Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

RANDOM DATA SPLITS CREATE ARTIFICIAL CONFIDENCE IN VIRTUAL SCREENING, MOLECULAR PROPERTY PREDICTION, AND DRUG–TARGET MODELING

Anders Nilsson[1]*, Emma Lindholm[2], James Cooper[3], Sven Larsson[4]

1. ***Department of Robust Model Validation and Data Splitting, Faculty of Pharmacy, Swedish University of Agricultural Sciences, Uppsala, Sweden.***
2. ***Department of Artificial Confidence and Leakage Detection, Faculty of Pharmaceutical Sciences, University of Copenhagen, Copenhagen, Denmark.***
3. ***Department of Virtual Screening Reliability, Faculty of Pharmacy, University of Leeds, Leeds, United Kingdom.***
4. ***Department of Drug–Target Modeling and Evaluation, Faculty of Pharmacy, University of Helsinki, Helsinki, Finland.***

ARTICLE INFO

Received:
07 October 2024
Received in revised form:
03 December 2024
Accepted:
04 December 2024
Available online:
28 December 2024

Keywords: Data splitting, Scaffold leakage, Virtual screening, Molecular-property prediction, Drug–target interaction, Distribution shift

ABSTRACT

Machine-learning models are increasingly used to prioritize compounds, estimate molecular properties, and infer drug–target relationships, yet their reported performance often depends more strongly on evaluation design than headline metrics reveal. Random data splitting remains common because it is simple, statistically efficient, and compatible with standardized model comparison. In chemically structured pharmaceutical datasets, however, random allocation frequently places closely related molecules, shared scaffolds, homologous targets, replicated assays, or records from the same experimental source on both sides of the training–test boundary. The resulting test set may therefore measure interpolation within familiar evidence rather than performance on the novel compounds, targets, or experimental conditions implied by a prospective deployment claim. This Original Validation-Standards Article develops a conceptual account of how chemical similarity, scaffold leakage, and hidden dependence generate artificial confidence across virtual screening, molecular-property prediction, and drug–target modeling. It proposes Deployment-Aligned Validation Credibility as a structured construct linking the intended pharmaceutical use to a dependence audit, a shift-matched split strategy, task-appropriate metrics, uncertainty assessment, external corroboration, and transparent reporting. The analysis distinguishes random interpolation from scaffold transfer, chemical-cluster transfer, temporal prediction, entity-disjoint evaluation, source-independent testing, and prospective experimental confirmation. It also emphasizes that no single performance measure can establish pharmaceutical usefulness, mechanistic validity, experimental reproducibility, or therapeutic value. The proposed construct is methodological rather than empirically validated and remains conditional on dataset provenance, assay comparability, similarity definitions, and the practical decision being supported. Its purpose is to make validation claims narrower, more interpretable, and more consistent with the distribution shifts that molecular models will encounter in actual pharmaceutical research.

To Cite This Article: Nilsson A, Lindholm E, Cooper J, Larsson S. Random Data Splits Create Artificial Confidence in Virtual Screening, Molecular Property Prediction, and Drug–Target Modeling. Pharmacophore. 2024;15(6):97-107. https://doi.org/10.51847/0p8QlYktAq

Introduction

Machine learning now contributes to compound prioritization, molecular representation, target prediction, pharmacological modeling, and other stages of pharmaceutical discovery, but the scientific meaning of its performance remains conditional on how the evaluation problem has been constructed. The increasing integration of computational models into drug-design workflows has expanded the range of decisions potentially influenced by algorithmic outputs, while critical assessments have cautioned that retrospective predictive success should not be treated as equivalent to demonstrated pharmaceutical impact [1–3]. A model may discriminate compounds accurately within a curated benchmark and still fail when confronted with unfamiliar scaffolds, newly generated assay data, previously unseen targets, or measurement practices that differ from those represented

Corresponding Author: Anders Nilsson; Department of Robust Model Validation and Data Splitting, Faculty of Pharmacy, Swedish University of Agricultural Sciences, Uppsala, Sweden. E-mail: anders.nilsson@slu.se.

during training. The central issue is therefore not whether predictive modeling has value, but whether the evidence used to support a performance claim adequately represents the novelty, dependence structure, and decision conditions of the intended use.

Benchmark initiatives have improved the comparability of molecular-learning methods by organizing datasets, prediction tasks, split procedures, and performance measures within common evaluation environments [4]. This standardization has been valuable for algorithm development, error diagnosis, and methodological communication. Nevertheless, a benchmark does not passively reveal an intrinsic property of a model. Its apparent difficulty is partly created by decisions about which records are included, how molecular or biological entities are partitioned, which endpoints are combined, how missing or conflicting measurements are handled, and which metric is designated as primary. A score should consequently be understood as the output of a model–data–split–metric system rather than as a context-free measure of predictive ability. When the partitioning scheme preserves strong relationships between training and test observations, a benchmark may reward recognition of familiar patterns while being interpreted as evidence of broader generalization.

Large-scale comparisons on bioactivity collections demonstrate the practical value of evaluating multiple algorithms under shared conditions, but they also illustrate why method ranking and deployment validity are distinct questions [5]. A common benchmark can indicate which model performs better within the benchmark's data-generating and partitioning assumptions. It cannot, by itself, establish that the same ranking will persist for new chemical series, later project compounds, independent laboratories, or targets lacking close representatives in the training data. This distinction is especially important in pharmaceutical datasets because observations are rarely independent in the statistical sense implied by random sampling. Medicinal-chemistry campaigns deliberately generate analogue series; bioactivity databases aggregate repeated measurements and related assays; proteins occur in homologous families; and molecular endpoints are often concentrated within particular chemical domains. These structures create scientifically meaningful dependence, but they also make conventional random holdouts deceptively easy.

This article addresses the resulting validation gap by developing the proposed construct of Deployment-Aligned Validation Credibility. The construct defines credibility as the degree to which an evaluation supports its stated pharmaceutical deployment claim after accounting for dependence between observations, the distribution shift represented by the split, the suitability of the selected performance measures, uncertainty under that shift, and corroboration on genuinely separate evidence. It does not replace established validation methods with a universal score or prescribe one split for every task. Instead, it organizes the relationship between the intended prediction problem and the evidence required to test it. The article argues that random splitting may be appropriate when the intended use is interpolation among observations drawn from a familiar distribution, but it becomes misleading when its results are extended to novel-scaffold discovery, future portfolio prediction, external assay transfer, or cold-start interaction modeling. The proposed contribution is conceptual and requires future empirical evaluation before it can be treated as an operational standard.

Why Random Splitting Remains Attractive and Misleading

Random splitting remains attractive because it is easy to implement, usually preserves class balance, distributes molecular diversity across partitions, and produces test sets large enough for stable model comparison. Standardized retrospective benchmarks can therefore make algorithmic differences appear clear and reproducible, even though chemical similarity, benchmark memorization, and limitations in the suitability of chemical and biological data may restrict the meaning of those differences [6–8]. Random allocation is also consistent with the conventional assumption that observations are exchangeable samples from one underlying distribution. In many general statistical applications, that assumption is reasonable. In pharmaceutical datasets, however, the observations have often been generated through sequential, clustered, and hypothesis-driven processes. Compounds from the same optimization series may share a core structure and differ only by local substitutions; assays may be repeated under related protocols; and targets may belong to closely connected biological families. Randomization distributes these relationships rather than removing them.

The resulting evaluation can be technically correct while scientifically misaligned. Descriptor-based and graph-based models, for example, may be compared fairly under the same random partition, yet the observed ranking remains conditional on the chemical and endpoint structure that the partition preserves [9]. A model that exploits local similarity effectively may perform particularly well when each test compound has a close training neighbour. That capability can be useful for interpolating within an established lead series, but it should not be described as general competence across chemical space. The misleading element is therefore not randomization itself. It is the expansion of a narrow in-distribution result into a claim about unseen scaffolds, later project data, independent assay systems, or unfamiliar drug–target combinations. Random splitting often answers the question, "Can the model predict withheld members of the distribution from which it learned?" It does not necessarily answer, "Can the model support the next pharmaceutical decision for which it is being proposed?"

Scaffold-aware or chronological evaluation can more closely approximate certain prospective prediction problems because it withholds structural families or later observations that were not available during model development [10]. These designs typically make the prediction task more difficult, but difficulty alone is not their purpose. Their value lies in representing a form of novelty that is closer to the intended use. A scaffold split may be relevant when the objective is to prioritize a new chemotype, whereas a temporal split may be more informative when a model will be trained on historical project data and applied to compounds synthesized later. Neither design is universally superior. A scaffold split can still allow chemically similar frameworks to cross partitions, and a temporal split may combine chemical novelty with changes in assay protocols,

project priorities, or data curation. Validation credibility therefore depends on whether the split represents the deployment problem, not on whether it produces the lowest performance estimate.

The proposed DAVC construct treats random-split performance as evidence of interpolation unless stronger separation has been demonstrated. This interpretation avoids two opposite errors. The first is to present high random-split performance as proof of broad pharmaceutical usefulness, particularly when close chemical or biological neighbours remain available during training. Benchmark memorization can create apparently strong classification without demonstrating transfer to genuinely distinct compounds [7]. The second error is to dismiss all random-split evaluation as scientifically worthless. Chemical and biological datasets vary in quality, purpose, and dependence structure, and an in-distribution estimate may be appropriate when the intended decision concerns additional observations from a stable and familiar domain [8]. What is required is an explicit statement of the claim being tested, the dependencies retained across partitions, and the claims that remain unsupported. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop why random splitting remains attractive and misleading within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Why random splitting remains attractive and misleading

Construct or evidence domain	Core question	Scientific basis	Role in the article	Failure or overclaiming risk	Boundary statement
Benchmark convenience	Why is random splitting routinely retained?	Random allocation is simple, computationally inexpensive, compatible with stratification, and convenient for common model comparisons.	Explains the persistence of random holdouts without treating their use as careless or irrational.	Convenience is interpreted as scientific realism.	Operational simplicity does not establish correspondence with future pharmaceutical data.
Exchangeability assumption	Are observations plausibly independent samples from one distribution?	Conventional random splitting assumes that train and test records are exchangeable apart from sampling variation.	Identifies the statistical assumption underlying the design.	Clustered medicinal-chemistry and assay data are treated as independent observations.	Exchangeability must be examined rather than presumed in chemically structured datasets.
In-distribution interpolation	What capability does a random split most directly examine?	Related compounds, targets, assays, and sources are commonly distributed across all partitions.	Defines the narrow claim that random evaluation can legitimately support.	Interpolation is reported as novel-scaffold, new-target, or external generalization.	Random-split results primarily support performance within the represented distribution.
Chemical-neighbour availability	How close are test compounds to training compounds?	Molecular datasets frequently contain analogue series and densely sampled local chemical regions.	Connects high apparent performance to chemically familiar test observations.	A model's retrieval of nearby analogues is interpreted as learning transferable pharmacological principles.	Neighbour-aware analysis is needed before broad generalization is claimed.
Representation and model interaction	Does the split favour specific representations or architectures?	Descriptor, fingerprint, sequence, and graph models exploit different forms of local and global similarity.	Prevents performance from being attributed to architecture independently of the evaluation design.	One model is declared universally superior on the basis of a favourable partition.	Model ranking is conditional on the dataset, representation, split, and metric.
Metric compression	Does one aggregate score conceal relevant failure modes?	Global discrimination or error measures average across easy and difficult chemical regions.	Supports the article's rejection of single-metric validation.	High average performance hides failures for novel scaffolds, rare targets, or local discontinuities.	Aggregate scores require subgroup, distance-sensitive, and uncertainty-aware interpretation.
Data suitability	Are the available labels and records appropriate for the intended claim?	Bioactivity databases combine heterogeneous assays, sources, endpoints, and curation practices.	Distinguishes data availability from evidence suitability.	Large data volume is treated as proof of representative or unbiased evidence.	A large dataset can remain unsuitable for a particular deployment question.
Claim–split alignment	Does the evaluation represent the novelty expected during use?	Different applications involve familiar analogues, unseen scaffolds, later compounds, new targets, or external measurements.	Provides the transition from random benchmarking to deployment-aligned validation.	A narrow benchmark estimate is generalized to every future setting.	Split strategy must be justified by the intended pharmaceutical use.

Chemical Similarity, Scaffold Leakage, and Hidden Dependence

Chemical similarity is not merely a nuisance variable in molecular modeling. It is one of the principal ways in which medicinal knowledge is organized and transferred. Closely related compounds often share physicochemical characteristics, synthetic routes, binding modes, and assay behaviour, making similarity an appropriate basis for analogue reasoning. The same structure, however, creates dependence between observations and can make a nominally held-out test set less independent than it appears. Analyses of structure-based screening datasets, attempts to construct less biased virtual-screening benchmarks, and similarity-aware partitioning methods collectively show that apparent train–test separation can fail when actives, inactives, molecules, or biological entities remain relationally coupled across partitions [11–13]. The validation problem is therefore not solved by

confirming that each row appears only once. Independence must be examined at the level of the scientific entities and relationships through which predictive information can travel.

Chemical similarity leakage occurs when test compounds have training neighbours close enough that prediction becomes largely an exercise in local interpolation or analogue recognition. Scaffold leakage is a more specific form in which molecules assigned to different partitions retain the same or closely related structural cores. Both concepts depend on representation. A fingerprint may describe two molecules as distant while a scaffold definition places them in the same series, or a generic-scaffold transformation may reveal a relationship that a more literal framework definition misses. Hidden dependence extends beyond molecular structure. Protein homology, repeated assay conditions, shared source publications, replicated measurements, batch effects, preprocessing fitted to the full dataset, and target-derived features can all connect training and test observations. Molecular-property performance consequently reflects interactions among data composition, representation, pretraining, model architecture, and evaluation design rather than an isolated model capability [14].

Scaffold separation should not be treated as an automatic certificate of chemical novelty. A scaffold algorithm may place closely related frameworks in different groups because ring systems, linker definitions, stereochemistry, or peripheral atoms are handled differently. Conversely, a chemically diverse scaffold holdout may be unnecessarily severe when the intended application is optimization within a known series. Bias analyses in virtual screening show that models can exploit structural regularities that distinguish benchmark actives from benchmark inactives without learning the intended protein–ligand relationship [11]. The construction of assay-grounded benchmark sets can reduce some artificial distinctions between active compounds and generated or property-matched decoys, but retrospective curation cannot remove every dependence or guarantee correspondence with a future screening library [12]. Similarity-aware partitioning methods provide a more explicit way to minimize cross-partition relationships, yet their outputs remain conditional on the molecular representation, similarity function, optimization objective, and metadata supplied to the partitioning procedure [13].

DAVC therefore proposes a dependence topology as the first substantive object of validation. A dependence topology is a map of the relationships that could allow information to cross the nominal training–test boundary, including molecular neighbours, shared scaffolds, target families, protein structural similarity, assay identity, source documents, laboratories, batches, timestamps, and preprocessing dependencies. The map is not intended to demonstrate causality or to prove that every identified relationship will inflate performance. Its function is to make the evaluation assumptions visible and to identify which forms of novelty have and have not been tested. Under this view, leakage is not a binary property but a claim-dependent condition. A molecular neighbour retained across partitions may be acceptable for a lead-series interpolation claim and unacceptable for a new-chemotype claim. Similarly, exposure to a related target may be relevant training information for family-level prediction but incompatible with a strict new-target evaluation. The scientific requirement is not maximal separation in every study; it is separation that is sufficient, transparent, and defensible for the stated deployment problem.

This dependence-centred interpretation also clarifies why no single similarity threshold can serve as a universal validation standard. The relevant distance depends on the endpoint, the molecular representation, the density of the chemical domain, the amount of available data, and the type of decision being supported. Structural similarity may correspond strongly to one physicochemical property but weakly to another, while activity cliffs demonstrate that locally similar molecules can still exhibit sharply different biological behaviour. A credible evaluation should therefore report both the nominal split rule and the resulting separation profile: scaffold overlap, nearest-neighbour relationships, target similarity, assay and source overlap, and any other dependence relevant to the task. These disclosures establish the foundation for examining how hidden dependence produces different consequences in virtual screening, molecular-property prediction, and drug–target modeling. They do not prove that the resulting model is experimentally useful, mechanistically valid, clinically meaningful, or ready for operational deployment.

Consequences for Virtual Screening, Property Prediction, and Interaction Modeling

The consequences of split-induced artificial confidence differ across pharmaceutical modeling tasks because each task places a different meaning on chemical and biological novelty. In structure-based virtual screening and binding-affinity prediction, a model may appear to learn transferable protein–ligand recognition while relying partly on repeated ligand chemotypes, related protein structures, shared binding sites, or dataset-specific characteristics. Deep neural models trained on protein–ligand complexes can show substantial retrospective performance yet remain frustrated by predictions beyond familiar structural relationships [15]. This problem affects both ranking and affinity estimation. A model that succeeds on complexes closely related to its training evidence may still fail to identify active compounds in a chemically distinct library, rank ligands for an unfamiliar target structure, or distinguish genuine intermolecular determinants from benchmark-specific correlations. Internal benchmark performance therefore supports only the degree of structural and chemical transfer actually represented by the partition.

Drug–target affinity and interaction models introduce a bipartite dependence structure because predictions are functions of both compound and target representations [16]. A pair-level random split can withhold a particular drug–target combination while allowing the same drug and the same target to appear independently in training with other interaction partners. Such an evaluation examines matrix completion among familiar entities rather than generalization to new pharmaceutical objects. This distinction matters because the apparent prediction problem may be described simply as drug–target modeling even though its deployment forms are scientifically different. Predicting an unmeasured interaction between a known drug and known target, inferring targets for a new compound, identifying ligands for a new target, and predicting an interaction when both entities are

unseen require progressively different evidence. A single performance measure averaged across these conditions obscures which form of transfer the model has demonstrated.

Cold-start evaluation makes this entity distinction explicit. New-drug, new-target, and double-cold settings are not interchangeable, and methods intended to transfer interaction knowledge must be tested under the specific form of entity novelty they claim to address [17]. Even an entity-disjoint partition may preserve hidden dependence if test compounds have close training analogues or if test targets are homologous to training proteins. The drug–target field therefore requires at least two concurrent separation audits: chemical independence for compounds and biological independence for targets. Pair independence alone is insufficient. When the deployment problem involves a new target family, sequence identity, domain composition, binding-site similarity, pathway relationships, and available structural information may all be relevant. When it involves a new compound, scaffold, analogue-series, and nearest-neighbour separation become central. The permissible claim must identify which entities were new and how their novelty was operationalized.

Molecular-property prediction is similarly vulnerable to global averages that hide chemically important failures. Activity cliffs demonstrate that structurally similar compounds can exhibit sharply different biological activities, exposing the limitations of models that assume a smooth relationship between molecular similarity and response [18]. A random split may contain many easy test compounds with close and label-consistent training neighbours, allowing favourable average error or discrimination even when the model fails on discontinuous regions most relevant to medicinal-chemistry decisions. For this reason, DAVC proposes that aggregate performance be accompanied by novelty-stratified analysis, nearest-neighbour distance, scaffold-level results, activity-cliff stress testing where applicable, and uncertainty assessment under the deployment-relevant shift. These assessments do not transform association into mechanism or prediction into experimental confirmation. They clarify where the model appears to interpolate reliably, where it extrapolates uncertainly, and where the evaluation provides no defensible evidence.

Table 2 organizes the evidence, constructs, and interpretation boundaries needed to develop consequences for virtual screening, property prediction, and interaction modeling within the article's central argument.

Table 2. Components, relationships, uncertainties, and validation needs within Consequences for virtual screening, property prediction, and interaction modeling

Component or layer	Inputs	Core function	Expected output	Uncertainty or limitation	Validation requirement
Ligand-based virtual screening	Molecular structures, fingerprints or learned representations, activity labels, assay provenance	Rank or classify candidate compounds for a specified target or assay	Prioritized compounds or activity probabilities	Close analogues may cross partitions; actives and inactives may contain benchmark-specific structural cues	Scaffold- or cluster-aware testing, similarity profiles, assay-level provenance, and external-library evaluation where claimed
Structure-based virtual screening	Protein structures, ligand structures or poses, docking configurations, interaction descriptors	Estimate compatibility or rank protein–ligand complexes	Scores, ranks, or predicted binding-related outcomes	Repeated protein families, ligand series, crystallographic biases, and pose-generation artifacts may inflate performance	Protein- and ligand-aware partitions, source-disjoint structures, pose-control analysis, and experimental follow-up
Molecular-property prediction	Molecular structures, physicochemical descriptors, assay-specific endpoints	Estimate continuous or categorical molecular properties	Property estimate, class probability, or uncertainty interval	Endpoint heterogeneity, limited domain coverage, activity cliffs, and local chemical discontinuities	Scaffold, cluster, temporal, and external-assay evaluation with novelty-stratified error analysis
Drug–target pair completion	Known drugs, known targets, observed interaction matrix	Infer missing interactions among familiar entities	Scores for previously unmeasured familiar-entity pairs	Pair independence may coexist with complete entity familiarity	Pair holdout with explicit disclosure that the claim concerns familiar-entity completion
New-drug prediction	Unseen compound, known targets, training interactions for other compounds	Infer targets or affinities for a compound lacking training interactions	Target ranking or affinity prediction	Close chemical analogues may remain in training; target side may be highly familiar	Compound-disjoint and similarity-controlled evaluation
New-target prediction	Known compounds, unseen target, training interactions for other targets	Identify candidate ligands or affinities for a target lacking training interactions	Compound ranking or interaction scores	Homologous proteins or shared binding-site structures may remain in training	Target-disjoint evaluation with sequence, family, or structure separation
Double-cold interaction prediction	Unseen compound and unseen target	Transfer knowledge to a pair in which neither entity has training interactions	Interaction or affinity estimate	Highest combined chemical and biological uncertainty; sparse direct evidence	Simultaneous compound- and target-disjoint testing with graded similarity reporting
Local-discontinuity stress testing	Matched molecular pairs, activity-cliff definitions, scaffold families	Detect errors hidden by aggregate averages	Subgroup error, cliff classification, or local ranking performance	Definitions of cliffs and similarity are representation- and endpoint-dependent	Predefined subgroup analysis and sensitivity to alternative similarity definitions
Uncertainty and abstention layer	Predictive probabilities, ensembles, intervals, distance measures	Identify predictions requiring caution or human review	Calibrated uncertainty, selective prediction, or abstention recommendation	Confidence can remain high under distribution shift; in-distribution calibration may not transfer	Calibration and selective-risk analysis on the same deployment-relevant holdout

Pharmaceutical interpretation boundary	Split design, metric bundle, provenance, uncertainty, decision context	Restrict conclusions to the evidence actually generated	Bounded statement of supported and unsupported uses	Predictive success can be confused with mechanism, synthesis feasibility, safety, or therapeutic value	Explicit claim statement, expert review, and experimental confirmation for downstream pharmaceutical assertions

Figure 1 presents the validation mirage split-screen, showing how the article's key components and boundaries are connected within consequences for virtual screening, property prediction, and interaction modeling.

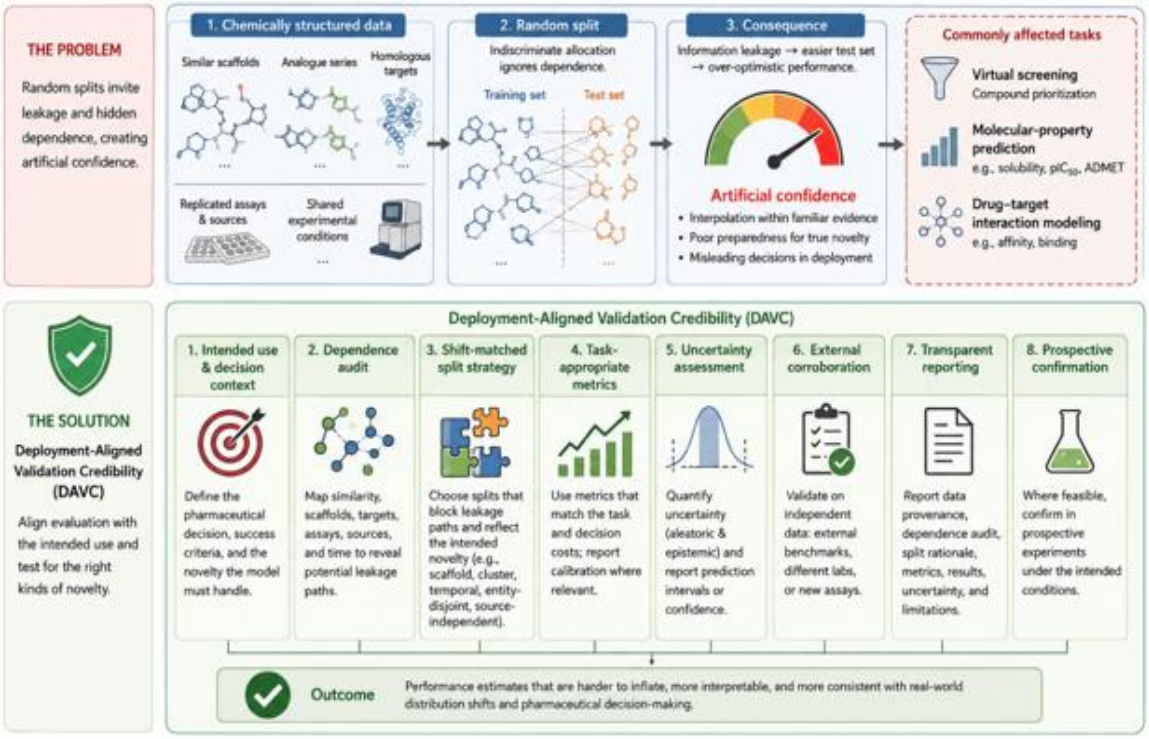

Figure 1. Validation Mirage Split-Screen.

The figure is an original conceptual synthesis that organizes the article's central contribution across random splitting remains attractive and misleading, chemical similarity, scaffold leakage, and hidden dependence, chemical similarity, scaffold leakage, hidden dependence, consequences for virtual screening, property prediction, and interaction modeling. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

A Hierarchy of Temporal, Scaffold, Cluster, and External Validation Designs

Validation designs can be organized according to the type of dependence they remove and the distribution shift they attempt to reproduce. Random splitting provides the lowest separation when records are independently assigned without chemical, biological, temporal, or source constraints. Scaffold splitting withholds designated molecular cores, while chemical clustering or similarity-threshold partitioning seeks to reduce broader structural proximity across partitions. Comparative work on chemical-data partitioning shows that scaffold and similarity-based approaches create different balances between structural separation, sample size, class distribution, and implementation cost [19]. These designs should not be interpreted as rungs in a universally superior sequence. Their relevance depends on whether the deployment question concerns familiar analogues, unfamiliar chemical families, future compounds, or evidence generated outside the original data source.

Cluster-based evaluation can be particularly useful when conventional scaffold definitions fail to capture the molecular relationships that matter to a model. Clustering in fingerprint or learned-representation space can separate structurally related regions while preserving a sufficiently large test set, and clustering-based holdouts have been proposed as more rigorous tests for some virtual-screening applications [20]. Nevertheless, cluster membership is not an objective property independent of methodological choice. It changes with molecular representation, distance measure, dimensionality reduction, clustering algorithm, and parameter selection. A model can therefore appear robust or fragile depending on how the chemical landscape has been partitioned. A credible cluster-based study should report these choices, characterize the resulting nearest-neighbour separation, and examine whether conclusions persist under reasonable alternative definitions.

Temporal validation introduces a different logic. Rather than defining novelty solely by structure, it trains on observations available before a cutoff and tests on later evidence. This can approximate how a pharmaceutical model is developed from historical records and applied to future compounds or assays. Temporal separation may also capture realistic changes in

chemical priorities, target programs, experimental practices, and organizational processes. Yet these changes make interpretation difficult: reduced performance may reflect chemical novelty, assay drift, altered curation, or a shift in the project portfolio. External or experimental corroboration provides stronger evidence that predictions transport beyond the original benchmark. Computational predictions that are subsequently investigated experimentally offer information unavailable from internal retrospective evaluation alone [21]. Such evidence remains conditional on experimental design, candidate selection, assay validity, and the range of compounds or targets tested.

Out-of-distribution evaluation should consequently be treated as a family of designs rather than as one standardized test. Molecular-property studies show that conclusions about robustness can vary according to whether unfamiliarity is defined through scaffolds, molecular similarity, chemical clusters, or another distributional criterion [22]. DAVC proposes a claim-indexed hierarchy: random or stratified splits for familiar-distribution interpolation; scaffold holdouts for designated core novelty; cluster or similarity-threshold holdouts for broader chemical separation; temporal splits for later evidence; entity-disjoint designs for new compounds or targets; source-disjoint testing for transport across databases, laboratories, or organizations; and prospective experiments for predictions fixed before new measurements. The hierarchy describes which deployment shift each design may interrogate. It does not imply that later designs automatically contain fewer biases, that one successful external dataset establishes universal generalization, or that prospective prediction establishes therapeutic or clinical value.

Matching Split Strategy to the Intended Deployment Problem

Split selection should begin with a precise statement of what will be new when the model is used. Drug–target validation guidelines emphasize that new-drug, new-target, and new-pair predictions require different partitioning logic and different controls for overlap [23]. The same principle applies across molecular pharmaceutical science. For a model used to rank unscreened analogues within an established lead series, retaining the known scaffold may be appropriate because local interpolation is the intended task. For a model advertised as discovering new chemotypes, that same design would be insufficient. A formulation of the deployment problem should identify the expected chemistry, target familiarity, assay system, source environment, time horizon, and decision supported. Without this statement, no split can be judged realistic or unrealistic because realism is relative to an intended use.

Virtual screening and lead optimization illustrate why one benchmark design cannot serve every pharmaceutical decision. Real-world compound-activity benchmarking distinguishes screening-oriented settings from optimization-oriented settings because they differ in assay organization, chemical-series structure, class prevalence, and the way predictions are used [24]. Virtual screening often prioritizes a small number of compounds from a large and chemically varied library, making early retrieval and novel-chemotype performance important. Lead optimization operates within a narrower chemical domain and may place greater emphasis on local ranking, continuous property improvement, and consistency among related analogues. A severe cluster or scaffold holdout can therefore be appropriate for one purpose and unnecessarily pessimistic for another. The correct question is not which split is hardest, but which split creates the evidence needed for the stated decision.

Metric selection must follow the same deployment logic. Decision-theoretic evaluation shows that a model's practical value depends on how predictions alter candidate selection, including the cost of testing, the value of successful compounds, and the consequences of false prioritization [25]. A global classification metric can be insensitive to the part of the ranking that determines which compounds are purchased or synthesized. A regression error may not indicate whether the model preserves the order of candidates near a medicinal-chemistry decision threshold. Calibration can matter when predictions are used to defer, abstain, or trigger expert review. DAVC therefore rejects a universal primary metric and instead requires a task-sensitive bundle that may include discrimination, early enrichment, precision–recall behaviour, ranking quality, continuous error, calibration, coverage, subgroup performance, and decision relevance. Metric diversity must remain disciplined: reporting many measures without a predeclared interpretation can create another form of selective confidence.

Validation must also be integrated into the complete pharmaceutical modeling workflow rather than added after model development. Good practice in pharmaceutical machine learning includes data preparation, endpoint definition, feature construction, model optimization, evaluation, interpretation, and reproducibility [26]. Leakage can enter at any of these stages. Feature normalization, representation learning, assay filtering, threshold selection, hyperparameter optimization, and model selection can all transmit information from test data when performed before partitions are frozen. DAVC therefore proposes a pre-model sequence: define the intended deployment problem; map the dependence topology; select a shift-matched split; freeze all test-facing preprocessing; predefine the metric bundle and error subgroups; and reserve external or prospective evidence where feasible. This sequence is proposed as a validation discipline, not as a validated certification process. Its effectiveness requires comparison with existing workflows and prospective testing across different pharmaceutical tasks.

Minimum Reporting and Benchmarking Standards

Minimum reporting should allow a reader to determine what was predicted, what was withheld, which relationships crossed the split, and which claims the evidence can support. DOME recommendations structure supervised biological machine-learning reporting around data, optimization, model, and evaluation [27]. Applied to molecular pharmaceutical modeling, this means disclosing dataset provenance, endpoint definitions, inclusion and exclusion decisions, duplicate handling, chemical standardization, target identity, assay relationships, split algorithms, random seeds, preprocessing boundaries, hyperparameter-selection data, metric definitions, and uncertainty procedures. A label such as "scaffold split" or "external test set" is not

sufficiently informative on its own. The resulting separation must be characterized through scaffold overlap, nearest-neighbour similarity, target relatedness, source independence, temporal ordering, and other task-relevant dependencies.

Reproducibility requires enough information and accessible material to reconstruct the analysis, including data versions, code, software environments, configuration settings, and the precise train–validation–test assignments [28]. In molecular research, reproduction may be limited by proprietary structures, confidential assay data, licensing restrictions, or security concerns. These constraints do not justify omission of all validation detail. Studies can report hashed identifiers, split-generation code, aggregate separation statistics, versioned preprocessing rules, or auditable procedures that preserve confidentiality. Reproducing a model on the same records establishes only one layer of reliability. It does not establish external validity, transport across laboratories, prospective usefulness, or mechanistic correctness. Reporting standards must therefore distinguish computational reproducibility from scientific generalization.

Transparent reporting must also protect against leakage introduced through iterative analysis. Consensus recommendations for machine-learning-based science emphasize explicit controls for data leakage, analytical flexibility, and incomplete reporting [29]. In molecular benchmarks, repeated inspection of test results can transform a nominal test set into an informal validation set, particularly when architectures, representations, endpoints, or split thresholds are repeatedly modified in response to performance. The number and role of evaluation sets should be stated, and final model selection should not depend on the evidence used to support the principal generalization claim. Benchmark comparisons should additionally report failed runs, unstable models, excluded datasets, and sensitivity to seeds or split realizations where these affect conclusions. Selective presentation of one favourable partition can create artificial confidence even when the nominal split strategy is appropriate.

Method comparison requires evidence that an observed advantage is stable, meaningful, and relevant to the pharmaceutical task. Protocols developed specifically for small-molecule machine learning emphasize repeated, practically interpretable comparisons rather than reliance on a single favourable result [30]. DAVC proposes that minimum benchmarking include a justified baseline set, identical data and preprocessing boundaries, repeated paired evaluation where feasible, uncertainty around comparative differences, novelty-stratified performance, and disclosure of the conditions under which model rankings change. The purpose is not to require every possible experiment. It is to prevent a small average advantage on an easy or unstable split from being presented as broad methodological superiority. **Table 3** organizes the evidence, constructs, and interpretation boundaries needed to develop minimum reporting and benchmarking standards within the article's central argument.

Table 3. Evaluation, implementation, and research priorities arising from Random Data Splits Create Artificial Confidence in Virtual Screening, Molecular Property Prediction

Evaluation dimension	What must be tested	Suitable evidence or assessment	Failure signal	Interpretive limitation	Decision relevance
Intended-use specification	Whether the evaluation represents the chemistry, biology, time, source, and decision expected at deployment	Predeclared context-of-use statement and claim–split concordance review	Broad deployment claim with no corresponding novelty definition	Context descriptions require scientific judgment	Determines which split and metrics are relevant
Chemical separation	Whether close analogues, scaffolds, or molecular clusters cross partitions	Scaffold overlap, nearest-neighbour similarity, cluster membership, analogue-series audit	Test compounds have near-identical training neighbours despite a novelty claim	Similarity depends on representation and threshold	Distinguishes interpolation from new-chemistry transfer
Target separation	Whether test targets are biologically independent of training targets	Sequence, family, domain, binding-site, or structural similarity assessment	"New target" belongs to a highly familiar target family	Biological relatedness cannot be reduced to one metric	Defines the strength of new-target and double-cold claims
Assay and source separation	Whether shared experiments, publications, laboratories, or databases connect partitions	Assay identifiers, document provenance, batch records, laboratory metadata	Replicated or closely related measurements appear across partitions	Provenance may be incomplete or inconsistent	Determines whether apparent transfer is independent of the original source
Temporal integrity	Whether all training information genuinely predates test evidence	Frozen cutoff, data-version history, chronological audit	Later labels or curation decisions influence training or preprocessing	Time combines multiple forms of drift	Supports future-portfolio rather than random interpolation claims
Preprocessing isolation	Whether normalization, feature selection, representation learning, and filtering were fit without test information	Pipeline audit and code review	Full-dataset preprocessing precedes splitting	Some unsupervised operations may still encode distribution information	Protects the independence of the final evaluation
Metric fitness	Whether reported measures match ranking, regression, classification, or selection decisions	Predefined metric bundle and decision rationale	One convenient global metric drives all conclusions	No metric alone captures pharmaceutical usefulness	Connects performance reporting to actual candidate decisions
Subgroup and distance stress testing	Whether failures concentrate in novel, sparse, or discontinuous regions	Scaffold-level errors, similarity bins, target groups, activity-cliff analysis	High average performance with collapse in deployment-critical subgroups	Subgroup definitions may be unstable in small datasets	Identifies where expert review or abstention is needed

Calibration and uncertainty	Whether confidence remains meaningful under the tested shift	Calibration, interval coverage, selective risk, abstention curves	Highly confident errors on unfamiliar chemistry or targets	In-distribution calibration may not transfer	Supports risk-aware prioritization rather than blind ranking
Stability of method comparison	Whether model rankings persist across seeds, partitions, and reasonable preprocessing choices	Repeated paired evaluations and uncertainty around differences	Superiority appears in only one favourable run or split	Repetition cannot rescue a conceptually misaligned split	Determines whether an improvement is credible enough to influence method choice
External corroboration	Whether performance transports beyond internal curation and source decisions	Independent database, laboratory, organization, or prospective test	External set retains hidden overlap or incompatible endpoints	Externality does not guarantee comparability or absence of bias	Strengthens—but does not complete—the evidence for prospective use
Reproducibility envelope	Whether another analyst can reconstruct data processing, splitting, training, and evaluation	Versioned data, code, environment, configurations, and split assignments	Undocumented procedures are required to recover the result	Reproducibility is not equivalent to validity	Enables scrutiny, correction, and cumulative comparison
Claim boundary	Whether conclusions remain within the evidence generated	Explicit supported, unsupported, and unresolved claim statement	Benchmark performance is equated with mechanism, experimental confirmation, safety, or therapeutic value	Boundary statements cannot replace future validation	Prevents computational evidence from being used beyond its scientific scope

Limitations and Boundary Conditions

The proposed DAVC construct is a conceptual validation standard rather than an empirically validated measurement instrument. Its components are derived from recurring problems in molecular benchmarking, leakage analysis, out-of-distribution evaluation, drug–target validation, and scientific reporting, but the construct has not been prospectively shown to improve compound selection, assay success, portfolio decisions, or translational outcomes. No universal threshold is proposed for chemical independence, target novelty, split severity, calibration, or external validity. Such thresholds are likely to depend on the endpoint, chemical domain, target family, assay technology, data density, and intended decision. DAVC should therefore be used to organize questions and constrain claims, not to assign an unvalidated numerical credibility score.

The evidence base also has important limitations. Many methodological findings are derived from public bioactivity and molecular-property collections whose curation, assay heterogeneity, and chemical coverage differ from proprietary pharmaceutical portfolios. External and temporal evaluations remain vulnerable to undocumented differences in measurement protocols and project selection. Scaffold, cluster, and similarity-based splits are representation-dependent, while target separation can vary according to sequence, structure, domain composition, or pharmacological function. Prospective experimental studies provide stronger temporal separation but may test small or selectively chosen candidate sets. Consequently, no validation design can fully reproduce every condition of future use, and a difficult split should not automatically be interpreted as realistic, fair, or scientifically superior.

Misuse can occur in both restrictive and permissive directions. A permissive interpretation would treat completion of the proposed reporting elements as proof of model quality, mechanistic validity, experimental reproducibility, clinical utility, or regulatory acceptability. DAVC cannot establish any of these outcomes. A restrictive interpretation would require maximal chemical, biological, temporal, and source separation for every task, even where the intended application is local interpolation within a stable domain. This could reject useful models for reasons unrelated to their context of use. The correct boundary is claim-specific: the validation design should remove the dependencies that would make the stated deployment claim circular, while retaining the information legitimately available in that deployment setting.

Research Agenda and Implementation Priorities

The first research priority is to test whether deployment-aligned evaluation predicts prospective performance more accurately than conventional random benchmarking. Such studies should compare random, scaffold, cluster, temporal, entity-disjoint, source-disjoint, and prospective designs using datasets with traceable provenance and clearly defined downstream decisions. Performance should be examined as a function of chemical and biological distance rather than only at one split threshold. For drug–target modeling, studies should separate familiar-pair completion, new-drug, new-target, and double-cold prediction. For molecular-property modeling, they should evaluate lead-series interpolation, scaffold transfer, later compounds, external assays, and activity-cliff subgroups. The objective should not be to identify one universally hardest split, but to determine which retrospective designs best anticipate each form of future use.

A second priority is infrastructure for dependence-aware data stewardship. Molecular records should be accompanied by machine-readable links among standardized structures, scaffold families, assay identifiers, target versions, source documents, experimental batches, laboratories, timestamps, and preprocessing histories. Benchmark repositories should distribute fixed partitions together with their separation profiles rather than only split labels. Reporting templates should capture the intended use, retained dependencies, excluded records, optimization boundaries, metric rationale, uncertainty procedure, and supported claim. Proprietary organizations may require privacy-preserving implementations, but confidential data should not prevent internal provenance auditing or transparent description of the evaluation logic.

Implementation should proceed in stages. An initial stage can require only an intended-use statement, a dependence audit, a justified split, and disclosure of separation statistics. A second stage can add repeated evaluation, novelty-stratified error

analysis, calibration under shift, and structured reporting of unsupported claims. A third stage can incorporate source-independent datasets, prospective predictions, and decision-based comparison with established selection practices. Across all stages, pharmaceutical scientists, medicinal chemists, assay specialists, structural biologists, pharmacologists, statisticians, and machine-learning researchers should jointly define what counts as meaningful novelty and error. Such staged use may help evaluate whether DAVC improves scientific communication and experimental prioritization, but implementation should remain explicitly provisional until comparative and prospective evidence is available.

Conclusion

Random data splitting can provide a valid estimate of interpolation within a familiar molecular and biological distribution, but it creates artificial confidence when that estimate is presented as evidence for novel-scaffold discovery, future-project performance, external assay transfer, or cold-start drug–target generalization. The proposed Deployment-Aligned Validation Credibility construct reframes validation as a relationship among the intended pharmaceutical use, hidden dependence, the distribution shift represented by the split, task-appropriate metrics, uncertainty under that shift, external corroboration, and transparent reporting. Its central contribution is not a new universal ranking of split strategies, but a requirement that each performance claim be bounded by the form of novelty actually tested. Chemical similarity, scaffold overlap, target relatedness, assay provenance, time, and source must therefore be treated as components of the validation problem rather than incidental dataset characteristics. The construct remains conceptual and cannot establish experimental success, mechanism, synthesizability, developability, safety, therapeutic value, clinical utility, regulatory acceptability, or deployment readiness. Its value will depend on future evidence showing that more explicit claim–split alignment produces better calibrated scientific conclusions and more reliable pharmaceutical decisions.

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. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3
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. 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
6. Lenselink EB, ten Dijke N, Bongers B, Papadatos G, van Vlijmen HWT, Kowalczyk W, et al. Beyond the hype: Deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. J Cheminform. 2017;9:45. doi:10.1186/s13321-017-0232-0
7. 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
8. 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
9. Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminform. 2021;13(1):12. doi:10.1186/s13321-020-00479-8
10. Yang K, Swanson K, Jin W, Coley C, 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
11. Sieg J, Flachsenberg F, Rarey M. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening. J Chem Inf Model. 2019;59(3):947-61. doi:10.1021/acs.jcim.8b00712
12. Tran-Nguyen VK, Jacquemard C, Rognan D. LIT-PCBA: An unbiased data set for machine learning and virtual screening. J Chem Inf Model. 2020;60(9):4263-73. doi:10.1021/acs.jcim.0c00155

13. Joeres R, Blumenthal DB, Kalinina OV. Data splitting to avoid information leakage with DataSAIL. Nat Commun. 2025;16(1):3337. doi:10.1038/s41467-025-58606-8
14. 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
15. Volkov M, Turk JA, Drizard N, Martin N, Hoffmann B, Gaston-Mathé Y, et al. On the frustration to predict binding affinities from protein-ligand structures with deep neural networks. J Med Chem. 2022;65(11):7946-58. doi:10.1021/acs.jmedchem.2c00487
16. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: Deep drug-target binding affinity prediction. Bioinformatics. 2018;34(17). doi:10.1093/bioinformatics/bty593
17. Nguyen TM, Nguyen T, Tran T. Mitigating cold-start problems in drug-target affinity prediction with interaction knowledge transferring. Brief Bioinform. 2022;23(4). doi:10.1093/bib/bbac269
18. 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
19. Simm J, Humbeck L, Zalewski A, Sturm N, Heyndrickx W, Moreau Y, et al. Splitting chemical structure data sets for federated privacy-preserving machine learning. J Cheminform. 2021;13(1):96. doi:10.1186/s13321-021-00576-2
20. Guo Q, Hernandez-Hernandez S, Ballester PJ. UMAP-based clustering split for rigorous evaluation of AI models for virtual screening on cancer cell lines. J Cheminform. 2025;17(1):94. doi:10.1186/s13321-025-01039-8
21. 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
22. Fooladi H, Vu TNL, Mathea M, Kirchmair J. Evaluating machine learning models for molecular property prediction: Performance and robustness on out-of-distribution data. J Chem Inf Model. 2025;65(19):9871-91. doi:10.1021/acs.jcim.5c00475
23. Tanoli Z, Schulman A, Aittokallio T. Validation guidelines for drug-target prediction methods. Expert Opin Drug Discov. 2025;20(1):31-45. doi:10.1080/17460441.2024.2430955
24. Tian T, Li S, Zhang Z, Chen L, Zou Z, Zhao D, et al. Benchmarking compound activity prediction for real-world drug discovery applications. Commun Chem. 2024;7(1):127. doi:10.1038/s42004-024-01204-4
25. Watson OP, Cortés-Ciriano I, Taylor AR, Watson JA. A decision-theoretic approach to the evaluation of machine learning algorithms in computational drug discovery. Bioinformatics. 2019;35(22):4656-63. doi:10.1093/bioinformatics/btz293
26. Talevi A, Morales JF, Hather G, Podichetty JT, Kim S, Bloomingdale PC, et al. Machine learning in drug discovery and development part 1: A primer. CPT Pharmacometrics Syst Pharmacol. 2020;9(3):129-42. doi:10.1002/psp4.12491
27. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, ELIXIR Machine Learning Focus Group, et al. DOME: Recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18(10):1122-7. doi:10.1038/s41592-021-01205-4
28. 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
29. Kapoor S, Cantrell EM, Peng K, Pham TH, Bail CA, Gundersen OE, et al. REFORMS: Consensus-based recommendations for machine-learning-based science. Sci Adv. 2024;10(18). doi:10.1126/sciadv.adk3452
30. Ash JR, Wognum C, Rodríguez-Pérez R, Aldeghi M, Cheng AC, Clevert DA, et al. Practically significant method comparison protocols for machine learning in small molecule drug discovery. J Chem Inf Model. 2025;65(18):9398-411. doi:10.1021/acs.jcim.5c01609