Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Artificial Intelligence in Computational Drug Discovery: Reproducibility, Validation, and Decision Stability

Hadeel T. Al-Sinjilawi
Department of Science, Faculty of Arts and Sciences, The World Islamic Science & Education University (WISE), Amman, Jordan
Corresponding Author: Hadeel T. Al-Sinjilawi, **E-mail**: hadeel.alsinjilawi@wise.edu.jo

| ABSTRACT

Abstract

Artificial intelligence (AI), machine learning, and molecular modeling have become integral components of modern computational drug-discovery workflows. But, as has been demonstrated in the industry, success in benchmark datasets does not always guarantee success in real screening projects, or in different experimental conditions. Given this limitation, the present review investigates the ability of AI-driven computational workflows to support a stable compound prioritization process in the face of uncertainty, but not just a higher retrospective score. Relevant literature published from 2018 to 2025 is reviewed, in addition to previous docking and virtual-screening studies, and community benchmarks, as well as reports of limitations under practical discovery conditions. The reviewed evidence indicates that unnoticed data leakage, validation approaches that do not mimic real-world utilization, sensitivity to docking protocols and preprocessing, stochastic variation across runs, and selection effects in large-library screening are some of the recurring factors that can lead to overconfidence. In real cases seemingly minor modifications, such as protonation-state assignments, definition of the docking-box, or random-seed selection, could make a significant difference in compound ranking that affects decisions in the lab. Validating the model in the deployment environment, performing replicate analyses, checking for rank agreement, validating the model through appropriate negative controls, and developing calibration methods are also important operational considerations for reliable prioritization. These findings highlight the growing importance of AI-assisted computational workflows that emphasize validation, reproducibility, and reliable decision support in modern drug-discovery research. Reproducibility is still important but the real utility of computational workflows is to help make decisions that are defensible in experimentation under realistic discovery conditions.

| KEYWORDS

Artificial Intelligence; Computational Drug Discovery; Machine Learning; Molecular Docking; Virtual Screening; Decision Support; Reproducibility; Validation

| ARTICLE INFORMATION

ACCEPTED: 01 June 2026 **PUBLISHED:** 13 July 2026 **DOI:** 10.32996/jcsts.2026.8.8.9

Introduction

Artificial intelligence (AI) and machine-learning technologies have transformed computational drug-discovery workflows by enabling large-scale virtual screening, molecular modeling, compound prioritization, and predictive decision support. However, from past samples of prospective screening campaigns, well-performing benchmarks have not always led to good decisions in actual discovery projects. The models are often highly accurate in their ability to predict the outcome when tested on similar chemical space, closely related chemical tasks, or well-known training sets. Frequently, however, the predictive performance of the same workflows drops when transferred to new biological targets, chemotypes that were not seen before, or when assay conditions are changed (Steshin, 2023). This has led to a greater focus on methods that better reflect real-life drug discovery practice. Hit ID and hit optimization, in particular, can be more useful than retrospective measures of success in determining which compounds to prioritize and move forward for experimental testing. A model that performs well in repeated screening

campaigns and is not specific to the initial screening environment might thus be more useful than one that has high benchmark performance but has not performed well outside of the environment it was originally tested in (Steshin, 2023). Meanwhile, the computational drug discovery community has come to understand that predictive accuracy is just part of a valuable modeling process. Robustness, reproducibility, transparency of reporting and documentation of model development and validation procedures have thus gained greater focus. Such initiatives are part of a wider movement from optimizing the performance of a metric to building trust in model-informed decision making (Kapoor et al., 2024).

Strong performance in benchmarking is not sufficient for AI-driven computational workflows to enable trustworthy decisions for drug discovery. In the end, workflows that seem to be working well in the past can yield very different results if used with new biological targets, new chemical space, or new assay conditions. Assessment methods that simulate the screening environment must document that a workflow can be used to make decisions in environments other than the controlled benchmark (Aittokallio, 2022). The importance of such evaluation is particularly obvious when conducting sequential screening campaigns, where subsequent screening compounds are dependent on the results of the previous ones. With Under practical screening constraints, active-learning methods can focus resources on the most informative candidates on the most informative candidates, reducing the wasteful screening of related chemotypes and making subsequent investigations more efficient (Reker & Schneider, 2015; Wang et al., 2024). The same trends are observed in recent studies of machine learning in drug discovery, where whole workflows, such as virtual screening, molecular docking, bioinformatics, and cheminformatics, are often emphasized over the evaluation of individual algorithms (Atasever, 2024).

There are two common limitations that many have attributed to the advances in computational drug development. First of all is the quality and availability of the data; both of these play an important role in the model performance. In practice, some of the best datasets are only accessible within certain industrial or institutional environments, and not to the broader research community. Unfortunately, access to such information is still restricted, which hinders independent examination and external validation. Consequently, models that are trained using highly curated internal datasets often do not generalize well to environments where they have not been trained. Reported performance can therefore be a measure of the quality of the data as well as the diversity and complexity of the drug-discovery programmes performed (Gadiya et al., 2023).

To address the growing requirement for machine-readable and interoperable information in R&D efforts for biopharmaceutical products, FAIR strategies have been recommended. They are particularly useful when data sets are reused across projects, when they are not always documented, or when they are not used in the same format, with parts of them missing, or due to corruption by a different curation process. It is therefore essential to have structured and traceable data management in place to guarantee continuity and accessibility when the procedures are transferred, extended or independently evaluated (Wise et al., 2019). A second major challenge is uncertainty in drug-discovery decision-making. The selection of compounds is often done as if the outputs of the ranking were final conclusions, while the resulting variation due to data quality, model behaviour, and experimental conditions is not taken into account. Not considering such uncertainty may lead to unwarranted confidence, particularly with few candidates proceeding to experimental testing (Mervin et al., 2021). What matters here is the usefulness of the computed models beyond simply being a better benchmark of the past, but rather for decision support. The importance of strong benchmark measures is reduced when there are few or infrequent interpretations for prioritized drugs in experimental follow-up. Thus, the quality of the process is best evaluated in terms of the ability of the process to deliver rankings that are believable in isolation, reproduced and validated by experimentation (Jiménez-Luna et al., 2021). This sentiment is echoed by the increasing attention being paid to uncertainty quantification in drug discovery decision making. If uncertainty is only added after the ranking process has been completed, it is unlikely that it will alter the selection of compounds since important choices have already been made. However, incorporating uncertainty at an early stage in the workflow also enables prioritization to account for performance as well as associated risk; this is crucial in the case of limited resources for experimental studies and where follow-up studies have measurable costs (Renz et al., 2019). This paper examines computational tools supporting decision making in early stage drug development which are repeatable, uncertainty management and perform in real screening scenarios. Some of the selected historical studies are reviewed as necessary, but the main focus of the literature is on those published from 2018 to 2025.

The focus is on the performance at the workflow level, including hit identification and lead optimisation, where a consistent compound prioritisation is often more useful than incremental gains in the predictive accuracy or computation time. Studies selected for the study are based on community norms, blind challenges and studies that reveal common pitfalls that lead to failure including data leakage, decoy bias and protocol design sensitivity. This type of evidence provides a more realistic measure of how the workflow is used in actual screening contexts than does studies that concentrate on gains only. Therefore, reports with higher benchmark scores which are not accompanied by measures of stability, repeatability, or decision making reliability are said to have a secondary value (Kapoor et al., 2024; Kapoor & Narayanan, 2023).

This narrative review was informed by common methodological issues that have been reported in computational drug-discovery studies. Blinded challenges, studies of external validation and reports focusing on workflow behavior in varying experimental conditions were preferred. The purpose was not to provide a comprehensive coverage of the literature, but to identify factors that affect the reproducibility and prioritization reliability of the discovery process in a practical setting.

AI-driven computational workflowss in Drug Discovery

AI-driven computational models are increasingly used as decision-support tools throughout modern drug-discovery workflows, where their practical value depends on the quality and reliability of the decisions they support rather than on isolated performance metrics alone. In the context of routine discovery processes, computational methods act as filters, progressively narrowing chemical space, providing a set of testable hypotheses and informing compound selection within realistic constraints (synthesis, computer resources, experimental capacity). In general, during mass screening campaigns, early prioritization procedures often determine whether the later stages remain viable and viable or the number of candidates to be pursued is too large. In such cases, the usefulness of every step of the workflow is quite apparent (Figure 1).

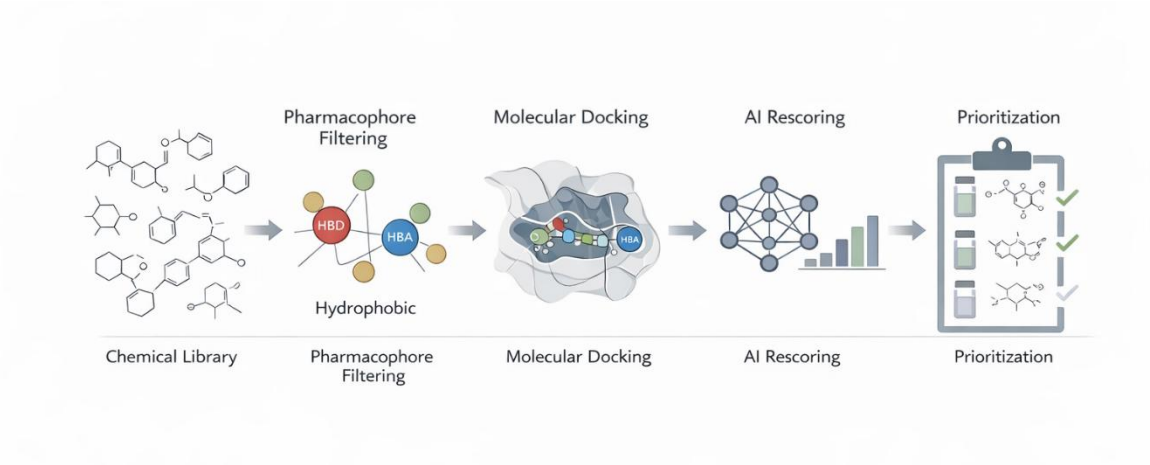

Figure 1. An AI-assisted computational drug-discovery pipeline in which pharmacophore screening, molecular docking, machine-learning-based rescoring, and prioritization strategies progressively reduce chemical space before experimental validation.

In this regard, the study will be primarily centered on computational techniques for chemical prioritization and decision support, including pharmacophore screening, molecular docking, scoring, uncertainty quantification and active-learning methodologies. Examples of areas such as reaction prediction and retrosynthetic planning cannot be covered in this discussion. References to synthesis are made only when emphasizing the change from virtual hits to candidates that can be tested in the lab and synthesized.

Pharmacophore models are often used as a first pass filter in the computational drug discovery process. A pharmacophore is the set of steric and electronic properties that are thought to be required for contact with a biological target and the resulting biological response (Kaserer et al., 2015). The aim of the large-scale virtual screening is to find compounds that can mimic the desired interaction pattern. Those molecules which don't satisfy the above requirements are discarded and those that do form a smaller subset of molecules that contain more likely active molecules than the original library (Kaserer et al., 2015). Common Pharmacophoric properties are hydrogen-bond donors, hydrogen-bond acceptors, hydrophobic groups, aromatic sites, ionizable groups and metal binding groups (Kaserer et al., 2015; Giordano et al., 2022). One of the common misconceptions is a pharmacophore match implies biological activity. The success of the match, in fact, is just that a compound has the essential features that allow it to pass a preliminary screening test. Further testing is needed such as molecular docking, ADME-tox property evaluation and experimental testing. When binding geometry, physicochemical properties or the assay-specific requirements are considered, many compounds that comply with the pharmacophore criteria are ultimately discarded. The refinement of the model is therefore an integral part of pharmacophore-based screening. Adjusted features, tolerances and exclusion criteria are often required based on screening results as the model often produces too many similar analogues and/or is overly specific, reducing chemical diversity. The characteristics of the resulting hit list are set as well according to the model's source. Ligand based pharmacophores are made up of aligned active molecules and can exhibit a preference for already known chemotypes. Pharmacophore models are constructed based on the interaction between a ligand and its target and describe the characteristics of the binding site itself (Kaserer et al., 2015).

Molecular docking can be divided into two tasks: pose generation and pose evaluation. While docking is still widely used for suggesting plausible binding modes and prioritizing compounds for further investigation, there are still issues with the scores used to prioritize docked poses (Kitchen et al., 2004). Docking is often employed as a screening tool in a standard screening campaign to prioritize computational and experimental efforts. Binding poses that are 'fair' are not necessarily 'accurate' estimates of binding affinity and 'attractive' poses could give false binding data (Guedes et al., 2018). The docking results should thus be regarded as a hypothesis that needs further investigations instead of the direct assessment of binding strength of a ligand. The behavior of the scoring function plays a significant role in the docking efficiency (Yang et al., 2022). In the practical screening process, ability to reproduce relative affinity trends is more significant than the individual images' quality. Even compounds associated with reasonable binding geometries can be ranked incorrectly, an issue which is often only identified during the course of additional experiments. Therefore, most researchers consider the ability to rank rather than to accurately predict absolute binding affinity to be a more valuable metric of docking utility, and this is certainly the case when the true binding affinities remain beyond the reach of current scoring functions (Martis & Téletchéa, 2025). However, there are still advantages in using reference datasets and structured evaluation exercises to investigate the behavior of docking procedures in controlled conditions. A Comparative Assessment of Scoring Functions (CASF) was developed to separate restrictions due to posture prediction and scoring restrictions (Su et al., 2019). This separation can help determine if the poor compound ranking is due to docking problems or is a problem with the scoring function itself. Performance of the controlled evaluation condition does not necessarily reflect on the success of the prospective screening, but the investigations are helpful in identifying problems in the method and in promoting workflow refinement (Figure 2).

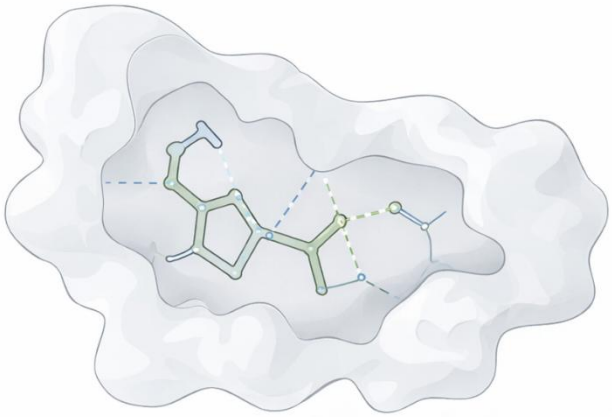

Figure 2. Docking-generated protein–ligand binding hypothesis.

There are also, often, additional computational layers added following molecular docking to enhance compound prioritization and compensate for the weaknesses of known scoring functions (Yang et al., 2022). Mostly, however, only the docking data is not enough because the accuracy of the ranking is still sensitive to the approximations used in the scoring functions, and conformational sampling is uncertain. For this reason, the first draft lists of candidates are often filtered and re-evaluated before a definitive selection is made (Atasever et al., 2024). An alternative approach widely adopted is the use of machine learning scoring systems as an aid to the conventional scoring methods (Mysinger et al., 2025). While good quality experimental protein structures are typically used to assess their performance, often the practical screening campaigns are made on modeled structures, alternative conformations, or proteins other than those used for technique development. This discrepancy introduces the challenge of ranking behavior and limits the extent to which a claimed performance is transferred to routine screening applications (Su et al., 2019). Another important advancement in computational drug discovery is the active-learning approaches. These methods can incorporate new experimental data during future screening cycles to guide the selection of a compound with the greatest information value, instead of the highest predicted score (Vellal et al., 2024). These strategies are particularly attractive when there is limited resources for experiments and a follow-up research is costly. De novo molecular design is also an area where generative approach has been popular. They have shown an ability to suggest new chemical structures, making a greater number of compounds available for research. However, there are still challenges in data quality, data validation and data synthetic feasibility. Thus these models are of value as a tool for proposing new classes of chemicals to expand the range of chemical possibilities, and must be tested in the lab before a particular candidate can be advanced (Mysinger et al., 2025; Vellal et al., 2024; Arús-Pous et al., 2020). Although they are useful, these workflows have many of the limitations of standard computational methods and can introduce new sources of uncertainty. Reproducible docking requires careful parameter control,

clear reporting, and iterative testing and validation, with potential variations in the result even if the same protocol is used, due to slight differences in the setup of the systems, search parameters, or the way the procedures are designed (Mysinger et al., 2025). Many docking methodologies are stochastic in nature, highlighting the need to carefully document the methodology used because different initializations and sampling methods can result in different docking poses and compound ranking. Machine-learning parts introduce new quality and design problems for data evaluation. The performance of the ML-based scoring methods can differ drastically depending on the dataset used and may be affected by the bias in the training data, which reduces the reliability when scoring in new situations (Artrith et al., 2021). Difficulties also arise when trying to implement a closed-loop active-learning workflow that involves repeatedly selecting molecules from a small part of chemical space that eventually results in diminishing diversity and increased model overconfidence (Vellal et al., 2024). One of the biggest constant threats to effective assessment is data loss. Information may be added to a workflow by pre-processing, feature engineering or the creation of a data set, all without a person's awareness, leading to overly optimistic estimates of the performance that may not be observed when the process is used in the future (Kapoor & Narayanan, 2023). Thus, in low data scenarios, where minor decisions in methodology can have a disproportionate effect on the results, workflow design, sampling strategy, and data quality can have a greater impact on practical performance than model architecture alone (Artrith et al., 2021). The real-world applications of computational drug-discovery procedures are most clearly seen as a conjunction of virtual predictions with experimental validation, rather than as stand-alone goals. Several investigations were then carried out by Taha and colleagues, including pharmacophore modeling, QSAR analysis, molecular docking and virtual screening, and then biological evaluation of the compounds in a broad spectrum of therapeutic applications. All of these reports can be found in a range of therapeutic areas, including GSK-3β, DPP-IV, HSL, neuraminidase, and PTP1B (Taha et al., 2015; Taha et al., 2017; Taha et al., 2018a; Taha et al., 2018b; Taha et al., 2021). The works represent realistic examples of the use of computational techniques for hit identification and lead discovery, provided that the modelling results are combined with other stages of the discovery process that rely on experimental evidence. They also make a key part of this review, which is the utility of the computational approaches is not only based on their predictive accuracy, but also on the overall reliability, transparency, and verification of the discovery process. For most workflows in computational chemistry, thus, the practical usefulness is best assessed by their robustness in prioritizing compounds when reliable discovery constraints are applied as opposed to a carefully selected bench set of retroactive compounds. Value is evident when prospective candidates continue to perform well even when their experimental capacity is small, their settings change, and it is evaluated repeatedly. If the decisions of priority are not feasible when workflows are applied in normal discovery situations, then improving the benchmark measures is not as important.

Common Sources of Error and Practical Controls in AI-Driven Computational Drug-Discovery Workflows

Table 1 covers key decision points in computational drug discovery workflows, as well as recurring causes of false confidence, feasible mitigation techniques, and evaluation methodologies that are more reflective of projected screening settings.

Workflow stage / decision point	Typical outputs used for decisions	Common failure modes that inflate confidence	Controls that reduce false confidence	Evaluation that matches practical use
Problem setup & dataset definition	Endpoint/label policy; assay context; inclusion rules	Noisy or inconsistent labels; proxy endpoints; biased negative definitions	State intended use; document label provenance; pre-specify negatives	Time-aware or assay-aware splits; external test sets where feasible
Library preparation & filtering	Filtered library; retained scaffolds	Over-filtering; benchmark-tuned rules; scaffold bias	Report thresholds; track removals by scaffold; keep diversity holdout	Report enrichment *and* diversity; compare to minimal-filter baseline
Pharmacophore screening	Hypothesis; hit list; fit scores	Treating fit as activity; chemotype lock-in	Validate vs actives/inactives; iterate with stated rules	EF@k/BEDROC; scaffold/time splits; replicate hypotheses
Docking (pose generation)	Poses; interaction patterns; pose clusters	Apparent certainty from one run; prep sensitivity	Replicate seeds; test prep variants; receptor ensembles when justified	Pose stability across runs; sensitivity analysis of prep choices
Docking (scoring/ranking)	Ranked list; top-k shortlist	Rank instability; protocol dependence; library-size artifacts	Consensus scoring; rank agreement checks; triage rules	Rank correlation across perturbations; EF@k/BEDROC; prospective mini-panels
Machine-learning	Rescored ranks;	Crystal-to-docked mismatch;	Harder splits;	Blinded challenges where feasible;

rescoring and scoring functions	predicted affinity; confidence (if provided)	shortcut learning; similarity leakage	evaluate on docked poses; ablations; calibration checks	external tests; calibration reporting
QSAR / property prediction (incl. ADMET)	Predicted property; probability/interval (if provided)	Random-split optimism; domain shift; miscalibration	Scaffold/time/cluster splits; applicability domain; calibration	Mean±SD across splits/seeds; calibration metrics; temporal validation
Active learning selection	Next-batch picks; acquisition score	Exploitative collapse; feedback bias; undefined stopping	Balance exploration; track diversity; define stopping	Cumulative hit rate + diversity; baseline comparison; per-round outcomes
Hit triage & experimental follow-up	Final shortlist (often dozens to <100); rationale	• Undocumented human rules; selection bias; visual pose bias	Pre-specify criteria; record rationales; orthogonal filters	Prospective mini-validation; hit rate with uncertainty; failure analysis
Reproducibility & reporting (cross-cutting)	Code/env/parameters; logs; datasets	Environment drift; hidden randomness; undocumented preprocessing	Pin versions; containers; runnable scripts + test data	Fresh-environment reruns; checklist-style reporting of rerun essentials

Notes: EF@k, enrichment factor at top-k; BEDROC, Boltzmann-enhanced discrimination of ROC; QSAR, quantitative structure–activity relationship; ADMET, absorption/distribution/metabolism/excretion/toxicity. The framework is applicable to molecular docking, machine-learning scoring functions, virtual screening, and AI-assisted drug-discovery workflows.

Retrospective enrichment analysis of curated data sets is not considered as informative as benchmark studies or blinded community challenges in this research. Narrative reviews are primarily used to identify patterns, limitations and typical reasons for failure rather than to substantiate claims of performance. Therefore, Table 1 should be viewed as a useful guide to associating key decision points in computational drug-discovery pipeline steps with common sources of false confidence and interventions that help stabilize compound prioritization across prospective screening. Workflow based techniques help to increase the repeatability of the workflow and clarify the single processing steps. Documentation of workflows allows investigators to follow the analytical decision making process, be able to repeat the analysis with similar conditions and to identify if conclusions are valid when the analysis is repeated (Pineau et al., 2021). But, being able to replicate is not just about having code available. A common failure is environment drift, particularly if software dependencies and library versions are not well documented or system parameters are not documented. Many of the computational operations published in the literature are found to be too complex to be executed without major modifications when considered on a large scale (Pineau et al., 2021). Another practical benefit of a computational screen is its ability to narrow down huge chemical libraries to a small subset of candidates prior to synthesis or further experimental testing. Enabling exploration of chemical space on a larger scale than can be done in the lab by using access to large virtual libraries shifts resources to a smaller number of compounds with a greater chance of success (Lyu et al., 2019). The region of chemical space finally explored depends on the screening strategy adopted, and is similar to comparative studies of virtual screening methods and their effect on compound selection (Su et al., 2019).

In recent years, AI-driven computational workflows for drug discovery have been increasingly integrated, and virtual screening, drug-target interaction prediction, molecular modeling and other computational methods are integrated into a decision-making pipeline (Atasever, 2024). Deep-learning-based molecular language models have improved our ability to explore huge and complex chemical spaces and propose structures with desired attributes (Mysinger et al., 2025). However, synthetic candidates also have to be subjected to the same downstream limitation as the created compounds, which include filtering, synthetic feasibility assessment and experimental validation. The same results have been observed in the automated structure-based de novo design approaches, where rapid generation of candidate structures is often possible, but synthetic accessibility and validation are important practical constraints (Díaz-Rovira et al., 2023). One of the most important steps in the drug discovery process still to be transferred from virtual prioritizing to experimental testing. Whether or not highly ranked candidates follow through to experimental evaluation is, of course, what determines the value of the computational forecasts. This means that the success of the overall workflow can depend a great deal on the compound selection processes and many potential virtual hit lists do not yield any candidates that make it to the testing phase (Lyu et al., 2019). An ever-present problem in computational drug discovery research is the difference between performance in the past and the potential utility in the future. As mentioned above, methods that are successful for publicly available datasets or carefully designed evaluation sets may not be as successful in special assay systems, new chemical space, and/or experimental conditions outside of those used in model building (Kapoor & Narayanan, 2023). This makes it difficult to use retrospective performance evaluations to predict future behaviour. Blinded community challenges are a more stringent test of generalization since they do not have information for model building.

Examples of such grand challenges include the Drug Design Data Resource (D3R), which studied pose prediction and affinity ranking for prospective evaluation conditions, and similar projects for predicting binding free energy (Gaieb et al., 2018; Gaieb et al., 2019). These investigations provide a more realistic view of the performance of computational operations when they are faced with new data. Pre-registration models for computational modelling investigations are also becoming more popular. Techniques such as Registered Modelling Reports promote the involvement of the investigator in establishing the criteria for evaluation, techniques for data partitioning, and the procedures for decision making prior to the start of the investigation. Pre-specification reduces analytical flexibility and the chances of overly positive post-hoc interpretations. Consequently, the reported performance tends to reflect study goals rather than changes in goals after the study has been conducted (Kapoor et al., 2024). Although there are known limitations, confidence stemming from benchmark performance is commonplace in computational drug development. While successful outcomes in controlled evaluation environments are not always replicated in the consistent situations and variability of real discovery processes (Kapoor & Narayanan, 2023) Changes in benchmark indicators are not sufficient evidence that a workflow would enable for stable or transportable decision making. The selection of the data sets used for retrospective evaluation, the techniques used to select the data, and assumptions about the method used can affect the results, while the selection of the measures used can obscure performance during early-recognition tasks that are particularly useful to virtual screening (Truchon & Bayly, 2007; Triballeau et al., 2005). On a wider scale, the evaluation artifacts and methodological biases may result in over-optimistic conclusions that do not stand up to subsequent evaluation under other test conditions (Wallach & Heifets, 2018). Less frequently reported are negative outcomes, unsuccessful campaigns and technical limitations encountered in experimental follow-up. Publication bias is a major problem; negative or inconclusive findings are underreported, for example (Dirnagl et al., 2022; Fanelli, 2012). Problems arise with the development of machine-learning scoring functions, too: inactive molecules are often poorly described and can exhibit significant differences from the active molecule that are used to train and test the model (Truchon & Bayly, 2007). There's one additional restriction: experimental capacity. The number of compounds that pass the virtual screening stage and are typically sent to the lab is very low, and, therefore, success rates among the top compounds in the virtual screen are only applicable to a small number of compounds. (Lyu et al., 2019) Some pitfalls and counterintuitive facts for computational drug discovery workflows. The migration of AI-driven computational workflows from a controlled evaluation environment to more realistic screening and experimental follow-up scenarios often results in false confidence. Although it may seem to operate as intended, there are several recurring reasons for instability that could affect compound prioritization. The different sources are summarized in Figure 3 and the emphasis is placed on practical measures to improve the stability of the decision making process.

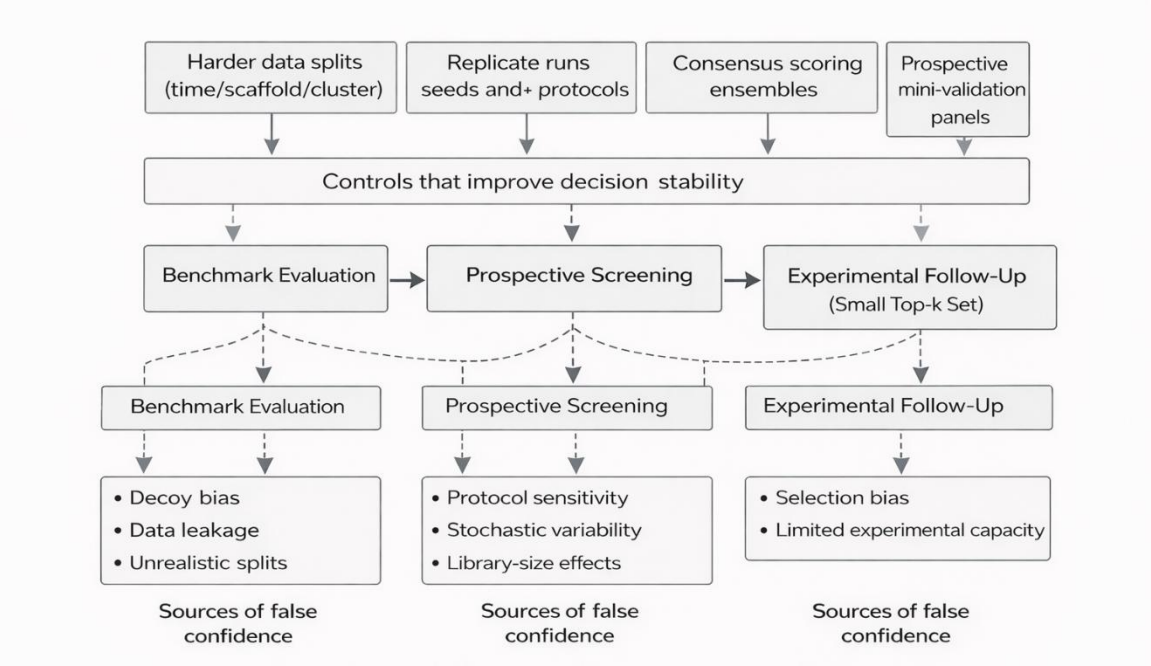

Figure 3. Sources of false confidence and practical controls that improve decision stability across AI-driven computational drug-discovery workflows.

As workflows shift from retrospective evaluation to prospective screening and experimental follow-up, the false confidence can grow as shown in figure 3. The diagram also highlights some of the more practical solutions that can help make decisions more stable, including more restrictive data-partitioning strategies, replication analyses, consensus approaches, or future mini-validation studies. The closer one gets to experimental testing, the more one becomes reliant on assessment design and

workflow setup for ranking outcomes rather than a true, consistent predictive signal. Small changes to the preprocessing methods, pre-processing parameters, or data partitioning could greatly change compound ranks, with implications for subsequent experimental design. Such effects are frequently only noticeable when workflows are repeated, independently analyzed, or implemented under different situations. This is further compounded by some common sources of instability such as data leakage, decoy bias, protocol sensitivity, stochastic variation, selection bias and limited experimental follow-up capability. These characteristics are representative of the moments when ranking may become brittle even if it seems like performance measures are adequate. In Figure 3, the procedures outlined lend themselves to reducing these risks and to making the prioritisation judgements more reliable in the case of prospective screening environments, whilst illustrating critical decision points within a computational drug discovery workflow and highlighting common sources of false confidence and appropriate methods to enhance prioritising reliability when there is limited capacity for experimental follow up (Kapoor et al., 2024; Kapoor & Narayanan, 2023; Pineau et al., 2021).

One of the more visible reasons for instability is molecular docking and protocol selection may affect the ranking behavior much more than the biological system itself. Search space is a good example. Even if the same structures and the same scoring method are applied, the size of the docking-box can differ and affect the compound ranking, giving the appearance of improved performance, but not necessarily altering the signal (Guedes et al., 2018). A new uncertainty comes into play when assigning protonation states and preparing the charge. The predicted interactions and ranking of compounds may vary, particularly for targets that have a large number of ionizable groups (Pantsar & Poso, 2018) when there are small differences in protonation.

In both experimental and computational studies, changes in protonation state have been demonstrated to affect affinity estimates and interactions (Onufriev & Alexov, 2013). A third source of uncertainty is score functions. A variety of aspects of interactions have been highlighted by force field-based, empirical and knowledge-based scoring methods, often leading to markedly different ranking results even though the same binding postures are given as equivalent. Therefore, the agreement in pose prediction does not imply the agreement in the affinity rating. The apparent stability may therefore be masking an underlying instability and only become apparent in repeated operations or when the parameters are changed. Docking scores and ranked compound lists must be used with caution, as small variations in methodology can make significant differences to downstream priority decisions.

Another type of instability arises through the stochastic processes in docking and machine-learning operations. Due to the stochastic nature of the conformational space exploration process used in molecular docking, it is possible that two runs of similar conditions can give different poses or ranking results (Mysinger et al., 2025). This heterogeneity is often masked when the results of one execution are reported, even if it has an impact on the compound prioritization. Ranking discrepancies can be identified in repeated studies which may impact the selection of individuals for experimental follow-up. Likewise, in machine-learning workflows. For chemical data, data partitioning methods can greatly affect the estimations of performance, particularly when structurally similar compounds are distributed across the training and test sets (Kapoor & Narayanan, 2023). Many research report a single training run, even though there can be a lot of difference between runs, depending on the random seed used for training. Other evaluation approaches have been suggested, such as temporal splitting, which will give a more realistic prediction of future performance (Sheridan, 2013). Multiple docking/multiple model training results in the generation of different sets of molecules with high ranks at each, showing the practical significance of these effects. If there are only a handful of candidates that are advanced to experimental testing, this uncertainty directly affects the subsequent judgement. Therefore, single-run performance estimation only gives a partial idea of process reliability and must be used together with the measures of variability and ranking stability.

The third source of misplaced confidence comes when screening is done on a large-library scale. The probability of observing high ranked molecules rises with the size of the library even if the predictive signal itself does not change (Lyu et al., 2019). At the same time, the number of artifactual high ranking options will increase the chances of picking something that "looks" like a success without actually being one. A practical constraint for screening campaigns is that it is usually only a small fraction of the compounds of interest that can be measured in practice. In such cases, statistics from the ranking system can only give a partial view of the quality of the final shortlist. The results can vary considerably depending on the different compound-selection procedures applied to the same sorted list, particularly if there is no explicit consideration of structural diversity (Besnard et al., 2012; Reker et al., 2014). Consequently, the selection strategy can be as important as the ranking method that produced it, and can have as much impact on the final candidate set as the method of ranking. Shortlists are often subjected to post-processing filters to remove false-positives and improve quality of the shortlist, which may also affect the compounds retained in the shortlist. Large-scale virtual screening has shown that an accurate ranking does not always lead to the selection of good candidates, especially when making decisions for a relatively small fraction of the ranked chemical space.

The other three sources for wrongful confidence are the design and implementation of negative controls. Poorly-defined inert compounds or when decoys are very different from the active compounds, workflows can seem more selective than they really

are (Artrith et al., 2021). When this occurs, models can learn surface dataset properties instead of binding determinants, leading to unrealistically high performance scores. Studies of some of the common benchmarking sets have found that apparent enrichment was the result of analogue bias or artifacts created in the construction of the decoys (Truchon & Bayly, 2007).

Negative datasets are essential in composition. Systematic bias is a common problem that can arise in model construction and evaluation as a result of the poorly described putative inactive chemical, and imbalance between the active and inactive molecules is common (Kapoor & Narayanan, 2023). It gets complicated with decoy design. It has been demonstrated that the quality measured for well-docked methods varies substantially using property-matched decays, suggesting that property selection can have a strong effect on the quality of the concluded methods (Chen et al., 2019). Reported performance should be considered within the context of the generation and dissemination of negative instances. A number of strategies have been suggested to reduce these effects, such as increasing the complexity of the sets used in the deconys, and attempts to decrease the similarity between the active chemicals and the negative controls (Chen et al., 2019). However, experimental counterexamples are still the most insightful evaluation of the reliability of the process. Compounds that are not high scoring can actually reveal defects in their binding that would not be seen during retrospective analysis (Lyu et al., 2019). If the discriminatory strength is weak, then the ranking results can seem convincing, even if there are not any robust negative controls.

An additional reason for false self-confidence is due to the evaluation and reporting process. While docking scores are known to be bad predictors of binding, visual inspection is still an important criterion in the judgment of compound selection, particularly when considering their docking score. The rules that are used to decide whether to accept or reject docking poses, however, are often not well-defined or fully advertised (Pantsar & Poso, 2018). This results in a variety of investigators selecting different candidates from the same output of the docking, which is hard to capture in published output and makes the results less reproducible. Therefore, the conclusions drawn might be just one of many possible explanations that can be derived from the available evidence. Add to this the publication bias. Negative, inconclusive or unsuccessful outcomes are less likely to be reported and this creates a literature that overemphasizes successful applications and that very little is learnt about the instability or failure of workflows (Dirnagl et al., 2022; Munafò et al., 2017). To compensate for these deficiencies, efforts to improve the reproducibility of published studies have been undertaken, reporting standards have been developed, and standardized checklists have been introduced to enhance documentation of computational workflows (Pineau et al., 2021). This helps to ensure that the data-processing procedures, the setting of parameters and details of execution are sufficiently specified for independent examination. Procedures that have not been properly documented may seem to be well established in published procedures, but may fail to work consistently if the procedures are repeated or carried out in slightly different circumstances.

Decision Stability in AI-Driven Drug Discovery Workflows

In AI-driven drug-discovery workflows, decision stability is often a more informative indicator of practical utility than isolated improvements in predictive performance metrics. In computational drug discovery, stability in computational drug discovery workflows means that the prioritisation of decisions does not vary significantly if repeated or if conditions change within reasonable limits. Examples of such changes include alternative random seeds, changes in docking methodology, protonation-state assignments, data partitioning algorithms, and differences in the makeup of the library (Pantsar & Poso, 2018). The results of docking and machine-learning experiments demonstrate that very small methods changes can result in changes in ranking behaviour, despite seemingly stable overall performance. Because of the stochastic search processes involved in repeated docking runs, these may yield different postures or ranking (Mysinger et al., 2025) and improvements in claimed accuracy do not necessarily lead to predictions which are of chemical significance or experimentally feasible (Kapoor et al., 2024). The results indicate that the behavior of the workflows needs to be tested across a range of scenarios, and not only a single performance estimation. By current practices for computational modeling, and in the machine learning of chemistry, it is good practice to strengthen confidence in a workflow through reporting, re-evaluation, and evidence for results being consistent across independent runs (Kapoor & Narayanan, 2023).

Consensus methods can enhance the confidence in compound prioritization when some methods return different rankings. It has consistently been found in consensus docking reviews that none of the docking software performs optimally for all the targets and screening contexts (Houston & Walkinshaw, 2013; Wang et al., 2016). The benefit of using multiple docking techniques is that they can increase the percentage of correct poses and provide more structural information for selecting best candidates than any single docking method (Houston & Walkinshaw, 2013). Ensemble docking extends this idea by including ensembles of data from multiple docking algorithms, multiple runs or alternate receptor conformations (Amaro et al., 2018). This is sometimes interpreted as sign of greater robustness of the studies, particularly when several competing runs of the same study are obtained. The same has been reported with the receptor-ensemble screening, in which more than one protein conformation is known to increase the extent of discrimination between binders and non-binders, especially when the protein has a flexible binding site (Amaro et al., 2018). These methods give a much more comprehensive view of possible binding interactions than screening against a designated receptor structure.

The use of consensus scoring has also been proposed as a method to minimize false positive and false negative scores by weighting agreement between the scoring functions over using a single scoring scheme to isolate the ones or zeros (Genheden & Ryde, 2015). Many of the same concepts are applied in machine learning: prediction uncertainty is calculated via ensemble methods and bayesian approaches (Hirschfeld et al., 2020). However, these estimations are very sensitive to calibration. Uncertainty numbers have been shown to not accurately represent actual predictive performance unless careful attention is given to the calibration procedures under similar conditions as intended use (Hirschfeld et al., 2020). Therefore, the uncertainty estimations need to be assessed carefully and tested empirically if possible.

It is important to evaluate decision stability directly, not from a single study or a visual assessment of the docking results. The primary question is if this trend persists in workflows repeated or if methodologically, reasonable changes are made, the high-ranking compounds are still prioritized. There are several statistical methods that can be used to assess the consistency of the ranking; however, early recognition criteria can provide useful estimates of the concentration of active chemicals towards the top of a screening list (Mysinger et al., 2025; Triballeau et al., 2005; Truchon & Bayly, 2007). The most useful practical test of a priority is often whether it is sufficiently stable to guide experimental decision making: from a drug discovery perspective. This is particularly true if only a few compounds can be considered for biological testing, in which case the precision of the final shortlist is more critical than any minor differences in overall performance measures (Scannell & Bosley, 2016).

Negative controls are more useful when used as a decision support tool than as an actual assessment. The ability for a workflow to detect meaningful signals in artifacts depends greatly on the selection of the negative examples and their inclusion. The composition of decoys can have a profound effect on perceived performance. Several studies have demonstrated that commonly used decoy sets could contain systematic biases that can be used by prediction models (Sieg et al., 2019; Chen et al., 2019). In addition to dataset generation, experimentally validated nonbinders provide a feasible method of distinguishing whether an extremely rated compound is a real opportunity for further exploration or a mere 'sure thing' but wrong answer (Lyu et al., 2019). This role becomes particularly vital during hit identification as the amount of experimental resources is limited and only a few hits can be advanced to testing. In this context, negative controls act as a filter to select untrustworthy individuals that will not be placed in the red list, thus freeing up other resources (Lyu et al., 2019).

Uncertainty-aware Interpretation bridges computer predictions and decision-making in drug discovery. Uncertainty is integral to early stage compound selection however confidence in model outputs is often treated informally or not considered at all (Hirschfeld et al., 2020). Measuring uncertainty can help investigators assess the risk of a prediction and more effectively allocate their experimental resources. This means that confidence estimations are used during the decision making process, instead of after the compound ranking has been finished (Renz et al., 2019). There are several approaches for deriving prediction uncertainty at the individual compound level such as the conformal prediction and related uncertainty quantification approaches (Norinder et al., 2014; Svensson et al., 2018). Key objective of these approaches is to not only provide confidence numbers but to also establish circumstances when predictions could be less accurate, thus requiring further investigation. This is particularly important in early drug discovery projects, where there may only be small amounts of data and limited experimental labels and additional assumptions in the model may make it more susceptible to noise (Hirschfeld et al., 2020). A practical use of uncertainty estimation is when it affects compound prioritization and experimental preparation. Compounds with high uncertainty prediction could be the potential candidates that need additional validation, while those with good ranking and a consistent level of confidence could be explored further.

A Practical Framework for AI-Assisted Drug Discovery in Resource-Limited Settings

AI-assisted computational workflows can provide substantial value in resource-limited research environments when their design emphasizes transparency, validation, reproducibility, and decision reliability. Detailed documentation of study design and data preparation and analysis is the first step toward reproducibility. Computational operations cannot be successfully replicated if important elements such as preparation, parameter settings, or execution steps are left out (Pineau et al., 2021). Practical implications extend beyond computational benchmarking to real-world pharmaceutical research and development. Workflows that appear robust under retrospective benchmark conditions may perform unpredictably when applied to novel biological targets or unfamiliar chemical space. Consequently, computational approaches that maintain stable compound prioritization across methodological perturbations are more likely to support productive experimental follow-up than workflows optimized solely for benchmark performance. In practical drug-discovery settings, prioritizing a smaller number of computationally well-supported candidates may provide greater value than generating extensive lists of uncertain predictions (Scannell & Bosley, 2016). Structured workflows and pipeline-based methods help to reduce ambiguity by ensuring that the analytical phases are explicit and traceable and investigations can be reproduced and transferred to new research settings (Sandve et al., 2013). A minimum of, reproducible research should contain sufficient detail to enable independent investigators to re-execute the research process, including software prerequisites, software installation instructions, access to representative data sets, and scripts to produce key results (Peng, 2011). The steps involved in data processing and data preparation should be given special

consideration because if the data is changed during these steps that change is likely to cause inconsistencies and irreproducible results (Kapoor & Narayanan, 2023). Similarly, it is crucial to use proper baseline comparisons. Simple and well-established alternatives can be compared to a workflow to determine if it is a worthwhile development or just another way of doing something that is complicated (Pineau et al., 2021).

It is important to consider the variability in the results of repeated computational screening when evaluating the results. Subtle changes to the methodology can affect the ranking of compounds and thus the selection of compounds for experimental follow-up (Mysinger et al., 2025). This is why it is not possible to make a judgement based solely on one analysis, particularly if only a few compounds are sent for analysis. The same holds true for molecular docking. Even if predicted binding modes are chemically plausible, ranking results can change after acceptable changes in the process variables (Mysinger et al., 2025). This means that the validity of prioritization is often tested in practice by comparing the prioritization between different sets of parameters, and not only the best ranking solution (Truchon & Bayly, 2007). Compounds that maintain their high ranking under several situations can be trusted more than those backed by a single analysis. The key challenge from a drug-discovery perspective is not that the same product can be consistently reproduced, but that priority decisions are sufficiently consistent to allow experimental planning and compound selection (Pineau et al., 2021).

Practical screening protocols should find a balance between rigor and the computational and experimental resources available. Many drug-discovery programs use multi-stage screening procedures to limit the number of compounds that need to be thoroughly evaluated. Before virtual screening is applied, often pharmacophore-based filtering is used to help narrow down the attention to drugs that possess necessary interaction requirements while also lowering the number of drugs included in the virtual screening set (Giordano et al., 2022). Another popular option is consensus-based approaches. The use of multiple docking techniques or scoring functions may be used to enhance the selection of potential candidates while minimizing the need for a single computational approach (Houston & Walkinshaw, 2013). Compounds can be prioritized again after the first rescore using additional rescore processes, particularly when there is a large number of candidates of similar structures in the initial scores (Mysinger et al., 2025). Model interpretations is also an important point. Projected performance should be considered along with the degree of confidence in the prediction when making prioritization decisions. This is particularly important if you have limited experimental capacity and only a maximum number of compounds can be advanced for testing. It is also important to consider computational cost, because the usefulness of a specific screening technique depends partly on its ability to be performed within the available resources, and with reliable compound selection (Lyu et al., 2019; Truchon & Bayly, 2007).

However, the usefulness of computational screening is ultimately determined by the compounds selected for experimental follow-up. Therefore, it is essential that the compound advancement is driven not just by the score, but by certain choice criteria. In many cases, hit selection can be more reliable if hit selection thresholds, ligand efficiency, chemical tractability, and the possibility for further optimization are considered (Hopkins et al., 2014). Methodological limitations should be noted before carrying out the test for compounds. Tools are becoming more important than biological activity in scoring, virtual screening, and related predictions (Jiménez-Luna et al., 2021).

Experimental validation is crucial in the discovery process. Follow-up assays should not only be used to validate activity but also for the removal of false positive, for establishing preliminary structure-activity relationships, and for the assessment of target engagement (Lyu et al., 2019). For a limited number of chemicals, the design of these tests would have a direct influence on the usefulness of the computational predictions. Specific reporting of selection criteria, control substances and uncertainty reported in the priority decision process may enhance the interpretation of screening results and facilitate cross-study comparison (Scannell & Bosley, 2016; Pineau et al., 2021; Kapoor et al., 2024).

The worth of computational workflows, in practice, is ultimately determined by the value of the judgements that they allow. If priorities are shifting or the progress towards experiments does not clearly reflect the improvements in the performance, such an overemphasis on performance may have limited value. It is much more successful to focus methodological rigor on stages where it is likely to affect compound selection and resource allocation, where increasing stability is likely to be considered beneficial in terms of downstream outcomes (Aittokallio, 2022). There have also been decision-theoretic arguments to suggest that the quality of choices can have a more significant impact on long-term research productivity than can the increase in throughput (Scannell & Bosley, 2016). In the drug-discovery setting, it is often better to have a smaller number of candidates that have a lot of supporting evidence than have a huge list of candidates with little evidence. Computational methods are still vital for reducing the cost and size of the initial screening, by focusing on a small number of drugs worthy of experimental testing (Pineau et al., 2021). The real significance of their value, however, is that they are used to support strong and defensible prioritization decisions within the context of realistic discovery limits.

Robust AI-Driven Decision Making for Drug Discovery

In recent years, there has been a focus on public reporting and comprehensive assessment of computational prediction methods. To enhance documentation of model development and evaluation procedures, bias sources and intended uses, several reporting and quality-assessment frameworks have been developed (Lyu et al., 2019; Collins et al., 2015; Moons et al., 2015; Wolff et al., 2019; Kapoor et al., 2024).

These projects are part of a wider trend in the increased transparency, reproducibility and understanding of computational predictions. In applications to drug discovery, the relevance of these guidelines is that they give more confidence that the results reported can be applied to the process of decision making beyond the setting of the original study. Documentation of the data sources, methods of evaluation, limitations and intended use clearly separates procedures that are likely to remain informative for future applications from those that are highly dependent on the particular data set or experimental parameters. While most of these reporting formats were developed for clinical prediction and intervention studies, they all have the concept of intended use, clear documentation of work, and a focus on bias evaluation, which remains relevant to computational drug discovery workflows. These frameworks are regarded as sources for transferrable ideas but not as standards set specifically for docking, virtual screening or hit-identification investigations and are thus not depicted in this manner.

For AI-driven computational workflows to be reliable, relevant and sufficient documentation of the results shall be provided, along with adequate methodological transparency (Pineau et al., 2021). Recent recommendations also emphasize routine stability testing in workflow development, particularly if the predictions are intended to drive further experimental decisions (Aittokallio, 2022). A related problem is calibration, because poor calibration of the confidence estimates may lead to poor decisions about compound selection (Hirschfeld et al., 2020).

Generally speaking, there is an increasing interest in evaluation systems that are more representative of the actual environment where the findings are made than in an idealized test environment. In computational drug development, trust in a workflow has come to be increasingly based on evidence that, rather than the performance estimated from a single evaluation exercise, prioritization is retained in a variety of datasets, settings, and repeated analyses. Large prospective studies offer important evidence, but are often impractical because of time, cost, and resource requirements for doing experiments. Prospective mini-validation is a more feasible technique for many drug-discovery efforts that involves advancing a small, preset set of priority compounds for experimental testing. The idea is not to make sweeping claims about performance but if the candidates can hold up under realistic laboratory conditions, they are of computational value. This approach aligns with suggestions to support experimental validation and validation techniques in order to confirm results and validate the relevance to biology, while minimizing reliance on computational predictions (Lyu et al., 2019).

These forms of validation help to identify if priority decisions lead to action and further opportunities. This approach can be enhanced with iterative processes, which involve using experimental feedback in subsequent cycles of compound selection. Active learning has been proposed as a way to enhance chemical space exploration and to prioritize testing of compounds that will provide the most useful data given the limited testing capacity (Vellal et al., 2024). In vivo and downstream biological validation of the research adds to this further importance of linking the computations with experimentally observed outcomes (Taha et al., 2021). Validation is best completed as part of the normal workflow, and through the use of resource constrained screening campaigns, rather than as an endpoint (Lyu et al., 2019). A helpful approach for gauging the continual value of workflow products as discovery progresses is to provide regular feedback between computational prioritization and experimental testing (Renz et al., 2019).

Reliable computational experimental drug discovery processes need transparent data handling. Data curation, filtering, and data pre-processing may affect subsequent prioritization and, if not well documented, the interpretation of findings reported (Pineau et al., 2021). The resulting documentation of these techniques makes it easier to believe that they accurately represent workflow behavior, without being constrained by documentation that was not originally planned. Community resources and publicly available datasets have made it easier to design and compare methods for a variety of drug discovery applications (Kapoor & Narayanan, 2023). However, performance on standardized datasets is not always representative of the actual screening process in the field, where the chemical diversity and variations in assays to be deployed can vary significantly (Truchon & Bayly, 2007).

Consequently, greater focus has been given to more application-specific evaluation methods, such as careful consideration of data partitioning methods and external validation methods (Kapoor & Narayanan, 2023). More comprehensive data-management projects have also led to the structuring of data and workflows to facilitate easier access to data, interoperability and reusability across projects (Kapoor & Narayanan, 2023). In practice, these ideas are most helpful in making things clearer and allowing a separation between computational results and their evaluation.

Any statements on workflow performance should be commensurate with the supporting evidence. Improvements that are evaluated, prioritized and can have positive consequences beyond the immediate assessment context are most convincing. Interpretability, repeatability and generalizability have been identified as challenges in the application of computational drug discovery methods, as reported in a number of reviews, and caution should be used when reporting on performance claims (Jiménez-Luna et al., 2021). Such issues have been discussed in other fields of biology, with special focus on transparency of the prediction algorithm, consistency of performance in real-world settings, and confidence in accuracy, not only (Kapoor et al., 2024; Pineau et al., 2021). Similar problems apply to physics-based methods; in some cases, enhancing the computational efficiency does not eliminate limitations on forecast accuracy or transferability (Guedes et al., 2018; Pantsar & Poso, 2018).

Next steps would be to routinely perform prospective mini-stability and mini-calibration studies for chemicals that are selected. The evidence is more robust to assess the usefulness of AI-driven computational workflows than the performance gains reported in isolated evaluation conditions and it can support the maintaining of the relevance of the prioritising decisions during experimental follow-up.

Computational methods will not help in drug discovery programs unless they help to identify candidates for experimental testing. In the process of hit-to-lead optimization, often many decisions are made in the absence of information and resources on the most efficient chemical modification, biological assessment, and synthesis. In such cases, processes that have a high overlap in terms of the number of potential candidates identified may be more beneficial than those processes that just refine the benchmark performance. In this perspective, the ability to reproduce and the stability of the decision do not amount to an end; they reduce the risk of moving on to drugs that ultimately do not pass the experimental testing phase, and promote the effective use of laboratory resources (Lyu et al., 2019; Scannell & Bosley, 2016).

Emerging Artificial Intelligence Technologies in Drug Discovery

Recent advances in artificial intelligence have expanded the scope of computational drug discovery beyond traditional virtual screening and molecular docking (Jiménez-Luna et al., 2021). Generative AI models can propose novel chemical structures with desired properties, while foundation models and large language models (LLMs) are increasingly being explored for knowledge extraction, hypothesis generation, and decision support in pharmaceutical research (Sanchez-Lengeling & Aspuru-Guzik, 2018; Zeng et al., 2022). Although these technologies show considerable promise, their practical value still depends on rigorous validation, reproducibility, calibration, and experimental confirmation (Pineau et al., 2021; Hirschfeld et al., 2020). As a result, future computational drug-discovery workflows will likely combine AI-driven molecular design with robust evaluation frameworks that support reliable decision-making under realistic discovery conditions (Kapoor et al., 2024; Aittokallio, 2022).

Conclusion

Artificial intelligence, machine learning, molecular docking, and related computational approaches have become routine components of modern drug-discovery workflows. The literature reviewed suggests that their practical value is best assessed by the quality and reliability of the decisions they support rather than by isolated improvements in benchmark performance metrics.

A series of such common areas of confidence were identified through the literature reviewed: data leakage, protocol sensitivity, stochastic variation, inadequate negative-control design, and evaluation approaches that fail to simulate intended use. In all of these contexts, there were consistent trends that the compound prioritization was a more informative, and useful, metric of practical value than retrospective improvements alone. Methods that continue to point to the same set of potential candidates given the same (or similar) evaluation conditions are more likely to aid experimental decision making than methods that get higher scores given a more narrowly defined set of evaluation conditions.

Transparency of reporting, validation strategies, careful use of negative controls, and routine experimental confirmation are also emphasized in the literature. Simple prospective validation studies, conducted as part of discovery, may be more informative regarding practical utility than the more complex, retrospective validation studies. Advances in computational drug discovery should be assessed not only by predictive accuracy, but also by the utility of the method for reliable compound selection and successful experimental follow-up in realistic discovery settings.

Data Availability Statement

No new datasets were generated or analyzed during this study. All information discussed in this review was obtained from publicly available sources cited in the reference list.

References

(1) Steshin, S. Lo-Hi: Practical ML Drug Discovery Benchmark. arXiv October 10, 2023. https://doi.org/10.48550/arXiv.2310.06399.

(2) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13* (6), 505–508. https://doi.org/10.1038/s41557-021-00716-z.

(3) Kapoor, S.; Cantrell, E. M.; Peng, K.; Pham, T. H.; Bail, C. A.; Gundersen, O. E.; Hofman, J. M.; Hullman, J.; Lones, M. A.; Malik, M. M.; Nanayakkara, P.; Poldrack, R. A.; Raji, I. D.; Roberts, M.; Salganik, M. J.; Serra-Garcia, M.; Stewart, B. M.; Vandewiele, G.; Narayanan, A. REFORMS: Consensus-Based Recommendations for Machine-Learning-Based Science. *Science Advances* **2024**, *10* (18), eadk3452. https://doi.org/10.1126/sciadv.adk3452.

(4) Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns* **2023**, *4* (9). https://doi.org/10.1016/j.patter.2023.100804.

(5) Pineau, J.; Vincent-Lamarre, P.; Sinha, K.; Lariviere, V.; Beygelzimer, A.; d'Alche-Buc, F.; Fox, E.; Larochelle, H. Improving Reproducibility in Machine Learning Research(A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research* **2021**, *22* (164), 1–20.

(6) Aittokallio, T. What Are the Current Challenges for Machine Learning in Drug Discovery and Repurposing? *Expert Opinion on Drug Discovery* **2022**, *17* (5), 423–425. https://doi.org/10.1080/17460441.2022.2050694.

(7) Reker, D.; Schneider, G. Active-Learning Strategies in Computer-Assisted Drug Discovery. *Drug Discovery Today* **2015**, *20* (4), 458–465. https://doi.org/10.1016/j.drudis.2014.12.004.

(8) Wang, L.; Zhou, Z.; Yang, X.; Shi, S.; Zeng, X.; Cao, D. The Present State and Challenges of Active Learning in Drug Discovery. *Drug Discovery Today* **2024**, *29* (6), 103985. https://doi.org/10.1016/j.drudis.2024.103985.

(9) Atasever, S. In Silico Drug Discovery: A Machine Learning-Driven Systematic Review. *Med Chem Res* **2024**, *33* (9), 1465–1490. https://doi.org/10.1007/s00044-024-03260-w.

(10) Gadiya, Y.; Ioannidis, V.; Henderson, D.; Gribbon, P.; Rocca-Serra, P.; Satagopam, V.; Sansone, S.-A.; Gu, W. FAIR Data Management: What Does It Mean for Drug Discovery? *Front. Drug Discov.* **2023**, *3*. https://doi.org/10.3389/fddsv.2023.1226727.

(11) Wise, J.; de Barron, A. G.; Splendiani, A.; Balali-Mood, B.; Vasant, D.; Little, E.; Mellino, G.; Harrow, I.; Smith, I.; Taubert, J.; van Bochove, K.; Romacker, M.; Walgemoed, P.; Jimenez, R. C.; Winnenburg, R.; Plasterer, T.; Gupta, V.; Hedley, V. Implementation and Relevance of FAIR Data Principles in Biopharmaceutical R&D. *Drug Discovery Today* **2019**, *24* (4), 933–938. https://doi.org/10.1016/j.drudis.2019.01.008.

(12) Mervin, L. H.; Johansson, S.; Semenova, E.; Giblin, K. A.; Engkvist, O. Uncertainty Quantification in Drug Design. *Drug Discovery Today* **2021**, *26* (2), 474–489. https://doi.org/10.1016/j.drudis.2020.11.027.

(13) Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. Artificial Intelligence in Drug Discovery: Recent Advances and Future Perspectives. *Expert Opinion on Drug Discovery* **2021**, *16* (9), 949–959. https://doi.org/10.1080/17460441.2021.1909567.

(14) Renz, P.; Adler, T.; Hochreiter, S.; Klambauer, G. Uncertainty Estimation Methods to Support Decision-Making in Early Phases of Drug Discovery. *Neural Information Processing Systems Foundation (NeurIPS 2019), 2019* **2019**.

(15) Baker, M. 1,500 Scientists Lift the Lid on Reproducibility. *Nature* **2016**, *533* (7604), 452–454. https://doi.org/10.1038/533452a.

(16) Kaserer, T.; Beck, K. R.; Akram, M.; Odermatt, A.; Schuster, D. Pharmacophore Models and Pharmacophore-Based Virtual Screening: Concepts and Applications Exemplified on Hydroxysteroid Dehydrogenases. *Molecules* **2015**, *20* (12), 22799–22832. https://doi.org/10.3390/molecules201219880.

(17) Giordano, D.; Biancaniello, C.; Argenio, M. A.; Facchiano, A. Drug Design by Pharmacophore and Virtual Screening Approach. *Pharmaceuticals* **2022**, *15* (5), 646. https://doi.org/10.3390/ph15050646.

(18) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat Rev Drug Discov* **2004**, *3* (11), 935–949. https://doi.org/10.1038/nrd1549.

(19) Guedes, I. A.; Pereira, F. S. S.; Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.* **2018**, *9*. https://doi.org/10.3389/fphar.2018.01089.

(20) Yang, C.; Chen, E. A.; Zhang, Y. Protein–Ligand Docking in the Machine-Learning Era. *Molecules* **2022**, *27* (14), 4568. https://doi.org/10.3390/molecules27144568.

(21) Martis, E. A. F.; Téletchéa, S. Ten Quick Tips to Perform Meaningful and Reproducible Molecular Docking Calculations. *PLOS Computational Biology* **2025**, *21* (5), e1013030. https://doi.org/10.1371/journal.pcbi.1013030.

(22) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **2019**, *59* (2), 895–913. https://doi.org/10.1021/acs.jcim.8b00545.

(23) Rastelli, G.; Pinzi, L. Refinement and Rescoring of Virtual Screening Results. *Front. Chem.* **2019**, *7*. https://doi.org/10.3389/fchem.2019.00498.

(24) Ballester, P. J. Selecting Machine-Learning Scoring Functions for Structure-Based Virtual Screening. *Drug Discovery Today: Technologies* **2019**, *32–33*, 81–87. https://doi.org/10.1016/j.ddtec.2020.09.001.

(25) Yu, J.; Li, X.; Zheng, M. Current Status of Active Learning for Drug Discovery. *Artificial Intelligence in the Life Sciences* **2021**, *1*, 100023. https://doi.org/10.1016/j.ailsci.2021.100023.

(26) Zeng, X.; Wang, F.; Luo, Y.; Kang, S.; Tang, J.; Lightstone, F. C.; Fang, E. F.; Cornell, W.; Nussinov, R.; Cheng, F. Deep Generative Molecular Design Reshapes Drug Discovery. *Cell Reports Medicine* **2022**, *3* (12), 100794. https://doi.org/10.1016/j.xcrm.2022.100794.

(27) Chen, P.; Chen, C.; Zheng, Y.; Chen, F.; Liu, Z.; Ren, S.; Song, H.; Liu, T.; Lu, Z.; Sun, H.; Kong, Y.; Yuan, H. Discovery of 2,3-Dihydro[1,4]Dioxino[2,3-g]Benzofuran Derivatives as Protease Activated Receptor 4 (PAR4) Antagonists with Potent Antiplatelet Aggregation Activity and Low Bleeding Tendency. *J. Med. Chem.* **2024**, *67* (7), 5502–5537. https://doi.org/10.1021/acs.jmedchem.3c02099.

(28) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*. https://doi.org/10.3389/fphar.2020.565644.

(29) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49* (20), 5912–5931. https://doi.org/10.1021/jm050362n.

(30) 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), 947–961. https://doi.org/10.1021/acs.jcim.8b00712.

(31) Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* **2018**, *58* (5), 916–932. https://doi.org/10.1021/acs.jcim.7b00403.

(32) van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **2022**, *62* (23), 5938–5951. https://doi.org/10.1021/acs.jcim.2c01073.

(33) Taha, M. O.; Bustanji, Y.; Al-Ghussein, M. A. S.; Mohammad, M.; Zalloum, H.; Al-Masri, I. M.; Atallah, N. Pharmacophore Modeling, Quantitative Structure–Activity Relationship Analysis, and in Silico Screening Reveal Potent Glycogen Synthase Kinase-3β Inhibitory Activities for Cimetidine, Hydroxychloroquine, and Gemifloxacin. *J. Med. Chem.* **2008**, *51* (7), 2062–2077. https://doi.org/10.1021/jm7009765.

(34) Al-Masri, I. M.; Mohammad, M. K.; Taha, M. O. Discovery of DPP IV Inhibitors by Pharmacophore Modeling and QSAR Analysis Followed by in Silico Screening. *ChemMedChem* **2008**, *3* (11), 1763–1779. https://doi.org/10.1002/cmdc.200800213.

(35) Taha, M. O.; Dahabiyeh, L. A.; Bustanji, Y.; Zalloum, H.; Saleh, S. Combining Ligand-Based Pharmacophore Modeling, Quantitative Structure−Activity Relationship Analysis and in Silico Screening for the Discovery of New Potent Hormone Sensitive Lipase Inhibitors. *J. Med. Chem.* **2008**, *51* (20), 6478–6494. https://doi.org/10.1021/jm800718k.

(36) Abu Hammad, A. M.; Afifi, F. U.; Taha, M. O. Combining Docking, Scoring and Molecular Field Analyses to Probe Influenza Neuraminidase-Ligand Interactions. *J Mol Graph Model* **2007**, *26* (2), 443–456. https://doi.org/10.1016/j.jmgm.2007.02.002.

(37) Taha, M. O.; Bustanji, Y.; Al-Bakri, A. G.; Yousef, A.-M.; Zalloum, W. A.; Al-Masri, I. M.; Atallah, N. Discovery of New Potent Human Protein Tyrosine Phosphatase Inhibitors via Pharmacophore and QSAR Analysis Followed by in Silico Screening. *J Mol Graph Model* **2007**, *25* (6), 870–884. https://doi.org/10.1016/j.jmgm.2006.08.008.

(38) Schaduangrat, N.; Lampa, S.; Simeon, S.; Gleeson, M. P.; Spjuth, O.; Nantasenamat, C. Towards Reproducible Computational Drug Discovery. *J Cheminform* **2020**, *12* (1), 9. https://doi.org/10.1186/s13321-020-0408-x.

(39) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229. https://doi.org/10.1038/s41586-019-0917-9.

(40) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360–365. https://doi.org/10.1126/science.aat2663.

(41) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **2020**, *60* (12), 5714–5723. https://doi.org/10.1021/acs.jcim.0c00174.

(42) Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M. J.; Denniston, A. K. Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension. *Nat Med* **2020**, *26* (9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x.

(43) Gaieb, Z.; Liu, S.; Gathiaka, S.; Chiu, M.; Yang, H.; Shao, C.; Feher, V. A.; Walters, W. P.; Kuhn, B.; Rudolph, M. G.; Burley, S. K.; Gilson, M. K.; Amaro, R. E. D3R Grand Challenge 2: Blind Prediction of Protein–Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J Comput Aided Mol Des* **2018**, *32* (1), 1–20. https://doi.org/10.1007/s10822-017-0088-4.

(44) Mobley, D. L.; Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annual Review of Biophysics* **2017**, *46* (Volume 46, 2017), 531–558. https://doi.org/10.1146/annurev-biophys-070816-033654.

(45) Nosek, B. A.; Ebersole, C. R.; DeHaven, A. C.; Mellor, D. T. The Preregistration Revolution. *Proceedings of the National Academy of Sciences* **2018**, *115* (11), 2600–2606. https://doi.org/10.1073/pnas.1708274114.

(46) Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. *PLOS ONE* **2019**, *14* (8), e0220113. https://doi.org/10.1371/journal.pone.0220113.

(47) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47* (2), 488–508. https://doi.org/10.1021/ci600426e.

(48) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, *52* (4), 867–881. https://doi.org/10.1021/ci200528d.

(49) DeVito, N. J.; Goldacre, B. Catalogue of Bias: Publication Bias. *BMJ Evidence-Based Medicine* **2019**, *24* (2), 53–54. https://doi.org/10.1136/bmjebm-2018-111107.

(50) Song, F.; Parekh, S.; Hooper, L.; Loke, Y. K.; Ryder, J.; Sutton, A. J.; Hing, C.; Kwok, C. S.; Pang, C.; Harvey, I. Dissemination and Publication of Research Findings: An Updated Review of Related Biases. *Health Technology Assessment* **2010**, *14* (8), 1–220. https://doi.org/10.3310/hta14080.

(51) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of High-Throughput Screening in Biomedical Research. *Nat Rev Drug Discov* **2011**, *10* (3), 188–195. https://doi.org/10.1038/nrd3368.

(52) Feinstein, W. P.; Brylinski, M. Calculating an Optimal Box Size for Ligand Docking and Virtual Screening against Experimental and Predicted Binding Pockets. *J Cheminform* **2015**, *7* (1), 18. https://doi.org/10.1186/s13321-015-0067-5.

(53) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J Comput Aided Mol Des* **2013**, *27* (3), 221–234. https://doi.org/10.1007/s10822-013-9644-8.

(54) Onufriev, A. V.; Alexov, E. Protonation and pK Changes in Protein–Ligand Binding. *Quarterly Reviews of Biophysics* **2013**, *46* (2), 181–209. https://doi.org/10.1017/S0033583513000024.

(55) Huang, S.-Y.; Grinter, S. Z.; Zou, X. Scoring Functions and Their Evaluation Methods for Protein–Ligand Docking: Recent Advances and Future Directions. *Phys. Chem. Chem. Phys.* **2010**, *12* (40), 12899–12908. https://doi.org/10.1039/C0CP00151A.

(56) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of Computational Chemistry* **2009**, *30* (16), 2785–2791. https://doi.org/10.1002/jcc.21256.

(57) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9* (2), 513–530. https://doi.org/10.1039/C7SC02664A.

(58) Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53* (4), 783–790. https://doi.org/10.1021/ci400084k.

(59) Sadybekov, A. V.; Katritch, V. Computational Approaches Streamlining Drug Discovery. *Nature* **2023**, *616* (7958), 673–685. https://doi.org/10.1038/s41586-023-05905-z.

(60) Gimeno, A.; Ojeda-Montes, M. J.; Tomás-Hernández, S.; Cereto-Massagué, A.; Beltrán-Debón, R.; Mulero, M.; Pujadas, G.; Garcia-Vallvé, S. The Light and Dark Sides of Virtual Screening: What Is There to Know? *International Journal of Molecular Sciences* **2019**, *20* (6). https://doi.org/10.3390/ijms20061375.

(61) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53* (7), 2719–2740. https://doi.org/10.1021/jm901137j.

(62) Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Deep Generative Models for 3D Linker Design. *J. Chem. Inf. Model.* **2020**, *60* (4), 1983–1995. https://doi.org/10.1021/acs.jcim.9b01120.

(63) Mysinger, M. M.; Carchia, M.; Irwin, John. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594. https://doi.org/10.1021/jm300687e.

(64) McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K. A Specific Mechanism of Nonspecific Inhibition. *J. Med. Chem.* **2003**, *46* (20), 4265–4272. https://doi.org/10.1021/jm030266r.

(65) Fischer, A.; Smieško, M.; Sellner, M.; Lill, M. A. Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. *J. Med. Chem.* **2021**, *64* (5), 2489–2500. https://doi.org/10.1021/acs.jmedchem.0c02227.

(66) Nosek, B. A.; Spies, J. R.; Motyl, M. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspect Psychol Sci* **2012**, *7* (6), 615–631. https://doi.org/10.1177/1745691612459058.

(67) Stodden, V.; McNutt, M.; Bailey, D. H.; Deelman, E.; Gil, Y.; Hanson, B.; Heroux, M. A.; Ioannidis, J. P. A.; Taufer, M. Enhancing Reproducibility for Computational Methods. *Science* **2016**, *354* (6317), 1240–1241. https://doi.org/10.1126/science.aah6168.

(68) Jaghoori, M. M.; Bleijlevens, B.; Olabarriaga, S. D. 1001 Ways to Run AutoDock Vina for Virtual Screening. *J Comput Aided Mol Des* **2016**, *30* (3), 237–249. https://doi.org/10.1007/s10822-016-9900-9.

(69) Blanes-Mira, C.; Fernández-Aguado, P.; Andrés-López, J. de; Fernández-Carvajal, A.; Ferrer-Montiel, A.; Fernández-Ballester, G. Comprehensive Survey of Consensus Docking for High-Throughput Virtual Screening. *Molecules* **2022**, *28* (1). https://doi.org/10.3390/molecules28010175.

(70) Poli, G.; Martinelli, A.; Tuccinardi, T. Reliability Analysis and Optimization of the Consensus Docking Approach for the Development of Virtual Screening Studies. *Journal of Enzyme Inhibition and Medicinal Chemistry* **2016**, *31* (sup2), 167–173. https://doi.org/10.1080/14756366.2016.1193736.

(71) Houston, D. R.; Walkinshaw, M. D. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. *J. Chem. Inf. Model.* **2013**, *53* (2), 384–390. https://doi.org/10.1021/ci300399w.

(72) Triches, F.; Triches, F.; Lino de Oliveira, C. Consensus Combining Outcomes of Multiple Ensemble Dockings: Examples Using dDAT Crystalized Complexes. *MethodsX* **2022**, *9*, 101788. https://doi.org/10.1016/j.mex.2022.101788.

(73) Bottegoni, G.; Rocchia, W.; Rueda, M.; Abagyan, R.; Cavalli, A. Systematic Exploitation of Multiple Receptor Conformations for Virtual Ligand Screening. *PLOS ONE* **2011**, *6* (5), e18845. https://doi.org/10.1371/journal.pone.0018845.

(74) Nhat Phuong, D.; Flower, D. R.; Chattopadhyay, S.; Chattopadhyay, A. K. Towards Effective Consensus Scoring in Structure-Based Virtual Screening. *Interdiscip Sci Comput Life Sci* **2023**, *15* (1), 131–145. https://doi.org/10.1007/s12539-022-00546-8.

(75) Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; Makarenkov, V.; Nahavandi, S. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Information Fusion* **2021**, *76*, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008.

(76) Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q. On Calibration of Modern Neural Networks. arXiv August 3, 2017. https://doi.org/10.48550/arXiv.1706.04599.

(77) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys Rev* **2017**, *9* (2), 91–102. https://doi.org/10.1007/s12551-016-0247-1.

(78) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801. https://doi.org/10.1021/jm0608356.

(79) Réau, M.; Langenfeld, F.; Zagury, J.-F.; Lagarde, N.; Montes, M. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Front. Pharmacol.* **2018**, *9*. https://doi.org/10.3389/fphar.2018.00011.

(80) Stein, R. M.; Yang, Y.; Balius, T. E.; O'Meara, M. J.; Lyu, J.; Young, J.; Tang, K.; Shoichet, B. K.; Irwin, J. J. Property-Unmatched Decoys in Docking Benchmarks. *J. Chem. Inf. Model.* **2021**, *61* (2), 699–714. https://doi.org/10.1021/acs.jcim.0c00598.

(81) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432* (7019), 862–865. https://doi.org/10.1038/nature03197.

(82) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat Rev Drug Discov* **2005**, *4* (8), 649–663. https://doi.org/10.1038/nrd1799.

(83) Hirschfeld, L.; Swanson, K.; Yang, K.; Barzilay, R.; Coley, C. W. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60* (8), 3770–3780. https://doi.org/10.1021/acs.jcim.0c00502.

(84) Shafer, G.; Vovk, V.; Uk, C. R. A. A Tutorial on Conformal Prediction.

(85) Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? arXiv October 5, 2017. https://doi.org/10.48550/arXiv.1703.04977.

(86) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10* (2), 370–377. https://doi.org/10.1039/C8SC04228D.

(87) Piccolo, S. R.; Frampton, M. B. Tools and Techniques for Computational Reproducibility. *Gigascience* **2016**, *5* (1), s13742-016-0135–4. https://doi.org/10.1186/s13742-016-0135-4.

(88) Grüning, B.; Chilton, J.; Köster, J.; Dale, R.; Soranzo, N.; Beek, M. van den; Goecks, J.; Backofen, R.; Nekrutenko, A.; Taylor, J. Practical Computational Reproducibility in the Life Sciences. *cels* **2018**, *6* (6), 631–635. https://doi.org/10.1016/j.cels.2018.03.014.

(89) Peng, R. D. Reproducible Research in Computational Science. *Science* **2011**, *334* (6060), 1226–1227. https://doi.org/10.1126/science.1213847.

(90) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci Data* **2016**, *3* (1), 160018. https://doi.org/10.1038/sdata.2016.18.

(91) Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; Meger, D. Deep Reinforcement Learning That Matters. arXiv January 30, 2019. https://doi.org/10.48550/arXiv.1709.06560.

(92) Yang, S.-Y. Pharmacophore Modeling and Applications in Drug Discovery: Challenges and Recent Advances. *Drug Discovery Today* **2010**, *15* (11), 444–450. https://doi.org/10.1016/j.drudis.2010.03.013.

(93) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacological Reviews* **2014**, *66* (1), 334–395. https://doi.org/10.1124/pr.112.007336.

(94) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand Efficiency: A Useful Metric for Lead Selection. *Drug Discovery Today* **2004**, *9* (10), 430–431. https://doi.org/10.1016/S1359-6446(04)03069-7.

(95) Pinzi, L.; Rastelli, G. Molecular Docking: Shifting Paradigms in Drug Discovery. *International Journal of Molecular Sciences* **2019**, *20* (18). https://doi.org/10.3390/ijms20184331.

(96) Hevener, K. E.; Zhao, W.; Ball, D. M.; Babaoglu, K.; Qi, J.; White, S. W.; Lee, R. E. Validation of Molecular Docking Programs for Virtual Screening against Dihydropteroate Synthase. *J. Chem. Inf. Model.* **2009**, *49* (2), 444–460. https://doi.org/10.1021/ci800293n.

(97) Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E. R. Small-Sample Precision of ROC-Related Estimates. *Bioinformatics* **2010**, *26* (6), 822–830. https://doi.org/10.1093/bioinformatics/btq037.

(98) Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* **2015**, *10* (3), e0118432. https://doi.org/10.1371/journal.pone.0118432.

(99) Freedman, L. P.; Cockburn, I. M.; Simcoe, T. S. The Economics of Reproducibility in Preclinical Research. *PLOS Biology* **2015**, *13* (6), e1002165. https://doi.org/10.1371/journal.pbio.1002165.

(100) Scannell, J. W.; Bosley, J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. *PLOS ONE* **2016**, *11* (2), e0147215. https://doi.org/10.1371/journal.pone.0147215.

(101) Collins, G. S.; Moons, K. G. M.; Dhiman, P.; Riley, R. D.; Beam, A. L.; Calster, B. V.; Ghassemi, M.; Liu, X.; Reitsma, J. B.; Smeden, M. van; Boulesteix, A.-L.; Camaradou, J. C.; Celi, L. A.; Denaxas, S.; Denniston, A. K.; Glocker, B.; Golub, R. M.; Harvey, H.; Heinze, G.; Hoffman, M. M.; Kengne, A. P.; Lam, E.; Lee, N.; Loder, E. W.; Maier-Hein, L.; Mateen, B. A.; McCradden, M. D.; Oakden-Rayner, L.; Ordish, J.; Parnell, R.; Rose, S.; Singh, K.; Wynants, L.; Logullo, P. TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. **2024**. https://doi.org/10.1136/bmj-2023-078378.

(102) Moons, K. G. M.; Damen, J. A. A.; Kaul, T.; Hooft, L.; Navarro, C. A.; Dhiman, P.; Beam, A. L.; Calster, B. V.; Celi, L. A.; Denaxas, S.; Denniston, A. K.; Ghassemi, M.; Heinze, G.; Kengne, A. P.; Maier-Hein, L.; Liu, X.; Logullo, P.; McCradden, M. D.; Liu, N.; Oakden-Rayner, L.; Singh, K.; Ting, D. S.; Wynants, L.; Yang, B.; Reitsma, J. B.; Riley, R. D.; Collins, G. S.; Smeden, M. van. PROBAST+AI: An Updated Quality, Risk of Bias, and Applicability Assessment Tool for Prediction Models Using Regression or Artificial Intelligence Methods. *BMJ* **2025**, *388*, e082505. https://doi.org/10.1136/bmj-2024-082505.

(103) Hernandez-Boussard, T.; Bozkurt, S.; Ioannidis, J. P. A.; Shah, N. H. MINIMAR (MINimum Information for Medical AI Reporting): Developing Reporting Standards for Artificial Intelligence in Health Care. *J Am Med Inform Assoc* **2020**, *27* (12), 2011–2015. https://doi.org/10.1093/jamia/ocaa088.

(104) Norgeot, B.; Quer, G.; Beaulieu-Jones, B. K.; Torkamani, A.; Dias, R.; Gianfrancesco, M.; Arnaout, R.; Kohane, I. S.; Saria, S.; Topol, E.; Obermeyer, Z.; Yu, B.; Butte, A. J. Minimum Information about Clinical Artificial Intelligence Modeling: The MI-CLAIM Checklist. *Nat Med* **2020**, *26* (9), 1320–1324. https://doi.org/10.1038/s41591-020-1041-y.

(105) Sandve, G. K.; Nekrutenko, A.; Taylor, J.; Hovig, E. Ten Simple Rules for Reproducible Computational Research. *PLOS Computational Biology* **2013**, *9* (10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285.

(106) Riley, R. D.; Snell, K. I.; Ensor, J.; Burke, D. L.; Harrell Jr, F. E.; Moons, K. G.; Collins, G. S. Minimum Sample Size for Developing a Multivariable Prediction Model: PART II - Binary and Time-to-Event Outcomes. *Statistics in Medicine* **2019**, *38* (7), 1276–1296. https://doi.org/10.1002/sim.7992.

(107) Van Calster, B.; McLernon, D. J.; van Smeden, M.; Wynants, L.; Steyerberg, E. W.; Bossuyt, P.; Collins, G. S.; Macaskill, P.; McLernon, D. J.; Moons, K. G. M.; Steyerberg, E. W.; Van Calster, B.; van Smeden, M.; Vickers, A. J.; On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: The Achilles Heel of Predictive Analytics. *BMC Med* **2019**, *17* (1), 230. https://doi.org/10.1186/s12916-019-1466-7.

(108) Chen, X.; Yan, C. C.; Zhang, X.; Zhang, X.; Dai, F.; Yin, J.; Zhang, Y. Drug–Target Interaction Prediction: Databases, Web Servers and Computational Models. *Brief Bioinform* **2016**, *17* (4), 696–712. https://doi.org/10.1093/bib/bbv066.

(109) Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating High-Throughput Virtual Screening through Molecular Pool-Based Active Learning. *Chem. Sci.* **2021**, *12* (22), 7866–7881. https://doi.org/10.1039/D0SC06805E.

(110) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of Early Drug Discovery. *British Journal of Pharmacology* **2011**, *162* (6), 1239–1249. https://doi.org/10.1111/j.1476-5381.2010.01127.x.

(111) Steyerberg, E. W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*; Statistics for Biology and Health; Springer International Publishing: Cham, 2019. https://doi.org/10.1007/978-3-030-16399-0.

(112) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat Rev Drug Discov* **2019**, *18* (6), 463–477. https://doi.org/10.1038/s41573-019-0024-5.

(113) Rabbani, N.; Kim, G. Y. E.; Suarez, C. J.; Chen, J. H. Applications of Machine Learning in Routine Laboratory Medicine: Current State and Future Directions. *Clinical Biochemistry* **2022**, *103*, 1–7. https://doi.org/10.1016/j.clinbiochem.2022.02.011.

(114) Xu, B.; Yang, G. Interpretability Research of Deep Learning: A Literature Survey. *Information Fusion* **2025**, *115*, 102721. https://doi.org/10.1016/j.inffus.2024.102721.

(115) King, E.; Aitchison, E.; Li, H.; Luo, R. Recent Developments in Free Energy Calculations for Drug Discovery. *Front. Mol. Biosci.* **2021**, *8*. https://doi.org/10.3389/fmolb.2021.712085.

(116) Wang, Y.; Takaba, K.; Chen, M. S.; Wieder, M.; Xu, Y.; Zhu, T.; Zhang, J. Z. H.; Nagle, A.; Yu, K.; Wang, X.; Cole, D. J.; Rackers, J. A.; Cho, K.; Greener, J. G.; Eastman, P.; Martiniani, S.; Tuckerman, M. E. On the Design Space between Molecular Mechanics and Machine Learning Force Fields. *Appl. Phys. Rev.* **2025**, *12* (2), 021304. https://doi.org/10.1063/5.0237876.

Scannell, J. W.; Bosley, J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. PLoS ONE 2016, 11 (2), e0147215.