

CADD-Driven Molecular Design: Computational Strategies for the Discovery and Optimization of Bioactive Molecules

Anitha Sadula*, Usha Rani Peddaboina*, Aruna Devi P*, Pendyala Meghana, Thahera Shaik, Sandhya Rani Moluguri, Pranathi Lakkakula

Department of Pharmacy, University College of Technology (A), Osmania University, Hyderabad, 500007, Telangana, India.

* Corresponding author

DOI: <https://doi.org/10.51244/IJRSI.2026.1309000152>

Received: 25 September 2026; Accepted: 30 September 2026; Published: 14 October 2026

ABSTRACT

Computer-Aided Drug Design (CADD) has become a widely adopted and, in many documented cases, demonstrably valuable component of pharmaceutical chemistry, enabling the rational, in silico modeling, prediction, and optimization of bioactive molecules before the commitment of synthetic and biological resources. Over the past four decades, CADD has evolved from simple two-dimensional structure-activity correlations into a mature, multidisciplinary science integrating structural biology, statistical mechanics, medicinal chemistry, pharmacokinetics, and, most recently, artificial intelligence. This review provides a comprehensive treatment of the theoretical foundations, computational methodologies, and practical applications that define modern CADD, beginning with its historical evolution and foundational concepts-structure-activity relationships, molecular interactions, descriptors, and energy minimization-before examining the two principal design philosophies: structure-based drug design (SBDD) and ligand-based drug design (LBDD). Core techniques including molecular docking, molecular dynamics simulation, binding free-energy calculation, pharmacophore modeling, and QSAR analysis are described with attention to their algorithms, assumptions, and practical execution. The review further explores computational ADME and toxicity prediction, lead identification and optimization, applications across anticancer, antimicrobial, antiviral, central nervous system, and cardiovascular drug discovery, and computational drug repurposing. A dedicated discussion addresses integration of CADD with experimental pipelines through the design-make-test-analyze cycle, and a substantial section is devoted to artificial intelligence and machine learning, including deep learning-based property prediction, AI-augmented virtual screening, and generative de novo molecular design. Emerging frontiers-quantum chemical calculations, multi-parameter optimization, digital twins, and integration of omics data-are surveyed alongside a critical appraisal of scoring-function accuracy, computational cost, and reproducibility challenges that continue to constrain the field. Throughout, established, well-replicated methodology is distinguished from emerging approaches whose evidence base remains preliminary, and a comparative summary of method-level advantages, limitations, computational cost, and validation status is provided to support critical, rather than purely descriptive, evaluation. Collectively, this review aims to give researchers, students, and practitioners an integrated understanding of how computational strategies are discovering and optimizing the bioactive molecules of today and tomorrow.

Keywords: Computer-Aided Drug Design (CADD); Molecular Docking; Molecular Dynamics Simulation; Structure-Based Drug Design; Ligand-Based Drug Design; QSAR; Pharmacophore Modeling; ADMET Prediction; Virtual Screening; Artificial Intelligence in Drug Discovery; Lead Optimization; De Novo Drug Design; Binding Free Energy; Drug Repurposing

INTRODUCTION

The discovery and development of a single new pharmaceutical agent remains one of the most complex, costly, and time-consuming endeavors in applied science. Contemporary estimates place the average cost of bringing a new molecular entity from discovery to market approval at well over two billion dollars, with a

typical timeline of ten to fifteen years [1]. Compounding this is a discouragingly high attrition rate: the overwhelming majority of compounds entering preclinical development never reach the clinic, and only a small fraction of those entering clinical trials are ultimately approved, most commonly due to unforeseen toxicity, poor pharmacokinetics, or lack of efficacy [2]. These statistics have driven pharmaceutical and academic research toward increasingly rational, hypothesis-driven strategies capable of anticipating such failures earlier in the discovery process.

Computer-Aided Drug Design (CADD) has emerged as the principal response to this challenge. CADD encompasses computational chemistry, structural biology, bioinformatics, statistical modeling, and, increasingly, artificial intelligence, applied collectively to rationalize, predict, and accelerate the identification and optimization of bioactive molecules [3]. Rather than relying solely on iterative, empirical trial-and-error synthesis and testing—a paradigm that dominated medicinal chemistry through much of the twentieth century—CADD allows researchers to formulate testable hypotheses about how a molecule will bind its target, how it will behave pharmacokinetically, and how it might be modified to improve potency, selectivity, or safety, before the laboratory bench.

The scale of the problem CADD addresses is worth appreciating quantitatively. Drug-like chemical space—synthetically feasible small molecules satisfying basic physicochemical criteria for oral bioavailability—has been estimated at 10^{23} to 10^{60} unique structures, vastly exceeding what could ever be synthesized and tested experimentally [4]. Computational methods make it possible to search, filter, and prioritize this space intelligently, focusing synthetic and biological effort on the small fraction of molecules with a reasonable probability of activity against a given target.

CADD is not a single technique but an integrated ecosystem of complementary approaches. At one end lie physics-based methods—molecular mechanics, docking, and molecular dynamics—which explicitly model the three-dimensional structure and energetics of molecular interactions. At the other lie statistical and data-driven methods—QSAR, pharmacophore modeling, and increasingly machine learning—which infer structure-activity patterns directly from empirical data without necessarily requiring a three-dimensional target structure. The choice and combination of approaches depends on the nature and quality of available data, whether a validated target structure exists, and the specific stage of the discovery pipeline being addressed.

This review guides the reader through the conceptual foundations of CADD, the major structure-based and ligand-based methodologies, their integration with computational ADMET prediction, their application across major therapeutic areas, and the evolving influence of artificial intelligence on the field, with emphasis throughout on how these techniques are integrated with one another and with experimental validation to form a coherent, iterative drug discovery strategy.

1.1 Literature Search and Review Methodology

This article is a narrative, educational review rather than a systematic review; its scope was defined a priori to provide comprehensive, textbook-style coverage of CADD methodology and application rather than to exhaustively pool primary literature on a single narrow question, as a systematic review or PRISMA-type meta-analysis would. Readers seeking a quantitative, systematic synthesis of a specific CADD sub-topic are directed to the specialized systematic reviews and benchmarking studies cited within the relevant sections below.

Within this narrative framework, a defined and reportable search process was followed. Literature was identified through structured searches of PubMed, Scopus, Web of Science, and Google Scholar, supplemented by targeted searches of the preprint server arXiv for recent machine-learning methodology, using combinations of terms including “computer-aided drug design,” “structure-based drug design,” “ligand-based drug design,” “molecular docking,” “molecular dynamics simulation,” “binding free energy,” “QSAR,” “pharmacophore modeling,” “virtual screening,” “ADMET prediction,” “de novo drug design,” “artificial intelligence drug discovery,” and “machine learning drug discovery,” each combined with relevant therapeutic-area terms (e.g., “anticancer,” “antiviral,” “antimicrobial”) where appropriate. The search was restricted to English-language, peer-reviewed journal articles and highly cited methodological or seminal papers published between 1988 and

2024, prioritizing high-impact original research, authoritative reviews (e.g., *Nature Reviews Drug Discovery*), and widely used methodological benchmarks. Conference abstracts, non-peer-reviewed sources (with one exception—a widely cited methodological preprint retained for its historical significance in early deep learning-based scoring, reference 62), and sources lacking sufficient methodological detail were excluded. From several hundred candidate sources, seventy were selected for citation based on relevance, methodological importance, citation impact, and representativeness of current consensus in each subtopic; this selection is therefore illustrative and representative rather than exhaustive, and omission of any individual study is not a judgment on its scientific merit. Where claims describe an effect as well established, this reflects convergence across multiple independent studies and authoritative reviews; where claims describe an approach as emerging, promising, or exploratory, this is stated explicitly, and the corresponding evidence should be understood as preliminary and not yet validated through independent replication or prospective clinical outcome.

ROLE OF CADD IN PHARMACEUTICAL CHEMISTRY

2.1 Evolution of Computer-Aided Drug Design

The conceptual origins of CADD trace to Corwin Hansch and colleagues in the early 1960s, who first demonstrated that biological activity could be quantitatively correlated with physicochemical substituent parameters such as lipophilicity (π), electronic effects (σ), and steric bulk (E_s)-giving rise to Hansch analysis, or classical QSAR [5]. This work showed that medicinal chemistry could move beyond empirical structural analogy toward a mathematically grounded, predictive framework.

The 1980s brought interactive computer graphics workstations capable of rendering three-dimensional molecular structures, coupled with maturing molecular mechanics force fields, and the first molecular docking algorithms such as DOCK, developed by Kuntz and colleagues [6]. The 1990s are widely regarded as a watershed decade for structure-based drug design, catalyzed by the successful design of HIV protease inhibitors: following elucidation of the three-dimensional crystal structure of HIV-1 protease, chemists rationally designed peptidomimetic inhibitors—including saquinavir, indinavir, and zidovudine—that transformed HIV infection from a near-universally fatal disease into a manageable chronic condition [7]. This achievement provided the pharmaceutical industry with unambiguous proof that structure-based computational design could deliver clinically transformative medicines, catalyzing massive investment in CADD infrastructure.

The 2000s saw high-throughput virtual screening capable of docking millions of compounds within practical computational timeframes, alongside improved scoring functions and routine three-dimensional QSAR (CoMFA, CoMSIA). The following decade brought exponential increases in computing power, GPU-accelerated molecular dynamics, and improved binding free-energy methods such as Free Energy Perturbation (FEP+), which achieved unprecedented accuracy for lead optimization [8]. Most recently, the field has entered what might be termed the “AI-CADD era”: AlphaFold’s breakthrough in predicting protein structure from sequence alone has dramatically expanded the pool of druggable targets amenable to structure-based design, and generative AI models for de novo molecular design, together with deep learning-based property and activity predictors, are now integrated directly into CADD workflows [8].

2.2 Importance of CADD in Modern Drug Discovery

CADD has transitioned from a peripheral tool to a central component of virtually every modern drug discovery program. At the target-validation and hit-identification stage, computational methods allow assessment of a putative target’s druggability—whether its structure possesses binding pockets capable of accommodating a drug-like ligand [9]. During hit identification, virtual screening allows libraries of millions of compounds to be computationally triaged, reducing the number requiring experimental testing; comparisons with random or diversity-based screening have consistently shown ten- to one-hundred-fold enrichment in active-compound hit rates [10]. During lead optimization, CADD provides mechanistic hypotheses for why structural modifications improve or diminish potency, selectivity, or pharmacokinetics, enabling more rational, efficient design cycles—now formalized as the design-make-test-analyze (DMTA) cycle, discussed in Section 13. Finally, CADD contributes to de-risking candidates by predicting pharmacokinetic liabilities, off-target interactions, and toxicity risks earlier in the pipeline, when remediation is far less costly [9].

2.3 Advantages and Limitations of CADD

The advantages of computational drug design are substantial. Because *in silico* experiments are rapid and comparatively inexpensive once infrastructure is established, researchers can explore vast chemical space—including hypothetical, unsynthesized compounds—without the burden of physical synthesis [11]. CADD provides mechanistic insight into binding, selectivity, and activity, generating interpretable hypotheses that guide synthetic strategy, and enables systematic, simultaneous consideration of multiple design objectives—potency, selectivity, solubility, metabolic stability—that would be difficult to optimize experimentally.

Nevertheless, CADD is subject to well-recognized limitations. Prediction accuracy is bounded by the quality of underlying physical models—force fields, scoring functions, and solvation models—all approximations of true quantum mechanical behavior. Docking scoring functions in particular correlate imperfectly with experimental binding affinities across structurally diverse series, owing to simplified treatment of entropy, desolvation, and protein flexibility [12]. Structure-based predictions also depend on the resolution and completeness of the experimental structures used; poor resolution, missing loops, or crystallization artifacts can propagate errors throughout the design process. Finally, CADD predictions ultimately require experimental validation, and discrepancies between computational and experimental results remain common, underscoring the need for a tightly integrated, iterative relationship between computation and experiment rather than treating CADD as a standalone replacement for laboratory science.

FUNDAMENTAL CONCEPTS OF CADD

3.1 Structure-Activity Relationship (SAR)

SAR is the foundational principle of rational drug design: biological activity is a direct, if complex, function of chemical structure. SAR analysis begins with systematic synthesis and testing of structurally related analogs, in which substituents, ring systems, stereocenters, or functional groups are methodically varied while activity is measured [13]. By correlating structural changes with shifts in potency, selectivity, or physicochemical properties, chemists identify which features are essential for activity, which are tolerated but non-essential, and which are detrimental.

A particularly informative extension is the “activity cliff,” in which a minor structural change—sometimes a single atom substitution—produces a disproportionately large change in activity, often revealing critical, highly specific ligand-target interactions [13]. SAR data also forms the essential training set for essentially all ligand-based computational methods described later in this review, including QSAR and pharmacophore modeling.

3.2 Molecular Interactions and Binding

Recognition of a small-molecule ligand by its biological target arises from noncovalent interactions operating cooperatively across the binding interface. Hydrogen bonds, formed between a donor (N-H or O-H) and an acceptor (carbonyl oxygen or nitrogen lone pair), are often the most geometrically directional and energetically significant contributors to binding specificity [14]. Electrostatic interactions, including salt bridges, can contribute substantially, particularly in solvent-exposed regions.

Van der Waals interactions, individually weak, become collectively significant with close shape complementarity, and the hydrophobic effect—favorable exclusion of water from nonpolar surfaces on complex formation—is frequently the dominant driving force for binding affinity [14]. More specialized interactions, including π - π stacking, cation- π interactions, and halogen bonding, are also increasingly recognized and routinely considered in modern structure-based design. A quantitative understanding of these interactions, and the ability to model them computationally, underlies virtually all structure-based CADD techniques.

3.3 Molecular Descriptors and Chemical Space

Molecular descriptors are numerical representations encoding the structural, topological, electronic, or physicochemical characteristics of a molecule. They range from simple constitutional descriptors (molecular

weight, atom counts, rotatable bonds), to topological indices from graph-theoretic representations of connectivity (e.g., Wiener or Balaban indices), to geometric and electronic descriptors from three-dimensional structures or quantum calculations (dipole moment, polar surface area, HOMO-LUMO gap) [15].

Molecular fingerprints-bit-vector representations encoding presence of substructural fragments-are widely used for similarity comparison and machine learning owing to their computational efficiency and interpretability.

The set of descriptor values for a molecule defines its coordinates within a high-dimensional “chemical space,” a framework allowing systematic comparison, clustering, and visualization of large compound libraries, identification of novel chemotypes, and assessment of library diversity [4,15]. Chemical space analysis underlies compound library design, diversity selection, and interpretation of QSAR and machine learning models throughout CADD.

3.4 Energy Minimization and Molecular Optimization

Before a structure can be used in docking or molecular dynamics, it must be optimized to a physically realistic, low-energy conformation through energy minimization. Molecular mechanics force fields-parameterized functions such as AMBER, CHARMM, OPLS, and GAFF-approximate potential energy as a sum of terms for bond stretching, angle bending, torsional rotation, and non-bonded electrostatic and van der Waals interactions [16].

Minimization algorithms such as steepest descent or conjugate gradient iteratively adjust atomic coordinates to reduce total potential energy, converging toward a stable conformation. Because pharmaceutically relevant molecules typically have multiple rotatable bonds, conformational searching-systematic torsional scanning, Monte Carlo methods, or MD-based sampling-is used to identify the global minimum, or more usefully the full ensemble of accessible conformations [16]. This underpins virtually every subsequent CADD technique, since a ligand’s bioactive conformation may differ substantially from its lowest-energy conformation in isolation.

MAJOR APPROACHES IN COMPUTER-AIDED DRUG DESIGN

4.1 Structure-Based Drug Design (SBDD)

SBDD leverages explicit knowledge of a target’s three-dimensional structure, typically from X-ray crystallography, solution NMR, or cryo-EM, to guide the rational design of geometrically and chemically complementary ligands [17]. By directly visualizing the shape, electrostatic character, and interaction points of a binding pocket, chemists can design or select molecules with a high probability of favorable binding based on explicit structural complementarity.

SBDD encompasses protein structure preparation, molecular docking, molecular dynamics simulation, and binding free-energy calculation, detailed in Section 5. It has been particularly transformative for well-characterized, readily crystallizable target classes such as proteases and kinases, with HIV protease inhibitor design remaining one of the most celebrated achievements in rational drug design [7]. Its principal limitation-dependence on a validated three-dimensional structure-has been substantially mitigated by advances in cryo-EM and computational structure prediction methods such as AlphaFold.

4.2 Ligand-Based Drug Design (LBDD)

Where a target’s three-dimensional structure is unknown, unresolved, or difficult to model with confidence-as is often the case for membrane receptors, intrinsically disordered proteins, or poorly characterized targets-LBDD offers a complementary strategy, proceeding from systematic analysis of known active and inactive ligands rather than the target structure itself [18]. By identifying structural, steric, and electronic features shared among active compounds, LBDD constructs predictive models-pharmacophores and QSAR equations-used to screen libraries, prioritize synthetic targets, or rationalize existing SAR data.

The two principal LBDD techniques, pharmacophore modeling and QSAR, are described in Section 6. LBDD does not require a resolved target structure, but its predictive power is constrained by the quality, diversity, and

quantity of training data; models built from limited or homogeneous ligand sets may generalize poorly to novel chemotypes (the applicability domain problem) [18].

4.3 Hybrid and Integrated Approaches

In practice, structure-based and ligand-based design are increasingly viewed as a spectrum rather than a strict dichotomy, and the most successful campaigns integrate both philosophies iteratively. A common hybrid strategy extracts key interaction features from a docked ligand pose to build a structure-based pharmacophore, subsequently used as a three-dimensional query for ligand-based virtual screening, combining mechanistic precision with computational efficiency [19].

Similarly, docking-derived poses are used to refine 3D-QSAR alignments, while QSAR-derived predictions can re-rank docking hits, improving overall screening enrichment. Consensus scoring, combining predictions from multiple independent structure- and ligand-based models, has also been shown to improve robustness relative to any single method [19]. This integrative philosophy is a defining characteristic of contemporary CADD practice.

STRUCTURE-BASED DRUG DESIGN TECHNIQUES

5.1 Protein Structure Preparation

The reliability of any structure-based analysis depends critically on the quality of protein structure preparation, a frequently underestimated step that dramatically influences downstream results. Raw crystallographic structures from the Protein Data Bank often contain artifacts that must be corrected: hydrogen atoms (generally unresolved by X-ray crystallography) must be added computationally; missing side chains or loop regions must be modeled; and crystallographic waters, ions, and cryoprotectants must be evaluated for retention or removal [20].

A particularly consequential step is assigning appropriate protonation states to titratable residues (histidine, aspartate, glutamate, lysine) at physiological pH, since protonation state can dramatically alter predicted binding interactions. Side-chain rotamers must be optimized to resolve steric clashes, and induced-fit conformational changes not captured in a static crystal structure must be considered [20]. Widely used tools include the Protein Preparation Wizard (Schrödinger) and open-source alternatives such as PDB2PQR and PROPKA. Rigorous structure preparation is a scientifically consequential step that substantially influences the accuracy of all subsequent structure-based predictions.

5.2 Molecular Docking

Molecular docking addresses two interrelated problems: predicting the preferred three-dimensional pose a ligand will adopt within a binding site, and estimating the associated binding affinity via a scoring function [21]. The process begins with generating candidate ligand conformations and orientations within the binding site, governed by a search algorithm.

Search strategies vary: genetic algorithms (AutoDock, GOLD) evolve populations of candidate poses through selection, crossover, and mutation; incremental construction algorithms (FlexX, Glide) anchor a rigid core fragment and progressively add flexible substituents; and systematic or stochastic methods, including Monte Carlo sampling, explore conformational space through randomized or exhaustive sampling [21,36]. Each candidate pose is evaluated by a scoring function estimating binding free energy or a related empirical score, allowing ranking of the most favorable poses. Docking can be performed in rigid-receptor mode or various degrees of flexible-receptor mode, the latter offering improved accuracy at substantially higher computational cost.

5.3 Molecular Dynamics (MD) Simulations

While docking provides a static snapshot, biological macromolecules undergo continuous conformational fluctuation. MD simulation addresses this by numerically integrating Newton's equations of motion for every

atom, using forces from a molecular mechanics force field, to generate a time-resolved trajectory [22].

A typical MD workflow solvates the prepared complex in explicit water, adds counter-ions for neutrality and physiological ionic strength, performs energy minimization, gradually equilibrates temperature and pressure under positional restraints, and then runs an unrestrained production simulation, often spanning nanoseconds to

microseconds [22]. MD provides insight unavailable from static structures or docking alone, including complex flexibility and stability, solvent-mediated interactions, induced-fit adaptations, and transient or cryptic binding pockets.

5.4 Binding Free-Energy Calculations

Whereas empirical docking scores provide rapid but approximate affinity estimates, rigorous free-energy calculation offers substantially improved accuracy at greater computational expense, typically reserved for later, resource-intensive lead optimization stages [23]. Free Energy Perturbation (FEP) and Thermodynamic Integration (TI), grounded in statistical mechanics, compute the free-energy difference between two chemical states by simulating intermediate “alchemical” states, often achieving errors within 1 kcal/mol of experimental values for congeneric series [23].

Faster, more approximate end-point methods-MM-PBSA and MM-GBSA-estimate binding free energy from a limited number of simulation snapshots combined with implicit solvation models, offering a favorable cost-accuracy balance for ranking larger compound sets, albeit with generally lower absolute accuracy than FEP or TI [23]. Selecting an appropriate free-energy method is an important strategic decision balancing desired accuracy against the number of compounds and available computational budget.

5.5 Structure-Based Virtual Screening

Structure-based virtual screening extends docking to the scale required for hit identification, computationally docking libraries of tens of thousands to billions of compounds to identify a small, enriched subset predicted to bind with high affinity [24]. Given the scale of modern virtual libraries such as ZINC and Enamine REAL, screening typically proceeds through a tiered protocol: inexpensive, high-throughput docking first filters the full library, followed by progressively more rigorous re-scoring or re-docking of the top-ranked subset [24].

Post-screening analysis typically incorporates pose-plausibility assessment, diversity selection, and drug-likeness and synthetic-accessibility filters before a prioritized list is submitted for experimental testing. Structure-based virtual screening has been particularly effective when combined with fragment-based screening data or applied to well-validated, high-confidence target structures.

LIGAND-BASED DRUG DESIGN TECHNIQUES

6.1 Pharmacophore Modeling

A pharmacophore is the ensemble of steric and electronic features necessary to ensure optimal interactions between a ligand and a target, sufficient to trigger or block its biological response [25]. Rather than describing a specific molecule, a pharmacophore model represents an abstract three-dimensional arrangement of feature types-hydrogen bond donors and acceptors, hydrophobic centers, aromatic ring centroids, and ionizable groups-with the geometric distances and angles relating them.

Pharmacophore models can be derived from structural alignment of known active ligands (purely ligand-based) or from analysis of protein-ligand complex structures (structure-based) [25]. Once established, a model can be used as a three-dimensional query to screen large databases, identifying molecules with a compatible spatial arrangement of features regardless of scaffold-“scaffold hopping.” Pharmacophore modeling is computationally efficient relative to docking and particularly valuable early in a campaign, when a validated target structure is unavailable or the goal is to diversify an existing hit series.

6.2 Quantitative Structure-Activity Relationship (QSAR)

QSAR constructs statistical models relating quantitative activity measures to molecular descriptors, enabling prediction of activity for untested compounds [26]. The workflow proceeds through defined stages: assembling a training set with reliably measured activity; calculating descriptors; selecting non-redundant, statistically relevant descriptors; building a predictive model using methods from multiple linear regression to random forests, support vector machines, or neural networks; and rigorously validating the model via cross-validation and external test sets [26].

A critical, frequently underappreciated aspect is assessing the model's applicability domain-the region of chemical space within which predictions are reliable-since QSAR models generally extrapolate poorly to structurally novel chemotypes outside this domain. Properly validated QSAR models are powerful, inexpensive tools for prioritizing virtual libraries, rationalizing SAR data, and guiding analog design.

6.3 2D-QSAR

Two-dimensional QSAR derives descriptors from a compound's two-dimensional molecular graph, without reference to three-dimensional conformation. Common 2D descriptors include constitutional parameters (molecular weight, atom or bond counts), topological indices (Wiener index, Randić connectivity index), and physicochemical property estimates such as calculated logP and topological polar surface area [27].

2D-QSAR's principal advantage is computational efficiency, independent of the ambiguous and expensive task of three-dimensional alignment, since descriptors are calculated rapidly and unambiguously from a molecule's connection table-making it well suited to rapid screening of large libraries. Its limitation is that two-dimensional descriptors cannot capture stereochemical differences or three-dimensional shape complementarity, potentially limiting accuracy for activity differences driven by stereochemistry or conformation [27].

6.4 3D-QSAR

Three-dimensional QSAR incorporates explicit structural information, typically aligning a series of molecules according to a proposed common binding conformation and sampling steric and electrostatic fields at grid points surrounding them. The most widely recognized technique, Comparative Molecular Field Analysis (CoMFA), calculates steric and electrostatic interaction energies between each aligned molecule and a probe atom at grid points, correlating the resulting field values with activity via partial least squares regression [28].

A related technique, CoMSIA, uses Gaussian-type distance functions rather than the Lennard-Jones and Coulombic potentials used in CoMFA, offering improved numerical stability and the ability to incorporate hydrophobic and hydrogen-bonding fields beyond CoMFA's steric and electrostatic fields [28]. The resulting three-dimensional contour maps offer chemically intuitive guidance on which regions around a scaffold favor or disfavor activity, directly informing synthetic design. The principal challenge for all 3D-QSAR approaches is the requirement for accurate molecular alignment to a hypothesized common bioactive conformation, introducing subjectivity, particularly for structurally diverse series lacking a well-defined common scaffold.

6.5 Similarity and Shape-Based Screening

Similarity searching relies on the "similar property principle": structurally similar molecules are statistically more likely to share biological activity. Two-dimensional similarity searching typically employs molecular fingerprints compared via metrics such as the Tanimoto coefficient, enabling rapid database searches for compounds related to a known active reference [29].

Shape-based screening extends this into three dimensions, comparing overall molecular shape and, in more sophisticated implementations, electrostatic field distribution against a reference query, without requiring explicit scaffold similarity. Tools such as ROCS perform rapid three-dimensional shape and color (pharmacophoric feature) alignment to identify compounds with a similar overall volume to a reference molecule, even with an entirely distinct scaffold [29]. This makes shape-based screening a powerful tool for scaffold hopping, identifying structurally novel series with potential advantages in intellectual property,

physicochemical properties, or selectivity, while retaining the shape complementarity believed to underlie the reference compound's activity.

COMPUTATIONAL METHODS FOR DRUG DISCOVERY

7.1 Molecular Modeling

Molecular modeling is the broad computational foundation for the more specialized CADD techniques in this review, encompassing construction, visualization, manipulation, and analysis of three-dimensional structures. A particularly important application is homology (comparative) modeling, predicting a protein's structure based on sequence similarity to homologous proteins of known structure [30].

Homology modeling involves sequence alignment to a suitable template, construction of structurally conserved regions, modeling of variable loops and side chains, and final refinement and validation using tools such as Ramachandran plot analysis [30]. While generally less accurate than experimental structures, homology models have historically enabled structure-based CADD for targets recalcitrant to experimental determination, such as many GPCRs and ion channels. The emergence of AlphaFold2 has substantially improved the accuracy of computational structure models, further expanding the scope of targets amenable to structure-based design.

7.2 Virtual Screening

Virtual screening refers to computational filtering of large compound libraries to identify a prioritized, enriched subset with elevated probability of desired activity, prior to experimental testing [31]. Approaches are classified by whether they rely on the target's three-dimensional structure (structure-based, Section 5.5) or on known-ligand properties (ligand-based, employing pharmacophore matching, similarity searching, and QSAR, Section 6).

Most contemporary campaigns employ a multi-stage, hierarchical filtering strategy: inexpensive filters (physicochemical property filters, structural alert removal, rapid similarity or pharmacophore pre-screening) precede more intensive methods such as docking and, ultimately, free-energy re-scoring of top candidates [31]. The objective is not perfect activity prediction but meaningful "enrichment"—a substantially higher hit rate among prioritized compounds than random or diversity-based selection would yield.

7.3 De Novo Drug Design

De novo drug design is a generative, rather than selective, approach in which entirely novel structures are algorithmically constructed to satisfy specified design criteria, rather than selected from an existing library [32]. Structure-based de novo algorithms identify favorable interaction points within a binding site and construct a scaffold connecting them, either by growing a molecule atom-by-atom or fragment-by-fragment from a starting point, or by linking multiple fragments occupying favorable sub-pockets [32].

Ligand-based de novo design generates structures satisfying a specified pharmacophore or optimizing predicted QSAR activity, without reference to target structure. A persistent challenge, particularly for earlier-generation algorithms, has been generating molecules that are theoretically valid but synthetically impractical; this has been substantially addressed through synthetic accessibility scoring and retrosynthetically aware generative algorithms, discussed further in Section 14.4.

7.4 ADME Prediction

Computational ADME prediction has become integral to modern drug discovery, allowing assessment of likely pharmacokinetic behavior at a very early stage, often before synthesis [33]. Models range from simple rule-based filters, such as Lipinski's Rule of Five (Section 7.6), to quantitative structure-property relationship (QSPR) models, to physiologically based pharmacokinetic (PBPK) models simulating time-dependent drug concentration across physiological compartments [33].

Early ADME prediction allows chemists to identify and address liabilities—poor solubility, excessive metabolic clearance, poor permeability—concurrently with potency optimization, rather than discovering them only after

costly in vivo studies. This “front-loaded” approach has been widely credited with reducing attrition due to pharmacokinetic failure.

7.5 Toxicity Prediction

In silico toxicity prediction encompasses computational methods designed to flag potential safety liabilities before or alongside experimental testing [34]. Structural alert-based approaches identify substructural fragments empirically associated with specific toxicological mechanisms—for example, certain nitroaromatic or epoxide groups linked to genotoxicity—providing a rapid, interpretable, though coarse-grained assessment.

More sophisticated approaches employ machine learning classifiers trained on curated toxicological databases to predict specific endpoints, including mutagenicity (benchmarked against the Ames test), hepatotoxicity, cardiotoxicity (particularly hERG channel inhibition, a major cause of drug-induced arrhythmia and late-stage attrition), and carcinogenicity [34]. These models are increasingly integrated into early hit and lead triage, allowing problematic chemotypes to be deprioritized before substantial resources are invested, consistent with the “3Rs” principle of animal testing.

7.6 Drug-Likeness and Bioavailability Prediction

Drug-likeness assesses whether a molecule’s physicochemical properties are broadly consistent with successful, orally bioavailable pharmaceuticals. The most influential filter, Lipinski’s Rule of Five, posits that poor absorption or permeation is more likely when molecular weight exceeds 500 Da, calculated logP exceeds 5, hydrogen bond donors exceed 5, or hydrogen bond acceptors exceed 10 [35].

Subsequent refinements, notably the Veber rules, added rotatable bond count and topological polar surface area, with molecules exceeding roughly 10 rotatable bonds or 140 Å² of polar surface area showing markedly reduced oral bioavailability in rodent models [35]. While criticized for excluding legitimate drug classes that violate one or more criteria (many natural products, macrocycles, biologics), these rule-based filters remain widely used for rapid, inexpensive triage during early library design, and have been extended by numerous more nuanced multi-parameter drug-likeness scoring schemes.

MOLECULAR DOCKING IN PHARMACEUTICAL CHEMISTRY

8.1 Principles of Molecular Docking

At its core, docking addresses molecular recognition: given a target binding site and a candidate ligand, what is the most energetically favorable pose, and how strong is the predicted interaction [21]? This divides into two coupled subproblems: sampling (generating a sufficiently diverse set of candidate poses) and scoring (accurately ranking these poses to identify the most physically plausible binding mode).

The docking search space is defined by the translational, rotational, and, for flexible ligands, internal torsional degrees of freedom, growing combinatorially with ligand size and flexibility—necessitating the sophisticated search algorithms described in Section 5.2 for tractable sampling [21].

8.2 Docking Algorithms and Scoring Functions

Scoring functions fall into three principal classes. Force-field-based functions calculate binding energy directly from molecular mechanics terms—van der Waals and electrostatic interactions—sometimes supplemented with simplified solvation and entropic corrections [36]. Empirical functions are derived through statistical fitting of a weighted sum of physically motivated interaction terms against a training set of experimental binding affinities, offering speed at the cost of interpretability and potential overfitting.

Knowledge-based scoring functions derive potentials from the statistical frequency of atom-pair contacts across large databases of experimentally determined complexes [36]. More recently, machine learning-based scoring functions, trained on large structural and affinity datasets, have shown improved correlation with experimental affinities in numerous benchmarks and are increasingly incorporated as a re-scoring or consensus layer (Section 14.3).

8.3 Protein-Ligand Interaction Analysis

Following pose generation, interaction analysis assesses biological plausibility and extracts medicinal chemistry insight, identifying hydrogen bonds (by donor-acceptor geometry and distance), hydrophobic contacts, π -stacking and cation- π interactions, and salt bridges [37].

Tools such as the Protein-Ligand Interaction Profiler (PLIP) automate detection and visualization of these interactions, generating interaction summaries and diagrams widely used in medicinal chemistry publications [37]. This analysis validates that a pose is consistent with known SAR and mutagenesis data, identifies interactions that might be strengthened through modification, and detects problematic interactions not adequately penalized by the scoring function alone.

8.4 Applications in Lead Identification and Optimization

Docking is applied across the structure-based pipeline. During hit identification, large-scale docking-based virtual screening (Section 5.5) prioritizes compounds from vast libraries for experimental testing [38]. During hit-to-lead and lead optimization, docking assesses proposed structural modifications' predicted impact on pose and affinity prior to synthesis, allowing prioritization among candidate analogs.

Docking is also used to rationalize observed SAR, for instance explaining why a substituent improves potency (an additional favorable interaction) or why a stereoisomer is inactive (an unfavorable steric clash), providing mechanistic insight informing subsequent design [38].

8.5 Limitations of Molecular Docking

Docking is subject to well-documented limitations. Most classical scoring functions treat the receptor as largely rigid, failing to capture induced-fit conformational changes and potentially excluding legitimately active compounds requiring an unrepresented receptor conformation [39].

Scoring functions across all classes struggle to capture entropic contributions-loss of ligand conformational entropy, release of ordered water-and precise desolvation thermodynamics, which can substantially contribute to binding free energy but are difficult to model within efficient high-throughput frameworks [39]. Consequently, docking scores correlate only modestly with experimental affinities, particularly across structurally diverse series, and docking is best regarded as a tool for pose prediction and qualitative enrichment rather than a precise predictor of absolute affinity-motivating the more rigorous free-energy methods of Section 5.4 for critical lead optimization decisions.

MOLECULAR DYNAMICS SIMULATIONS

9.1 Principles of MD Simulation

MD simulation is grounded in classical Newtonian mechanics, treating each atom as a point mass subject to forces from a force field's potential energy function [22]. At each discrete time step (typically one to two femtoseconds, constrained by the fastest atomic vibrations), net force on every atom is calculated and used to update positions and velocities via a numerical integrator, most commonly a Verlet variant.

Repeated iteration over millions of steps generates a continuous trajectory. Modern MD is almost universally performed with explicit solvent and physiological ion concentrations, under periodic boundary conditions, typically at constant temperature and pressure (NPT ensemble) [22]. GPU-accelerated MD software and specialized supercomputing hardware such as Anton have enabled simulation timescales from nanoseconds into the microsecond and millisecond regime, expanding the range of observable conformational transitions and binding events.

9.2 Protein-Ligand Stability Analysis

A primary MD application is assessing the conformational stability of a predicted complex over simulated time, providing dynamic validation of static docking predictions [40]. A ligand that remains stably bound, maintaining key predicted

interactions, provides greater confidence than the docking score alone; conversely, drift, loss of interactions, or dissociation suggests the initial prediction may be inaccurate [40]. This stability assessment has become a standard validation step following virtual screening, helping filter docking false positives before committing to synthesis and testing.

9.3 RMSD, RMSF, Radius of Gyration and SASA

Quantitative MD analysis relies on standard structural metrics. Root-mean-square deviation (RMSD) quantifies average positional displacement relative to a reference structure over simulation time; RMSD plateauing at a low, stable value after equilibration indicates a stable complex, whereas continuously increasing or fluctuating values suggest instability [41].

Root-mean-square fluctuation (RMSF) quantifies time-averaged per-residue flexibility, identifying mobile loops versus rigid elements-valuable for understanding induced-fit and allosteric mechanisms. Radius of gyration measures overall protein compactness, with significant increases suggesting partial unfolding, while solvent-accessible surface area (SASA) quantifies solvent exposure, informing changes in compactness and solvation-related thermodynamics [41].

9.4 Principal Component Analysis and Free-Energy Landscapes

Because MD trajectories generate enormous high-dimensional positional data, dimensionality reduction is essential to extract meaningful information. Principal Component Analysis (PCA) identifies the dominant collective modes of atomic motion accounting for the largest proportion of conformational variance, separating significant large-scale transitions from unimportant local fluctuations [42].

Projecting the trajectory onto dominant principal components allows construction of a free-energy landscape, revealing distinct, thermodynamically stable conformational substates (local minima or “basins”) and the energetic barriers separating them [42]. This has proven valuable for characterizing conformational heterogeneity in flexible binding sites, identifying cryptic or allosteric pockets, and rationalizing how ligand binding may select for a particular protein conformation.

9.5 Applications in Drug Discovery

Beyond stability assessment, MD-based conformational sampling refines docking poses that may represent only local energy minima, and induced-fit docking protocols incorporating short MD-based receptor relaxation have improved pose prediction relative to rigid-receptor docking [43]. MD is also used to explore allosteric regulatory mechanisms-how ligand binding distant from the active site propagates conformational changes-informing the design of allosteric modulators for targets lacking a druggable orthosteric site.

As noted in Section 5.4, MD trajectories provide the conformational sampling required for rigorous free-energy calculations, linking MD directly to lead optimization decision-making. Extended MD combined with enhanced sampling techniques such as metadynamics or replica exchange has also been used to characterize complete binding and unbinding pathways, providing kinetic insight that complements the thermodynamic information traditionally emphasized in CADD and that is increasingly recognized as an independent determinant of in vivo drug efficacy and duration of action [43].

COMPUTATIONAL ADME AND TOXICITY PREDICTION

10.1 Absorption Prediction

Computational absorption prediction relies primarily on aqueous solubility and membrane permeability models, the two properties that jointly determine oral drug absorption per the Biopharmaceutics Classification System (BCS), which categorizes compounds into four classes based on high or low solubility and permeability [44]. Solubility prediction ranges from simple regression equations (e.g., the General Solubility Equation) to machine learning models trained on curated datasets, while permeability prediction correlates calculated descriptors-lipophilicity, polar surface area, hydrogen bonding capacity-with permeability measured across assays such as Caco-2 or PAMPA [44]. These models let chemists address solubility or permeability liabilities early, through structural modification or formulation strategies.

10.2 Distribution and Blood-Brain Barrier Prediction

Computational distribution models predict how a drug partitions between plasma and tissue compartments, heavily influenced by plasma protein binding and tissue-specific partitioning related to lipophilicity and ionization [45]. Of particular importance for CNS drug discovery is blood-brain barrier (BBB) permeability prediction, since the BBB excludes most systemically administered small molecules from reaching therapeutic brain concentrations.

Specialized BBB classifiers, trained on curated brain-to-plasma ratio datasets, incorporate descriptors related to lipophilicity, size, polar surface area, hydrogen bonding, and susceptibility to efflux transporters such as P-glycoprotein [45]. Accurate BBB prediction is essential both for CNS-targeted therapeutics, where good brain penetration is desired, and for peripherally acting drugs designed to avoid central nervous system side effects.

10.3 Metabolism Prediction

Computational metabolism prediction addresses two questions: identifying sites of metabolism most susceptible to enzymatic biotransformation, chiefly via cytochrome P450 (CYP) enzymes, and predicting whether a candidate will act as a substrate, inhibitor, or inducer of specific CYP isoforms, critical for anticipating metabolic stability and drug-drug interaction risk [46].

Site-of-metabolism tools combine quantum mechanical or semi-empirical bond-reactivity calculations with steric and electronic accessibility models from CYP active-site structures, allowing chemists to anticipate metabolic “soft spots” and block metabolism through modifications such as fluorine substitution or deuteration [46]. CYP inhibition prediction models are typically machine learning classifiers trained on curated data across major clinically relevant isoforms (CYP3A4, CYP2D6, CYP2C9).

10.4 Excretion Prediction

Computational excretion prediction estimates clearance and elimination half-life, which together determine dosing frequency and duration of therapeutic drug concentrations [47]. Renal clearance models incorporate descriptors related to molecular size, charge, and plasma protein binding, which determine glomerular filtration and tubular reabsorption or secretion, while hepatobiliary clearance draws on the metabolic stability and CYP substrate predictions of Section 10.3.

Integrated PBPK modeling frameworks increasingly combine these individual ADME predictions to simulate complete time-dependent drug concentration profiles, allowing prediction of human pharmacokinetics and informing first-in-human dose selection before clinical data are available [47].

10.5 In-Silico Toxicity Assessment

Comprehensive in silico toxicity assessment increasingly incorporates predictive models addressing organ-specific toxicity (hepato-, nephro-, cardiotoxicity), genetic toxicity, reproductive and developmental toxicity, and endocrine disruption, often integrated within platforms such as Derek Nexus and the OECD QSAR Toolbox [34]. These combine expert-derived structural alerts, QSAR models, and machine learning classifiers for a multi-layered, weight-of-evidence assessment.

Regulatory agencies, including the FDA and EMA, have increasingly formalized computational toxicology within submission processes, particularly for genotoxic impurity assessment under frameworks such as ICH M7-reflecting growing confidence in well-validated in silico toxicity prediction as a complement to, though not a replacement for, experimental testing.

CADD IN LEAD IDENTIFICATION AND OPTIMIZATION

11.1 Hit Identification

Hit identification combines computational and experimental methods to identify a tractable set of chemically diverse compounds with measurable, reproducible target activity [48]. Computational contributions include structure- and ligand-based virtual screening (Sections 5.5 and 7.2) and computational analysis of fragment-based screening data, in which weakly binding fragments are computationally elaborated or linked into higher-affinity lead-like compounds.

Effective hit identification requires triage beyond potency alone, including assessment of chemical tractability, absence of problematic liabilities (reactive groups, promiscuous aggregation-prone chemotypes), and confirmation of a plausible binding mode, often established through orthogonal biophysical techniques and, where possible, X-ray or cryo-EM structural confirmation [48].

11.2 Hit-to-Lead Optimization

The hit-to-lead phase focuses on computationally guided exploration of chemical space surrounding a hit series, aiming to improve potency by one to two orders of magnitude while establishing preliminary SAR and addressing the most severe physicochemical or ADMET liabilities [49]. Computational SAR analysis, informed by iterative docking, MD-based binding mode confirmation, and early QSAR modeling, guides prioritization of analog series for synthesis.

11.3 Lead Optimization

Lead optimization is the most resource-intensive phase, in which a validated series is iteratively refined to simultaneously satisfy sub-nanomolar to low-nanomolar potency, high selectivity, favorable physicochemical and ADMET properties, and acceptable preliminary safety margins [50]. The full suite of CADD techniques—rigorous free-energy calculations, extensive MD-based stability and selectivity analysis, comprehensive ADMET prediction, and multi-parameter optimization—is typically applied intensively before a compound advances to preclinical candidate selection.

11.4 Prediction of Binding Affinity

Accurate binding affinity prediction is central to lead identification and optimization, and, as discussed in Sections 5.2 and 5.4, method choice reflects a trade-off between speed and accuracy [23]. Empirical docking scores provide rapid, approximate ranking suitable for large-scale hit identification, while rigorous alchemical methods such as FEP+ provide improved accuracy for fine-grained ranking of closely related analogs, accelerating optimization by reducing required synthetic cycles.

11.5 Optimization of Pharmacokinetic Properties

Because clinical efficacy depends on target potency and on achieving adequate drug concentrations at the site of action, lead optimization must simultaneously address pharmacokinetics alongside potency and selectivity [51]. This requires the computational ADME and toxicity prediction methods of Section 10, integrated within multi-parameter optimization frameworks (Section 15.3) that help teams navigate trade-offs—for example, between potency and lipophilicity, or metabolic stability and solubility—systematically rather than through chemical intuition alone.

12. APPLICATIONS OF CADD IN PHARMACEUTICAL CHEMISTRY

12.1 Anticancer Drug Discovery

CADD has been instrumental in discovering and optimizing targeted anticancer therapeutics, most notably small-molecule protein kinase inhibitors, now among the largest and most clinically successful categories of

modern oncology drugs [52]. Structure-based design has been particularly effective here because the kinase ATP-binding site is well characterized structurally, allowing rational design of inhibitors exploiting subtle differences between closely related isoforms [52].

Beyond kinase inhibitors, CADD has contributed to PARP inhibitors exploiting synthetic lethality in BRCA-mutant cancers, and to inhibitors of protein-protein interactions such as MDM2-p53-a target class historically considered “undruggable” but successfully addressed through structure-based fragment growing and de novo design.

12.2 Antimicrobial Drug Discovery

The escalating antimicrobial resistance crisis has intensified interest in computational approaches to accelerate novel antibacterial discovery, against a backdrop of diminished industry investment in this area [53]. CADD has identified novel inhibitors of essential bacterial targets-cell wall biosynthesis enzymes, DNA gyrase and topoisomerase IV, ribosomal machinery-and screened for compounds overcoming resistance mechanisms such as efflux pumps or target-site mutations.

Structure-based design has also been extensively applied to beta-lactamase inhibitor optimization, addressing a clinically significant resistance mechanism through molecules maintaining potency against evolving enzyme variants.

12.3 Antiviral Drug Discovery

Antiviral discovery has historically been among the most productive CADD application areas, beginning with HIV protease and reverse transcriptase inhibitor design (Section 2.1) [7] and extending through influenza neuraminidase inhibitors and, most recently, the rapid structure-based design of SARS-CoV-2 main protease (Mpro) inhibitors during COVID-19 [54]. Rapid public release of high-resolution SARS-CoV-2 structures enabled worldwide structure-based virtual screening and de novo design campaigns, contributing to the historically rapid development of agents such as nirmatrelvir, illustrating mature CADD infrastructure’s capacity to respond to emerging infectious disease threats [54].

12.4 CNS Drug Discovery

CNS drug discovery presents distinctive challenges centered on achieving adequate blood-brain barrier penetration while avoiding excessive interaction with closely related receptor subtypes, a selectivity challenge particularly acute for GPCR and ion channel targets [45]. CADD contributes through the BBB permeability models of Section 10.2, structure-based design increasingly enabled by cryo-EM structures of previously intractable membrane protein targets, and ligand-based pharmacophore and QSAR models improving selectivity across related receptor subtypes.

12.5 Cardiovascular Drug Discovery

CADD has contributed to major cardiovascular drug classes, including structure-based design of ACE inhibitors-exemplified by captopril, designed by structural analogy to a known peptide inhibitor with subsequent computational refinement-and direct thrombin and factor Xa inhibitors for anticoagulant therapy [55]. Contemporary efforts increasingly focus on PCSK9 inhibitors for cholesterol management and hERG channel selectivity optimization, given the importance of avoiding cardiotoxic liabilities across all therapeutic areas (Section 7.5).

12.6 Drug Repurposing

Computational drug repurposing seeks new therapeutic applications for existing approved drugs, attractive given the reduced development cost and timeline for candidates with established human safety profiles [56]. Approaches include structure-based virtual screening of approved-drug libraries against novel targets, ligand-based similarity searching, and increasingly network-based and machine learning methods integrating

genomic, transcriptomic, and clinical data to identify non-obvious drug-disease associations [56]. Repurposing gained prominence during COVID-19, when numerous approved drugs were rapidly screened for antiviral activity, and continues to grow across oncology, rare diseases, and neglected tropical diseases.

INTEGRATION OF CADD WITH EXPERIMENTAL DRUG DISCOVERY

13.1 Computational Prediction → Synthesis → Biological Evaluation

The design-make-test-analyze (DMTA) cycle is the formalized, iterative framework through which computational predictions are translated into experimentally validated candidates across essentially all modern discovery organizations [57]. In “design,” computational methods generate hypotheses about improved potency, selectivity, or pharmacokinetics; in “make,” chemists synthesize prioritized compounds; in “test,” compounds are evaluated in biochemical, cellular, and pharmacokinetic assays; and in “analyze,” results assess whether objectives were met and recalibrate computational models for the next cycle [57]. Cycle efficiency has become a key performance metric, with substantial effort directed toward accelerating each stage through automation of synthesis (flow chemistry, parallel synthesis) and testing (miniaturized, automated assay platforms).

13.2 Computational Prioritization of Synthetic Candidates

Given that synthetic capacity is a persistent, often rate-limiting bottleneck, computational prioritization of which analogs to actually synthesize is a critical value-adding function within the DMTA cycle [58]. Rather than synthesizing every conceivable variation suggested by an SAR hypothesis, prioritization incorporating predicted potency, selectivity, physicochemical properties, and synthetic tractability focuses limited resources on the subset most likely to yield meaningful improvement.

13.3 Experimental Validation of Computational Predictions

No computational prediction can be considered fully reliable without appropriate experimental validation, and a rigorous CADD-driven program must incorporate systematic testing of computational hypotheses at every stage [59]. Biochemical and cellular functional assays validate or refute predicted potency rankings, while biophysical techniques such as surface plasmon resonance (SPR), isothermal titration calorimetry (ITC), and, where feasible, X-ray or cryo-EM structural determination directly validate or invalidate predicted binding modes [59]. This continuous validation loop confirms individual predictions and provides feedback data to iteratively improve computational model accuracy.

13.4 Role of CADD in Reducing Time and Cost of Drug Discovery

The cumulative effect of integrated computational and experimental strategies has been a demonstrable, if difficult to precisely quantify, reduction in time and cost across the discovery pipeline [9]. By reducing the number of compounds requiring synthesis and testing, addressing liabilities earlier when remediation is inexpensive, and accelerating hypothesis-driven medicinal chemistry, CADD is widely regarded as a valuable contributor to improving pharmaceutical R&D productivity. It is important to note, however, that this contribution is difficult to isolate quantitatively from concurrent improvements in synthetic chemistry, assay technology, and target biology; most supporting evidence derives from retrospective case studies rather than prospective, controlled comparison, so the precise magnitude of CADD’s independent effect on R&D productivity remains an active and unresolved subject of debate, and claims of cost or time savings from any individual program should be interpreted as case-specific rather than generalizable.

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN CADD

The application of AI to CADD spans a spectrum of technological maturity, and this section deliberately distinguishes methods with a substantial, replicated evidence base from those that remain exploratory. Classical machine learning for QSAR and ADMET prediction (Section 14.1) and deep learning-based property

prediction (Section 14.2) are now well established, with peer-reviewed, prospectively validated applications across numerous organizations. AI-augmented virtual screening (Section 14.3) has moved from exploratory to

increasingly routine use. Generative de novo design (Section 14.4), by contrast, remains earlier in its validation trajectory: while several AI-designed candidates have entered clinical trials, robust, independently replicated evidence that generative design outperforms conventional medicinal chemistry in overall program success rate is still limited, a distinction revisited in Section 14.5.

14.1 Machine Learning-Based Drug Discovery

Classical machine learning predates the current deep learning era by decades, with random forests, support vector machines, and naive Bayesian classifiers extensively applied to QSAR modeling, virtual screening, and ADMET prediction since at least the early 2000s [60]. These approaches generally operate on pre-computed, human-engineered descriptors or fingerprints, learning a statistical mapping to the property of interest, and have consistently outperformed simpler linear regression QSAR for complex, nonlinear structure-activity relationships [60]. Their interpretability, efficiency, and modest data requirements make them attractive for smaller project-specific datasets where deep learning's large data requirements are unavailable.

14.2 Deep Learning in Molecular Property Prediction

Deep learning has achieved substantial, well-documented performance improvements for molecular property prediction, particularly through architectures operating directly on molecular graph representations [61]. Graph neural networks (GNNs) represent molecules as graphs, with atoms as nodes and bonds as edges, and learn to iteratively aggregate information via message passing, automatically learning task-relevant representations rather than relying on hand-engineered descriptors [61].

This end-to-end representation learning has proven valuable where relevant structural determinants may not be well captured by traditional descriptors, with GNN-based models achieving state-of-the-art performance on benchmarks spanning solubility, toxicity, and binding affinity prediction. Transformer-based architectures, adapted from natural language processing to operate on SMILES strings, have similarly performed strongly, particularly when pre-trained on large unlabeled chemical datasets before fine-tuning on smaller, task-specific data, mitigating deep learning's data-hungry nature.

14.3 AI-Based Virtual Screening

AI has been integrated into virtual screening in several complementary forms. Deep learning-based scoring functions, trained on large structural and affinity datasets, are increasingly used as standalone predictors or, more commonly, as a re-scoring layer applied to top-ranked docking poses, with numerous studies showing improved correlation with experimental affinity relative to classical scoring functions [62].

Beyond re-scoring, machine learning models that rapidly and approximately predict docking scores or affinity directly from structure, without the expensive docking calculation itself, have enabled "ultra-large-scale" virtual screening evaluating libraries of hundreds of millions to billions of molecules-several orders of magnitude beyond conventional docking, substantially expanding the scope of hit identification [62].

14.4 Generative AI and De Novo Drug Design

Generative AI has expanded the toolkit for de novo molecular design (Section 7.3) by learning the underlying statistical distribution of chemically valid, drug-like structures from large training datasets and generating novel structures sampled from this distribution [63]. Variational autoencoders (VAEs) encode structures into a continuous latent representation from which novel structures can be decoded, enabling optimization within this latent space; generative adversarial networks (GANs) employ competitive training between a generator producing realistic structures and a discriminator distinguishing them from real examples [63].

Reinforcement learning frameworks, rewarding a generative model for producing molecules scoring favorably on specified objectives (predicted affinity, drug-likeness, synthetic accessibility), suit multi-objective de novo

design. More recent transformer- and diffusion-based generative architectures, including three-dimensional structure-conditioned models designing molecules directly within a specified binding pocket, represent the current frontier and have shown increasingly promising experimental validation in recent studies.

14.5 Opportunities and Challenges of AI-CADD

AI's integration into CADD presents substantial opportunities alongside actively debated challenges [64]. AI methods can identify complex, nonlinear structure-activity relationships that may elude physics-based or classical statistical approaches, dramatically accelerate screening of vast libraries, and enable generative design capabilities extending beyond traditional selection-based virtual screening.

Significant challenges persist, however. Model performance remains fundamentally dependent on the quality, size, and diversity of training data, and many pharmaceutically relevant datasets remain small, noisy, or biased toward particular chemotypes, limiting generalizability [64]. Interpretability of complex deep learning models remains a concern where mechanistically interpretable insight is often as valuable as accurate prediction, motivating research into explainable AI for molecular property prediction. Finally, the risk of models exploiting spurious dataset-specific artifacts rather than genuine, generalizable relationships—an AI-context “activity cliff” or applicability domain problem—continues to necessitate rigorous, prospective experimental validation before AI-CADD predictions can be confidently relied upon for critical decisions.

To summarize the field's maturity: machine learning-based QSAR/ADMET prediction and deep learning-based property prediction are established, evidence-supported tools with a substantial peer-reviewed track record. AI-based re-scoring and ultra-large-scale virtual screening are increasingly well supported but still being benchmarked against physics-based methods across diverse target classes. Generative de novo design, digital twins (Section 15.4), and fully autonomous closed-loop discovery platforms (Section 15.2) should be regarded as emerging and promising but not yet definitively validated at the level of independently replicated improvement in clinical success rates; claims of their clinical impact in the current literature, including in this review, should be read as illustrative of an active research direction rather than settled scientific consensus.

EMERGING TRENDS IN COMPUTER-AIDED DRUG DESIGN

15.1 Quantum Chemical Calculations and DFT

Quantum mechanical methods, most prominently Density Functional Theory (DFT), provide a more rigorous description of molecular electronic structure than classical force fields, explicitly accounting for electronic polarization, charge transfer, and covalent bond formation or breaking [65]. Their substantially greater computational cost has historically restricted application to small systems or specific sub-problems—covalent inhibitor reaction mechanisms, accurate partial atomic charges for classical simulations, or metalloenzyme active sites—but continuing hardware advances and efficient linear-scaling and fragment-based DFT methods are progressively expanding the scope of systems amenable to direct quantum treatment. Hybrid quantum mechanics/molecular mechanics (QM/MM) methods, applying quantum treatment specifically to the relevant active-site region while treating the remainder classically, represent an important and increasingly widely used intermediate strategy [65].

15.2 AI-Driven Molecular Design

A broader trend beyond the specific AI-CADD techniques of Section 14 is the progressive integration of AI across essentially every pipeline stage, from target identification through preclinical candidate selection, within increasingly automated, “closed-loop” discovery platforms [66]. These combine generative molecular design, automated synthesis planning and robotic execution, and rapid experimental testing within a continuously iterating cycle that compresses the traditional DMTA timeline. Several biotechnology companies have publicly reported advancing AI-designed candidates into human clinical trials; these early reports are encouraging, but the sample of publicly disclosed programs remains small, few have reported late-stage clinical efficacy outcomes, and independent, peer-reviewed comparison of AI-driven programs against matched conventional

programs is still largely lacking-the paradigm should be regarded as promising but not yet conclusively validated.

15.3 Multi-Parameter Optimization

As drug discovery has matured, there is increasing recognition that successful candidates must simultaneously satisfy many competing design criteria-potency, selectivity, solubility, permeability, metabolic stability, and safety margins-rather than being optimized sequentially [51]. Formal multi-parameter optimization (MPO) approaches, ranging from simple weighted-sum desirability functions to Pareto-optimal multi-objective evolutionary algorithms, allow teams to systematically navigate these trade-offs, identifying compounds with the best available overall balance rather than optimizing any single parameter to the detriment of others-a strategy increasingly recognized as essential to reducing late-stage attrition from unaddressed secondary liabilities.

15.4 Digital Twins and Advanced Molecular Simulations

The “digital twin” concept in pharmaceutical research envisions increasingly comprehensive, integrated computational models-spanning molecular, cellular, and ultimately whole-organism or patient-specific physiological scales-simulating drug response with realism exceeding any individual technique described elsewhere in this review [67]. Although still at a comparatively early stage, digital twin approaches, combining molecular-scale CADD with systems pharmacology, physiologically based pharmacokinetic simulation, and patient-specific genomic and clinical data, represent an ambitious long-term direction with potential for genuinely personalized prediction of drug efficacy and safety.

15.5 Integration of Omics Data with CADD

The increasing availability of genomic, transcriptomic, proteomic, and metabolomic (“omics”) data is progressively transforming target identification and validation, historically among the highest-risk stages of the discovery pipeline [68]. Computational integration of large-scale omics datasets-genome-wide association studies, disease-specific expression signatures, proteomic profiling-increasingly informs identification of novel druggable targets with genetically validated disease relevance, while pharmacogenomic data informs patient stratification for trial design and precision medicine-based treatment selection. Formal integration of these systems-level analyses with molecular-level CADD represents an important, rapidly developing frontier bridging target identification and molecular-level design.

CHALLENGES AND LIMITATIONS OF CADD

16.1 Quality of Structural Data

The reliability of structure-based CADD predictions is fundamentally bounded by the quality of the underlying target structure [12]. Low-resolution crystallographic structures may contain significant positional uncertainty; structures may be missing functionally important flexible loops; and crystallization conditions may induce non-physiological conformational artifacts. Even highly accurate predicted structures, including those from AlphaFold, generally represent a single static conformation and may not capture the flexibility relevant to ligand binding-underscoring the continued importance of careful structural assessment.

16.2 Docking and Scoring-Function Limitations

The persistent limitations of docking scoring functions (Section 8.5) remain among the most significant, actively researched challenges in CADD [39]. Despite decades of development and increasing use of machine learning-based scoring, no scoring function has achieved consistently reliable, quantitatively accurate binding affinity prediction across the full diversity of target classes and chemical scaffolds encountered in real-world discovery, and appropriate selection, validation, and target-specific recalibration remain important practical considerations.

16.3 Computational Cost

Despite continuing advances in GPU acceleration and cloud infrastructure, the most rigorous CADD methods-extended MD, alchemical free-energy calculations, and quantum mechanical methods-remain computationally expensive relative to the compound-set scale medicinal chemistry teams often wish to evaluate, necessitating strategic matching of computational method to pipeline stage [23]. This continues to motivate research into more efficient approximate methods, including machine learning-accelerated free-energy prediction, aiming to approach rigorous physics-based accuracy at reduced cost.

16.4 Prediction-Experiment Discrepancies

Despite CADD's substantial value, discrepancies between computational predictions and experimental results remain common, arising from the cumulative effect of methodological limitations discussed throughout this review [69]. Such discrepancies, while sometimes frustrating for project timelines, are valuable learning opportunities: retrospective analysis of prediction failures frequently reveals previously unappreciated structural, energetic, or biological factors that can improve future model accuracy, underscoring the iterative, self-correcting relationship between computational prediction and experimental validation that characterizes mature CADD practice.

16.5 Comparative Summary of Major CADD Methods

Because the preceding sections describe each method largely in isolation, Table 1 provides a consolidated, comparative summary intended to support critical, method-selection-oriented evaluation rather than purely descriptive reading. The table summarizes, for each major technique, its principal advantage, principal limitation, relative computational cost, current validation status, and the pipeline stage at which it is most appropriately applied.

Table 1. Comparative overview of major CADD methods.

Method	Key Advantage	Key Limitation	Relative Computational Cost	Validation Status	Typical Pipeline Stage
Molecular docking	Rapid pose and affinity estimation across large libraries	Limited receptor flexibility; weak entropy/desolvation treatment [39]	Low-Moderate	Well established; scoring accuracy target-dependent	Hit identification, virtual screening
Molecular dynamics simulation	Captures dynamic behavior, induced fit, complex stability	High computational cost; timescale limitations [22]	High	Well established for stability/mechanism; less standardized for affinity ranking	Pose validation, mechanism, lead optimization
FEP / TI (alchemical free energy)	High quantitative accuracy for congeneric series [23]	Very high cost; requires expert setup; limited to closely related analogs	Very High	Well established for lead optimization within pharma; less validated across scaffold-hopping	Late-stage lead optimization
MM-PBSA / MM-GBSA	Faster than FEP with reasonable relative ranking	Lower absolute accuracy than FEP/TI; sensitive to protocol choices [23]	Moderate-High	Established but variable accuracy across systems	Mid-stage lead optimization
Pharmacophore modeling	Fast; enables scaffold	Model quality depends on	Low	Well established	Hit identification,

	hopping; no target structure needed	alignment/training set quality [25]			scaffold hopping
2D/3D-QSAR	Interpretable, quantitative prediction; low cost	Restricted applicability domain; alignment sensitivity (3D-QSAR) [26,28]	Low-Moderate	Well established for congeneric series	Hit-to-lead, SAR rationalization
Structure-based virtual screening	Scales to very large libraries; explicit structural rationale	Dependent on structure quality and scoring accuracy [12,24]	Moderate-High (library-dependent)	Well established	Hit identification
Computational ADMET prediction	Early, low-cost liability flagging	Model accuracy variable by endpoint and chemical space [33,34]	Low	Established for common endpoints; variable for rarer endpoints	Hit-to-lead through candidate selection
Classical ML (QSAR/ADMET)	Interpretable; effective on moderate-sized datasets [60]	Reliant on descriptor quality; limited extrapolation	Low-Moderate	Well established	Throughout pipeline
Deep learning property prediction	Learns representations automatically; strong benchmark performance [61]	Data-hungry; interpretability challenges [64]	Moderate-High	Established for property prediction; ongoing validation for novel chemotypes	Hit-to-lead, ADMET, screening
Generative de novo design (AI)	Explores chemical space beyond existing libraries; multi-objective optimization [63]	Synthetic accessibility and prospective validation still maturing [64]	Moderate-High	Emerging; limited independently replicated prospective evidence	Hit identification, scaffold innovation
Quantum mechanical / DFT / QM-MM	Highest physical accuracy; essential for covalent/metal systems [65]	Very high cost; limited to small systems or active-site regions	Very High	Well established for specific mechanistic questions	Mechanistic studies, covalent inhibitor design

16.6 Reproducibility and Validation

Reproducibility of computational predictions across software platforms, parameter settings, and research groups is an increasingly recognized challenge [70]. Variability in force field parameterization, docking algorithm implementation, and structure preparation protocol can lead to meaningfully different predictions for the same scientific question, and the field has placed increasing emphasis on transparent methodology reporting, standardized benchmark datasets, and, where feasible, consensus approaches combining multiple independent methods to improve robustness.

FUTURE PERSPECTIVES

CADD's trajectory points toward a future in which physics-based simulation and data-driven AI are combined synergistically rather than applied as competing paradigms. Hybrid approaches using machine learning to accelerate physics-based calculations-machine learning-accelerated free-energy prediction or ML-derived force field parameters approaching quantum mechanical accuracy at classical cost-represent a particularly promising near-term direction [66]. Continued improvement of protein structure prediction, building on AlphaFold's success, is expected to further expand the population of targets amenable to structure-based design, including challenging membrane protein and intrinsically disordered protein targets.

Generative AI-based de novo design is likely to become an increasingly central component of hit identification and lead optimization, supplementing and, in some applications, reducing reliance on screening pre-existing libraries. Deeper integration of multi-omics data with molecular-level CADD is expected to progressively improve target identification and validation, while advances in PBPK modeling and, more speculatively, digital twin technologies may eventually enable patient-specific prediction of drug response. Continued advances in computational hardware, including specialized AI accelerators and potentially practical quantum computing, may further expand the scale and accuracy of the physics-based methods that remain foundational to the entire CADD enterprise [67,68].

CONCLUSION

Computer-Aided Drug Design has undergone a remarkable evolution over the past several decades, progressing from simple, empirically derived structure-activity correlations to a sophisticated, integrated discipline encompassing structure-based design, ligand-based design, molecular dynamics simulation, comprehensive computational ADMET and toxicity prediction, and, increasingly, transformative artificial intelligence and machine learning methodologies. Throughout this evolution, CADD has consistently demonstrated its capacity to rationalize complex structure-activity relationships, dramatically narrow the chemical space that must be experimentally explored, and identify and address pharmacokinetic and toxicological liabilities earlier than experimental methods alone would allow. Despite persistent challenges relating to scoring function and force field accuracy, the computational cost of the most rigorous methods, and the ongoing necessity of experimental validation, the continued, tightly integrated combination of computational prediction with iterative experimental testing-embodied in the design-make-test-analyze cycle that now defines modern pharmaceutical research-represents a durably effective and continuously improving strategy for discovering and optimizing safe, effective bioactive molecules. As artificial intelligence, structural biology, and computational hardware continue to advance in tandem, CADD is poised to play an increasingly central role in addressing unmet medical needs across the spectrum of human disease. This potential, however, remains contingent on the discipline's continued commitment to rigorous experimental validation: no computational prediction discussed in this review, regardless of its underlying method's sophistication, should be regarded as a substitute for appropriate biochemical, biophysical, pharmacokinetic, and toxicological confirmation, and responsible application of CADD requires consistent, transparent communication of uncertainty and a clear distinction between well-validated, established methodology and emerging, still-maturing approaches.

REFERENCES

1. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J Health Econ.* 2016;47:20-33.
2. Paul SM, Mytelka DS, Dunwiddie CT, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov.* 2010;9(3):203-214.
3. Sliwoski G, Kothiwale S, Meiler J, Lowe EW. Computational methods in drug discovery. *Pharmacol Rev.* 2014;66(1):334-395.
4. Reymond JL. The chemical space project. *Acc Chem Res.* 2015;48(3):722-730.
5. Hansch C, Fujita T. ρ - σ - π analysis: a method for the correlation of biological activity and chemical structure. *J Am Chem Soc.* 1964;86(8):1616-1626.

6. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. *J Mol Biol.* 1982;161(2):269-288.
7. Wlodawer A, Vondrasek J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu Rev Biophys Biomol Struct.* 1998;27:249-284.
8. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596(7873):583-589.
9. Macalino SJ, Gosu V, Hong S, Choi S. Role of computer-aided drug design in modern drug discovery. *Arch Pharm Res.* 2015;38(9):1686-1701.
10. Song CM, Lim SJ, Tong JC. Recent advances in computer-aided drug design. *Brief Bioinform.* 2009;10(5):579-591.
11. Leelananda SP, Lindert S. Computational methods in drug discovery. *Beilstein J Org Chem.* 2016;12:2694-2718.
12. Warren GL, Andrews CW, Capelli AM, et al. A critical assessment of docking programs and scoring functions. *J Med Chem.* 2006;49(20):5912-5931.
13. Kubinyi H. QSAR and 3D QSAR in drug design. *Drug Discov Today.* 1997;2(11):457-467.
14. Bissantz C, Kuhn B, Stahl M. A medicinal chemist's guide to molecular interactions. *J Med Chem.* 2010;53(14):5061-5084.
15. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics.* Wiley-VCH; 2009.
16. Ponder JW, Case DA. Force fields for protein simulations. *Adv Protein Chem.* 2003;66:27-85.
17. Anderson AC. The process of structure-based drug design. *Chem Biol.* 2003;10(9):787-797.
18. Acharya C, Coop A, Polli JE, Mackerell AD. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr Comput Aided Drug Des.* 2011;7(1):10-22.
19. Yang SY. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov Today.* 2010;15(11-12):444-450.
20. Sastry GM, 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.
21. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov.* 2004;3(11):935-949.
22. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. *Nat Struct Biol.* 2002;9(9):646-652.
23. Cournia Z, Allen B, Sherman W. Relative binding free energy calculations in drug discovery. *J Chem Inf Model.* 2017;57(12):2911-2937.
24. Lyne PD. Structure-based virtual screening: an overview. *Drug Discov Today.* 2002;7(20):1047-1055.
25. Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA. Glossary of terms used in medicinal chemistry. *Pure Appl Chem.* 1998;70(5):1129-1143.
26. Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: where have you been? Where are you going to? *J Med Chem.* 2014;57(12):4977-5010.
27. Roy K, Kar S, Das RN. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment.* Academic Press; 2015.
28. Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). *J Am Chem Soc.* 1988;110(18):5959-5967.
29. Rush TS, Grant JA, Mosyak L, Nicholls A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem.* 2005;48(5):1489-1495.
30. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol.* 1993;234(3):779-815.
31. Walters WP, Stahl MT, Murcko MA. Virtual screening-an overview. *Drug Discov Today.* 1998;3(4):160-178.
32. Schneider G, Fechner U. Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov.* 2005;4(8):649-663.
33. Van de Waterbeemd H, Gifford E. ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov.* 2003;2(3):192-204.

34. Raies AB, Bajic VB. In silico toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdiscip Rev Comput Mol Sci*. 2016;6(2):147-172.
35. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery. *Adv Drug Deliv Rev*. 1997;23(1-3):3-25.
36. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*. 1997;267(3):727-748.
37. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein-ligand interaction profiler. *Nucleic Acids Res*. 2015;43(W1):W443-W447.
38. Meng XY, Zhang HX, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des*. 2011;7(2):146-157.
39. Chen YC. Beware of docking! *Trends Pharmacol Sci*. 2015;36(2):78-95.
40. Hollingsworth SA, Dror RO. Molecular dynamics simulation for all. *Neuron*. 2018;99(6):1129-1143.
41. Martínez L. Automatic identification of mobile and rigid substructures in molecular dynamics simulations and fractional structural fluctuation analysis. *PLoS One*. 2015;10(3):e0119264.
42. David CC, Jacobs DJ. Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol Biol*. 2014;1084:193-226.
43. De Vivo M, Masetti M, Bottegoni G, Cavalli A. Role of molecular dynamics and related methods in drug discovery. *J Med Chem*. 2016;59(9):4035-4061.
44. Amidon GL, Lennernäs H, Shah VP, Crison JR. A theoretical basis for a biopharmaceutic drug classification. *Pharm Res*. 1995;12(3):413-420.
45. Clark DE. In silico prediction of blood-brain barrier permeation. *Drug Discov Today*. 2003;8(20):927-933.
46. Kirchmair J, Göller AH, Lang D, et al. Predicting drug metabolism: experiment and/or computation? *Nat Rev Drug Discov*. 2015;14(6):387-404.
47. Wang J, Urban L. The impact of early ADME profiling on drug discovery and development strategy. *Drug Discov World*. 2004;5(4):73-86.
48. Bajorath J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discov*. 2002;1(11):882-894.
49. Keserü GM, Makara GM. Hit discovery and hit-to-lead approaches. *Drug Discov Today*. 2006;11(15-16):741-748.
50. Nicolaou CA, Brown N. Multi-objective optimization methods in drug design. *Drug Discov Today Technol*. 2013;10(3):e427-e435.
51. Segall MD. Multi-parameter optimization: identifying high quality compounds with a balance of properties. *Curr Pharm Des*. 2012;18(9):1292-1310.
52. Noble ME, Endicott JA, Johnson LN. Protein kinase inhibitors: insights into drug design from structure. *Science*. 2004;303(5665):1800-1805.
53. Silver LL. Challenges of antibacterial discovery. *Clin Microbiol Rev*. 2011;24(1):71-109.
54. Jin Z, Du X, Xu Y, et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature*. 2020;582(7811):289-293.
55. Sneader W. The discovery and synthesis of captopril. In: *Drug Discovery: A History*. Wiley; 2005.
56. Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov*. 2019;18(1):41-58.
57. Plowright AT, Johnstone C, Kihlberg J, et al. Hypothesis driven drug design: improving quality and effectiveness of the design-make-test-analyse cycle. *Drug Discov Today*. 2012;17(1-2):56-62.
58. Schneider G. Automating drug discovery. *Nat Rev Drug Discov*. 2018;17(2):97-113.
59. Renaud JP, Chung CW, Danielson UH, et al. Biophysics in drug discovery: impact, challenges and opportunities. *Nat Rev Drug Discov*. 2016;15(10):679-698.
60. Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today*. 2018;23(8):1538-1546.
61. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. *Proc 34th Int Conf Mach Learn*. 2017:1263-1272.
62. Wallach I, Dzamba M, Heifets A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv:1510.02855*. 2015.

63. Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci*. 2018;4(2):268-276.
64. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov*. 2019;18(6):463-477.
65. Ryde U, Söderhjelm P. Ligand-binding affinity estimates supported by quantum-mechanical methods. *Chem Rev*. 2016;116(9):5520-5566.
66. Chan HCS, Shan H, Dahoun T, Vogel H, Yuan S. Advancing drug discovery via artificial intelligence. *Trends Pharmacol Sci*. 2019;40(8):592-604.
67. Björnsson B, Borrebaeck C, Elander N, et al. Digital twins to personalize medicine. *Genome Med*. 2020;12(1):4.
68. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. 2017;18(1):83.
69. Ferreira LG, Dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. *Molecules*. 2015;20(7):13384-13421.
70. Nicholls A. What do we know and when do we know it? *J Comput Aided Mol Des*. 2008;22(3-4):239-255.