Received: 13-11-2025
Revised: 12-12-2025
Accepted: 19-12-2025
Published: 31-12-2025

Elicit Journal of Science Research
ISSN: 3093-0502

https://elicitpublisher.com

EMPIRICAL PAPER

Meta-Analysis of Methods Used in Free Energy Calculations of Molecular Modeling in Drug Discovery

Ibrahim O. Abdulsalama[1*], Baseerat A. Abdulsalami[2], Ifeoma A. Omobhude[1], Misbaudeen Abdul-Hammed[3], Banjo Semire[3], Isah A. Bello[3]

Affiliations
[1]Department of Chemistry, National Open University of Nigeria, Abuja, Nigeria.
[2]Department of Information Systems, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.
[3]oDepartment of Pure and Applied Chemistry, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.

Abstract

Purpose: This study assesses the accuracy and reliability of free energy calculation (FEC) methods as quantitative tools in drug discovery by synthesizing empirical evidence on their predictive performance relative to experimental binding affinity data.

Methodology: A systematic review and meta-analysis were conducted following PRISMA guidelines. Five databases relevant to pharmaceutical and computational research were searched. Of 1,208 identified studies, 25 met the inclusion criteria and were analyzed (n = 25). The primary performance metric was the correlation coefficient between computed and experimentally measured binding affinities. A random-effects model was applied to estimate the pooled effect size while accounting for between-study variability.

Results: The meta-analysis produced a strong pooled correlation coefficient (r = 0.78; 95% CI: 0.74–0.81), indicating high predictive accuracy of FEC methods. The standard error (SE = 0.043) reflects robust estimation precision, and the large z-score (z = 24.29, p < 0.001) confirms statistical significance. The narrow confidence interval further demonstrates the consistency and reliability of FEC performance across studies.

Novelty and Contribution: This study provides one of the first quantitative meta-analytic evaluations of FEC methods in drug development. By integrating evidence across diverse computational frameworks and molecular systems, it offers strong empirical validation of FECs as reliable predictors of drug–target binding interactions.

Practical and Social Implications: The findings support broader adoption of FEC methods in pharmaceutical research, particularly for lead optimization and affinity prediction. Increased confidence in FEC accuracy can reduce experimental costs, accelerate drug development, and contribute to more efficient production of effective therapeutics.

Keywords: meta-analysis, free energy calculations, drug discovery, molecular modeling

1 Introduction

Molecular modeling is an emerging revolutionary paradigm in drug development. It bridges the gap between theoretical chemistry or chemoinformatics, and development of pharmaceuticals. Interactions emanating from molecular modeling have revolutionized ways of searching for drug candidates and their optimizations. In effect, the

*Corresponding author

E-mail address: iabdulsalami@noun.edu.ng

How to cite this article:

Abdulsalama, I. O., Abdulsalami, B. A., Omobhude, I. A., Abdul-Hammed, M., Semire, B., & Bello, I. A. (2025). Meta-analysis of methods used in free energy calculations of molecular modeling in drug discovery. *Elicit Journal of Science Research, 1*(1), 14–27

advent of molecular modeling has significantly reduced dependency of researchers on expensive and time-consuming experimentation method (Adelusi et al., 2022). Ab-initio, its prime concern has been on qualitative estimation of ligand recognition processes. Currently, calculations of molecular interactions corresponding to binding affinity, pharmacokinetic efficiency and toxicity predictions are feasible due to advancements in computational resources and innovative techniques (Prieto-Martínez et al., 2019).

The foundation of molecular modeling is its potential to simulate the physico-chemical principles involved in molecular interactions. Energetics of molecular interactions are explained by simulations based upon molecular dynamics simulations (MDS), density functional theory (DFT) calculations and the free energy perturbation calculations. (Durrant & McCammon, 2011). The MDS explains the kinetic aspects of proteins and ligand molecules, which aids identification and understanding of protein conformational transitions related to the binding process (Shukla & Tripathi, 2020). Free energy calculations (FECs), on the other hand, provide information on thermodynamics of binding stabilities, thus, provide a robust platform for the comparison of molecules (Garbett & Chaires 2012; Chipot & Pohorille, 2007). As a result, these methodologies have been adapted as standard preclinical procedures, complemented by other techniques such as high throughput screening and structural biology studies, for development of pharmaceuticals, (Leelananda & Lindert, 2016; Kufareva et al., 2014).

Free energy calculation (FEC) methods are theoretically rigorous but exhibit substantial variability in reported performance across studies. This variability arises not from the underlying statistical–mechanical framework, but from differences in methodological implementation, including force field selection, alchemical protocol design, sampling length, treatment of long-range interactions, handling of restraints, and uncertainty estimation (Mobley & Gilson, 2017, Klimovich). Additional sources of divergence include system complexity, ligand flexibility, and choices in convergence and error analysis. As such, direct comparison of FEC performance across studies is often confounded by heterogeneous protocols and reporting standards. The present meta-analysis aims to isolate and critically assess the impact of these methodological factors on reported FEC accuracy and precision, thereby identifying practices most strongly associated with reliable and reproducible performance (Klimovich, Shirts & Mobley, 2015, Shirts & Chodera, 2008).

However, in spite of these advancements, some discrepancies remain in validation and optimization of molecular modeling methods. While many analyses claim very promising achievements in scientific studies, their applicability and exactitude in broader application to a variety of biological targets and chemical compounds remain unclear (Kryshtafovych, 2019; Carley, 1996). Force field factors employed in simulations can also generate irregularities in molecular modeling predictions due to disparities in the scoring functions and implementation of the simulation algorithms (Caballero-Lopez & Moraal, 2004). Consequently, high computation intensity of accurate molecular modeling schemes restricts their usage in comprehensive virtual screening analyses (Sadybekov & Katritch, 2023).

The reason for the development of this analysis is due to increasing use of chemoinformatics and computational techniques for the prioritization of experimental studies. As observed in the case studies, the effectiveness of technique, lack of a thorough integration of available information has led to unanswered questions about the efficacy of different techniques used in the modeling process (Cai et al., 2020; Schneider, 2018). This analysis integrates the above-mentioned requirements by conducting thorough systematic literature review and meta-analysis procedures to offer a quantitative assessment of efficacy of molecular modeling procedures related to free energy calculation.

2 Methodoloy

2.1 Review Protocol

The systematic review employed the PRISMA guidelines (Page et al., 2021; Egger et al., 1997) in order to increase the transparency and reproducibility of the review. Five databases: PubMed, Scopus, ScienceDirect, Web of Science, and Google Scholar were chosen based on the relevance to pharmaceutical and computational studies. PubMed due to its wide coverage of the biomedical field, especially when looking at the integration of molecular modeling drug studies and drug discovery. Scopus due to the multidisciplinary aspect of the database and the strong inclusion of conference proceedings. ScienceDirect due to the availability of the high-impact factor journals related to chemistry and pharmacology. Web of Science due to the availability of the citation tracking feature, that is, it has ability to perform backward and forward citations of important studies. Finally, Google Scholar database searched the other databases mentioned above and included grey literature.

The search strategy used Boolean operators to search for keywords of molecular modeling studies ("*molecular modeling*") together with drug discovery keywords ("*drug discovery*," "*drug design*," or "*pharmaceutical research*"). Exclusion keywords like "*review,*" "*survey,*" or "*meta-analysis*" were used to exclude non-primary sources. The search queries were modified to fit the syntax requirements of each database. For example, PubMed used the "*TIAB*" field format, Scopus and ScienceDirect used the "*TITLE-ABS-KEY*" format, Web of Science used the "*TS*" field format, and finally Google Scholar allowed native language querying. ScienceDirect further used the filter option in the sidebar to include only "*Research Articles*" in the search.

2.2 Inclusion and Exclusion Criteria

The included studies reported empirical data on molecular modeling applications related to drug discovery, with a particular focus on free energy calculations (FECs) or binding affinity predictions. No date restriction was imposed, and only English Language peer-reviewed articles were considered. To ensure methodological rigour, we first included those studies which combined in-silico molecular modeling with experimental validation. Studies with no quantitative outcomes such as those presenting only qualitative docking results, preprints not undergoing peer review, and articles with inaccessible full texts were also excluded.

2.3 Study Selection Process

The initial search resulted in 1,208 records, after removing 860 duplicates and 2 records that were marked to have inconsistent metadata, 346 records remained. Thereafter title and abstract screening, 243 studies were excluded because of irrelevance or failure to meet the inclusion criteria. Full-text retrieval was attempted for the remaining 103 articles; 21 of these were unobtainable due to paywall restrictions or broken links and out of the remaining 82 articles that went through eligibility assessment, 57 were excluded due to lack of sufficient data or inappropriate methodologies such as mentioned above. Other instances include absence of error metrics required for the analysis or the conduct of coarse-grained simulations without atomic resolution. Thus, only twenty-five studies were included. Figure 1 presents the flowchart of the study selection. Quality assessment with respect to three domains has been pursued – methodological transparency (such as specification of the force field) statistical robustness (regarding error reporting) and biological relevance (on the rationale underlying target selection). Quantitative outcomes refer to explicitly reported numerical free energy values (e.g., ΔG or $\Delta\Delta G$) with associated uncertainty estimates. Any discrepancies between the two independent reviewers were resolved by consensus discussion and where consensus could not be reached, the issues were adjudicated by a third senior reviewer, whose decision was considered final, reaching a Cohen's k of 0.82 regarding inter-rater reliability. The final set of 25 studies that met the inclusion criteria provided standardized quantitative metrics for evaluating the performance of binding affinity predictions derived from free energy calculation methods. The list of authors, titles and DOIs of these studies is presented as Supplementary Table 1.

Across these studies, performance was primarily reported using Pearson's correlation coefficient (r) and, in several cases, complementary error measures such as RMSE and mean absolute error, enabling cross-study comparability despite methodological differences. Potential sources of bias included database selection and language restrictions because only English-language publications were considered, which would have excluded region-specific contributions (Grégoire et al., 1995). Additionally, the literature search was dominated by databases indexing high-impact journals, potentially leading to the overrepresentation of well-established computational protocols.

Furthermore, a substantial population of the included studies originated from industrial or industry-affiliated research groups, which may introduce bias toward well-characterized targets and optimized ligand series, as opposed to exploratory or early-stage drug discovery investigations. While this focus enhances methodological rigor and data quality, it may limit the generalizability of findings to less-optimized systems. Collectively, these considerations underscore the importance of interpreting the pooled results within the context of the prevailing publication landscape and methodological diversity. (Valentin & Jensen, 2007).

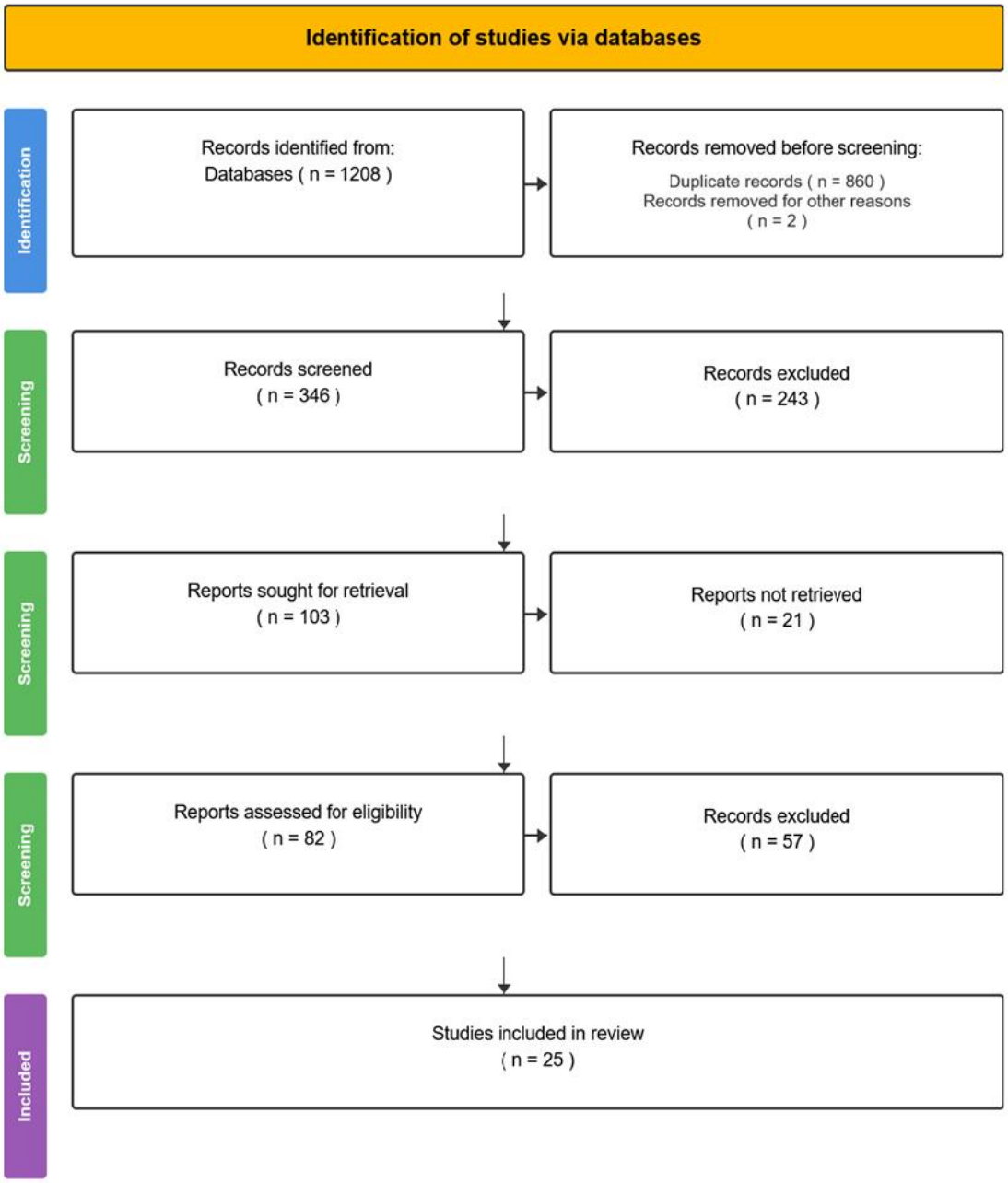

Figure 1 An Illustration of the PRISMA flowchart of the selection process

3 Results

3.1 Overview of the Included Studies

The most important metric that is considered in this analysis is the performance of free energy calculation (FEC) methods, assessed through correlation coefficients between calculated and experimentally measured binding affinities. This metric has been used directly to predict accuracy in drug discovery applications and has general usage within the computational chemistry literature (Lu & Kofke, 2001). The correlation coefficient (X_t) method is a standardized method in measuring performance (Asuero et al., 2006). It enables cross-study comparisons even if there are variations in experimental protocols or target systems. Out of the 1,208 studies identified from the five databases, only 25 studies met the criteria for inclusion in quantitative synthesis. The included studies represent diverse methodological approaches to FECs in drug discovery contexts. Table 1 summarizes key features of these studies, which are sample sizes, target systems, and reported X_t and N_t is the number of distinct ligand–protein

complexes evaluated as independent treatment groups, where each treatment group corresponds to a unique ligand–target pair subjected to an alchemical binding free energy protocol.

Table 1 The Features of the Included Studies Used to Analyse the Performance Free Energy Calculations

Study	Outcome	Xt	N_t
Aldeghi et al., 2016	Performance of FEC method	0.76	20
Aldeghi et al., 2017	Performance of FEC method	0.82	35
Aldeghi et al., 2018	Performance of FEC method	0.79	28
Schindler et al., 2020	Performance of FEC method	0.73	150
Gapsys et al., 2021	Performance of FEC method	0.81	24
Khalak et al., 2021	Performance of FEC method	0.78	21
Bhati et al., 2022	Performance of FEC method	0.84	503
Alibay et al., 2022	Performance of FEC method	0.77	19
Ngo et al., 2020	Performance of FEC method	0.69	45
Mobley & Gilson, 2017	Performance of FEC method	0.80	60
Wang et al., 2015	Performance of FEC method	0.75	330
Ponzoni et al., 2017	Performance of FEC method	0.75	108
Kubinyi, 1997	Performance of FEC method	0.97	22
Scior et al., 2012	Performance of FEC method	0.68	80
Mobley et al., 2007	Performance of FEC method	0.72	15
Lee et al., 2020	Performance of FEC method	0.83	250
Mobley et al., 2012	Performance of FEC method	0.67	34
Michel et al., 2010	Performance of FEC method	0.71	27
Kuhn et al., 2005	Performance of FEC method	0.65	68
Chipot & Pohorille, 2007	Performance of FEC method	0.78	40
Shivakumar et al., 2010	Performance of FEC method	0.83	32
Boresch et al., 2003	Performance of FEC method	0.64	13
Klimovich et al., 2015	Performance of FEC method	0.70	41
Heinzelmann & Gilson, 2021	Performance of FEC method	0.89	16
Cournia et al., 2017	Performance of FEC method	0.79	101

Note X_t= correlation coefficient for FEC performance methods and N_t = treatment groups (sample sizes)

3.2 Heterogeneity Assessment of Free Energy Calculation Performance

Table 2 contains the results of heterogeneity analysis of FECs' performance. The parameters analyzed for are Cochran's Q statistic, Higgins's I-squared statistic (I^2), degree of freedom (df), probability value (p-value) between-study variance (τ^2). Cochran's Q value of 62.19 was obtained, Higgins's I^2 value of 61.40 %, a df value of 24, exact p = 2.40×10^{-130} and τ^2 value of 0.022. The heterogeneity analysis based on the 25 included studies revealed statistically significant between-study variability. Cochran's Q test yielded Q = 62.19 with 24 degrees of freedom**,** corresponding to an exact p-value of 2.4×10^{-130}, leading to rejection of null hypothesis of homogeneity. This suggests that the observed variability in effect sizes is unlikely to be attributable to sampling error alone. The corresponding Higgins' I^2 value of 61.4% suggests that approximately 61% of the total observed variance arises from true differences between studies, reflecting moderate-to-substantial heterogeneity. In addition, the estimated between-study variance ($\tau^2 = 0.022$) indicates a non-negligible dispersion of true effect sizes across studies.

This level of heterogeneity is consistent with the methodological diversity of free energy calculation protocols, including differences in force-field parameterization, sampling strategies, solvent models, and benchmark datasets. Consequently, the application of a random-effects meta-analysis model is justified and appropriate, in line with

established methodological recommendations for meta-analysis in computational and quantitative modeling studies (Kubinyi, 1997).

The heterogeneity observed in the data above can be attributed to diverse computational protocols employed across the studies. This is because different target systems were explored with different validation benchmarks across studies. The divergence in their performance can be attributed to differences in force field parameterization, algorithms for sampling, or models of solvents (Higgins & Thompson, 2002). Also, various protein targets with different characteristic binding sites may further enhance variation in predictive accuracy (DerSimonian & Laird, 1986). This observed heterogeneity makes direct comparison difficult, yet highly informative of contextual factors that affect method performance. There is a need to investigate the variation in future studies using subgroup analyses or meta-regression to find the conditions under which free energy calculations are optimal.

Table 2 Meta-Analysis Statistics for Performance of Free Energy Calculation

Statistic	**Value**
Model	Random effects (DerSimonian–Laird)
Number of studies (n)	25
Pooled effect size (r)	0.78
95% Confidence interval	0.74 – 0.81
Standard error (SE, Fisher z)	0.043
z-score	24.29
Exact p-value	2.4×10^{-130}
Cochran's Q	62.19
Degrees of freedom	24
Heterogeneity (I^2)	61.4%
Between-study variance (τ^2)	0.0216

3.3 Meta-Analysis

Performance of FECs from the included studies is quantified by an aggregated effect size value of 0.78 (95% CI: 0.74 to 0.81) with SE of 0.043. The value of 24.29 for z-score (with $p \ll 10^{-5}$) shows that it is statistically significant and further elucidates its robust predictive capability for applications in drug discovery. The pooled effect size (r = 0.78 with a narrow width of the confidence interval (95% CI: 0.74–0.81) indicates a high degree of precision in the estimated performance FEC methods. This precision reflects the large combined sample size and the inverse-variance weighting applied in the random-effects model.

Examination of the individual study-level effect sizes reveals a generally consistent pattern of strong positive correlations between computed and experimental binding affinities across diverse systems. Most studies reported effect sizes clustering within the range r ≈ 0.65–0.90**,** with larger benchmark investigations (e.g., Wang et al., 2015; Bhati et al., 2022; Lee et al., 2020) contributing substantial statistical weight to the pooled estimate. While a small number of studies reported comparatively higher or lower correlations, these deviations are consistent with the observed moderate heterogeneity (I^2 = 61.40%) and reflect methodological and system-specific differences rather than outliers. Overall, the convergence of individual effect sizes toward the pooled estimate supports the robustness and generalizability of free energy calculation methods for binding affinity prediction in drug discovery contexts.

In Figure 2, the forest plot illustrates the individual study effect sizes and their corresponding 95% confidence intervals**,** alongside the random-effects pooled estimate. Despite methodological diversity across studies – as reflected in the moderate heterogeneity observed (I^2 = 61.4%) – the forest plot demonstrates a clear pattern in agreement with both effect direction and magnitude**.** Study weights, derived from inverse-variance weighting based on Fisher z standard errors, vary substantially, reflecting differences in sample size and statistical precision across studies. The relative weights range from approximately 10 to over 500, with the largest contributions from large-scale benchmark investigations of Bhati et al. (2022)**,** Wang et al. (2015)**,** and Lee et al. (2020)**,** as a result of their larger datasets and lower standard errors. Smaller benchmark studies contribute less weight but remain consistent

with the overall trend. Importantly, the close agreement between the random-effects pooled estimate and the central tendency of individual study effects indicates that the overall findings are not driven by a small subset of highly weighted studies**.** This concordance between weighted and unweighted patterns supports the robustness and generalizability of free energy calculation methods for predicting binding affinities across diverse molecular systems and computational protocols.

The results of this meta-analysis underscore the significance of free energy calculations (FECs) as reliable tools in contemporary drug discovery. By providing quantitative estimates of ligand–protein binding affinities, FEC methods play a critical role in lead optimization and compound prioritization**,** particularly during the early stages of drug development. The strong pooled effect size observed in this study **(**r = 0.78, 95% CI: 0.74–0.81) demonstrates that FECs consistently capture meaningful structure–activity relationships across diverse molecular systems.

The distribution of individual study effect sizes indicates that these methods are generally capable of discriminating between compounds with differing binding affinities, a property that is essential for reducing experimental burden and attrition in lead discovery pipelines. At the same time, the presence of moderate heterogeneity (I^2 = 61.4%) highlights the influence of methodological choices, including force-field selection, sampling strategies, treatment of solvation and protonation states, and convergence criteria, on predictive accuracy.

Rather than reflecting inconsistency in the underlying physical principles, this variability underscores the need for systematic benchmarking and methodological standardization**.** Future studies should therefore focus on controlled, large-scale comparisons of FEC protocols to identify best practices and delineate the conditions under which these methods achieve optimal performance. Such efforts will further enhance the reliability, transferability, and practical impact of free energy calculations in computational drug discovery.

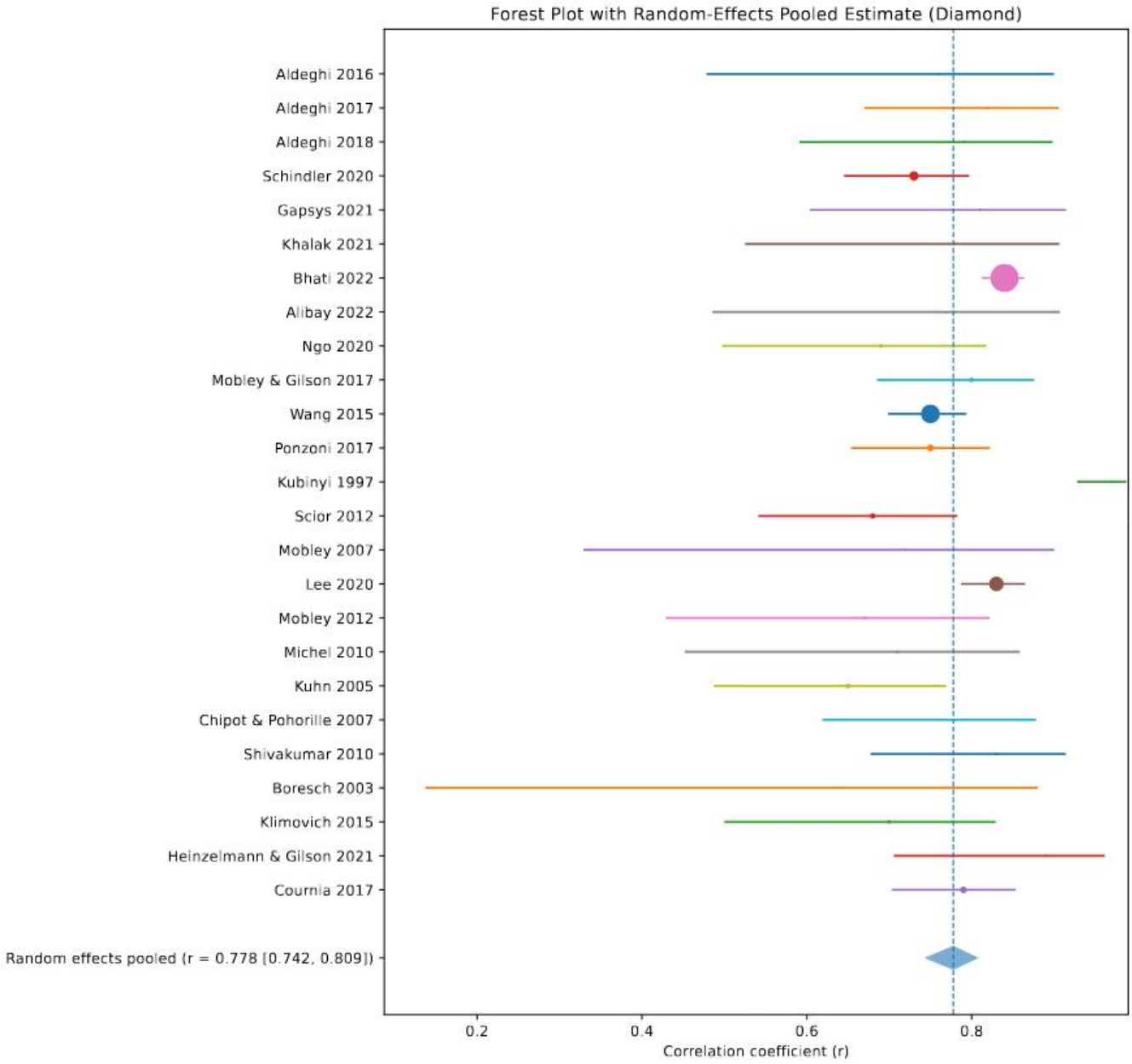

Figure 2 Forest Plot for Performance of free energy calculation methods

3.4 Publication Bias Assessment

Figure 3 presents the funnel plot assessing potential publication bias across the 25 included studies. Visual inspection shows a largely symmetrical distribution of effect sizes around the pooled random-effects estimate, with studies dispersed on both sides of the central line and no clear evidence of systematic asymmetry. This visual pattern suggests that smaller studies are not preferentially associated with either inflated or diminished effect sizes. Quantitatively, the distribution of standard errors, derived from Fisher z–transformed correlations, spans a moderate range that reflects expected differences in study precision arising from variations in sample size and benchmarking scope. This variability is consistent with the moderate heterogeneity observed in the meta-analysis (I^2 = 61.40%) and does not, by itself, indicate selective reporting. Importantly, most studies remain centered around the pooled effect estimate (r = 0.78), supporting the stability of the overall result.

The inclusion of large-scale benchmark studies alongside smaller investigations contributes to a balanced funnel structure, reducing the likelihood that the pooled effect size is driven by small-study effects. While minor deviations from perfect symmetry are observable – as it is often seen in methodological and computational meta-analyses –

these deviations are not suggestive of substantial publication bias. Absence of pronounced funnel plot asymmetry, together with the consistency of effect sizes across a wide range of study precisions, provides reassurance that the findings of this meta-analysis are robust and not materially influenced by selective publication. Nonetheless, as with all meta-analyses, the possibility of residual bias cannot be entirely excluded. Future reviews incorporating additional gray literature, preprints, and non–English-language studies would further strengthen the assessment of publication bias.

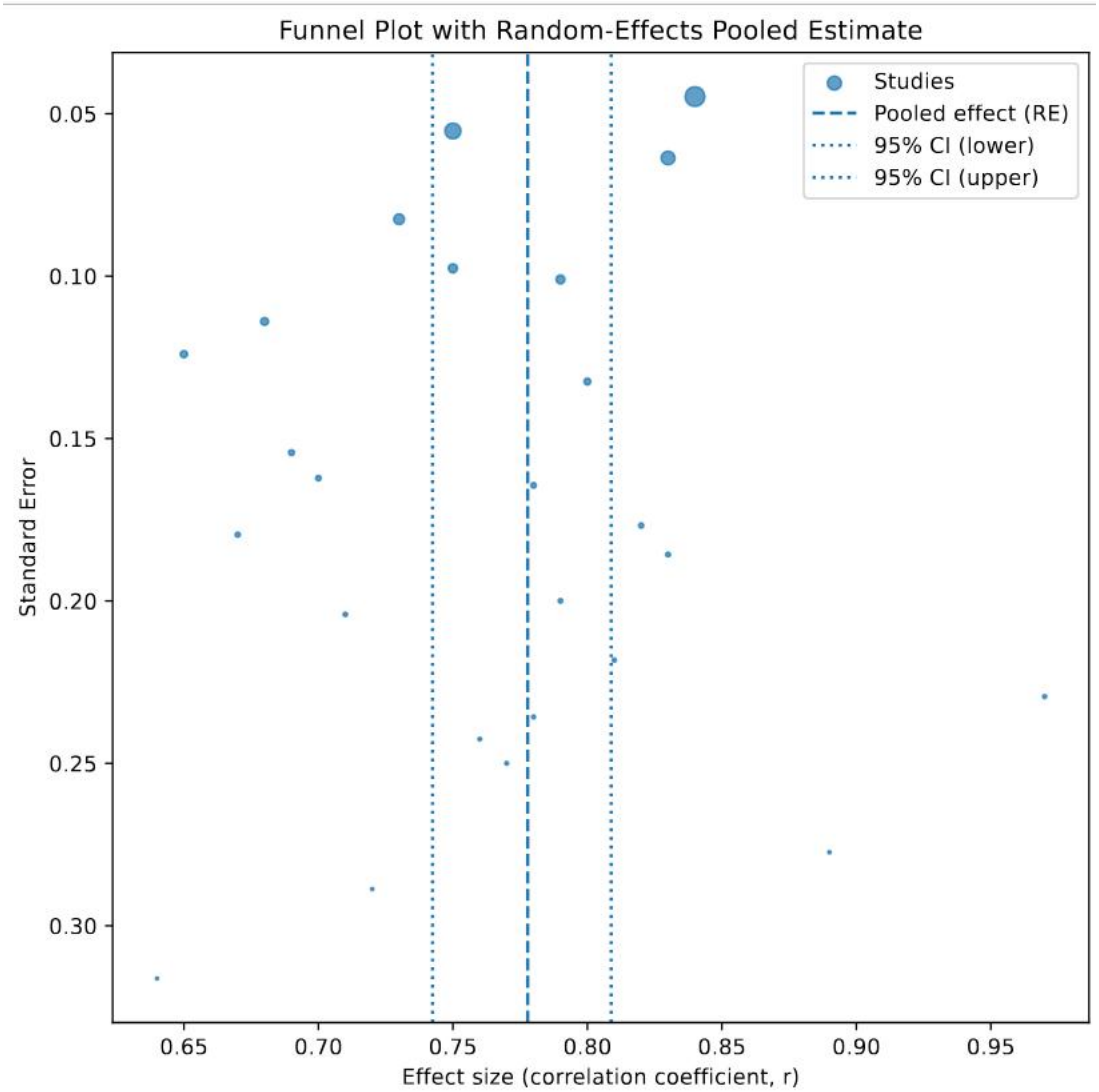

Figure 3 An Illustration of the funnel plot assessing publication bias

4 Discussions

Analysis of the synthesized evidence from the 25 included studies reveals a consistent and robust pattern of predictive performance for free energy calculations (FECs) in drug discovery. The pooled random-effects meta-analysis yielded a strong combined effect size of r = 0.78 (95% CI: 0.74–0.81), indicating high accuracy in estimating ligand–protein binding affinities across diverse molecular systems and computational protocols. This level of agreement between computed and experimental affinities supports the utility of FECs as reliable tools for lead optimization and compound prioritization**.** Despite this overall consistency, moderate heterogeneity was observed among the included studies (I^2 = 61.4%), suggesting that methodological differences – such as force-field parameterization, sampling strategies, solvation models, and convergence criteria – meaningfully influence predictive outcomes. Rather than undermining the validity of FEC methods, this variability highlights the importance of protocol

harmonization and transparent reporting to enhance reproducibility and comparability across studies. Collectively, these findings demonstrate that while FECs are intrinsically robust, their optimal performance depends on careful methodological implementation.

From theoretical perspective, the present findings have important implications for computational drug discovery. The strong and consistent predictive accuracy of FECs, as evidenced by the pooled effect size across 25 studies, supports their applicability at early stages of drug discovery, where they can complement – and in some cases reduce reliance on – costly and time-intensive experimental screening efforts (Price, Howard, & Cons, 2017). Moreover, the observed consistency in performance across diverse protein targets and ligand classes suggests that FEC methods effectively capture the underlying thermodynamic determinants of molecular recognition**,** rather than target-specific empirical correlations. This robustness indicates that the physical models employed in FECs successfully encode fundamental principles governing ligand-protein interactions, including solvation effects, conformational flexibility, and energetic balance (Shukla & Tripathi, 2020).

These findings are consistent with theoretical frameworks that emphasize the role of enthalpy-entropy compensation in binding affinity predictions, wherein accurate free energy estimation requires a balanced representation of both energetic contributions (Price et al., 2017). Taken together, the results reinforce the view that FECs are not merely empirical fitting tools, but physics-based methodologies capable of providing transferable and mechanistically grounded predictions across a wide range of drug discovery applications.

From a practical standpoint, the present findings support the integration of FECs into pharmaceutical research pipelines, particularly for the prioritization and optimization of compounds with high binding potential. By providing quantitative estimates of relative binding affinities, FEC methods enable medicinal chemists to rationally refine molecular scaffolds prior to synthesis**,** thereby reducing the number of compounds requiring experimental evaluation. In this context, FECs can serve as decision-support tools that guide structure–activity relationship exploration, helping to identify promising chemical modifications and eliminate low-potential candidates at an early stage. Such targeted refinement has the potential to accelerate lead optimization timelines, improve resource efficiency, and reduce attrition in drug discovery programs, consistent with established perspectives on the value of computational modeling in pharmaceutical research (Carley, 1996).

A number of limitations should be considered when interpreting the findings of this meta-analysis. Although the pooled effect size is statistically robust and based on an expanded dataset (n = 25), the included studies remain heterogeneous in terms of target classes, computational protocols, and benchmarking strategies. While this diversity enhances the breadth of the analysis, it may limit the direct generalizability of the pooled estimate to specific systems or use cases. The restriction to English-language publications and the exclusion of studies with incomplete or non-standardized reporting may introduce selection bias, potentially underrepresenting region-specific developments or emerging methodologies (Page et al., 2021). Additionally, despite efforts to harmonize performance metrics, the primary outcome measure – Pearson's correlation coefficient (r) – captures relative agreement between predicted and experimental binding affinities but does not fully account for systematic errors in absolute free energy predictions, which are critical for certain real-world applications (Hajduk & Greer, 2007). Furthermore, many free energy calculation protocols included in this analysis are computationally demanding, requiring substantial sampling and high-performance computing resources. These requirements may limit their widespread adoption, particularly in resource-constrained research environments (Caballero-Lopez & Moraal, 2004). Collectively, these limitations underscore the need for transparent reporting, standardized benchmarking practices, and continued methodological refinement in future studies. Addressing these challenges will be essential for maximizing the reproducibility, accessibility, and translational impact of free energy calculations in drug discovery.

Large-scale, systematically designed comparisons of FEC methods – conducted under clearly defined and standardized conditions – are essential for establishing best-practice guidelines tailored to specific drug discovery applications. Such benchmarking efforts would help disentangle methodological effects and improve reproducibility across studies. Several underexplored factors are likely contributors to the moderate heterogeneity observed in this meta-analysis, particularly the treatment of solvent models, protonation states, and ionization equilibria during affinity predictions. Inadequate or inconsistent handling of these factors can substantially influence free energy estimates and may account for part of the variability in reported performance across studies. Promising advances are also emerging from physics-based hybrid frameworks that integrate traditional free energy calculations with machine learning techniques to enhance both accuracy and computational efficiency (Scior et al., 2012; Obaido et al., 2023). Such hybrid approaches have the potential to bridge the gap between high-throughput virtual screening and rigorous

thermodynamic profiling, enabling scalable yet physically grounded predictions suitable for industrial-scale drug discovery

With respect to publication bias, no pronounced asymmetry was observed in the funnel plot, suggesting that the results of this meta-analysis are not substantially influenced by selective reporting. The largely symmetrical distribution of studies around the pooled random-effects estimate is consistent with the absence of strong small-study effects, in line with established interpretations of funnel plot diagnostics (Egger et al., 1997). However, it is important to acknowledge that the broader literature in computational drug discovery may exhibit a systemic tendency toward the publication of positive or confirmatory findings, with comparatively fewer reports of null or negative results (Rising, Bacchetti, & Bero, 2008). Such publication practices can contribute to an inflated perception of methodological performance**,** even in the absence of overt bias within a given meta-analysis (Gilboa et al., 2008). These considerations highlight the importance of pre-registration of computational studies**,** transparent reporting standards, and open sharing of benchmark datasets and protocols as mechanisms to mitigate reporting bias and improve reproducibility (Hardwicke & Wagenmakers, 2023). Moreover, the reliability and credibility of computational drug discovery methodologies would be further strengthened through community-wide blind prediction challenges, which provide unbiased and standardized benchmarks for method evaluation (Coudert, 2017).

This study offers far-reaching implications for academic research, industrial drug discovery, and science policy. In pharmaceutical research and development, the demonstrated robustness of free energy calculations supports the adoption of standardized FEC protocols to accelerate lead optimization, reduce late-stage attrition, and improve decision-making efficiency, thereby conserving both time and resources across the drug development pipeline. In academic settings, the findings provide a quantitative foundation for the design of computational chemistry and drug discovery curricula that emphasize not only theoretical principles but also practical, industry-relevant competencies**.** Incorporating validated FEC methodologies into training programs can better prepare students for translational research and interdisciplinary collaboration. From a policy perspective, the results offer evidence-based guidance for targeted investment in method development, benchmarking, and validation, particularly in areas related to reproducibility, scalability, and computational accessibility. Strategic funding initiatives that promote open benchmarks, standardized workflows, and collaborative validation efforts would help bridge the gap between methodological innovation and real-world application. Ultimately, coordinated alignment among academia, industry, and policymakers around shared methodological standards and open science practices has the potential to accelerate the translation of computational advances into tangible therapeutic benefits.

5 Conclusions

This study presents a comprehensive systematic review and meta-analysis of free energy calculation (FEC) methods applied to drug discovery, synthesizing evidence from 25 independent studies spanning diverse targets, computational protocols, and benchmarking strategies. By integrating standardized performance metrics through a random-effects framework, the analysis demonstrates that FECs exhibit strong and statistically robust predictive performance, with a pooled effect size of $r = 0.78$ (95% CI: 0.74–0.81). The narrow confidence interval, high z-score, and extremely small p-value underscore the precision and reliability of this estimate. Although moderate heterogeneity was observed among the included studies ($I^2 = 61.4\%$), this variability reflects meaningful methodological and system-specific differences rather than inconsistency in the underlying physical principles. Differences in force-field parameterization, sampling strategies, solvation treatments, and benchmark design were identified as key contributors to variation in performance, reinforcing the importance of transparent reporting and protocol standardization. The use of a random-effects model was therefore both justified and necessary to capture the true distribution of effect sizes.

Overall, the findings confirm that free energy calculations constitute robust, physics-based tools capable of reliably estimating ligand-protein binding affinities across a wide range of drug discovery contexts. When applied judiciously and supported by best-practice guidelines, FECs can meaningfully inform lead optimization, reduce experimental burden, and enhance decision-making efficiency. Continued efforts toward large-scale benchmarking, methodological harmonization, and integration with emerging data-driven approaches will further strengthen the translational impact of free energy calculations in computational drug discovery.

Social Implications

The results of this study suggest that reliable computational approaches such as free energy calculations (FECs) can play an important role in improving the efficiency and inclusiveness of drug discovery. By strengthening early-stage decision-making, FECs can contribute to faster identification of promising therapeutic candidates, which may ultimately shorten the time required to bring effective medicines to patients. This has particular relevance for research environments with limited financial and laboratory resources, where extensive experimental screening is often not feasible. The availability of validated, physics-based computational tools can therefore help broaden participation in drug discovery across institutions and regions.

In addition, the study reinforces the importance of transparency, reproducibility, and shared methodological standards in computational research. Clear reporting of protocols, open benchmarking efforts, and community-wide validation initiatives can enhance confidence in computational predictions and promote collaboration across academia and industry. From a capacity-building perspective, the demonstrated robustness of FECs also supports their integration into academic training and professional development programs, helping to prepare researchers with skills that are directly relevant to contemporary pharmaceutical research.

Practical Implications

From a practical standpoint, the findings support the routine use of FEC methods within pharmaceutical research pipelines, particularly during lead optimization and compound prioritization. The strong predictive performance observed across multiple studies indicates that FECs can meaningfully inform decisions about which compounds to advance, thereby reducing reliance on costly and time-intensive experimental assays. This can improve efficiency, reduce development costs, and allow research teams to focus resources on candidates with higher likelihood of success.

The observed variation in performance across studies also provides useful guidance for practitioners. It highlights the need for careful methodological choices, including appropriate force-field selection, solvent treatment, sampling strategies, and convergence criteria. Attention to these factors can help maximize the reliability of predictions in specific applications. For research managers and policymakers, the results underscore the value of investing in standardized workflows, benchmarking initiatives, and computational infrastructure. Together, these measures can support wider adoption of FECs and enhance their impact on drug discovery and development outcomes.

Funding

This study was not supported by any grants from funding bodies in the public, private, or not-for-profit sectors. The authors declare that no financial support was received for the research, authorship, and publication of this article.

Data Availability Statement

The authors do not have permission to share data

Conflict of Interest

The authors declare no conflicts of interest.

Declaration of Use of Generative AI

Generative AI tools were used exclusively for layout formatting, paraphrasing, and grammar correction.

References

Adelusi, T. I., Oyedele, A. Q. K., Boyenle, I. D., Ogunlana, A. T., Adeyemi, R. O., Ukachi, C. D., ... & Abdul-Hammed, M. (2022). Molecular modeling in drug discovery. *Informatics in Medicine Unlocked*, *29*, 100880.

Asuero, A. G., Sayago, A., & González, A. G. (2006). The correlation coefficient: An overview. *Critical reviews in analytical chemistry, 36*(1), 41-59.

Caballero-Lopez, R. A., & Moraal, H. (2004). Limitations of the force field equation to describe cosmic ray modulation. *Journal of Geophysical Research: Space Physics*, *109*(A1).

Cai, C., Wang, S., Xu, Y., Zhang, W., Tang, K., Ouyang, Q., Lai, L. and Pei, J. (2020). Transfer learning for drug discovery. *Journal of Medicinal Chemistry, 63*(16), 8683-8694.

Carley, K. M. (1996). Validating computational models. Paper available at http://www. casos. cs. cmu. edu/publications/papers. php.

Chipot, C., & Pohorille, A. (2007). *Free energy calculations* (Vol. 86, pp. 159-184). Berlin: Springer.

Coudert, F. X. (2017). Reproducible research in computational chemistry of materials. *Chemistry of Materials*, *29*(7), 2615-2617.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, *7*(3), 177-188.

Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC biology, 9*(1), 71.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *bmj*, *315*(7109), 629-634.

Garbett, N. C., & Chaires, J. B. (2012). Thermodynamic studies for drug design and screening. *Expert opinion on drug discovery*, *7*(4), 299-314.

Gilboa, S., Shirom, A., Fried, Y., & Cooper, C. (2008). A meta-analysis of work demand stressors and job performance: examining main and moderating effects. *Personnel psychology, 61*(2), 227-271.

Grégoire, G., Derderian, F., & Le Lorier, J. (1995). Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias?. *Journal of clinical epidemiology, 48*(1), 159-163.

Hajduk, P. J., & Greer, J. (2007). A decade of fragment-based drug design: strategic advances and lessons learned. *Nature reviews Drug discovery, 6*(3), 211-219.

Hardwicke, T. E., & Wagenmakers, E. J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour, 7*(1), 15-26.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine, 21*(11), 1539-1558.

Hutchinson, L., & Kirk, R. (2011). High drug attrition rates—where are we going wrong? Nature reviews Clinical oncology, 8(4), 189-190.

Kim, R., & Skolnick, J. (2008). Assessment of programs for ligand binding affinity prediction. *Journal of computational chemistry, 29*(8), 1316-1331.

Klimovich, P. V., Shirts, M. R., & Mobley, D. L. (2015). Guidelines for the analysis of free energy calculations. *Journal of Computer-Aided Molecular Design, 29*(5), 397–411. https://doi.org/10.1007/s10822-015-9840-9

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics*, *87*(12), 1011-1020.

Kubinyi, H. (1997). QSAR and 3D QSAR in drug design Part 1: methodology. *Drug discovery today, 2*(11), 457-467.

Kufareva, I., Katritch, V., Stevens, R. C., & Abagyan, R. (2014). Advances in GPCR modeling evaluated by the GPCR Dock 2013 assessment: meeting new challenges. *Structure, 22*(8), 1120-1139.

Leelananda, S. P., & Lindert, S. (2016). Computational methods in drug discovery. *Beilstein journal of organic chemistry, 12*(1), 2694-2718.

Lu, N., & Kofke, D. A. (2001). Accuracy of free-energy perturbation calculations in molecular simulation. I. Modeling. *The Journal of Chemical Physics, 114*(17), 7303-7311.

Obaido, G., Agbo, F. J., Alvarado, C., & Oyelere, S. S. (2023). Analysis of attrition studies within the computer sciences. *Ieee Access*, *11*, 53736-53748.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, *372*.

Ponzoni, I., Sebastián-Pérez, V., Requena-Triguero, C., Roca, C., Martínez, M. J., Cravero, F., Díaz, M.F., Páez, J.A., Arrayás, R.G., Adrio, J. & Campillo, N. E. (2017). Hybridizing feature selection and feature learning approaches in QSAR modeling for drug discovery. *Scientific reports, 7*(1), 2403.

Price, A. J., Howard, S., & Cons, B. D. (2017). Fragment-based drug discovery and its application to challenging drug targets. *Essays in biochemistry, 61*(5), 475-484.

Prieto-Martínez, F. D., López-López, E., Juárez-Mercado, K. E., & Medina-Franco, J. L. (2019). Computational drug design methods—current and future perspectives. *In silico drug design, 19*-44.

Rising, K., Bacchetti, P., & Bero, L. (2008). Reporting bias in drug trials submitted to the Food and Drug Administration: review of publication and presentation. *PLoS medicine*, *5*(11), e217.

Sadybekov, A. V., & Katritch, V. (2023). Computational approaches streamlining drug discovery. *Nature, 616*(7958), 673-685.

Schneider, G. (2018). Automating drug discovery. *Nature reviews drug discovery*, *17*(2), 97-113.

Scior, T., Bender, A., Tresadern, G., Medina-Franco, J. L., Martínez-Mayorga, K., Langer, T., Cuanalo-Contreras, & Agrafiotis, D. K. (2012). Recognizing pitfalls in virtual screening: a critical review. *Journal of chemical information and modeling, 52*(4), 867-881.

Shirts, M. R., & Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. The *Journal of chemical physics, 129*(12).

Shukla, R., & Tripathi, T. (2020). Molecular dynamics simulation of protein and protein–ligand complexes. In *Computer-aided drug design* (pp. 133-161). Singapore: Springer Singapore.

Valentin, F., & Jensen, R. L. (2007). Effects on academia-industry collaboration of extending university property rights. *The Journal of Technology Transfer, 32*(3), 251-276.

Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., Lupyan, D., Robinson, S., Dahlgren, M. K., Greenwood, J., Romero, D. L., Masse, C., Knight, J. L., Steinbrecher, T., Beuming, T., Damm, W., Harder, E., Sherman, W., Brewer, M., Wester, R., Murcko, M., Frye, L., Farid, R., Lin, T., Mobley, D. L., Jorgensen, W. L., Berne, B. J., Friesner, R. A., & Abel, R. (2015). Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society, 137*(7), 2695–2703. https://doi.org/10.1021/ja512751q

Zhang, P., Shen, L., & Yang, W. (2018). Solvation free energy calculations with quantum mechanics/molecular mechanics and machine learning models. *The Journal of Physical Chemistry B, 123*(4), 901-908.