemical-Space Coverage, Split Strategy, External Validation, Robustness Testing, and Uncertainty

Benchmarking component	Core evaluation question	Required evidence or test	Potential output	Failure risk	Validation requirement	Reproducibility need	Framework role
Chemical structure quality	Are molecular records chemically valid and consistently standardized?	Valence, aromaticity, charge, salt, tautomer, mixture, and identifier checks	Validated, corrected, quarantined, or excluded record	Invalid structures distort descriptors, similarity, and scaffold assignment	Independent cheminformatics review of consequential corrections	Versioned standardization rules, software, and exception logs	Dataset-entry gate

Stereochemistry annotation	Are stereocentres and geometric isomers represented unambiguously where relevant?	Isomeric representation checks, undefined-centre flags, and source verification	Complete, partial, conflicting, or absent stereochemical annotation	Distinct natural products are merged or treated as equivalent	Manual verification for high-priority or stereochemistry-sensitive compounds	Preservation of original and standardized stereochemical records	Natural product identity control
Compound source provenance	Can each compound be traced to an organism, extract, publication, or contributing database?	Provenance-chain, identifier, taxonomy, and circular-import audit	Verified, indirect, conflicting, or unknown provenance	Misidentified, synthetic, or repeatedly imported records are treated as independent natural products	Primary-source confirmation when origin affects interpretation	Persistent identifiers and auditable source histories	Provenance gate
Bioactivity annotation	Are activity labels, units, qualifiers, and endpoint meanings compatible?	Unit harmonization, qualifier parsing, replicate assessment, censoring review, and endpoint stratification	Harmonized label with an evidence grade	Incompatible measurements become misleading training labels	Assay-aware curation and sensitivity analysis	Retention of raw values, transformed values, and processing scripts	Label-quality gate
Assay metadata	Is the biological context sufficient to interpret the reported endpoint?	Assay type, organism, cell line, target construct, protocol, controls, time, and readout review	Complete, partially complete, or unusable assay record	Contextually different assays are pooled as equivalent	Assay-expert review and stratified evaluation where necessary	Machine-readable metadata and versioned assay mappings	Biological-context gate
Target annotation	Is the target identity current, specific, and biologically appropriate?	Identifier mapping, species, isoform, complex, family, and ontology checks	Verified target, target family, or unresolved annotation	Incorrect target mapping produces invalid labels and mechanistic interpretations	Expert review for mechanism-sensitive tasks	Versioned ontology and target-mapping tables	Target-context gate
Data completeness	Are the fields required for the intended benchmark task sufficiently populated?	Field-level missingness and missingness-pattern analysis	Completeness profile and exclusion or imputation plan	Selective missingness creates hidden bias	Complete-case and sensitivity analyses where appropriate	Publication of missing-data rules and field definitions	Dataset-fitness gate
Data consistency	Do structures, identifiers, labels, targets, assays, and sources agree across records?	Cross-field and cross-source consistency checks	Conflict report and documented resolution status	Contradictory records contaminate training and testing	Independent reconciliation of decision-critical conflicts	Retention of conflict history and resolution logic	Integrity gate
Duplicate detection	Are exact, stereochemical, salt, assay, scaffold, and near-duplicate	Canonical structure, stereochemistry-aware identifier, scaffold,	Duplicate clusters and deduplicated dataset	Replicates or close analogues cross split boundaries and	Evaluation before and after deduplication	Publication of clustering rules and retained-record logic	Leakage-prevention control

	records identified?	similarity, and record-linkage tests		inflate performance			
Label noise	How stable are labels across replicates, assays, laboratories, and sources?	Replicate discordance, inter-assay variability, outlier review, and evidence grading	Noise estimate or evidence-weighted label	Models learn curation artifacts or inconsistent assay behavior	Noise-aware sensitivity analysis and manual audit sampling	Traceable label provenance and correction history	Reliability control
Class imbalance	Are minority outcomes represented and evaluated adequately?	Distribution analysis by split, target, scaffold, source, and assay	Imbalance profile and adjusted evaluation plan	Aggregate accuracy conceals minority-class failure	Appropriate metrics, stratification, and sensitivity analysis	Documentation of resampling, weighting, and threshold procedures	Metric-selection control
Data leakage	Has test information entered training, preprocessing, feature construction, or model selection?	Duplicate, scaffold, target, temporal, source, normalization, and preprocessing audits	Leakage-free status or severity report	Inflated performance and false generalization	Pipeline reconstruction and reevaluation after material leakage	Training-fold fitting of preprocessing with complete logs	Mandatory validity gate
Chemical-space coverage	Does the dataset represent the natural product chemistry addressed by the intended use?	Descriptor, scaffold, stereochemical, class, and diversity mapping	Coverage map and underrepresented-region profile	Findings are extrapolated beyond the evidence base	Comparison with independent intended-use chemical space	Publication of descriptors and mapping procedures	Scope-definition layer
Applicability domain	Can supported and extrapolative predictions be distinguished?	Distance, density, leverage, ensemble, calibration, or related domain analyses	In-domain, borderline, or outside-domain designation	Unsupported predictions appear equally credible	Prospective testing of domain rules on independent data	Predefined domain method and thresholds	Use-restriction gate
Training–test split logic	Does the partition answer the claimed generalization question?	Comparison of random, scaffold, target, source, temporal, or hybrid splits	Task-aligned primary partition and sensitivity partitions	An easy split measures interpolation rather than intended novelty	Split rationale fixed before final model selection	Release of identifiers and allocation rules for each partition	Benchmark-design core
Scaffold split where appropriate	Are related chemotypes prevented from dominating both development and evaluation sets?	Task-appropriate scaffold grouping and cluster sensitivity analysis	Scaffold-held-out performance profile	Analogue-series memorization is mistaken for chemical generalization	Comparison with random and stricter similarity-based partitions	Publication of scaffold definitions and tie-handling rules	Chemical-novelty test
External validation set	Does the model transfer to independently sourced and curated data?	Source-separated data, frozen model, harmonized endpoint, and	External performance, calibration, and failure profile	Internal validation is mistaken for external utility	No tuning on the final external dataset	Preserved model version and external-set provenance	Generalizability gate

		independent curation					
Out-of-distribution evaluation	How does the model behave on unfamiliar chemistry, targets, assays, sources, or time periods?	Chemical-, target-, assay-, source-, and temporal-shift tests	Distribution-shift performance and abstention profile	Silent failure occurs under novelty or context change	Multiple shift tests relevant to intended use	Released selection rules and OOD definitions	Stress-testing layer
Robustness testing	Are findings stable across seeds, resampling, representations, and reasonable perturbations?	Repeated runs, bootstrap or resampling, hyperparameter sensitivity, series removal, and perturbation analysis	Result distribution and instability flags	A favourable single run is presented as reliable performance	Predetermined repetitions and robustness analyses	Recorded environments, configurations, seeds, and failed runs	Stability gate
Uncertainty estimation	Does reported uncertainty correspond to observed error, calibration, and novelty?	Calibration analysis, proper scoring rules, interval coverage, error-retention, and OOD testing	Calibrated uncertainty or abstention condition	Confident errors influence candidate prioritization	Calibration separate from final testing and external verification	Published uncertainty method, calibration data, and use thresholds	Decision-support gate
Reproducibility	Can an independent group rerun the complete evaluation?	Data versions, preprocessing code, model configuration, environment, seeds, and analysis plan	Reproduced, partially reproduced, or nonreproduced benchmark result	Undocumented choices prevent verification	Independent rerun where legally and technically possible	Archived code, data documentation, environments, and change logs	Cross-cutting credibility gate

Mechanistic validity

Mechanistic validity concerns whether an AI output is consistent with a biologically credible and experimentally testable explanation rather than whether the model merely reproduces statistical associations. For target-prediction systems, benchmark design should reflect the intended use, the novelty of evaluated compounds, the construction of negative examples, and the biological specificity of target labels [15]. Target-context plausibility should consider whether the predicted protein, isoform, complex, or target family is expressed and functionally relevant in the disease, tissue, organism, or experimental system under consideration. A favourable target-prediction result should therefore be interpreted as evidence supporting a target hypothesis, not as confirmation that a natural product engages that target under biologically relevant conditions.

Compound–target relationships require evidence beyond similarity, feature attribution, or model confidence. Computational predictions can be strengthened through orthogonal biochemical, biophysical, cellular, or target-engagement experiments designed to discriminate direct activity from correlated or nonspecific effects [16]. Model explanations may assist error analysis by identifying structural regions, descriptors, or learned features associated with a prediction, but such explanations do not independently establish causal molecular interactions or pharmacological mechanisms [17]. Explanation quality should instead be evaluated through stability, faithfulness, chemical plausibility, consistency with known structure–activity relationships, and expert review.

Mechanistic benchmarks should also test whether models transfer across target and biological contexts. Evaluation involving unseen targets or target-conditioned prediction can help determine whether a model captures transferable drug–target relationships rather than memorizing target-specific ligand patterns [18]. At a broader level, mechanism-of-action assessment may integrate chemical structure, target evidence, pathway information, phenotypic responses, transcriptomic signatures, proteomic changes, metabolomic profiles, and perturbational data [19]. These evidence types should not be collapsed into an unqualified composite mechanism score; instead, each should retain its provenance, biological context, uncertainty, and relationship to alternative hypotheses.

For natural products, multi-target activity and structurally diverse chemistry may make mechanistic evaluation

particularly dependent on evidence integration. Models trained on target similarity or existing drug–target knowledge may help prioritize natural compounds for experimental study, but their predictions remain constrained by the completeness and bias of known target relationships [20]. Mechanistic validity should therefore include compound–target evidence, pathway-context plausibility, network sensitivity analysis, omics consistency, assay relevance, achievable exposure where supported, ADMET compatibility, toxicity counter-screening, cross-model comparison, and experimental perturbation. Even when these components align, benchmark success should be described as strengthening a mechanistic hypothesis rather than proving mechanism.

Translational utility

Translational utility concerns whether benchmark outputs can support a meaningful and appropriately bounded drug discovery decision. Chemical and biological datasets contain biases, missing observations, assay dependencies, historical selection effects, and inconsistent evidentiary standards that can restrict the transfer of AI findings into experimental or development settings [21]. Consequently, translational benchmarking should ask whether a model identifies feasible follow-up experiments, improves prioritization relative to transparent baselines, recognizes uncertainty and failure conditions, and retains value when evaluated on independently sourced compounds, assays, targets, or time periods.

Implementation within drug development requires more than a technically successful prediction. AI outputs should be connected to validated scientific workflows, explicit decision points, available experimental capabilities, and defined responsibilities for reviewing and acting on results [22]. Experimental follow-up relevance may be assessed by determining whether a prediction yields a falsifiable hypothesis, an accessible compound or source material, a suitable assay, appropriate positive and negative controls, and a clear criterion for interpretation. Candidate prioritization value should similarly be evaluated against simpler baselines and practical constraints rather than assumed from model complexity.

Translational assessment also requires multidisciplinary integration. Experimental pharmacology, medicinal chemistry, natural product chemistry, toxicology, formulation science, bioinformatics, and clinical expertise may each reveal limitations not captured by a computational benchmark [23]. Safety-aware evaluation should include endpoint-specific ADMET and toxicity evidence, applicability limitations, potential off-target effects, exposure considerations, and uncertainty rather than relying on a single aggregate developability label [24]. Natural product source authenticity, taxonomic identification, voucher information, extraction conditions, chemical reproducibility, supply, sustainability, stability, solubility, purity, and formulation feasibility can further determine whether a computationally prioritized compound is suitable for experimental advancement.

Prospective testing provides a stronger translational assessment than repeated optimization on familiar retrospective data because it examines whether computational prioritization leads to experimentally relevant candidates under a frozen decision process [25]. Nevertheless, successful experimental follow-up would validate only the tested prediction and context; it would not establish general clinical usefulness, therapeutic efficacy, or regulatory readiness. The framework therefore includes explicit no-therapeutic-claim and no-clinical-utility boundaries. Translational benchmarking may support research prioritization and experimental resource allocation, but clinical utility requires separate fit-for-purpose clinical evidence, safety evaluation, and expert oversight.

Table 2 maps mechanistic validity and translational utility criteria required for meaningful AI benchmarking in natural product drug discovery.

Table 2. Mechanistic Validity and Translational Utility Criteria for AI Benchmarking in Natural Product Drug Discovery: Target Plausibility, Pathway Relevance, Bioassay Context, Omics Consistency, ADMET and Safety Relevance, Experimental Follow-Up, Candidate Prioritization, External Generalizability, and Claim Boundaries

Evaluation criterion	Core question	Evidence needed	Benchmarking function	Potential output	Risk of overinterpretation	Validation requirement	Claim boundary
Target-context plausibility	Is the predicted target biologically relevant in the intended system?	Expression, localization, disease association, species, isoform, complex, and functional context	Assess correspondence between prediction and biological setting	Plausible, uncertain, or context-incompatible target hypothesis	Database presence is treated as target validity	Independent biological review and context-specific evidence	Supports a target hypothesis; does not validate the target

Compound–target evidence	Is there credible evidence that the compound engages the proposed target?	Direct binding, biochemical activity, cellular engagement, competition, or orthogonal assays	Grade the strength and directness of interaction evidence	Direct, indirect, predicted-only, or conflicting evidence	Similarity or attribution is treated as binding proof	Orthogonal experimental confirmation	No target-engagement claim from prediction alone
Pathway-context plausibility	Could target modulation reasonably influence the proposed biological pathway?	Curated pathway evidence, directionality, cell context, redundancy, and feedback information	Evaluate pathway consistency and alternatives	Coherent, incomplete, or contradictory pathway hypothesis	Pathway membership is interpreted as causal control	Functional or perturbational validation	Enrichment does not establish pathway causality
Network pharmacology evidence	Are multi-target relationships supported rather than driven by network density?	Edge provenance, confidence levels, alternative databases, null models, and sensitivity analysis	Examine network stability and database dependence	Stable, uncertain, or database-sensitive subnetwork	Hub proteins generate false mechanistic centrality	Null-model testing and independent evidence review	Network topology does not prove pharmacological action
Omics consistency	Do molecular response profiles agree with the proposed mechanism?	Quality-controlled transcriptomic, proteomic, metabolomic, or phenotypic data matched for context, dose, and time	Test directional concordance, specificity, and replication	Consistent, mixed, or inconsistent response pattern	Correlation is interpreted as mechanism	Replication, controls, and perturbation experiments	Omics agreement supports but does not establish mechanism
Bioassay relevance	Does the assay measure a biologically meaningful and sufficiently specific effect?	Assay principle, controls, interference testing, cellular context, readout, and endpoint proximity	Grade assay relevance and specificity	Mechanistically proximal, supportive, nonspecific, or uninterpretable endpoint	Reporter interference or cytotoxicity is treated as target activity	Orthogonal assays and counter-screens	Nonspecific activity cannot validate mechanism
Exposure relevance	Could the compound reach an effective concentration in the relevant system?	Potency, free concentration, permeability, metabolism, stability, distribution, and tissue exposure	Compare predicted activity with plausible exposure	Exposure-supported, uncertain, or implausible hypothesis	Nominal in vitro concentration is treated as effective exposure	Pharmacokinetic or experimentally justified exposure analysis	No in vivo relevance claim without exposure support

		where available					
ADMET relevance	Do disposition properties permit the proposed biological action and follow-up?	Absorption, distribution, metabolism, excretion, transporter, interaction, and physicochemical evidence	Identify disposition-related limitations	Compatible, limiting, or unresolved ADMET profile	Predicted properties are treated as measured developability	Experimental confirmation for decision-critical endpoints	Supports prioritization only
Toxicity relevance	Could activity reflect general stress, reactivity, or nonspecific toxicity?	Cytotoxicity, selectivity, reactive-group, off-target, organ-toxicity, and interference evidence	Distinguish pharmacological activity from toxicity-confounded effects	Selective, toxicity-confounded, or unresolved effect	Toxic responses are described as therapeutic mechanisms	Counter-screens and safety-relevant assays	Toxicity-confounded evidence cannot establish mechanism
Mechanistic hypothesis consistency	Are compound, target, pathway, phenotype, exposure, and safety claims mutually coherent?	Structured evidence chain with provenance, contradictions, and alternative explanations	Audit cross-level consistency	Coherent hypothesis, evidence gap, contradiction, or competing hypothesis	A coherent narrative is treated as causal proof	Predefined falsification criteria and experimental testing	Coherence justifies testing, not confirmation
Experimental perturbation need	Does intervention on the proposed node alter the predicted response?	Genetic perturbation, pharmacological inhibition, rescue, competition, or dose-response evidence	Discriminate among causal hypotheses	Supported, refuted, or unresolved mechanism	Observational association is treated as intervention evidence	Controlled perturbation with appropriate comparators	Mechanistic claims require intervention evidence
Cross-model consistency	Do materially different models converge for independent reasons?	Different representations, algorithms, datasets, calibration methods, and error analyses	Identify robust agreement and shared dependencies	Independent convergence or correlated-model artifact	Agreement among similar models is treated as independent proof	Dependency-aware comparison	Consensus does not establish truth
Experimental follow-up relevance	Does the prediction identify a feasible and discriminating next experiment?	Testable hypothesis, compound availability, assay suitability, controls, and decision criteria	Assess experimental actionability	High-, conditional-, or low-priority follow-up	Predictions are ranked without a practical validation path	Prospective laboratory and expert review	Supports experiment selection only

Candidate prioritization value	Does the model improve prioritization relative to transparent alternatives?	Baseline comparisons, independent holdout data, uncertainty, novelty, and resource constraints	Evaluate incremental decision value	Evidence-qualified candidate tier	Complex rankings are assumed to be useful	Prospective or independent evaluation	Ranking does not establish efficacy
Safety relevance	Are foreseeable toxicity and off-target risks considered before advancement?	Toxicity evidence, alerts, selectivity, exposure, uncertainty, and risk context	Apply safety-aware prioritization	Proceed, investigate risk, or deprioritize	Potent but hazardous candidates are favoured	Orthogonal safety testing	No safety claim from computational output alone
Developability relevance	Can the candidate plausibly be isolated, synthesized, characterized, formulated, and supplied?	Source availability, complexity, stability, solubility, purity, scale, and formulation evidence	Identify practical development constraints	Feasible, conditional, or high-risk developability profile	Biological ranking ignores practical failure modes	Chemistry, sourcing, and formulation review	Feasibility does not prove therapeutic value
Natural product source relevance	Is the source authentic, traceable, reproducible, and responsibly obtainable?	Taxonomy, voucher, geography, extraction, abundance, batch, and provenance data	Assess source integrity and reproducibility	Verified, conditional, conflicting, or unsuitable source	Database origin is treated as authenticated material	Source verification and batch confirmation	Provenance records do not replace source authentication
Assay transferability	Does model-supported activity persist across protocols, laboratories, and biological systems?	Cross-assay, cross-site, and context-matched replication data	Evaluate endpoint transferability	Stable, context-dependent, or nontransferable result	A single assay is treated as universal evidence	Independent replication	Assay success does not establish organism-level activity
External generalizability	Does the model retain utility on independently sourced chemistry and biological contexts?	Frozen-model testing on source-, target-, assay-, temporal-, or chemistry-separated datasets	Evaluate transfer beyond development data	External utility profile and failure map	Internal validation is presented as broad applicability	Independent external validation	No broad-use claim from internal testing

Implementation feasibility	Can the benchmark and its outputs be used and maintained in the intended setting?	Compute, expertise, data rights, infrastructure, update requirements, and governance	Assess operational practicality	Feasible, conditional, or impractical implementation	Success depends on hidden resources or originating-team expertise	Pilot evaluation and maintenance review	Feasibility is not deployment approval
Reproducibility	Can another group reproduce the workflow and conclusions?	Data versions, code, environment, seeds, configurations, logs, and analysis plan	Enable independent rerunning and audit	Reproduced, partially reproduced, or not reproduced	Undocumented choices prevent verification	Independent reproduction where feasible	Reproducibility does not prove biological validity
Expert review	Do relevant experts judge the result coherent, actionable, and properly qualified?	Structured multidisciplinary assessment and documented rationale	Integrate evidence and identify hidden assumptions	Advance, revise, gather evidence, hold, or reject	Automation bias or undocumented opinion	Conflict disclosure, explicit criteria, and recorded dissent	Expert review does not replace empirical validation
No therapeutic-claim boundary	Are efficacy and safety claims restricted to the available evidence?	Evidence-level and manuscript-claim audit	Prevent inappropriate therapeutic interpretation	Permitted research-prioritization wording	Benchmark success is described as therapeutic validation	Editorial and expert review	No therapeutic efficacy or safety claim
No clinical-utility boundary	Are diagnosis, treatment, or patient-care claims prevented without clinical evidence?	Intended-use statement and clinical-evidence audit	Control clinical interpretation	Nonclinical designation or clinical-validation requirement	Research outputs are presented as clinical tools	Fit-for-purpose clinical validation	No clinical utility or patient-care claim

Proposed benchmarking framework

The proposed framework begins with an explicit benchmark-objective statement defining the prediction task, natural product discovery context, intended users, anticipated decision, evaluation unit, and prohibited claims. Uncertainty evaluation should be incorporated at this planning stage because confidence estimates require dedicated assessment of calibration, error awareness, and behaviour outside familiar chemical space [26]. Uncertainty may also help prioritize compounds for experimental follow-up when its relationship to predictive reliability and novelty has been evaluated rather than assumed [27]. The objective statement should determine the required data fields, biological granularity, comparison baselines, split strategy, external validation design, acceptable abstention conditions, and level of expert oversight.

The second stage is a dataset-integrity audit covering molecular identity, chemical structures, stereochemistry, compound provenance, bioactivity labels, assay metadata, target and pathway annotations, ADMET and toxicity evidence, missingness, duplicates, conflicts, label noise, and class imbalance. A benchmark datasheet should describe why the dataset was assembled, how records were

collected and transformed, what populations and chemical regions are represented, which uses are intended or inappropriate, and how versions will be maintained [28]. Records that fail critical identity, provenance, or endpoint-interpretability checks should be corrected, segregated, or excluded before model development. The resulting benchmark should retain raw and standardized representations, source histories, correction logs, and machine-readable metadata.

The third stage characterizes chemical space and establishes the evaluation design. This includes scaffold distributions, structural classes, stereochemical diversity, target coverage, assay coverage, source distribution, applicability domains, and potential distribution shifts. Data leakage controls should examine exact and near duplicates, related scaffolds, shared target families, repeated assay records, preprocessing leakage, temporal overlap, and source overlap. Split strategies should be aligned with the intended claim, using scaffold, target, source, temporal, or hybrid partitions where appropriate. Baselines and candidate models should then be evaluated under equivalent data, preprocessing, tuning, and reporting conditions, followed by repeated seeds, resampling, controlled perturbations, activity-cliff analysis, out-of-distribution testing, and external validation. Repeated evaluation can expose instability that would remain concealed by a single partition or run [29].

The fourth stage integrates mechanistic validity, translational utility, uncertainty, bias, reproducibility, and expert review. Validation intensity should increase with the consequence and intended use of the prediction, while benchmark success should remain distinct from clinical validation or regulatory acceptance [30]. Models intended only for exploratory prioritization may require transparent uncertainty, external testing, and expert review, whereas outputs proposed for consequential development decisions would require progressively stronger experimental, safety, operational, and context-specific evidence. A reproducibility package should include data versions, preprocessing logic, split assignments, source code where appropriate, software environments, random seeds, configurations, calibration procedures, model documentation, error analyses, and a record of failed or excluded runs.

Table 3 presents the proposed benchmarking framework for AI in natural product drug discovery.

Table 3. Proposed Benchmarking Framework for AI in Natural Product Drug Discovery: Benchmark Objective, Dataset Integrity Audit, Chemical-Space Characterization, Label Quality, Data Leakage Control, Split Strategy, Robustness Testing, External Validation, Mechanistic Validity, Translational Utility, Reproducibility, Expert Review, Feedback, and Decision Boundaries

Framework stage	Core purpose	Required input	Benchmarking function	Potential output	Validation need	Failure risk	Feedback mechanism	Decision boundary
Benchmark objective definition	Specify the exact evaluation task and intended decision	Prediction task, natural product context, endpoint, users, comparison, and intended use	Convert broad aims into testable benchmark questions	Benchmark specification and claim boundaries	Multidisciplinary review before model development	Post hoc objectives favour desired findings	Narrow or revise the objective when evidence is insufficient	No claim beyond the predefined task and context
Dataset integrity audit	Determine whether records are fit for benchmarking	Structures, identifiers, stereochemistry, provenance, labels, assays, targets, pathways, ADMET, and toxicity data	Conduct validity, completeness, consistency, duplicate, and conflict checks	Integrity report, correction log, and exclusion record	Independent review of consequential corrections	Hidden corruption invalidates later comparisons	Return unresolved records to curation	Benchmark stops when critical identity or label defects remain

Chemical-space characterization	Define the chemistry and biological contexts represented	Standardized structures, descriptors, scaffolds, classes, targets, assays, and intended-use compounds	Map coverage, novelty, diversity, and underrepresented regions	Chemical-space coverage and gap profile	Comparison with independent intended-use space	Narrow coverage is generalized beyond evidence	Add representative data or restrict claims	Outside-domain predictions require qualification or abstention
Label and assay quality assessment	Establish what each endpoint measures and how reliably	Raw activity data, units, qualifiers, replicates, protocols, controls, and metadata	Harmonize labels and stratify assay evidence	Evidence-graded labels and assay strata	Sensitivity analysis and assay-expert review	Incompatible assays generate misleading labels	Recurate, relabel, stratify, or exclude records	No pooled evaluation without defensible endpoint equivalence
Data leakage control	Prevent test information from influencing training or model selection	Structures, scaffolds, targets, dates, sources, duplicates, and pipeline code	Audit overlap, preprocessing, feature fitting, normalization, and repeated evidence	Leakage report and corrected evaluation pipeline	Complete reanalysis after material leakage	Inflated estimates are interpreted as generalization	Rebuild splits, features, and preprocessing	Material leakage invalidates the affected result
Split strategy selection	Match the partition to the intended form of generalization	Objective, dataset structure, novelty requirements, and deployment scenario	Select random, scaffold, target, source, temporal, or hybrid partitions	Primary split and predefined sensitivity splits	Freeze partition logic before final optimization	Convenient splits answer an easier question	Escalate to stricter partitions when intended use changes	Claims remain limited to the tested split regime
Baseline model comparison	Determine whether complexity provides incremental value	Transparent classical and simple neural baselines	Compare models under equivalent data, tuning, splits, and metrics	Relative performance and complexity assessment	Repeated evaluation across tasks and seeds	Weak baselines exaggerate sophistication	Add stronger comparators or revise claims	No superiority claim without fair comparison
Robustness testing	Examine stability under plausible analytical and chemical changes	Frozen pipeline, perturbation plan, alternative representations, seeds, and resampling design	Conduct repeated runs, activity-cliff tests, label perturbation, series removal, and sensitivity analyses	Stability distribution and failure profile	Predetermined repetitions and analysis procedures	A favourable run is described as reliable performance	Revise data, model, or use restrictions	Unstable results cannot pass the benchmark gate

Out-of-distribution evaluation	Characterize behaviour under chemical, target, assay, source, or temporal novelty	Defined shift sets and applicability-domain procedure	Test performance, uncertainty, calibration, and abstention under distribution change	OOD failure map and use restrictions	Multiple shift types relevant to intended use	Silent extrapolation produces confident errors	Recalibrate, retrain, abstain, or narrow intended use	Unsupported OOD predictions are not actionable
External validation	Test transfer to independently sourced evidence	Frozen model and independently curated external data	Evaluate discrimination, calibration, coverage, and errors externally	External generalization profile	No tuning on the final external set	External data become part of iterative development	Replace contaminated validation data with an untouched set	Internal success cannot establish external utility
Mechanistic validity assessment	Determine whether predictions support a biologically credible hypothesis	Compound–target, target-context, pathway, network, omics, assay, exposure, ADMET, and toxicity evidence	Apply evidence-strength, contradiction, and perturbation checks	Mechanistic plausibility dossier	Orthogonal experiments and expert pharmacology review	Explanations or correlations are presented as mechanism	Revise hypotheses and design falsification experiments	No mechanism-proof claim
Translational utility assessment	Determine whether outputs support a practical discovery decision	Candidate, safety, developability, source, assay, external, and implementation evidence	Assess experimental actionability and incremental decision value	Research-prioritization recommendation with limitations	Prospective testing and workflow evaluation	Numerical outputs lack experimental or operational meaning	Incorporate experimental and user feedback	No therapeutic or clinical-utility claim
Uncertainty assessment	Determine whether confidence reflects observed reliability	Predictions, calibration set, novelty measures, and error data	Assess calibration, coverage, error retention, and abstention	Calibrated uncertainty and decision restrictions	External verification of calibration and thresholds	Confidence scores conceal unsupported extrapolation	Recalibrate, abstain, or revise the model	High or unvalidated uncertainty prevents consequential use
Bias assessment	Identify systematic performance differences across chemistry and biological contexts	Source, scaffold, class, target, assay, provenance, and subgroup metadata	Compare coverage and error patterns across relevant strata	Bias profile and affected-use restrictions	Independent analysis of consequential disparities	Aggregate performance masks subgroup failure	Rebalance data or restrict intended use	Unassessed material bias prevents broad claims

ISSN (Online) 2249-6084 (Print) 2250-1029

Reproducibility package	Enable exact rerunning and independent scrutiny	Data versions, code, environments, configurations, seeds, logs, and analysis plan	Reproduce preprocessing, training, evaluation, and reporting	Executable or fully specified research package	Independent rerun where feasible	Selective reporting and environment drift prevent verification	Archive versions and correct discrepancies	Nonreproducible central findings cannot support advancement
Expert review	Integrate chemical, pharmacological, toxicological, natural product, and operational expertise	Complete benchmark dossier and unresolved issues	Conduct structured multidisciplinary adjudication	Advance, revise, collect evidence, hold, or reject	Documented criteria, conflicts, and dissent	Automation bias overrides missing evidence	Return deficiencies to the relevant framework stage	Expert review cannot waive critical validation
Feedback updating	Incorporate new evidence without contaminating historical tests	Experimental results, new data, corrected labels, errors, and expert feedback	Update datasets and models under version control and rebenchmark	New benchmark version and change log	Locked historical tests or newly independent evaluation sets	Repeated reuse causes adaptive benchmark overfitting	Rotate external sets and update documentation	Updated systems require renewed validation
No clinical-claim boundary	Prevent technical results from being interpreted as clinical readiness	Intended-use statement, evidence level, and claim taxonomy	Audit outputs, manuscripts, and decisions for prohibited claims	Research-only designation or escalation requirement	Separate clinical validation when clinical use is proposed	Benchmark success is converted into treatment or deployment claims	Block, revise, or reclassify unsupported outputs	No efficacy, safety, clinical utility, deployment, or regulatory-readiness claim

Figure 1 illustrates the proposed benchmarking framework for AI in natural product drug discovery, connecting data integrity, model robustness, mechanistic validity, translational utility, reproducibility, expert review, and feedback updating.

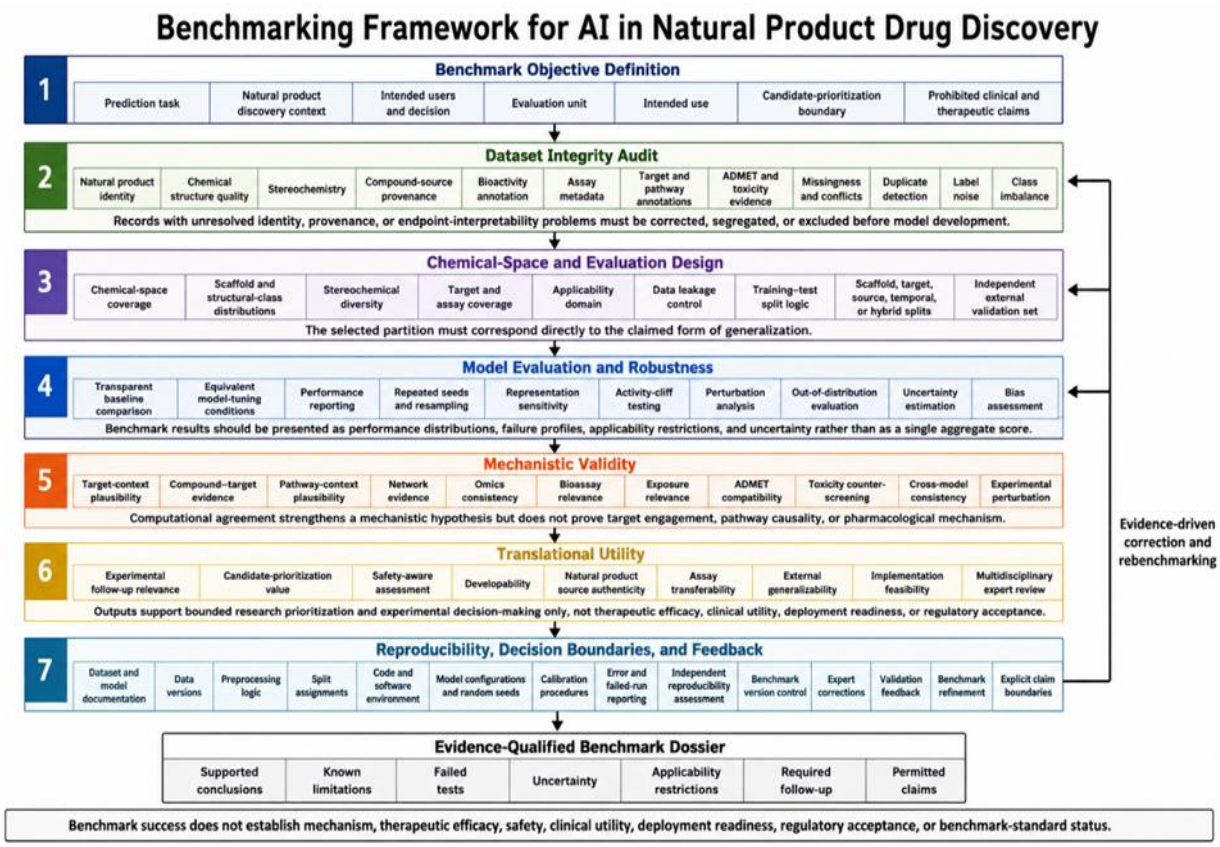

Figure 1. Benchmarking Framework for AI in Natural Product Drug Discovery

Alt-text description

A conceptual benchmarking framework showing AI evaluation for natural product drug discovery through dataset integrity, chemical-space assessment, leakage control, split strategy, robustness testing, external validation, mechanistic validity, translational utility, reproducibility, expert review, and feedback loops.

The framework is intended to operate as an iterative evidence-governance architecture rather than a linear scoring exercise. Failure at any stage should trigger a documented response: structural errors return to curation; narrow chemical coverage leads to claim restriction or data expansion; leakage requires reconstruction of the evaluation pipeline; poor robustness prompts model or dataset revision; uncertain target or pathway evidence leads to mechanistic experiments; weak experimental feasibility leads to reprioritization; and reproducibility failures require artifact correction before conclusions are retained. Final outputs should therefore consist of an evidence-qualified benchmark dossier containing supported conclusions, known limitations, uncertainty, failed tests, applicability restrictions, required follow-up, and explicit decision boundaries.

CONCLUSION

This article proposes a multidimensional benchmarking framework for AI in natural product drug discovery that integrates data integrity, model robustness, mechanistic validity, translational utility, uncertainty, bias assessment, reproducibility, expert review, feedback updating, and explicit validation boundaries. The framework shifts evaluation away from isolated predictive metrics toward a staged assessment of whether the underlying chemical and biological data are trustworthy, whether model behaviour is stable and generalizable, whether mechanistic interpretations are biologically plausible and experimentally testable, and whether outputs can support feasible research decisions. Its role is to improve comparison, transparency, limitation detection, and experimental prioritization. It does not convert technical benchmark performance into proof of target engagement, mechanism, therapeutic efficacy, safety, clinical utility, deployment readiness, regulatory acceptance, or benchmark-standard status.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

[1] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5

[2] Chen Y, Kirchmair J. Cheminformatics in natural product-based drug discovery. Mol Inform. 2020;39(12). doi:10.1002/minf.202000171

[3] Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009

[4] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A

[5] Sorokina M, Steinbeck C. Review on natural products databases: Where to find data in 2020. J Cheminform. 2020;12(1):20. doi:10.1186/s13321-020-00424-9

[6] Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: COlleCtion of Open Natural prodUcTs database. J Cheminform. 2021;13(1):2. doi:10.1186/s13321-020-00478-9

[7] Kim HW, Wang M, Leber CA, Nothias LF, Reher R, Kang KB, et al. NPClassifier: A deep neural network-based structural classification tool for natural products. J Nat Prod. 2021;84(11):2795-807. doi:10.1021/acs.jnatprod.1c00399

[8] Nainala VC, Rajan K, Kanakam SRS, Sharma N, Weißenborn V, Schaub J, et al. COCONUT 2.0: A comprehensive overhaul and curation of the collection of open natural products database. Nucleic Acids Res. 2025;53(D1). doi:10.1093/nar/gkae1063

[9] Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403

[10] Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441-51. doi:10.1039/C8SC00148K

[11] Yang K, Swanson K, Jin W, Coley CW, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88. doi:10.1021/acs.jcim.9b00237

[12] van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model. 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073

[13] Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F. A systematic study of key elements underlying molecular property prediction. Nat Commun. 2023;14(1):6395. doi:10.1038/s41467-023-41948-6

[14] Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18(10):1132-5. doi:10.1038/s41592-021-01256-7

[15] Mathai N, Chen Y, Kirchmair J. Validation strategies for target prediction methods. Brief Bioinform. 2020;21(3):791-802. doi:10.1093/bib/bbz026

[16] Cichonska A, Ravikumar B, Parri E, Timonen S, Pahikkala T, Airola A, et al. Computational-experimental approach to drug-target interaction mapping: A case study on kinase inhibitors. PLoS Comput Biol. 2017;13(8). doi:10.1371/journal.pcbi.1005678

[17] Wellawatte GP, Gandhi HA, Seshadri A, White AD. A perspective on explanations of molecular prediction models. J Chem Theory Comput. 2023;19(8):2149-60. doi:10.1021/acs.jctc.2c01235

[18] Svensson E, Hoedt PJ, Hochreiter S, Klambauer G. HyperPCM: Robust task-conditioned modeling of drug-target interactions. J Chem Inf Model. 2024;64(7):2539-53. doi:10.1021/acs.jcim.3c01417

[19] Trapotsi MA, Hosseini-Gerami L, Bender A. Computational analyses of mechanism of action (MoA): Data, methods and integration. RSC Chem Biol. 2022;3(2):170-200. doi:10.1039/D1CB00069A

[20] Periwal V, Bassler S, Andrejev S, Gabrielli N, Patil KR, Typas A, et al. Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs. PLoS Comput Biol. 2022;18(4). doi:10.1371/journal.pcbi.1010029

[21] Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037

[22] Mak KK, Pichika MR. Artificial intelligence in drug development: Present status and future prospects. Drug Discov Today. 2019;24(3):773-80. doi:10.1016/j.drudis.2018.11.014

[23] Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021;26(1):80-93. doi:10.1016/j.drudis.2020.10.010

[24] Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, et al. ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res. 2021;49(W1). doi:10.1093/nar/gkab255

[25] Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038-40. doi:10.1038/s41587-019-0224-x

[26] Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60(8):3770-80. doi:10.1021/acs.jcim.0c00502

[27] Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW. Evidential deep learning for guided molecular property prediction and discovery. ACS Cent Sci. 2021;7(8):1356-67. doi:10.1021/acscentsci.1c00546

[28] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86-92. doi:10.1145/3458723

[29] Barberis A, Aerts HJWL, Buffa FM. Robustness and reproducibility for AI learning in biomedical sciences: RENOIR. Sci Rep. 2024;14(1):1933. doi:10.1038/s41598-024-51381-4

[30] Niazi SK. A risk-tiered validation framework for artificial intelligence in drug discovery: From reproducibility to clinical translation. Int J Mol Sci. 2026;27(10):4349. doi:10.3390/ijms27104349