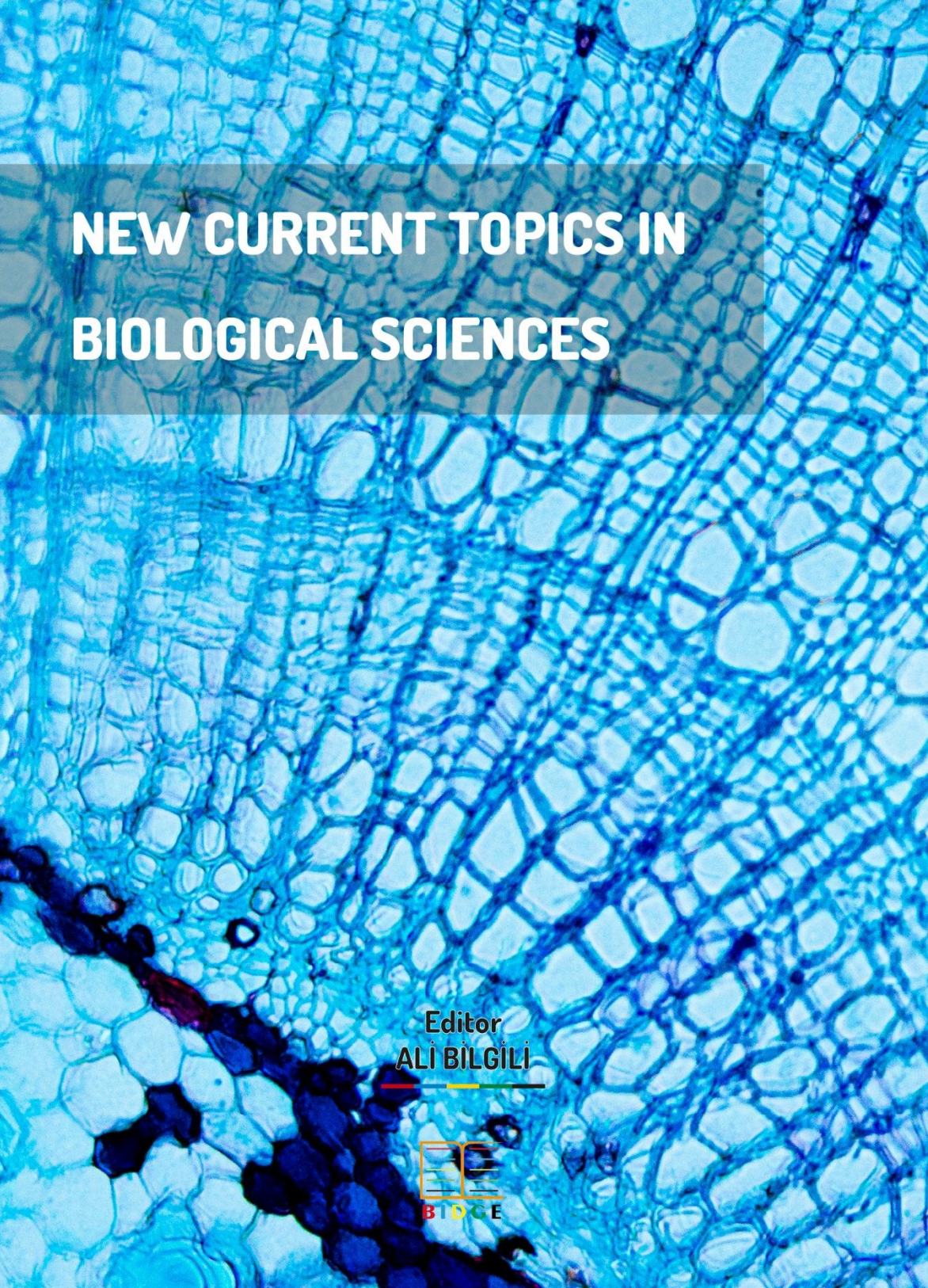
NEW CURRENT TOPICS IN BIOLOGICAL SCIENCES



Editor
ALİ BİLGİLİ

BİDGE

BIDGE Publications

NEW CURRENT TOPICS IN BIOLOGICAL SCIENCES

Editor: PROF. DR. ALİ BİLGİLİ

ISBN: 978-625-372-639-3

Page Layout: Gözde YÜCEL

1st Edition:

Publication Date: 26.03.2025

BIDGE Publications,

Certificate No: 71374

www.bidgeyayinlari.com.tr - bidgeyayinlari@gmail.com

Krc Bilişim Ticaret ve Organizasyon Ltd. Şti.

Güzeltepe Mahallesi Abidin Daver Sokak Sefer Apartmanı No: 7/9 Çankaya / Ankara

İÇİNDEKİLER

COMPUTER-AIDED DRUG DISCOVERY IN BIOTECHNOLOGY

PINAR SIYAH[1]

1.1 Introduction to Computer-Aided Drug Discovery (CADD)

Computer-Aided Drug Discovery (CADD) refers to the application of computational methods and tools to facilitate and optimize the drug discovery process. This interdisciplinary field integrates principles from chemistry, biology, physics, and computer science to design and analyze potential therapeutic compounds with enhanced efficiency and precision. The origins of in silico pharmacology can be traced back to the early 1960s, marking its formative stage, when Hansch and colleagues pioneered the development of quantitative structure–activity relationship (QSAR) models. By systematically analyzing data derived from molecular descriptors in relation to the physical, chemical, and biological properties of compounds, they introduced a computational framework aimed at predicting molecular bioactivity(A. C. Kaushik et al., 2018; Kostova, 2024a; Poongavanam & Ramaswamy, 2024). The subsequent development of molecular mechanics and molecular

[1] Department of Biochemistry, School of Pharmacy, Bahcesehir University, Istanbul, Turkey, pinar.siyah@med.bau.edu.tr, ORCID:0000-0003-1192-9416

dynamics simulations during the 1970s and 1980s marked a significant advancement in the field. The introduction of computational power allowed researchers to simulate biomolecular interactions with increasing accuracy. By the late 1990s, CADD tools were integrated into drug development pipelines in both academic and industrial settings, particularly for structure-based drug design (SBDD) and ligand-based drug design (LBDD) applications(Niazi & Mariam, 2024). The integration of biotechnological advances, such as genomics, proteomics, and high-throughput screening (HTS), into CADD methodologies has further transformed the landscape of drug discovery(Askari et al., 2023; Ouma et al., 2024). With the advent of next-generation sequencing (NGS) and structural biology techniques like cryo-electron microscopy (cryo-EM), researchers can now access high-resolution molecular structures of drug targets, thus enhancing the accuracy of predictive models(Ouma et al., 2024; Verkhivker et al., 2023). CADD methodologies have diversified into several distinct categories, each tailored to address specific stages of the drug discovery pipeline. Structure-based drug design (SBDD) utilizes the three-dimensional structures of target proteins to identify or optimize ligands through molecular docking and virtual screening techniques. Conversely, ligand-based drug design (LBDD) leverages known active molecules to predict new candidates by analyzing structural and physicochemical similarities (Bhunia et al., 2021; Yadav et al., 2022). Pharmacophore modeling, fragment-based drug discovery, and molecular dynamics simulations are additional strategies that have gained prominence in modern CADD practices (Sabe et al., 2021). The integration of artificial intelligence (AI) and machine learning (ML) into CADD has further revolutionized drug discovery by enabling the development of predictive models capable of learning from large datasets(Askari et al., 2023). Deep learning algorithms, such as convolutional neural networks (CNNs) and

recurrent neural networks (RNNs), are increasingly applied to predict molecular properties, drug-target interactions, and potential adverse effects with remarkable accuracy (Askr et al., 2023; J.-L. Yu et al., 2022).

1.3 CADD's Role in Modern Drug Development

The contemporary drug development landscape heavily relies on CADD to streamline and optimize various stages of the discovery pipeline(Rajkishan et al., 2021). CADD techniques contribute significantly to target identification, lead discovery, lead optimization, and preclinical testing, ultimately reducing the time and financial costs associated with traditional drug development methods(Xiang et al., 2012).

1.3.1. Target Identification and Validation: Target identification is the foundational step in drug discovery, wherein potential molecular targets involved in disease pathogenesis are identified. Bioinformatics tools and network pharmacology approaches are commonly employed to uncover novel targets by analyzing gene expression profiles, protein-protein interaction networks, and systems biology data(Berger & Iyengar, 2009; Hasan et al., 2020; Hopkins, 2008). Following identification, target validation is achieved through experimental techniques such as RNA interference (RNAi) and CRISPR/Cas9 gene editing, complemented by in silico methods like molecular docking and pathway analysis.

1.3.2. Lead Discovery and Optimization: Once a target is validated, lead compounds with potential therapeutic activity are identified through virtual screening of large chemical libraries. Molecular docking techniques show ligand-target interactions to predict binding affinities and interaction profiles, providing valuable insights into molecular recognition mechanisms(Owoloye et al., 2022; Paggi et al., 2024; Stanzione et al., 2021; Zhang et al., 2024a). Lead optimization subsequently involves modifying the chemical

structure to enhance potency, selectivity, and drug-like properties, guided by predictive models of pharmacokinetics and toxicity.

1.3.3. Preclinical and Clinical Development: In silico modeling plays a critical role in preclinical testing by predicting pharmacokinetic parameters such as absorption, distribution, metabolism, excretion, and toxicity (ADMET). Physiologically-based pharmacokinetic (PBPK) models and quantitative systems pharmacology (QSP) models facilitate the translation of preclinical findings to human physiology, thereby informing dose selection and clinical trial design(Rowland Yeo et al., 2024; Scheuher et al., 2023). One of the primary benefits of CADD lies in its ability to accelerate the timeline of drug discovery. Traditional drug discovery processes can take over a decade and require substantial financial investments, often exceeding $3.5 billion(Fernald et al., 2024). CADD tools, such as virtual screening and molecular docking, enable researchers to rapidly evaluate vast chemical libraries in silico, thereby identifying promising lead compounds for subsequent experimental validation(Hsieh et al., 2023; Verma & Pathak, 2022). Moreover, CADD enhances cost-effectiveness by minimizing the need for extensive laboratory testing. In silico approaches can predict pharmacokinetic and pharmacodynamic (PK/PD) properties, toxicity profiles, and potential off-target interactions before proceeding to in vitro and in vivo assays. Techniques like molecular dynamics simulations and free energy calculations provide detailed insights into ligand-receptor interactions, supporting rational drug design efforts(Pieroni et al., 2023; Procacci, 2021; Sadeghi et al., 2025; Syriopoulou et al., 2021). The success of CADD in modern drug development is exemplified by several FDA-approved drugs that were discovered or optimized using computational techniques. For instance, the HIV protease inhibitors saquinavir and indinavir were developed through structure-based drug design strategies that utilized X-ray crystallography data and molecular docking

simulations(Santos et al., 2025; Shanmuga Sundari et al., 2024). Furthermore, CADD has been instrumental in the development of small molecule inhibitors for kinases implicated in cancer. The discovery of imatinib, a selective BCR-ABL tyrosine kinase inhibitor for chronic myeloid leukemia, exemplifies the power of structure-based design supported by molecular docking and free energy calculations(Lexa & Carlson, 2012). In conclusion, CADD has evolved from rudimentary computational models to sophisticated, integrative approaches that leverage advancements in biotechnology and computational power. Its critical role in enhancing efficiency, cost-effectiveness, and success rates of drug discovery underscores its importance in the future of pharmaceutical innovation.

1.4 Key Techniques in Computer-Aided Drug Discovery (CADD)

1.4.1 Molecular Docking

Molecular docking is a computational technique that predicts the preferred orientation of a ligand when bound to a target protein, such as an enzyme or receptor, forming a stable complex. This technique simulates ligand-protein interactions to assess binding affinity and predict potential drug candidates with therapeutic potential. The docking process involves scoring functions that estimate the binding energy, which correlates with the strength and specificity of the interaction(de Angelo et al., 2025). The primary goal of molecular docking is to identify the optimal binding pose of a ligand within a target's active site by considering geometric and chemical complementarity. Docking algorithms employ search algorithms like genetic algorithms, simulated annealing, or particle swarm optimization to explore potential binding configurations(Fadahunsi et al., 2024; Kaza et al., 2024; Westhead et al., 1997). Molecular docking can be broadly classified into two

main types: rigid docking and flexible docking. In rigid docking, both the ligand and the receptor are treated as rigid bodies, with no allowance for conformational flexibility. While computationally less intensive, rigid docking may overlook crucial molecular interactions that depend on structural flexibility(Lexa & Carlson, 2012). Flexible docking accounts for the conformational changes of either the ligand, receptor, or both, providing a more realistic representation of biological interactions. Techniques such as induced-fit docking and ensemble docking are commonly used to incorporate flexibility into the docking process(Amaro et al., 2018; Chaudhury & Gray, 2008; Nabuurs et al., 2007). Docking algorithms are also categorized into global and local search methods. Global search algorithms explore the entire binding surface, while local search algorithms focus on specific regions identified through experimental data or prior knowledge of the target structure. Successful docking studies require high-quality structural data obtained from X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or cryo-electron microscopy(Plewczynski et al., 2011; Waszkowycz et al., 2011). Advanced docking techniques include consensus docking, ensemble docking, and hybrid docking, which combine elements from multiple docking methods to enhance predictive accuracy. Additionally, covalent docking, which accounts for the formation of covalent bonds between the ligand and target, is particularly relevant for enzyme inhibitors targeting catalytic residues(De Cesco et al., 2017; Lonsdale & Ward, 2018; Oyedele et al., 2023). Applications of molecular docking span various therapeutic domains, including oncology, infectious diseases, neurodegenerative disorders, and cardiovascular diseases. For instance, docking studies have facilitated the discovery of novel inhibitors targeting the SARS-CoV-2 main protease, providing valuable leads for COVID-19 drug development(G. Sharma et al., 2025; Tan et al., 2022; Yang et al., 2025).

1.4.2 Molecular Dynamics (MD) Simulations

Molecular dynamics simulations model the movement of atoms and molecules over time, providing insights into the structural, dynamic, and thermodynamic properties of biological systems. MD simulations track atomic positions by numerically solving Newton's equations of motion, offering a detailed understanding of biomolecular behavior at atomic resolution(Guvench, 2022; Huggins et al., 2019; Karplus & Petsko, 1990; Meier et al., 2013). MD simulations contribute to various stages of drug discovery, including docking validation, binding free energy calculations, and conformational analysis. Simulations can reveal transient binding sites, allosteric mechanisms, and conformational changes(Limongelli, 2020).

Key components of MD simulations include:

1. **Force Fields**: Force fields define the mathematical equations describing interatomic forces, including bonded (bond lengths, angles, dihedrals) and non-bonded interactions (electrostatics, van der Waals forces). Popular force fields include AMBER, CHARMM, GROMOS, and OPLS(Guvench & MacKerell, 2008; M. D. Smith et al., 2015).
2. **Solvation Models**: Biological processes occur in aqueous environments, necessitating the inclusion of explicit or implicit solvent models to capture solute-solvent interactions accurately(Van der Spoel et al., 2022).
3. **Simulation Protocols**: MD simulations often follow specific protocols, including energy minimization, equilibration, and production runs, conducted under constant temperature, pressure, or volume conditions(Brown & Clarke, 1984; Siyah, 2024b).

4. **Enhanced Sampling Techniques**: Methods like metadynamics, umbrella sampling, and replica exchange molecular dynamics (REMD) extend conventional MD by accelerating the exploration of conformational space, particularly for large biomolecules with complex energy landscapes(Bussi & Laio, 2020; D. Ghosh et al., 2024; M. Kumar et al., 2024).

Advanced applications of MD simulations include drug resistance mechanisms, protein folding studies, and the characterization of allosteric modulation. For instance, MD simulations have provided insights into the mechanism of action of allosteric inhibitors targeting the epidermal growth factor receptor (EGFR) in non-small cell lung cancer(Çoban, 2024a; Wan et al., 2019).

1.4.3 Quantitative Structure-Activity Relationship (QSAR)

QSAR models predict the biological activity of chemical compounds based on their structural features, facilitating the rational design of new drug candidates. QSAR analysis involves correlating chemical descriptors with biological activity through statistical and machine learning methods, providing insights into structure-activity relationships(Çoban, 2024b; Shahlaei, 2013). Quantitative Structure–Activity Relationship (QSAR) methodologies are typically classified into three main categories: 2D, 3D, and 4D QSAR. 2D QSAR relies on molecular descriptors derived from two-dimensional representations of chemical structures, including parameters such as hydrophobicity, electronic properties, and topological indices. 2D QSAR: Relies on molecular descriptors derived from two-dimensional structures, including hydrophobicity, electronic properties, and topological indices. Linear regression, partial least squares (PLS), and support vector machines (SVM) are

commonly employed in 2D QSAR models(S. Das et al., 2023). 3D QSAR: Incorporates three-dimensional structural information, such as molecular conformations and spatial distributions of electrostatic and steric properties. Comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) are widely applied in 3D QSAR studies(Banerjee et al., 2023). 4D QSAR: Extends 3D QSAR by incorporating time-dependent information to capture molecular flexibility and dynamic interactions with target proteins. Pharmacophore-based 4D QSAR models have been employed to identify inhibitors for G-protein-coupled receptors (GPCRs) and kinases(Tanrikulu & Schneider, 2008). Modern QSAR approaches increasingly utilize machine learning techniques like random forests, gradient boosting, and deep neural networks to improve predictive accuracy and generalizability. QSAR models have been instrumental in identifying inhibitors for kinases, GPCRs, and other therapeutic targets(Tropsha et al., 2024; Tsou et al., 2020). Applications of QSAR span diverse therapeutic areas, including antimicrobial resistance, cancer, neurodegenerative diseases, and metabolic disorders. Notable examples include the development of tyrosine kinase inhibitors for chronic myeloid leukemia, where QSAR models guided the optimization of imatinib and related compounds(Ciaffaglione et al., 2022). In summary, molecular docking, molecular dynamics simulations, and QSAR represent cornerstone techniques in CADD. Each method offers unique insights into drug-target interactions, contributing to more efficient, cost-effective, and rational drug discovery processes. The continuous integration of advanced computational tools and biotechnological innovations promises to further enhance the predictive power and applicability of CADD in modern pharmaceutical research.

1.5 Importance of Computer-Aided Drug Discovery (CADD) in Drug Discovery

Cost and Time Efficiency

The process of drug discovery is traditionally lengthy and resource-intensive, with an average development time exceeding a decade and costs often exceeding $3.5 billion, including expenditures on failed drug candidates(Fernald et al., 2024). Computer-Aided Drug Discovery (CADD) has emerged as a transformative tool to mitigate these challenges by integrating computational methods into various stages of the drug development pipeline, significantly reducing both time and cost. CADD accelerates the initial phases of drug discovery by facilitating the identification and optimization of lead compounds through virtual screening and molecular docking. These techniques allow for the in silico evaluation of thousands to millions of chemical compounds, prioritizing candidates with high binding affinity and favorable pharmacokinetic profiles for subsequent experimental validation(Niazi & Mariam, 2023). For instance, virtual screening studies have identified promising inhibitors for key drug targets, such as HIV protease and SARS-CoV-2 main protease, without the need for costly high-throughput screening (HTS) experiments(A. Singh, 2024). In addition to reducing time in the lead discovery phase, CADD contributes to time efficiency during lead optimization by predicting how structural modifications affect activity, selectivity, and pharmacokinetics. Molecular dynamics (MD) simulations and free energy calculations provide insights into ligand-target interactions at the atomic level, guiding the rational design of more potent and selective drug candidates(Parise et al., 2024). These predictions, when combined with machine learning algorithms, enable the rapid iteration of chemical structures, significantly shortening the optimization timeline. Cost efficiency is another major advantage of CADD, primarily due to the reduced reliance on

resource-intensive experimental procedures. Traditional drug discovery requires extensive chemical synthesis and biological testing to identify viable candidates. CADD tools, by contrast, allow for the virtual screening of large chemical libraries, minimizing the number of compounds synthesized and tested experimentally(Siddiqui et al., 2025). This reduction in laboratory workload directly translates to cost savings, as evidenced by the use of structure-based drug design (SBDD) in the development of HIV protease inhibitors, which significantly decreased the number of compounds requiring synthesis(Arthur et al., 2024; Siddiqui et al., 2025). Moreover, CADD platforms are increasingly integrated with high-throughput and ultra-large-scale screening pipelines, leveraging distributed computing resources such as cloud-based infrastructures. This approach has been instrumental in projects like the COVID Moonshot, where thousands of compounds were evaluated through distributed virtual screening efforts, accelerating the identification of potential therapeutics within months(Achdout et al., 2020; Von Delft et al., 2023).

Reducing Drug Failure Rates

The high attrition rates in drug development, particularly during clinical trials, remain a significant challenge. Historically, less than 10% of drug candidates entering clinical trials ultimately receive regulatory approval, with efficacy and safety issues being the primary reasons for failure(Sun et al., 2022). CADD plays a pivotal role in addressing these challenges by improving early-stage candidate selection and providing mechanistic insights into drug-target interactions. One of the primary ways CADD reduces failure rates is through enhanced target validation. Accurate target identification and validation are crucial for therapeutic success, as drugs designed against poorly characterized or irrelevant targets are likely to fail in later stages(Pun et al., 2023). Bioinformatics and systems biology tools, integrated with CADD techniques, enable the

identification of druggable targets by analyzing gene expression profiles, protein interaction networks, and disease-associated mutations.

Predictive modeling of pharmacokinetics and pharmacodynamics (PK/PD) properties also contributes to higher success rates. In silico methods like quantitative structure-activity relationship (QSAR) modeling, physiologically-based pharmacokinetic (PBPK) modeling, and molecular docking help predict a compound's absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles early in the discovery process(E. P. Chen et al., 2024; D. Kaushik & Kaushik, 2024a). By identifying potential liabilities such as poor solubility, off-target interactions, or metabolic instability, these tools guide the selection of candidates with favorable drug-like properties, thereby reducing the likelihood of clinical-stage failures. Molecular dynamics simulations provide further insights into the structural basis of drug resistance, an issue commonly encountered in the development of antibiotics, antivirals, and anticancer agents. For example, MD studies of HIV protease have elucidated the molecular mechanisms underlying resistance mutations, guiding the development of next-generation inhibitors with improved efficacy against resistant viral strains(Ali et al., 2010; A. K. Ghosh et al., 2008). Additionally, CADD facilitates the design of drugs with reduced toxicity profiles by predicting off-target effects and potential adverse reactions. Computational approaches such as inverse docking and polypharmacology modeling assess the interaction of candidate compounds with multiple biological targets, helping to identify and mitigate potential side effects(Kabir & Muth, 2022; Peters, 2024; Rigby, 2024). This proactive identification of toxicity risks aligns with regulatory guidelines emphasizing early safety assessments, ultimately contributing to higher clinical trial success rates. The integration of artificial intelligence (AI) and machine learning (ML)

with CADD has further improved predictive accuracy. Deep learning models trained on large-scale datasets of drug-target interactions, chemical structures, and clinical outcomes are now capable of identifying novel drug candidates with unprecedented speed and reliability(Abbasi et al., 2021; M. Sharma et al., 2025). These models have been applied to repurpose existing drugs for new therapeutic indications, such as the identification of baricitinib as a potential treatment for COVID-19 within weeks of the pandemic's onset(Janik et al., 2021; Saber-Ayad et al., 2021; D. P. Smith et al., 2021). In summary, CADD has become an indispensable component of modern drug discovery by enhancing cost and time efficiency, improving the predictive accuracy of drug-target interactions, and reducing clinical failure rates. Its continued evolution, driven by advancements in computational power, biotechnological innovations, and machine learning techniques, promises to further accelerate the development of safe and effective therapeutics across diverse disease areas. The integration of Computer-Aided Drug Discovery (CADD) with emerging technologies has significantly transformed drug design and development. Innovations in artificial intelligence (AI), machine learning (ML), and quantum computing have enhanced the predictive power, accuracy, and efficiency of CADD methodologies. This chapter explores the synergistic relationship between CADD and these advanced technologies, highlighting their applications, benefits, and the scientific advancements driving this integration.

1.6 AI and Machine Learning in CADD

Artificial intelligence (AI) and machine learning (ML) have revolutionized numerous scientific fields, including drug discovery. AI encompasses a broad range of computational techniques that simulate human intelligence to solve complex problems, while ML, a subset of AI, focuses on developing algorithms that learn from data and improve predictions over time. The application of AI/ML in

CADD has significantly accelerated drug discovery by automating processes, identifying complex patterns, and optimizing molecular design.

AI and ML techniques contribute to CADD in various stages:

1. **Target Identification and Validation**:
    - AI algorithms analyze multi-omics data, including genomics, proteomics, and transcriptomics, to identify potential drug targets with high disease relevance.
    - Deep learning models have been used to identify novel targets for cancer therapies by analyzing patterns of gene expression and mutation frequencies(Alharbi & Vakanski, 2023; Yuan et al., 2016).
2. **Virtual Screening and Molecular Docking**:
    - AI-driven virtual screening employs predictive models to screen millions of compounds against potential targets, significantly reducing the search space for experimental validation.
    - Deep docking algorithms have been developed to improve docking accuracy by leveraging convolutional neural networks (CNNs) trained on structural and activity data(Gentile et al., 2020; Hassan-Harrirou et al., 2020; Zhang et al., 2024b).
3. **Lead Optimization**:
    - ML models predict structure-activity relationships (SAR) by learning from past drug discovery projects, guiding chemical modifications to improve potency, selectivity, and pharmacokinetic properties.
    - Generative adversarial networks (GANs) and variational autoencoders (VAEs) are employed to

design novel molecules with optimized properties(Martinelli, 2022).

4. **ADMET Prediction**:
    - Predicting absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles is crucial for assessing drug-like properties.
    - AI-based models, such as DeepTox and ADMETlab, utilize large datasets to accurately forecast pharmacokinetic behavior and potential toxicity(Dong et al., 2018; Mayr et al., 2016; Sinha et al., 2023).

Key AI/ML Techniques in CADD:

Artificial intelligence (AI) and machine learning (ML) have become integral to modern computer-aided drug design, offering powerful tools to enhance various stages of the drug discovery pipeline. Artificial intelligence (AI) and machine learning (ML) have become integral to modern computer-aided drug design, offering powerful tools to enhance various stages of the drug discovery pipeline. Supervised learning techniques are widely employed to predict molecular activity, toxicity profiles, and pharmacokinetic properties using labeled datasets, thereby enabling more accurate and efficient compound prioritization. In contrast, unsupervised learning methods are utilized to identify hidden structures or patterns within complex biological and chemical datasets, supporting applications such as target identification, chemical clustering, and biomarker discovery. Moreover, reinforcement learning has emerged as a promising approach in de novo drug design, where models are trained to iteratively generate and optimize novel molecular structures by maximizing predefined reward functions (Olivecrona et al., 2017; Wang et al., 2022). Collectively, these AI/ML strategies

facilitate data-driven decision-making and accelerate the identification of promising therapeutic candidates.

Challenges and Future Directions:

Despite its success, AI/ML integration in CADD faces challenges related to data quality, model interpretability, and generalization across diverse chemical spaces. Ongoing research aims to develop more transparent algorithms, improve data curation processes, and incorporate domain-specific knowledge into predictive models(Alizadehsani et al., 2024; Han et al., 2023).

1.7 Quantum Computing in CADD

Quantum computing represents a paradigm shift in computational capabilities, promising to overcome limitations associated with classical computing in molecular simulations. Quantum computers leverage principles of quantum mechanics, such as superposition, entanglement, and tunneling, to perform calculations that would be infeasible on classical hardware(Yazdi, 2024).

Applications of Quantum Computing in Drug Discovery:

1. **Molecular Dynamics Simulations**: Quantum algorithms simulate molecular interactions at unprecedented resolution by solving the Schrödinger equation more efficiently than classical methods. Quantum-enhanced simulations have been applied to study enzyme-substrate interactions and conformational changes in complex biomolecules(N. B. Singh, 2024).
2. **Drug-Target Interaction Modeling**: Quantum computing provides more accurate predictions of drug-target binding by accounting for quantum effects in molecular interactions,

such as hydrogen bonding and electronic polarization(Li et al., 2024).

3. **Optimization of Drug Candidates**: Quantum optimization algorithms, like the Quantum Approximate Optimization Algorithm (QAOA), solve complex optimization problems related to molecular design, enabling the discovery of optimal drug candidates with desired properties(Salloum et al., n.d.).
4. **Quantum Machine Learning (QML)**: QML combines quantum computing with ML techniques to analyze large-scale chemical datasets more efficiently. Recent studies have demonstrated the application of QML in predicting molecular properties, such as solubility and bioactivity, with enhanced performance compared to classical models(Tripathi, 2025).

Challenges and Future Directions:

Quantum computing in CADD is still in its infancy, with significant challenges related to hardware limitations, error rates, and algorithm development. Current quantum devices, referred to as Noisy Intermediate-Scale Quantum (NISQ) systems, require further advancements in qubit stability and error correction before realizing their full potential in drug discovery(Pyrkov et al., 2023a).

Synergistic Integration of AI/ML and Quantum Computing with CADD

The convergence of AI/ML and quantum computing with CADD represents a promising frontier for drug discovery. Quantum-enhanced machine learning models can accelerate the development of predictive algorithms, while AI techniques can assist in optimizing quantum algorithms for specific drug discovery tasks(Pyrkov et al., 2023b; Vasanthakumar & Singh, 2024).

Collaborative efforts between academia, industry, and government agencies are fostering the development of hybrid computational frameworks that integrate classical, AI-driven, and quantum-based approaches. These frameworks are expected to address existing challenges in drug design, such as polypharmacology, drug resistance, and personalized medicine. In conclusion, the integration of AI, ML, and quantum computing with CADD has significantly advanced drug discovery capabilities. These technologies provide novel insights into molecular behavior, optimize drug design processes, and accelerate the identification of promising therapeutic candidates. Ongoing research and technological advancements are poised to further enhance the efficiency, accuracy, and success rates of drug discovery endeavors.

1.8 CADD's Applications in Drug Discovery

The applications of Computer-Aided Drug Discovery (CADD) encompass a vast array of domains in pharmaceutical research and development. This chapter provides an in-depth exploration of these applications, highlighting the methodologies, techniques, and outcomes associated with CADD implementation across different stages of drug discovery.

1.8.1 Target Identification and Validation

Target identification is the foundational step in drug discovery, involving the selection of specific molecular targets implicated in disease mechanisms. CADD facilitates this process by integrating bioinformatics tools, systems biology approaches, and cheminformatics techniques to analyze genetic, proteomic, and metabolomic data. Bioinformatics platforms like STRING and BioGRID assist researchers in constructing protein interaction networks and identifying disease-associated targets(Athar et al., 2024a, 2024b; Xiong et al., 2024). Additionally, molecular docking and molecular dynamics (MD) simulations are employed to validate

the target's druggability by assessing ligand-binding capabilities(Shoaib et al., 2023). The increasing availability of genome-wide association studies (GWAS) data has enabled the identification of novel targets linked to complex diseases like cancer, Alzheimer's, and cardiovascular diseases. Machine learning algorithms have further refined target identification by uncovering hidden patterns in multi-omics datasets. Tools like DeepTarget utilize convolutional neural networks (CNNs) to predict druggable targets based on structural and functional features of proteins(Lee et al., 2016).

1.8.2 Lead Identification through Virtual Screening

Lead identification is the process of selecting chemical compounds with potential therapeutic activity. CADD tools facilitate this task by enabling the virtual screening of large chemical libraries, predicting interactions with the target site based on molecular docking, pharmacophore modeling, and similarity analysis(B. Chen et al., 2025; Naithani & Guleria, 2024). Virtual screening strategies can be broadly classified into structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS). SBVS relies on the three-dimensional structure of the target protein, using docking algorithms to simulate potential ligand interactions. In contrast, LBVS employs known active compounds as templates to identify structurally similar candidates. High-throughput virtual screening (HTVS) has become a cornerstone in antiviral drug discovery, as demonstrated by the rapid identification of potential inhibitors against SARS-CoV-2(Olatunde, 2024). Moreover, AI-driven approaches like Chemprop and DeepChem have significantly enhanced the efficiency of virtual screening processes(Noor et al., 2024; Sivula et al., 2023).

1.8.3 Lead Optimization

Lead optimization refines initial hits to enhance potency, selectivity, and pharmacokinetic properties. CADD techniques such as free energy perturbation (FEP), quantum mechanical (QM) calculations, and molecular dynamics simulations guide the structural modifications of lead compounds(Cavasotto et al., 2018; de Oliveira et al., 2023). Molecular docking and pharmacophore modeling help identify essential molecular features for activity, while QSAR models predict the impact of chemical modifications on bioactivity(Kostova, 2024b; Xu et al., 2024). Automated tools like AutoDock, Schrödinger, and MOE streamline this optimization process. In oncology research, for instance, the optimization of kinase inhibitors often involves analyzing hydrogen bonding patterns, hydrophobic interactions, and molecular dynamics simulations to predict potential resistance mutations(Banavath et al., 2014).

1.8.4 ADMET Prediction and Toxicity Profiling

The accurate prediction and profiling of Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties are crucial components in the drug discovery process. ADMET predictions help minimize late-stage drug development failures by providing insights into how a compound behaves within a biological system. Computer-Aided Drug Discovery (CADD) has revolutionized this domain by integrating advanced computational tools, including Quantitative Structure-Activity Relationship (QSAR) models, molecular docking, physiologically-based pharmacokinetic (PBPK) simulations, and machine learning algorithms to assess these properties accurately(D. Kaushik & Kaushik, 2024b). ADMET considerations gained prominence during the 1990s as high-throughput screening techniques flooded the pharmaceutical industry with candidate molecules, many of which

failed during clinical trials due to suboptimal pharmacokinetics or toxicity issues(Zhou & Theil, 2015). Since then, CADD-based tools have evolved to predict ADMET properties early in the discovery pipeline, reducing time, costs, and attrition rates.

Computational Techniques for ADMET Prediction

QSAR models establish correlations between chemical structure and biological activity. These models use molecular descriptors—such as hydrophobicity, electronic properties, and topological indices—to predict ADMET parameters(Patel et al., 2014; Popelier & Smith, 2006). QSAR techniques have successfully identified hepatotoxicity risks in drug candidates by analyzing structural features associated with liver damage(Kelleci Çelik & Karaduman, 2023). Molecular docking simulates the binding of drug candidates to target proteins, providing insights into absorption mechanisms and potential off-target interactions. Tools such as AutoDock and Glide predict how drugs interact with transporters, enzymes, and receptors involved in ADMET processes(Siyah et al., 2021). Physiologically-Based Pharmacokinetic (PBPK) models simulate the distribution and metabolism of drugs across tissues and organs using physiological and biochemical parameters. Simcyp and GastroPlus are commonly employed PBPK tools that predict pharmacokinetic profiles across diverse populations(Demeester et al., 2023). Advanced machine learning techniques, including random forests, support vector machines, and deep neural networks, have significantly improved ADMET prediction accuracy. DeepTox, for instance, uses molecular fingerprints and graph-based neural networks to predict toxicological endpoints with high precision(Cavasotto & Scardino, 2022).

Key ADMET Parameters in Drug Development

Absorption determines the extent and rate at which a drug enters systemic circulation. CADD tools predict oral bioavailability by simulating interactions with intestinal transporters like P-glycoprotein (P-gp). Lipinski's Rule of Five remains a cornerstone for assessing drug permeability and solubility(Kiani & Jabeen, 2019; Siyah, n.d.). Distribution describes how a drug disperses throughout the body, often dictated by protein binding, membrane permeability, and tissue affinity. Computational models predict the volume of distribution (Vd) by analyzing molecular weight, polarity, and ionization constants(Holt et al., 2019). Metabolism involves the biotransformation of drugs into metabolites, primarily mediated by cytochrome P450 enzymes (CYPs). CADD-based docking and QSAR models predict metabolic stability and potential drug-drug interactions by simulating CYP interactions(Mikov et al., 2017; M. Zhao et al., 2021). Excretion refers to the elimination of drug molecules and metabolites from the body. PBPK models simulate renal clearance by predicting interactions with transporters like organic anion transporter 1 (OAT1)(Hsueh et al., 2018). Toxicity profiling assesses potential adverse effects and safety risks. In silico toxicity models evaluate hepatotoxicity, cardiotoxicity, nephrotoxicity, and genotoxicity using large-scale toxicity databases. The Tox21 initiative, for example, has provided datasets enabling the development of machine learning models for toxicity prediction(Manzoor, 2025).

Applications of ADMET Predictions

ADMET predictions streamline candidate selection by prioritizing compounds with favorable pharmacokinetic properties and minimal toxicity risks(Akinola et al., 2025). Pharmaceutical companies use CADD-based ADMET models to generate data for regulatory submissions to agencies like the FDA and EMA. These

models assist in identifying potential safety concerns, guiding preclinical and clinical trial designs(A. Sharma et al., 2025). ADMET modeling supports precision medicine by predicting drug responses in patient subpopulations with distinct genetic, physiological, and environmental characteristics(V. Kumar & Roy, 2025).

Challenges and Limitations

1. **Data Quality and Curation**: Inaccurate or incomplete datasets may compromise model reliability.
2. **Model Interpretability**: Deep learning models often operate as "black boxes," complicating mechanistic interpretations.
3. **Inter-Individual Variability**: Biological variability across populations challenges the generalizability of predictive models.

The future of ADMET prediction lies in integrating multi-omics data, quantum computing, and explainable AI techniques. Enhanced collaboration between academia, industry, and regulatory bodies will drive advancements in this domain, ultimately improving drug development outcomes.

Conclusion

ADMET prediction and toxicity profiling represent critical components of modern drug discovery. CADD tools, through the integration of QSAR models, molecular docking, PBPK simulations, and machine learning algorithms, provide valuable insights into drug behavior and safety profiles. As computational techniques evolve, ADMET predictions will continue to enhance drug discovery efficiency, reduce development costs, and improve clinical success rates.

1.9 Drug Repurposing and Repositioning

Drug repurposing, also referred to as drug repositioning, is a strategy that identifies new therapeutic uses for existing drugs, whether they are approved or investigational compounds. This approach significantly expedites the drug development process by bypassing the early stages of drug discovery, such as target identification and lead optimization. Computer-Aided Drug Discovery (CADD) plays an integral role in this domain by leveraging computational tools to predict drug-target interactions, analyze molecular similarities, and uncover novel applications for existing pharmacological agents(N. R. Das & Jena, 2024; Tiwari & Singh, 2022). The concept of drug repurposing has long been recognized in biomedical research, referring to the strategy of identifying new therapeutic uses for existing or previously approved drugs(Mishra et al., 2024; Tanoli et al., 2025). The successful repositioning of thalidomide from a sedative to a treatment for multiple myeloma is a prominent example of this strategy(Palumbo et al., 2008a, 2008b). The advent of CADD has revolutionized this field by providing a systematic, cost-effective, and time-efficient methodology for discovering new applications for known drugs.

Methodologies in Drug Repurposing

CADD employs several computational strategies to identify potential repurposing candidates:

1. **Molecular Docking**: Molecular docking techniques simulate the interaction between a drug molecule and a biological target, assessing binding affinity and interaction stability. Virtual screening through docking studies has led to the identification of potential SARS-CoV-2 inhibitors by repurposing drugs such as remdesivir and baricitinib(Almulhim et al., 2025).

2. **Ligand-Based Drug Design (LBDD)**: LBDD identifies compounds with structural similarity to known active molecules, predicting potential off-target effects and therapeutic applications(Nandi et al., 2024).
3. **Structure-Based Drug Design (SBDD)**: SBDD utilizes three-dimensional structural information of biological targets to identify repurposing candidates. Computational tools like AutoDock and Schrödinger facilitate the virtual screening of extensive compound libraries against known target structures(Siyah, 2024a).
4. **Pharmacophore Modeling**: Pharmacophore models represent the spatial arrangement of features essential for molecular recognition. Pharmacophore-based screening has successfully identified repositioned drugs for neurodegenerative disorders by targeting amyloid-beta aggregates(Maniam & Maniam, 2024).
5. **Network Pharmacology and Systems Biology**: Network-based approaches analyze drug-target interactions within biological pathways, revealing potential off-target effects and alternative therapeutic targets. Tools like STITCH and Cytoscape enable the visualization and analysis of drug-disease networks(Joshi et al., 2024).

Applications in Various Therapeutic Areas

CADD-based drug repurposing has significantly contributed to infectious disease research. The rapid identification of baricitinib as a potential COVID-19 treatment exemplifies the power of computational techniques in emergency situations(Ozaybi et al.,

2024). By utilizing AI-driven virtual screening, researchers analyzed the interaction of approved drugs with SARS-CoV-2 proteins, leading to the repurposing of several antiviral agents.

Cancer therapeutics have greatly benefited from drug repurposing strategies. For instance, the repositioning of metformin, originally developed for type 2 diabetes, has shown promise as an adjuvant in cancer therapy due to its impact on cellular metabolism and proliferation pathways(Gadducci et al., 2016). Computational analyses of cancer genomics data have identified multiple potential repurposing candidates across various malignancies.

Neurodegenerative diseases, characterized by complex and multifactorial pathologies, have become a focal point for drug repurposing. Donepezil, an acetylcholinesterase inhibitor for Alzheimer's disease, was initially explored for its effects on cognitive impairment in other neurological conditions using in silico pharmacophore modeling(X. Zhao et al., 2024).

The repurposing of JAK inhibitors, such as tofacitinib, for autoimmune disorders like rheumatoid arthritis has been guided by CADD-based analyses of cytokine signaling pathways. Molecular docking studies have revealed their potential application in psoriasis and inflammatory bowel diseases.

Advantages of CADD in Drug Repurposing

Repurposing drugs with established safety profiles reduces the need for extensive preclinical testing, significantly lowering development costs(Kiriiri et al., 2020). Computational screening accelerates the identification of candidates, particularly during pandemic situations when rapid therapeutic solutions are needed(H. Yu et al., 2020). CADD identifies off-target interactions and potential adverse effects early in the discovery process, improving clinical trial success rates(Iwaloye et al., 2023).

Despite its benefits, CADD-based drug repurposing faces several challenges. Accurate predictions require comprehensive, high-quality datasets encompassing chemical structures, bioactivity profiles, and clinical outcomes. Machine learning models may struggle to generalize across different chemical and biological contexts if trained on limited datasets(S. Singh & Sunoj, 2023). The multifactorial nature of many diseases complicates the identification of new therapeutic targets and drug mechanisms.

Future Perspectives

The integration of CADD with emerging technologies like artificial intelligence, quantum computing, and multi-omics data analysis is expected to enhance the efficiency and accuracy of drug repurposing efforts. Collaborative platforms like OpenTargets and COVID Moonshot exemplify the potential of community-driven approaches in accelerating drug discovery(Ackloo et al., 2022; Ekert et al., 2020).

Conclusion

Drug repurposing, powered by CADD, offers a promising avenue for addressing unmet medical needs across various therapeutic areas. By leveraging computational tools to explore novel applications of existing drugs, researchers can expedite the development of effective treatments, optimize resource allocation, and mitigate the risks associated with de novo drug discovery. The ongoing evolution of CADD methodologies, in conjunction with advancements in AI and big data analytics, will continue to drive innovation in this field.

REFERENCES

Abbasi, K., Razzaghi, P., Poso, A., Ghanbari-Ara, S., & Masoudi-Nejad, A. (2021). Deep learning in drug target interaction prediction: current and future perspectives. *Current Medicinal Chemistry*, *28*(11), 2100–2113.

Achdout, H., Aimon, A., Bar-David, E., & Morris, G. M. (2020). COVID moonshot: open science discovery of SARS-CoV-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *BioRxiv*.

Ackloo, S., Al-Awar, R., Amaro, R. E., Arrowsmith, C. H., Azevedo, H., Batey, R. A., Bengio, Y., Betz, U. A. K., Bologa, C. G., & Chodera, J. D. (2022). CACHE (Critical Assessment of Computational Hit-finding Experiments): A public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nature Reviews Chemistry*, *6*(4), 287–295.

Akinola, L. K., Abdalla, M., Abubakar, H. S., Abubakar, K. A., Rabiu, S. M., Rufai, M., & Umar, N. Z. (2025). In silico discovery of new ornithine decarboxylase inhibitor with potential for human African trypanosomiasis treatment. *Discover Chemistry*, *2*(1), 37.

Alharbi, F., & Vakanski, A. (2023). Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering*, *10*(2), 173.

Ali, A., Bandaranayake, R. M., Cai, Y., King, N. M., Kolli, M., Mittal, S., Murzycki, J. F., Nalam, M. N. L., Nalivaika, E. A., & Özen, A. (2010). Molecular basis for drug resistance in HIV-1 protease. *Viruses*, *2*(11), 2509–2535.

Alizadehsani, R., Oyelere, S. S., Hussain, S., Jagatheesaperumal, S. K., Calixto, R. R., Rahouti, M., Roshanzamir, M., & De Albuquerque, V. H. C. (2024). Explainable artificial intelligence for drug discovery and development-a comprehensive survey. *IEEE Access*.

Almulhim, M., Ghasemian, A., Memariani, M., Karami, F., Yassen, A. S. A., Alexiou, A., Papadakis, M., & Batiha, G. E.-S. (2025). Drug repositioning as a promising approach for the eradication of emerging and re-emerging viral agents. *Molecular Diversity*, 1–21.

Amaro, R. E., Baudry, J., Chodera, J., Demir, Ö., McCammon, J. A., Miao, Y., & Smith, J. C. (2018). Ensemble Docking in Drug Discovery. *Biophysical Journal*, *114*(10), 2271–2278. https://doi.org/https://doi.org/10.1016/j.bpj.2018.02.038

Arthur, D., Dogara, G. B., Babafem, A. A., & Akan, J. C. (2024). Computational Modeling of Chemical Inhibitors Targeting the HIV-1 Main Protease: A Review. *Medicinal and Medical Chemistry*, *1*(2), 71–83.

Askari, S., Ghofrani, A., & Taherdoost, H. (2023). Transforming Drug Design: Innovations in Computer-Aided Discovery for Biosimilar Agents. In *BioMedInformatics* (Vol. 3, Issue 4). https://doi.org/10.3390/biomedinformatics3040070

Askr, H., Elgeldawi, E., Aboul Ella, H., Elshaier, Y. A. M. M., Gomaa, M. M., & Hassanien, A. E. (2023). Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review*, *56*(7), 5975–6037. https://doi.org/10.1007/s10462-022-10306-1

Athar, M., Manhas, A. N. U., Rana, N., & Irfan, A. (2024a). Computational and bioinformatics tools for understanding disease mechanisms. *Biocell*, *48*(6), 935–944.

Athar, M., Manhas, A. N. U., Rana, N., & Irfan, A. (2024b). Computational and bioinformatics tools for understanding disease mechanisms. *Biocell*, *48*(6), 935–944.

Banavath, H. N., Sharma, O. P., Kumar, M. S., & Baskaran, R. (2014). Identification of novel tyrosine kinase inhibitors for drug resistant T315I mutant BCR-ABL: a virtual screening and molecular dynamics simulations study. *Scientific Reports*, *4*(1), 6948.

Banerjee, S., Baidya, S. K., Adhikari, N., & Jha, T. (2023). 3D-QSAR studies: CoMFA, CoMSIA, and topomer CoMFA methods. In *Modeling inhibitors of matrix metalloproteinases* (pp. 32–53). CRC Press.

Berger, S. I., & Iyengar, R. (2009). Network analyses in systems pharmacology. *Bioinformatics*, *25*(19), 2466–2472. https://doi.org/10.1093/bioinformatics/btp465

Bhunia, S. S., Saxena, M., & Saxena, A. K. (2021). Ligand- and Structure-Based Virtual Screening in Drug Discovery. In A. K. Saxena (Ed.), *Biophysical and Computational Tools in Drug Discovery* (pp. 281–339). Springer International Publishing. https://doi.org/10.1007/7355_2021_130

Brown, D., & Clarke, J. H. R. (1984). A comparison of constant energy, constant temperature and constant pressure ensembles in molecular dynamics simulations of atomic liquids. *Molecular Physics*, *51*(5), 1243–1252.

Bussi, G., & Laio, A. (2020). Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, *2*(4), 200–212. https://doi.org/10.1038/s42254-020-0153-0

Cavasotto, C. N., Adler, N. S., & Aucar, M. G. (2018). Quantum chemical approaches in structure-based virtual screening and lead optimization. *Frontiers in Chemistry*, *6*, 188.

Cavasotto, C. N., & Scardino, V. (2022). Machine learning toxicity prediction: latest advances by toxicity end point. *ACS Omega*, *7*(51), 47536–47546.

Chaudhury, S., & Gray, J. J. (2008). Conformer Selection and Induced Fit in Flexible Backbone Protein–Protein Docking Using Computational and NMR Ensembles. *Journal of Molecular Biology*, *381*(4), 1068–1087. https://doi.org/https://doi.org/10.1016/j.jmb.2008.05.042

Chen, B., Liu, S., Xia, H., Li, X., & Zhang, Y. (2025). Computer-Aided Drug Design in Research on Chinese Materia Medica: Methods, Applications, Advantages, and Challenges. *Pharmaceutics*, *17*(3), 315.

Chen, E. P., Dutta, S., Ho, M.-H., & DeMartino, M. P. (2024). Model-based virtual PK/PD exploration and machine learning approach to define PK drivers in early drug discovery. *Journal of Medicinal Chemistry*, *67*(5), 3727–3740.

Ciaffaglione, V., Consoli, V., Intagliata, S., Marrazzo, A., Romeo, G., Pittalà, V., Greish, K., Vanella, L., Floresta, G., & Rescifina, A. (2022). Novel tyrosine kinase inhibitors to target chronic myeloid leukemia. *Molecules*, *27*(10), 3220.

Çoban, G. (2024a). Structure-based virtual screening and molecular dynamics simulations for detecting novel candidates for allosteric inhibition of EGFRT790M. *Journal of Biomolecular Structure and Dynamics*, *42*(2), 571–597. https://doi.org/10.1080/07391102.2023.2194425

Çoban, G. (2024b). Structure-based virtual screening and molecular dynamics simulations for detecting novel candidates for allosteric inhibition of EGFRT790M. *Journal of Biomolecular Structure and Dynamics*, *42*(2), 571–597. https://doi.org/10.1080/07391102.2023.2194425

Das, N. R., & Jena, G. K. (2024). Bioinformatics and CADD approaches in drug discovery. In *Biochemical and Molecular Pharmacology in Drug Discovery* (pp. 313–321). Elsevier.

Das, S., Amin, S. A., Gayen, S., & Jha, T. (2023). 2D-QSAR Studies: Regression and Classification-Based QSAR Studies. In *Modeling Inhibitors of Matrix Metalloproteinases* (pp. 3–31). CRC Press.

de Angelo, R. M., de Sousa, D. S., da Silva, A. P., Chiari, L. P. A., da Silva, A. B. F., & Honorio, K. M. (2025). Molecular Docking: State-of-the-Art Scoring Functions and Search Algorithms. In *Computer-Aided and Machine Learning-Driven Drug Design: From Theory to Applications* (pp. 163–198). Springer.

De Cesco, S., Kurian, J., Dufresne, C., Mittermaier, A. K., & Moitessier, N. (2017). Covalent inhibitors design and discovery. *European Journal of Medicinal Chemistry*, *138*, 96–114. https://doi.org/https://doi.org/10.1016/j.ejmech.2017.06.019

de Oliveira, C., Leswing, K., Feng, S., Kanters, R., Abel, R., & Bhat, S. (2023). FEP protocol builder: optimization of free energy perturbation protocols using active learning. *Journal of Chemical Information and Modeling*, *63*(17), 5592–5603.

Demeester, C., Robins, D., Edwina, A. E., Tournoy, J., Augustijns, P., Ince, I., Lehmann, A., Vertzoni, M., & Schlender, J. F. (2023). Physiologically based pharmacokinetic (PBPK) modelling of oral drug absorption in older adults–an AGePOP review. *European Journal of Pharmaceutical Sciences*, *188*, 106496.

Dong, J., Wang, N.-N., Yao, Z.-J., Zhang, L., Cheng, Y., Ouyang, D., Lu, A.-P., & Cao, D.-S. (2018). ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *Journal of Cheminformatics*, *10*, 1–11.

Ekert, J. E., Deakyne, J., Pribul-Allen, P., Terry, R., Schofield, C., Jeong, C. G., Storey, J., Mohamet, L., Francis, J., & Naidoo, A. (2020). Recommended guidelines for developing, qualifying, and implementing complex in vitro models (CIVMs) for drug discovery. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, *25*(10), 1174–1190.

Fadahunsi, A. A., Uzoeto, H. O., Okoro, N. O., Cosmas, S., Durojaye, O. A., & Odiba, A. S. (2024). Revolutionizing drug discovery: an AI-powered transformation of molecular docking. *Medicinal Chemistry Research*, *33*(12), 2187–2203. https://doi.org/10.1007/s00044-024-03253-9

Fernald, K. D. S., Förster, P. C., Claassen, E., & van de Burgwal, L. H. M. (2024). The pharmaceutical productivity gap – Incremental decline in R&D efficiency despite transient improvements. *Drug Discovery Today*, *29*(11), 104160. https://doi.org/https://doi.org/10.1016/j.drudis.2024.104160

Gadducci, A., Biglia, N., Tana, R., Cosio, S., & Gallo, M. (2016). Metformin use and gynecological cancers: a novel treatment option emerging from drug repositioning. *Critical Reviews in Oncology/Hematology*, *105*, 73–83.

Gentile, F., Agrawal, V., Hsing, M., Ton, A.-T., Ban, F., Norinder, U., Gleave, M. E., & Cherkasov, A. (2020). Deep docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Central Science*, *6*(6), 939–949.

Ghosh, A. K., Chapsal, B. D., Weber, I. T., & Mitsuya, H. (2008). Design of HIV protease inhibitors targeting protein backbone: an effective strategy for combating drug resistance. *Accounts of Chemical Research*, *41*(1), 78–86.

Ghosh, D., Biswas, A., & Radhakrishna, M. (2024). Advanced computational approaches to understand protein

aggregation. *Biophysics Reviews*, *5*(2), 021302. https://doi.org/10.1063/5.0180691

Guvench, O. (2022). Atomic-Resolution Experimental Structural Biology and Molecular Dynamics Simulations of Hyaluronan and Its Complexes. In *Molecules* (Vol. 27, Issue 21). https://doi.org/10.3390/molecules27217276

Guvench, O., & MacKerell, A. D. (2008). Comparison of protein force fields for molecular dynamics simulations. *Methods in Molecular Biology*, *443*. https://doi.org/10.1007/978-1-59745-177-2_4

Han, R., Yoon, H., Kim, G., Lee, H., & Lee, Y. (2023). Revolutionizing medicinal chemistry: the application of artificial intelligence (AI) in early drug discovery. *Pharmaceuticals*, *16*(9), 1259.

Hasan, M. R., Paul, B. K., Ahmed, K., & Bhuyian, T. (2020). Design protein-protein interaction network and protein-drug interaction network for common cancer diseases: A bioinformatics approach. *Informatics in Medicine Unlocked*, *18*, 100311.

Hassan-Harrirou, H., Zhang, C., & Lemmin, T. (2020). RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. *Journal of Chemical Information and Modeling*, *60*(6), 2791–2802.

Holt, K., Nagar, S., & Korzekwa, K. (2019). Methods to predict volume of distribution. *Current Pharmacology Reports*, *5*, 391–399.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, *4*(11), 682–690. https://doi.org/10.1038/nchembio.118

Hsieh, C.-J., Giannakoulias, S., Petersson, E. J., & Mach, R. H. (2023). Computational chemistry for the identification of lead compounds for radiotracer development. *Pharmaceuticals*, *16*(2), 317.

Hsueh, C., Hsu, V., Zhao, P., Zhang, L., Giacomini, K. M., & Huang, S. (2018). PBPK modeling of the effect of reduced kidney function on the pharmacokinetics of drugs excreted renally by organic anion transporters. *Clinical Pharmacology & Therapeutics*, *103*(3), 485–492.

Huggins, D. J., Biggin, P. C., Dämgen, M. A., Essex, J. W., Harris, S. A., Henchman, R. H., Khalid, S., Kuzmanic, A., Laughton, C. A., Michel, J., Mulholland, A. J., Rosta, E., Sansom, M. S. P., & van der Kamp, M. W. (2019). Biomolecular simulations: From dynamics and mechanisms to computational assays of biological activity. In *Wiley Interdisciplinary Reviews: Computational Molecular Science* (Vol. 9, Issue 3). https://doi.org/10.1002/wcms.1393

Iwaloye, O., Ottu, P. O., Olawale, F., Babalola, O. O., Elekofehinti, O. O., Kikiowo, B., Adegboyega, A. E., Ogbonna, H. N., Adeboboye, C. F., & Folorunso, I. M. (2023). Computer-aided drug design in anti-cancer drug discovery: What have we learnt and what is the way forward? *Informatics in Medicine Unlocked*, *41*, 101332.

Janik, E., Niemcewicz, M., Podogrocki, M., Saluk-Bijak, J., & Bijak, M. (2021). Existing drugs considered as promising in COVID-19 therapy. *International Journal of Molecular Sciences*, *22*(11), 5434.

Joshi, C. P., Baldi, A., Kumar, N., & Pradhan, J. (2024). Harnessing network pharmacology in drug discovery: an integrated approach. *Naunyn-Schmiedeberg's Archives of Pharmacology*, 1–15.

Kabir, A., & Muth, A. (2022). Polypharmacology: The science of multi-targeting molecules. *Pharmacological Research*, *176*, 106055.

Karplus, M., & Petsko, G. A. (1990). Molecular dynamics simulations in biology. In *Nature* (Vol. 347, Issue 6294). https://doi.org/10.1038/347631a0

Kaushik, A. C., Kumar, A., Bharadwaj, S., Chaudhary, R., & Sahi, S. (2018). *Bioinformatics techniques for drug discovery: applications for complex diseases*. Springer.

Kaushik, D., & Kaushik, M. K. (2024a). *Textbook of Computer Aided Drug Development*. Shashwat Publication.

Kaushik, D., & Kaushik, M. K. (2024b). *Textbook of Computer Aided Drug Development*. Shashwat Publication.

Kaza, V. S., Anisha, P. R., & Reddy, C. K. K. (2024). Optimizing Drug Discovery: Molecular Docking with Glow-Worm Swarm Optimization. In *Next-Generation Cybersecurity: AI, ML, and Blockchain* (pp. 369–417). Springer.

Kelleci Çelik, F., & Karaduman, G. (2023). Machine learning-based prediction of drug-induced hepatotoxicity: An OvA-QSTR approach. *Journal of Chemical Information and Modeling*, *63*(15), 4602–4614.

Kiani, Y. S., & Jabeen, I. (2019). Exploring the chemical space of cytochrome p450 inhibitors using integrated physicochemical parameters, drug efficiency metrics and decision tree models. *Computation*, *7*(2), 26.

Kiriiri, G. K., Njogu, P. M., & Mwangi, A. N. (2020). Exploring different approaches to improve the success of drug discovery and development projects: a review. *Future Journal of Pharmaceutical Sciences*, *6*, 1–12.

Kostova, I. (2024a). “Rational Design and Synthesis of Bioactive Molecules.” In *International Journal of Molecular Sciences* (Vol. 25, Issue 18, p. 9927). MDPI.

Kostova, I. (2024b). “Rational Design and Synthesis of Bioactive Molecules.” In *International Journal of Molecular Sciences* (Vol. 25, Issue 18, p. 9927). MDPI.

Kumar, M., Tripathi, M. K., & Kaur, P. (2024). Molecular Dynamics and Its Significance in Drug Discovery. In M. A. Marti, A. G. Turjanski, & D. Fernández Do Porto (Eds.), *Structure-Based Drug Design* (pp. 149–175). Springer International Publishing. https://doi.org/10.1007/978-3-031-69162-1_6

Kumar, V., & Roy, K. (2025). Embracing the changes and challenges with modern early drug discovery. *Expert Opinion on Drug Discovery*, *just-accepted*.

Lee, B., Baek, J., Park, S., & Yoon, S. (2016). deepTarget: end-to-end learning framework for microRNA target prediction using deep recurrent neural networks. *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 434–442.

Lexa, K. W., & Carlson, H. A. (2012). Protein flexibility in docking and surface mapping. *Quarterly Reviews of Biophysics*, *45*(3), 301–343.

Li, W., Yin, Z., Li, X., Ma, D., Yi, S., Zhang, Z., Zou, C., Bu, K., Dai, M., & Yue, J. (2024). A hybrid quantum computing pipeline for real world drug discovery. *Scientific Reports*, *14*(1), 16942.

Limongelli, V. (2020). Ligand binding free energy and kinetics calculation in 2020. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *10*(4), e1455.

Lonsdale, R., & Ward, R. A. (2018). Structure-based design of targeted covalent inhibitors. *Chemical Society Reviews*, *47*(11), 3816–3830.

Maniam, S., & Maniam, S. (2024). Screening techniques for drug discovery in alzheimer's disease. *ACS Omega*, *9*(6), 6059–6073.

Manzoor, R. (2025). Toxicity in the Twenty-First Century (Tox-21): A Multidisciplinary Approach for Toxicity Prediction of Marine Organic Pollutants. In *Recent Trends in Marine Toxicological Assessment* (pp. 383–401). Springer.

Martinelli, D. D. (2022). Generative machine learning for de novo drug discovery: A systematic review. *Computers in Biology and Medicine*, *145*, 105403.

Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, *3*, 80.

Meier, K., Choutko, A., Dolenc, J., Eichenberger, A. P., Riniker, S., & Van Gunsteren, W. F. (2013). Multi-resolution simulation of biomolecular systems: A review of methodological issues. In *Angewandte Chemie - International Edition* (Vol. 52, Issue 10). https://doi.org/10.1002/anie.201205408

Mikov, M., Đanić, M., Pavlović, N., Stanimirov, B., Goločorbin-Kon, S., Stankov, K., & Al-Salami, H. (2017). The role of drug metabolites in the inhibition of cytochrome P450 enzymes. *European Journal of Drug Metabolism and Pharmacokinetics*, *42*(6), 881–890.

Mishra, A. S., Vasanthan, M., & Malliappan, S. P. (2024). Drug repurposing: A leading strategy for new threats and targets. *ACS Pharmacology & Translational Science*, *7*(4), 915–932.

Nabuurs, S. B., Wagener, M., & de Vlieg, J. (2007). A Flexible Approach to Induced Fit Docking. *Journal of Medicinal Chemistry*, *50*(26), 6507–6518. https://doi.org/10.1021/jm070593p

Naithani, U., & Guleria, V. (2024). Integrative computational approaches for discovery and evaluation of lead compound for drug design. *Frontiers in Drug Discovery*, *4*, 1362456.

Nandi, S., Bhaduri, S., Das, D., Ghosh, P., Mandal, M., & Mitra, P. (2024). Deciphering the lexicon of protein targets: a review on multifaceted drug discovery in the era of artificial intelligence. *Molecular Pharmaceutics*, *21*(4), 1563–1590.

Niazi, S. K., & Mariam, Z. (2023). Computer-aided drug design and drug discovery: a prospective analysis. *Pharmaceuticals*, *17*(1), 22.

Niazi, S. K., & Mariam, Z. (2024). Computer-Aided Drug Design and Drug Discovery: A Prospective Analysis. In *Pharmaceuticals* (Vol. 17, Issue 1). https://doi.org/10.3390/ph17010022

Noor, F., Junaid, M., Almalki, A. H., Almaghrabi, M., Ghazanfar, S., & Tahir ul Qamar, M. (2024). Deep learning pipeline for accelerating virtual screening in drug discovery. *Scientific Reports*, *14*(1), 28321.

Olatunde, S. (2024). *Quantum Mechanical Evaluation and Molecular Modeling of SARS-CoV-2 Main Drivers: A Computational Study on Putative Inhibitors for Drug Discovery*.

Olivecrona, M., Blaschke, T., Engkvist, O., & Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, *9*, 1–14.

Ouma, R. B. O., Ngari, S. M., & Kibet, J. K. (2024). A review of the current trends in computational approaches in drug design and

metabolism. *Discover Public Health*, *21*(1), 108. https://doi.org/10.1186/s12982-024-00229-3

Owoloye, A. J., Ligali, F. C., Enejoh, O. A., Musa, A. Z., Aina, O., Idowu, E. T., & Oyebola, K. M. (2022). Molecular docking, simulation and binding free energy analysis of small molecules as Pf HT1 inhibitors. *PloS One*, *17*(8), e0268269.

Oyedele, A.-Q. K., Ogunlana, A. T., Boyenle, I. D., Adeyemi, A. O., Rita, T. O., Adelusi, T. I., Abdul-Hammed, M., Elegbeleye, O. E., & Odunitan, T. T. (2023). Docking covalent targets for drug discovery: stimulating the computer-aided drug design community of possible pitfalls and erroneous practices. *Molecular Diversity*, *27*(4), 1879–1903. https://doi.org/10.1007/s11030-022-10523-4

Ozaybi, M. Q. B., Madkhali, A. N. M., Alhazmi, M. A. M., Faqihi, H. M. A., Alanazi, M. M., Siraj, W. H. Y., Zalah, A. H. A., Khormi, M. M. A., Al Salem, A. M. A., & Mashragi, T. Q. M. (2024). The Role of Artificial Intelligence in Drug Discovery and Development. *Egyptian Journal of Chemistry*, *67*(13), 1541–1547.

Paggi, J. M., Pandit, A., & Dror, R. O. (2024). The art and science of molecular docking. *Annual Review of Biochemistry*, *93*.

Palumbo, A., Facon, T., Sonneveld, P., Blade, J., Offidani, M., Gay, F., Moreau, P., Waage, A., Spencer, A., & Ludwig, H. (2008a). Thalidomide for treatment of multiple myeloma: 10 years later. *Blood, The Journal of the American Society of Hematology*, *111*(8), 3968–3977.

Palumbo, A., Facon, T., Sonneveld, P., Blade, J., Offidani, M., Gay, F., Moreau, P., Waage, A., Spencer, A., & Ludwig, H. (2008b). Thalidomide for treatment of multiple myeloma: 10 years later. *Blood, The Journal of the American Society of Hematology*, *111*(8), 3968–3977.

Parise, A., Cresca, S., & Magistrato, A. (2024). Molecular dynamics simulations for the structure-based drug design: targeting small-GTPases proteins. *Expert Opinion on Drug Discovery*, *19*(10), 1259–1279.

Patel, H. M., Noolvi, M. N., Sharma, P., Jaiswal, V., Bansal, S., Lohan, S., Kumar, S. S., Abbot, V., Dhiman, S., & Bhardwaj, V. (2014). Quantitative structure–activity relationship (QSAR) studies as strategic approach in drug discovery. *Medicinal Chemistry Research*, *23*, 4991–5007.

Peters, J.-U. (2024). *Polypharmacology: Strategies for Multi-Target Drug Discovery*. John Wiley & Sons.

Pieroni, M., Madeddu, F., Di Martino, J., Arcieri, M., Parisi, V., Bottoni, P., & Castrignanò, T. (2023). MD–ligand–receptor: a high-performance computing tool for characterizing ligand–receptor binding interactions in molecular dynamics trajectories. *International Journal of Molecular Sciences*, *24*(14), 11671.

Plewczynski, D., Łaźniewski, M., Augustyniak, R., & Ginalski, K. (2011). Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of Computational Chemistry*, *32*(4), 742–755.

Poongavanam, V., & Ramaswamy, V. (2024). *Computational drug discovery: methods and applications*. John Wiley & Sons.

Popelier, P. L. A., & Smith, P. J. (2006). QSAR models based on quantum topological molecular similarity. *European Journal of Medicinal Chemistry*, *41*(7), 862–873.

Procacci, P. (2021). Methodological uncertainties in drug-receptor binding free energy predictions based on classical molecular dynamics. *Current Opinion in Structural Biology*, *67*, 127–134. https://doi.org/https://doi.org/10.1016/j.sbi.2020.08.001

Pun, F. W., Ozerov, I. V, & Zhavoronkov, A. (2023). AI-powered therapeutic target discovery. *Trends in Pharmacological Sciences*, *44*(9), 561–572.

Pyrkov, A., Aliper, A., Bezrukov, D., Lin, Y.-C., Polykovskiy, D., Kamya, P., Ren, F., & Zhavoronkov, A. (2023a). Quantum computing for near-term applications in generative chemistry and drug discovery. *Drug Discovery Today*, *28*(8), 103675.

Pyrkov, A., Aliper, A., Bezrukov, D., Lin, Y.-C., Polykovskiy, D., Kamya, P., Ren, F., & Zhavoronkov, A. (2023b). Quantum computing for near-term applications in generative chemistry and drug discovery. *Drug Discovery Today*, *28*(8), 103675.

Rajkishan, T., Rachana, A., Shruti, S., Bhumi, P., & Patel, D. (2021). Computer-aided drug designing. *Advances in Bioinformatics*, 151–182.

Rigby, S. P. (2024). Uses of Molecular Docking Simulations in Elucidating Synergistic, Additive, and/or Multi-Target (SAM) Effects of Herbal Medicines. *Molecules*, *29*(22), 5406.

Rowland Yeo, K., Gil Berglund, E., & Chen, Y. (2024). Dose optimization informed by PBPK modeling: state-of-the art and future. *Clinical Pharmacology & Therapeutics*, *116*(3), 563–576.

Sabe, V. T., Ntombela, T., Jhamba, L. A., Maguire, G. E. M., Govender, T., Naicker, T., & Kruger, H. G. (2021). Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *European Journal of Medicinal Chemistry*, *224*, 113705. https://doi.org/https://doi.org/10.1016/j.ejmech.2021.113705

Saber-Ayad, M., Hammoudeh, S., Abu-Gharbieh, E., Hamoudi, R., Tarazi, H., Al-Tel, T. H., & Hamid, Q. (2021). Current status of baricitinib as a repurposed therapy for COVID-19. *Pharmaceuticals*, *14*(7), 680.

Sadeghi, M., Miroliaei, M., & Ghanadian, M. (2025). In Silico Approaches for Rational Drug Design and Potential Enzyme Inhibitors Discovery: A Mini-Review. *Journal of Advanced Biomedical Sciences*, *15*(1), 37–53.

Salloum, H., Lukin, R., & Mazzara, M. (n.d.). Quantum computing in drug discovery: A review of quantum annealing and gate-based approaches. *International Conference on Computational Optimization*.

Santos, V. C., Diniz, L. A., & Ferreira, R. S. (2025). Molecular Modeling Strategies in Drug Design, Development, and Discovery Targeting Proteases. In *Computer-Aided and Machine Learning-Driven Drug Design: From Theory to Applications* (pp. 435–461). Springer.

Scheuher, B., Ghusinga, K. R., McGirr, K., Nowak, M., Panday, S., Apgar, J., Subramanian, K., & Betts, A. (2023). Towards a platform quantitative systems pharmacology (QSP) model for preclinical to clinical translation of antibody drug conjugates (ADCs). *Journal of Pharmacokinetics and Pharmacodynamics*, 1–19.

Shahlaei, M. (2013). Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study. *Chemical Reviews*, *113*(10), 8093–8103. https://doi.org/10.1021/cr3004339

Shanmuga Sundari, M., Thotakura, S. A., Dharmana, M., Gadela, P., & Ammangatambu, M. M. (2024). Process and Applications of Structure-Based Drug Design. *Artificial Intelligence and Machine Learning in Drug Design and Development*, 321–368.

Sharma, A., Narang, A., Kumar, N., Rana, R., Megha, Pooja, Dhir, M., Gulati, H. K., Jyoti, & Khanna, A. (2025). CADD based designing and biological evaluation of novel triazole based

thiazolidinedione coumarin hybrids as antidiabetic agent. *Scientific Reports*, *15*(1), 4302.

Sharma, G., Kumar, N., Sharma, C. S., Alqahtani, T., Tiruneh, Y. K., Sultana, S., Rolim Silva, G. V., de Lima Menezes, G., Zaki, M. E. A., & Nobre Oliveira, J. I. (2025). Identification of promising SARS-CoV-2 main protease inhibitor through molecular docking, dynamics simulation, and ADMET analysis. *Scientific Reports*, *15*(1), 2830.

Sharma, M., Bhatia, A., Akhil, Dutta, A. K., & Alsubai, S. (2025). Optimizing hybrid deep learning models for drug-target interaction prediction: A comparative analysis of evolutionary algorithms. *Expert Systems*, *42*(2), e13683.

Shoaib, T. H., Abdelmoniem, N., Mukhtar, R. M., Alqhtani, A. T., Alalawi, A. L., Alawaji, R., Althubyani, M. S., Mohamed, S. G. A., Mohamed, G. A., & Ibrahim, S. R. M. (2023). Molecular docking and molecular dynamics studies reveal the anticancer potential of medicinal-plant-derived lignans as MDM2-P53 interaction inhibitors. *Molecules*, *28*(18), 6665.

Siddiqui, B., Yadav, C. S., Akil, M., Faiyyaz, M., Khan, A. R., Ahmad, N., Hassan, F., Azad, M. I., Owais, M., & Nasibullah, M. (2025). Artificial Intelligence in Computer-Aided Drug Design (CADD) Tools for the Finding of Potent Biologically Active Small Molecules: Traditional to Modern Approach. *Combinatorial Chemistry & High Throughput Screening*.

Singh, A. (2024). Artificial intelligence for drug repurposing against infectious diseases. *Artificial Intelligence Chemistry*, *2*(2), 100071.

Singh, N. B. (2024). *A Handbook of Quantum Mechanics in Drug Discovery*. NB Singh.

Singh, S., & Sunoj, R. B. (2023). Molecular machine learning for chemical catalysis: prospects and challenges. *Accounts of Chemical Research*, *56*(3), 402–412.

Sinha, K., Ghosh, N., & Sil, P. C. (2023). A review on the recent applications of deep learning in predictive drug toxicological studies. *Chemical Research in Toxicology*, *36*(8), 1174–1205.

Sivula, T., Yetukuri, L., Kalliokoski, T., Kasnanen, H., Poso, A., & Pöhner, I. (2023). Machine learning-boosted docking enables the efficient structure-based virtual screening of giga-scale enumerated chemical libraries. *Journal of Chemical Information and Modeling*, *63*(18), 5773–5783.

Siyah, P. (n.d.). In Silico Prediction of EGFR Inhibitors from Thiophene Derivatives. *Black Sea Journal of Engineering and Science*, *7*(6), 5–6.

Siyah, P. (2024a). Advanced Computational Pipeline for FAK Inhibitor Discovery: Combining Multiple Docking Methods with MD and QSAR for Cancer Therapy. *Computation*, *12*(11), 222.

Siyah, P. (2024b). In Silico Prediction of EGFR Inhibitors from Thiophene Derivatives. *Black Sea Journal of Engineering and Science*, *7*(6), 1131–1138. https://doi.org/10.34248/bsengineering.1537989

Siyah, P., Akgol, S., Durdagi, S., & Kocabas, F. (2021). Identification of viral OTU-like plasmodium parasite proteases and development of antimalarial DUB inhibitors. *Available at SSRN 3773786*.

Smith, D. P., Oechsle, O., Rawling, M. J., Savory, E., Lacoste, A. M. B., & Richardson, P. J. (2021). Expert-augmented computational drug repurposing identified baricitinib as a treatment for COVID-19. *Frontiers in Pharmacology*, *12*, 709856.

Smith, M. D., Rao, J. S., Segelken, E., & Cruz, L. (2015). Force-Field Induced Bias in the Structure of Aβ21-30: A Comparison of OPLS, AMBER, CHARMM, and GROMOS Force Fields. *Journal of Chemical Information and Modeling*, *55*(12). https://doi.org/10.1021/acs.jcim.5b00308

Stanzione, F., Giangreco, I., & Cole, J. C. (2021). Use of molecular docking computational tools in drug discovery. *Progress in Medicinal Chemistry*, *60*, 273–343.

Sun, D., Gao, W., Hu, H., & Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, *12*(7), 3049–3062.

Syriopoulou, A., Markopoulos, I., Tzakos, A. G., & Mavromoustakos, T. (2021). Ligand–receptor interactions and drug design. *Protein-Ligand Interactions and Drug Design*, 89–104.

Tan, B., Joyce, R., Tan, H., Hu, Y., & Wang, J. (2022). SARS-CoV-2 main protease drug design, assay development, and drug resistance studies. *Accounts of Chemical Research*, *56*(2), 157–168.

Tanoli, Z., Fernández-Torras, A., Özcan, U. O., Kushnir, A., Nader, K. M., Gadiya, Y., Fiorenza, L., Ianevski, A., Vähä-Koskela, M., & Miihkinen, M. (2025). Computational drug repurposing: approaches, evaluation of in silico resources and case studies. *Nature Reviews Drug Discovery*, 1–22.

Tanrikulu, Y., & Schneider, G. (2008). Pseudoreceptor models in drug design: bridging ligand-and receptor-based virtual screening. *Nature Reviews Drug Discovery*, *7*(8), 667–677.

Tiwari, A., & Singh, S. (2022). Computational approaches in drug designing. In *Bioinformatics* (pp. 207–217). Elsevier.

Tripathi, R. K. P. (2025). 7 Advancing Drug. *Quantum Computing: The Future of Information Processing*, 111.

Tropsha, A., Isayev, O., Varnek, A., Schneider, G., & Cherkasov, A. (2024). Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nature Reviews Drug Discovery*, *23*(2), 141–155.

Tsou, L. K., Yeh, S.-H., Ueng, S.-H., Chang, C.-P., Song, J.-S., Wu, M.-H., Chang, H.-F., Chen, S.-R., Shih, C., & Chen, C.-T. (2020). Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery. *Scientific Reports*, *10*(1), 16771.

Van der Spoel, D., Zhang, J., & Zhang, H. (2022). Quantitative predictions from molecular simulations using explicit or implicit interactions. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *12*(1), e1560.

Vasanthakumar, G. U., & Singh, M. (2024). 3 Quantum machine learning in healthcare: diagnostics and drug discovery. *Quantum Machine Learning: Quantum Algorithms and Neural Networks*, 39.

Verkhivker, G., Alshahrani, M., Gupta, G., Xiao, S., & Tao, P. (2023). From Deep Mutational Mapping of Allosteric Protein Landscapes to Deep Learning of Allostery and Hidden Allosteric Sites: Zooming in on “Allosteric Intersection” of Biochemical and Big Data Approaches. In *International Journal of Molecular Sciences* (Vol. 24, Issue 9). https://doi.org/10.3390/ijms24097747

Verma, S., & Pathak, R. K. (2022). Discovery and optimization of lead molecules in drug designing. In *Bioinformatics* (pp. 253–267). Elsevier.

Von Delft, A., Hall, M. D., Kwong, A. D., Purcell, L. A., Saikatendu, K. S., Schmitz, U., Tallarico, J. A., & Lee, A. A. (2023). Accelerating antiviral drug discovery: lessons from COVID-19. *Nature Reviews Drug Discovery*, *22*(7), 585–603.

Wan, S., Ruohong, Y., Ying, J., Zhonghuang, L., Jiajie, Z., & and Wu, X. (2019). Insight into binding mechanisms of EGFR allosteric inhibitors using molecular dynamics simulations and free energy calculations. *Journal of Biomolecular Structure and Dynamics*, *37*(16), 4384–4394. https://doi.org/10.1080/07391102.2018.1552197

Wang, M., Wang, Z., Sun, H., Wang, J., Shen, C., Weng, G., Chai, X., Li, H., Cao, D., & Hou, T. (2022). Deep learning approaches for de novo drug design: An overview. *Current Opinion in Structural Biology*, *72*, 135–144.

Waszkowycz, B., Clark, D. E., & Gancia, E. (2011). Outstanding challenges in protein–ligand docking and structure-based virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *1*(2), 229–259.

Westhead, D. R., Clark, D. E., & Murray, C. W. (1997). A comparison of heuristic search algorithms for molecular docking. *Journal of Computer-Aided Molecular Design*, *11*, 209–228.

Xiang, M., Cao, Y., Fan, W., Chen, L., & Mo, Y. (2012). Computer-aided drug design: lead discovery and optimization. *Combinatorial Chemistry & High Throughput Screening*, *15*(4), 328–337.

Xiong, J., Wang, X., Dou, F., Cao, H., Hong, T., Li, Z., & Jing, X. (2024). Construction and Application of Protein-Protein Interaction Knowledge Graph. *International Symposium on Artificial Intelligence and Robotics*, 44–53.

Xu, Y., Yang, L., Li, J., Li, N., Hu, L., Zuo, R., & Jin, S. (2024). Determination of the binding affinities of OPEs to integrin αvβ3 and elucidation of the underlying mechanisms via a competitive binding assay, pharmacophore modeling, molecular

docking and QSAR modeling. *Journal of Hazardous Materials*, *466*, 133650.

Yadav, V., Reang, J., Vinita, & Tonk, R. K. (2022). Chapter 3 - Ligand-based drug design (LBDD). In M. Rudrapal & C. Egbuna (Eds.), *Computer Aided Drug Design (CADD): From Ligand-Based Methods to Structure-Based Approaches* (pp. 57–99). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-323-90608-1.00009-5

Yang, Q., Zong, K., Zhao, X., Zhang, F., Li, F., & Li, X. (2025). Design, Synthesis, Evaluation, and Molecular Dynamics Simulation of SARS-CoV-2 Mpro Inhibitors. *ACS Medicinal Chemistry Letters*.

Yazdi, M. (2024). Application of Quantum Computing in Reliability Analysis. In *Advances in Computational Mathematics for Industrial System Reliability and Maintainability* (pp. 139–154). Springer.

Yu, H., Li, C., Wang, X., Duan, J., Yang, N., Xie, L., Yuan, Y., Li, S., Bi, C., & Yang, B. (2020). Techniques and strategies for potential protein target discovery and active pharmaceutical molecule screening in a pandemic. *Journal of Proteome Research*, *19*(11), 4242–4258.

Yu, J.-L., Dai, Q.-Q., & Li, G.-B. (2022). Deep learning in target prediction and drug repositioning: Recent advances and challenges. *Drug Discovery Today*, *27*(7), 1796–1814. https://doi.org/https://doi.org/10.1016/j.drudis.2021.10.010

Yuan, Y., Shi, Y., Li, C., Kim, J., Cai, W., Han, Z., & Feng, D. D. (2016). DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. *BMC Bioinformatics*, *17*, 243–256.

Zhang, X., Shen, C., Zhang, H., Kang, Y., Hsieh, C.-Y., & Hou, T. (2024a). Advancing ligand docking through deep learning: challenges and prospects in virtual screening. *Accounts of Chemical Research*, *57*(10), 1500–1509.

Zhang, X., Shen, C., Zhang, H., Kang, Y., Hsieh, C.-Y., & Hou, T. (2024b). Advancing ligand docking through deep learning: challenges and prospects in virtual screening. *Accounts of Chemical Research*, *57*(10), 1500–1509.

Zhao, M., Ma, J., Li, M., Zhang, Y., Jiang, B., Zhao, X., Huai, C., Shen, L., Zhang, N., & He, L. (2021). Cytochrome P450 enzymes and drug metabolism in humans. *International Journal of Molecular Sciences*, *22*(23), 12808.

Zhao, X., Hu, Q., Wang, X., Li, C., Chen, X., Zhao, D., Qiu, Y., Xu, H., Wang, J., & Ren, L. (2024). Dual-Target Inhibitors Based on Acetylcholinesterase: Novel Agents for Alzheimer's Disease. *European Journal of Medicinal Chemistry*, 116810.

Zhou, H., & Theil, F.-P. (2015). *ADME and translational pharmacokinetics/pharmacodynamics of therapeutic proteins: applications in drug discovery and development*. John Wiley & Sons.

APPLICATIONS OF CADD IN DRUG DISCOVERY

RAGHAD SHARBAJI[1]

In the previous chapter, we ventured through the foundational concepts that shape the world of Computer-Aided Drug Discovery (CADD). We examined how this powerful fusion of computational science and biotechnology has radically transformed the drug discovery process, driving it into an era where precision and speed are paramount. The theoretical aspects of CADD laid the groundwork for understanding its role in modern drug development, but it is in its practical application where its true potential is realized.

This chapter shifts focus from the broader strokes of CADD to its hands-on, real-world contributions in drug discovery. The techniques discussed here are not just abstract concepts but are actively employed to bring life-saving drugs from the laboratory bench to the clinic. We begin by exploring the role of molecular docking, a method that helps us understand how potential drug molecules bind to their target proteins. These docking simulations provide critical insights that guide the design of compounds with the

[1] Department of Biomedical Engineering, Bahçesehir University, Istanbul 34353, Turkey

highest likelihood of success. However, molecular docking is just one tool in a far more intricate toolbox.

Molecular dynamics simulations further enhance our understanding by allowing us to watch molecules in motion, simulating their behavior over time and under varying conditions. These simulations provide a dynamic, evolving picture of drug behavior in a biological system, refining predictions about a drug's stability, interactions, and potential efficacy in living organisms.

Another vital aspect of drug design explored here is Quantitative Structure-Activity Relationship (QSAR) modeling. QSAR allows researchers to predict the biological activity of drug candidates based on their chemical structure, a crucial step before moving into experimental testing. However, what truly sets this chapter apart is the integration of cutting-edge, more nuanced computational techniques, transforming how we predict and refine the biological properties of drug candidates.

As we move deeper into the chapter, we will uncover how advances in computational approaches are not just improving individual methods, but revolutionizing the entire drug development pipeline. New paradigms in computational modeling have emerged, pushing boundaries and offering unprecedented accuracy. Through these methods, we not only accelerate the process of drug discovery but also refine our ability to predict how molecules will behave in the complex environment of a living organism.

In short, this chapter brings the evolving landscape of drug discovery to life by showing how computational tools are now intricately woven into every stage of drug development. We will uncover how these tools help us go from theory to application, guiding decision-making at each critical juncture. The power of simulation, prediction, and computational analysis promises to not only streamline the development of new therapeutics but to also

make it more precise, efficient, and tailored to the needs of patients worldwide.

From Docking Predictions to In Vitro Experimental Validation

Since first being developed in the 1980s, Molecular docking techniques play a major role in the design and development of novel medications by predicting the experimental binding mechanism, affinity, and location of a small molecule (ligand) within the binding site of the target receptor (macromolecule). However, these computational predictions must be tested in real-world settings to confirm their accuracy and biological relevance (Shamim et al., 2024). Validating docking results is essential to ensure that the predicted interactions between the ligand and the protein target are accurate and reliable. In vitro applications bridge this gap, employing laboratory assays to test docking-derived candidates for binding strength and biological activity.

Validation can be approached through both experimental and computational methods. From the experimental point of view, the most direct way to validate a docking pose prediction is by determining the experimental protein structure in complex with the ligand, typically through X-ray crystallography or NMR spectroscopy. Also, assays such as surface plasmon resonance (SPR), isothermal titration calorimetry (ITC), and enzyme inhibition studies can provide experimental evidence for the binding affinity and kinetics of the ligand-protein interaction, which can be compared to docking predictions. Another experimental technique is site-directed mutagenesis that can be used to alter specific amino acids within the binding site. The effects of these mutations on ligand binding can validate the importance of predicted interactions(Aguiar, 2024; Ferreira, 2015).

The process of testing docking predictions in vitro typically begins with selecting high-scoring compounds from virtual screens,

based on metrics like binding energy (e.g., kcal/mol) or docking scores. These candidates are synthesized or sourced, then subjected to a suite of biophysical, biochemical, and structural assays to measure binding affinity and efficacy.

Surface plasmon resonance (SPR) assesses affinity by quantifying association (Ka) and dissociation (Kd) constants in real time, testing whether a ligand binds as predicted. Isothermal titration calorimetry (ITC) measures thermodynamic parameters- enthalpy (ΔH) and entropy (ΔS)- to confirm pose stability. Efficacy is evaluated via functional assays, such as enzyme inhibition studies (IC50) or cell-based assays (cytotoxicity, receptor activation). Structural techniques like X-ray crystallography and NMR spectroscopy further validate docking by resolving the atomic-level binding mode or mapping interactions in solution, respectively. Site-directed mutagenesis tests specific residue contributions by altering docking-predicted contact points and reassessing binding or function. These in vitro results refine docking models, improving scoring functions and ensuring only promising candidates progress, minimizing trial-and-error.

Case studies of successful drugs highlight docking's impact when validated in vitro. Imatinib (Gleevec), developed for chronic myelogenous leukemia (CML), exemplifies this synergy. Docking predicted imatinib binds the BCR-ABL kinase ATP pocket, stabilizing an inactive conformation (PDB ID: 1IEP). Kinase assays confirmed an IC50 in the nanomolar range (e.g., ~25–100 nM), proving potent inhibition (Schindler et al., 2000; Buchdunger et al., 1996). X-ray crystallography of the imatinib-BCR-ABL complex (PDB ID: 1IEP) revealed hydrogen bonds with Met318 and Thr315, confirming the docking pose and structural basis of affinity. These results drove imatinib to FDA approval in 2001, showcasing docking's role in accelerating discovery (Schindler et al., 2000).

Zanamivir (Relenza), an influenza neuraminidase inhibitor, offers another success. Docking predicted binding to the neuraminidase active site via hydrogen bonds with Arg118 and Asp151 (Gupta et al., 2011; von Itzstein et al., 1993; Smith et al., 2001). Enzyme assays confirmed a Ki of approximately 0.1 nM, by antiviral efficacy in cell assays (Cheer & Wagstaff, 2002). X-ray crystallography validated these interactions, while NMR studies mapped dynamic binding in solution, reinforcing affinity. Approved in 1999, zanamivir Exemplifies the in vitro confirmation of docking predictions (Smith et al., 2001; Cheer & Wagstaff, 2002).

Docking studies identified boceprevir as a potential inhibitor of SARS-CoV-2 main protease (Mpro, PDB ID: 6LU7). Enzyme assays confirmed its inhibitory activity, with an IC_{50} in the low micromolar range, supported by fluorescence-based inhibition tests (Alugubelli et al., 2022). Mutagenesis studies on Mpro catalytic residues (e.g., His41) confirmed their role in inhibitor binding, affecting IC_{50} values (Krismer et al., 2024). Studies in high-impact journals, including ScienceDirect, highlighted docking's role in rapid antiviral discovery during pandemics (Gong et al., 2025; Mo et al., 2024).

Drug repositioning leverages molecular docking to reduce costs and risks. A recent study screened organohalogen drugs from the CMC database using D3DOCKxb, a docking tool with a halogen bonding scoring function, targeting B-Raf V600E. Of 25% of marketed drugs that are organohalogens, three were tested via ELISA assays, with rafoxanide (IC50: 0.07 μM) and closantel (IC50: 1.90 μM) showing potency comparable to vemurafenib (IC50: 0.17 μM). Single point mutagenesis confirmed halogen bonding predicted by D3DOCKxb, highlighting its superiority over standard docking, as seen in prior repositioning successes (Li et al., 2016).

Boceprevir is a notable example of a drug candidate identified through molecular docking and successfully validated in vitro and is now an FDA-approved protease inhibitor used for the treatment of Hepatitis C Virus (HCV). At the onset of Boceprevir's development, high-throughput screening (HTS) failed to identify viable lead compounds against the HCV NS3-NS4A serine protease, a crucial enzyme for viral replication. Consequently, researchers at Schering-Plough Research Institute shifted to a structure-based drug design (SBDD) approach, leveraging molecular docking and X-ray crystallography to optimize inhibitors targeting the protease's active site. Docking studies guided the design of α-ketoamide inhibitors, which exhibited strong inhibitory activity by interacting with the catalytic triad (Ser139, His57, Asp81) of NS3-NS4A. Lead compound 16 demonstrated high potency (Ki = 1.9 nM), and further structure-activity relationship (SAR) optimization led to compound 17 (Ki = 25 nM, EC90 = 0.4 μM). Subsequent refinements resulted in the discovery of Boceprevir (SCH503034), which showed Ki = 14 nM and EC90 = 0.35 μM (Talele, Khedkar, & Rigby, 2010; Turk, 2006; Yan et al., 1998).

These examples show docking as a hypothesis generator, validated by in vitro assays. SPR, ITC, X-ray, and NMR test affinity, addressing docking's static limitations, while functional assays and mutagenesis ensure efficacy. Imatinib, zanamivir, SARS-CoV-2 main protease,organohalogen drugs, and Boceprevir demonstrate how these methods bridge computation to practical outcomes.

In Vivo Applications:

Molecular docking predicts how drug candidates bind to target proteins, providing hypotheses about affinity, specificity, and potential efficacy. While in vitro assays (e.g., IC50, enzyme inhibition) confirm these predictions in controlled settings, in vivo testing in animal models evaluates how these compounds perform in

complex biological systems. This step assesses pharmacokinetics (PK), absorption, distribution, metabolism, excretion (ADME), pharmacodynamics (PD), toxicity, and therapeutic efficacy, bridging computational design to clinical relevance. For instance, Imatinib's docking against Bcr-Abl kinase predicted ATP-pocket binding, validated in mice with leukemia xenografts showing >90% tumor reduction at 50 mg/kg/day (E Buchdunger et al., 2000). Another example is Zanamivir, an influenza neuraminidase (NA) inhibitor, which was initially identified through molecular docking as a strong NA binder (Mark von Itzstein et al., 1993). *In vivo* studies in mice and ferrets confirmed its antiviral efficacy, reducing viral titers by 2–3 logs at doses of 0.1–1 mg/kg when administered intranasally (Ryan et al., 1994). However, due to poor oral bioavailability (<5%), inhaled delivery was pursued. This strategy was validated in phase III clinical trials, where 10 mg BID significantly shortened symptom duration by 1–2.5 days ($p < 0.001$) in influenza-positive patients and high-risk groups (2.5 days, $p = 0.015$). Additionally, Zanamivir reduced influenza-related complications ($p < 0.001$) and antibiotic use ($p = 0.042$) (Dunn & Goa, 1999). These results bridged *in vivo* efficacy to clinical approval in 1999, demonstrating how docking predictions can successfully guide therapeutic translation.

In vivo studies in animal models serve as a necessary validation step, ensuring that computationally identified drug candidates demonstrate efficacy, pharmacokinetics, and safety within a living system before advancing to human trials. This stage is critical in translating theoretical findings into practical medical treatments. One of the key contributions of *in vivo* testing is confirming whether a computationally predicted drug candidate actually exerts the desired biological effect. For example, the anti-influenza drug Zanamivir was initially identified through structure-based docking as a neuraminidase (NA) inhibitor (Gupta et al., 2011). While docking results suggested strong target binding, the

drug's real potential was demonstrated in animal studies. When tested in mice and ferrets, intranasal administration of Zanamivir at doses of 0.1–1 mg/kg significantly reduced viral titers by 2–3 logs, establishing its antiviral activity in a living system (Ryan et al., 1994). These results were critical in justifying its progression to clinical trials, where it later showed efficacy in reducing influenza symptoms and complications in infected patients (Dunn & Goa, 1999). Without this intermediate validation step, computational predictions alone would not have been sufficient to support its clinical use.

Beyond efficacy, *in vivo* studies provide crucial insights into pharmacokinetics and bioavailability factors that cannot be accurately determined through docking alone. A notable example is the case of Zanamivir's poor oral bioavailability, which was discovered during animal testing (Dunn & Goa, 1999). Despite its strong in silico binding affinity, less than 5% of the orally administered drug reached systemic circulation, necessitating the development of an inhaled formulation instead (Moscona, 2005). Such findings highlight the importance of *in vivo* testing in refining drug administration strategies to ensure clinical success. Similarly, the B-Raf V600E inhibitor vemurafenib, which was identified through docking-based screening, underwent extensive animal studies before reaching human trials. These studies confirmed its tumor-suppressive effects and optimized dosing regimens, ultimately contributing to its FDA approval for metastatic melanoma (Bollag et al., 2012).

Another critical role of *in vivo* testing is in evaluating off-target effects and toxicity. Computational docking can predict interactions at the molecular level but cannot fully account for the complexity of biological systems, where unexpected side effects may arise. The case of thalidomide exemplifies the necessity of animal studies in drug development. Initially introduced as a

sedative, thalidomide was later found to cause severe birth defects, a discovery that reinforced the need for rigorous *in vivo* safety assessments before human trials (Kim & Scialli, 2011). In modern drug development, computational and experimental approaches work together to mitigate such risks.

Recent advances in drug repositioning also illustrate how *in vivo* studies validate docking predictions. A study using the docking tool D3DOCKxb screened organohalogen drugs for potential B-Raf V600E inhibitors. Among the identified candidates, rafoxanide and closantel were tested using ELISA assays and showed IC50 values of 0.07 μM and 1.90 μM, respectively, comparable to the clinically approved drug vemurafenib (Li et al., 2016). The transition from docking-based identification to enzymatic assays and eventual animal testing underscores how computational predictions must be rigorously validated before clinical translation.

In summary, *in vivo* studies are a crucial step in bridging the gap between computational docking predictions and real-world clinical applications. By assessing efficacy, pharmacokinetics, and safety in a biological system, these studies ensure that promising drug candidates are not only theoretically viable but also practically effective. While computational tools continue to improve drug discovery efficiency, they remain one piece of a larger validation process that ultimately determines whether a compound can become a safe and effective treatment.

Integrating AI in Molecular Docking

Molecular docking has long been a cornerstone of computational drug discovery, providing insights into ligand-protein interactions and facilitating virtual screening for potential drug candidates. Traditionally, docking methods relied on physics-based and empirical scoring functions to estimate binding affinities. These

approaches, while instrumental in early drug discovery efforts, faced several intrinsic challenges.

One of the primary limitations of conventional docking methods is their reliance on rigid or semi-flexible models, which often fail to accurately capture the dynamic nature of protein-ligand interactions. Proteins, especially those with flexible binding sites, undergo conformational changes upon ligand binding, a phenomenon known as induced fit. Traditional docking algorithms struggle to model such structural adaptability, leading to inaccuracies in pose prediction and binding affinity estimations. Furthermore, conventional docking approaches do not adequately account for solvent effects, entropy contributions, and water-mediated interactions, all of which are critical in determining the true biological activity of a ligand. The need for more accurate predictions has driven researchers to explore artificial intelligence (AI)-based techniques that can overcome these limitations and refine molecular docking outcomes.

AI has revolutionized molecular docking by introducing machine learning (ML) and deep learning (DL) models that leverage large datasets of experimentally validated protein-ligand interactions. Unlike traditional scoring functions, which are predefined and rigid, AI-driven methods continuously learn from structural databases, improving their predictive accuracy over time. Among the most promising AI-based docking tools are DeepDock and DeepChem, both of which employ deep neural networks to refine docking pose selection and binding affinity predictions.

DeepDock utilizes convolutional neural networks (CNNs) trained on thousands of protein-ligand complexes to enhance docking accuracy. By analyzing three-dimensional spatial features of molecules, DeepDock identifies critical interactions such as hydrogen bonding, hydrophobic interactions, and electrostatic

forces, which conventional docking methods often overlook. A study by Torng and Altman (2019) demonstrated that DeepDock significantly outperformed classical docking tools such as AutoDock Vina in predicting correct ligand binding poses, particularly in proteins with flexible active sites. Unlike traditional methods, which rely on rigid docking and empirical scoring functions, DeepDock dynamically adjusts its predictions based on learned patterns from experimentally validated structures. This allows for a more nuanced understanding of protein-ligand interactions and reduces false positives in virtual screening campaigns.

DeepChem takes a different approach by representing molecules as graphs, where atoms serve as nodes and chemical bonds as edges. This representation allows DeepChem to model non-linear interactions between functional groups and capture subtle binding effects that traditional docking algorithms fail to recognize. Unlike rigid docking methods that treat molecules as static 3D structures, DeepChem incorporates reinforcement learning to improve its predictive accuracy. A study by Ramsundar et al., 2017 showed that DeepChem could more accurately predict molecular binding affinities than conventional scoring functions, demonstrating its effectiveness in screening vast chemical libraries for promising drug candidates. This graph-based approach enables the model to better account for solvation effects, conformational changes, and entropic contributions, leading to more biologically relevant docking predictions.

GraphDelta employs graph convolutional networks (GCNs) to predict binding free energies by analyzing protein-ligand interactions at the atomic level. Unlike traditional docking approaches that rely on predefined scoring functions, GraphDelta learns from experimental binding affinity data to refine energy calculations. A study by Stepniewska-Dziubinska et al. (2018) demonstrated that GraphDelta outperformed classical scoring

functions such as GlideScore and MM-GBSA in accurately predicting experimental binding affinities. By incorporating molecular graph representations, GraphDelta improves docking accuracy by capturing subtle atomic-level interactions that other models might overlook.

DeltaDock introduces generative adversarial networks (GANs) into molecular docking, offering a novel approach to pose refinement. Traditional docking methods often generate multiple binding poses, but ranking these poses correctly remains a challenge. DeltaDock addresses this by employing a GAN-based architecture that generates docking poses and refines them before applying a deep learning-based scoring function. This reduces false positives and enhances the identification of true binding candidates. The use of adversarial training improves the robustness of docking predictions, making DeltaDock a promising tool for high-throughput virtual screening (Yan et al., 2024).

The integration of AI in molecular docking has led to significant improvements in accuracy, but the field is still evolving. AI-driven models have demonstrated remarkable success in refining pose selection and predicting binding affinities, yet challenges remain. One of the most critical advancements is the ability of AI models to generalize across different protein families. Unlike traditional docking, which often requires case-specific parameter tuning, AI-based models can be trained on diverse datasets to recognize universal binding patterns. This has made docking more robust and applicable to a wider range of drug targets, including challenging systems such as membrane proteins and intrinsically disordered proteins.

However, while AI-based methods enhance prediction accuracy, they are not infallible. Current models still struggle with highly flexible ligands and proteins with extreme conformational

variability. Additionally, while AI-based scoring functions have improved over classical methods, they still rely on training data derived from experimentally determined structures, which may not always reflect the full complexity of biological systems.

Despite the progress made, several challenges persist in AI-driven molecular docking:

1. **Data Quality and Bias**: AI models are only as good as the data they are trained on. Structural databases such as PDBbind and ChEMBL contain high-quality protein-ligand complexes, but biases in these datasets can lead to overfitting. Ensuring diverse and representative training data remains a significant challenge.
2. **Induced Fit and Allosteric Effects**: While AI models like DeepDock and DeepChem account for some degree of receptor flexibility, they do not fully capture induced fit effects and allosteric modulations. Future developments will likely integrate AI with molecular dynamics simulations to better address these issues.
3. **Explainability and Interpretability**: One of the major criticisms of AI-driven docking is the lack of interpretability. Traditional docking methods provide clear scoring functions based on well-defined physical principles, whereas deep learning models operate as "black boxes." Developing explainable AI models will be crucial in gaining broader acceptance in the drug discovery community.
4. **Computational Cost**: While AI-driven docking offers higher accuracy, deep learning models require substantial computational resources for training and

inference. Optimizing these models for efficiency without compromising accuracy is an ongoing area of research.

To address these challenges, researchers are exploring hybrid approaches that combine AI with physics-based methods. For example, integrating AI-based scoring functions with molecular dynamics (MD) simulations can provide a more realistic representation of protein-ligand interactions by incorporating dynamic conformational changes. Similarly, generative adversarial networks (GANs) are being employed to refine docking poses and reduce false positives in virtual screening.

Looking ahead, the next generation of AI-driven docking systems will likely incorporate transfer learning and multi-modal AI models that leverage multiple sources of biological data, including omics and cheminformatics. These advancements have the potential to revolutionize structure-based drug design, making the process more efficient and predictive than ever before.

The integration of AI in molecular docking represents a paradigm shift in computational drug discovery. By leveraging deep learning techniques such as CNNs and graph-based neural networks, AI-driven models have significantly improved the accuracy of docking predictions, addressing long-standing challenges in receptor flexibility, solvent effects, and scoring function limitations. While challenges remain—ranging from data biases to computational costs—the continued evolution of AI in molecular docking promises to expand the horizons of drug discovery, paving the way for more precise and efficient therapeutic development.

From Molecular Dynamics Simulations to In Vitro Experimental Validation

Molecular dynamics (MD) simulations have emerged as a powerful computational tool in drug discovery, providing critical insights into the dynamic behavior of biomolecular systems. Unlike static docking studies, MD simulations account for receptor flexibility, solvent effects, and conformational changes over time, offering a more accurate representation of drug interactions. These simulations refine predictions of ligand binding, stability, and biological activity, thereby bridging the gap between computational modeling and experimental validation (AlRawashdeh & Barakat, 2023). This section explores how MD simulations contribute to in vitro and in vivo drug discovery applications and highlights case studies where MD simulations have successfully predicted drug efficacy.

MD simulations play a crucial role in drug discovery by refining predictions of a drug's stability, binding affinity, and overall activity within biological systems. By mimicking physiological conditions, they allow scientists to study the interactions between small molecules and macromolecules such as proteins, nucleic acids, and lipid membranes. This approach enhances our understanding of molecular mechanisms, optimizes lead compounds, and reduces the need for extensive trial-and-error experimentation in wet lab settings (Liu et al., 2018).

In vitro applications of MD simulations can significantly improve the accuracy of binding affinity predictions. For example, MD simulations can track the interactions between a drug and its target at an atomic level, helping to determine how strong the binding is and how stable the drug-target complex will be in solution. Studies have demonstrated that MD simulations provide valuable insights into the role of water molecules in mediating protein-ligand

interactions, which is crucial for understanding the full binding process (Loschwitz et al., 2020; Rudling et al., 2018). For instance, MD simulations have been used to predict the binding affinity of HIV protease inhibitors, helping to optimize their efficacy by exploring various drug-target interactions and water-mediated binding events By simulating the dynamics of these interactions, researchers can identify key residues involved in the binding process and predict how mutations in the target might affect drug efficacy (Badaya & Sasidhar, 2020) .

Another significant in vitro application of MD simulations is in the study of drug permeability and interactions with biological membranes. Many drugs, especially those targeting membrane proteins, need to traverse cellular membranes to reach their target sites. MD simulations allow researchers to study how drug molecules interact with lipid bilayers and membrane proteins, helping to predict their permeability, transport mechanisms, and potential toxicity (Kutzner et al., 2011). For example, MD simulations have been applied to understand how small molecules interact with bacterial membrane proteins, providing insights into how drugs can be designed to penetrate bacterial membranes more effectively (Martinotti et al., 2020; Matamoros-Recio et al., 2021). This knowledge is especially important in the development of antibiotics, where overcoming membrane barriers is critical for drug effectiveness

MD simulations are also a valuable tool for investigating mechanisms of drug resistance. By simulating the interactions between drugs and mutated target proteins, researchers can understand how resistance mutations alter the binding dynamics and stability of drug-target complexes. This information can guide the development of next-generation drugs designed to overcome these resistance mechanisms. For example, MD simulations have been used to study drug resistance in HIV-1 protease and identify how

certain mutations impact inhibitor binding (Badaya & Sasidhar, 2020). Similarly, studies on antibiotic resistance mechanisms often involve MD simulations to explore how mutations in bacterial enzymes affect drug binding and how new compounds might be designed to counteract these mutations (Beer et al, 2024; Latallo et al, 2017)

In Vivo Applications:

Beyond in vitro assessments, MD simulations contribute to preclinical studies by optimizing the pharmacokinetic properties of drug candidates. One of the key applications in vivo is predicting ADME (absorption, distribution, metabolism, and excretion) properties, which are essential for assessing drug bioavailability and toxicity. This section delves into the use of MD simulations in enhancing the ADME properties of drug molecules by providing insights into their interactions with biological macromolecules, cell membranes, and metabolic enzymes. By doing so, MD simulations have the potential to reduce drug development costs, shorten timelines, and improve the safety and efficacy profiles of therapeutic agents before they proceed to clinical trials.

The first step in a drug's journey through the body is its absorption. Drug molecules need to cross biological barriers such as the intestinal epithelium and the blood-brain barrier (BBB) to reach their target sites. The permeability of these barriers is influenced by factors such as molecular size, polarity, charge, and interaction with membrane lipids. Traditional experimental methods like Caco-2 cell assays or in vivo studies provide valuable data but can be time-consuming and resource-intensive. MD simulations, however, allow for detailed molecular insights into drug-membrane interactions, facilitating predictions of absorption without the need for experimental testing.

MD simulations of the gastrointestinal (GI) tract model how drug molecules interact with lipid bilayers, particularly the intestinal epithelium, by simulating the partitioning and diffusion processes. For example, simulations have been used to study the permeability of hydrophobic drugs, such as those in the statin class (e.g., atorvastatin), to identify key factors influencing their absorption rates (Róg, Girych, & Bunker, 2021). Through these simulations, researchers were able to identify the critical role of drug-lipid interactions and the effect of molecular shape on permeability, leading to the optimization of these drugs for better absorption profiles.

In the case of the blood-brain barrier, MD simulations have been pivotal in studying the interactions between drug candidates and the tight junctions of endothelial cells that form the barrier. For example, a study on the CNS-active drug, lidocaine, showed how its molecular properties influenced its permeability across the BBB. By simulating the interactions between lidocaine molecules and lipid bilayers mimicking the BBB, researchers could predict its transport efficiency and identify modifications to increase its brain penetration (Zapata-Morin et al., 2014; Saeedi et al., 2017).

Once a drug is absorbed, its distribution throughout the body depends on factors such as blood flow, protein binding, and the drug's ability to cross various tissue barriers. MD simulations offer valuable insights into how drugs interact with plasma proteins and cellular membranes, which are key determinants of distribution.

One application of MD simulations in distribution studies is the simulation of drug-protein binding, particularly binding to serum albumin, a major plasma protein that affects drug bioavailability. A study on the binding of the anti-inflammatory drug ibuprofen to human serum albumin (HSA) used MD simulations to investigate the specific binding sites and the drug's conformational changes

upon binding (Evoli et al., 2016). By understanding the nature of these interactions, researchers can predict how long a drug will remain in circulation and how it might distribute across various tissues.

MD simulations have also been used to explore tissue-specific drug distribution. For example, researchers have applied MD simulations to model the distribution of antitumor agents in cancer tissues, including the interaction between drug molecules and the tumor vasculature. In these studies, MD simulations have helped optimize drug formulations to enhance drug delivery to tumor sites while minimizing systemic toxicity. A prominent example is the design of nanoparticle-drug conjugates, where MD simulations help predict the interaction between nanoparticles and cell membranes to improve the selective delivery of chemotherapeutic agents (Mundada et al., 2025; Salahshoori et al., 2024).

The metabolism of drugs is primarily mediated by enzymes such as the cytochrome P450 (CYP450) family, which is responsible for the oxidation of many drug molecules. MD simulations have become an essential tool for studying how drugs interact with these enzymes, allowing for predictions of metabolic pathways, enzyme-substrate interactions, and the potential for drug-drug interactions.

In preclinical drug development, understanding how a drug will be metabolized by CYP enzymes can provide critical information about its half-life, potential toxicity, and interactions with other drugs. MD simulations allow for the modeling of enzyme-substrate binding at an atomic level, which can reveal the structural basis of enzyme specificity and identify potential hotspots for drug interaction.

A well-known example is the study of the metabolism of the widely prescribed statin, simvastatin, by CYP3A4. MD simulations have been employed to simulate the binding of simvastatin to the

active site of CYP3A4, providing insights into the enzyme-substrate interactions and the subsequent metabolic transformation (Jiang et al., 2018; Jiang, 2013). These simulations have helped identify specific residues in the enzyme that are crucial for simvastatin binding, paving the way for the design of drugs with improved metabolic stability. In addition to drug-metabolizing enzymes, MD simulations are increasingly being used to study the potential for drug-drug interactions. A well-known example is the use of MD simulations to study the interaction of midazolam, a commonly used sedative, with CYP3A4. This interaction has been thoroughly investigated to understand how genetic variations in the CYP3A4 enzyme and drug-drug interactions affect midazolam metabolism (Denisov et al., 2021).

The final stage in a drug's journey is its elimination from the body, which occurs primarily through renal excretion or biliary clearance. MD simulations can provide valuable information about the transport mechanisms involved in drug excretion, especially when drugs are eliminated via membrane-bound transporters.

Adefovir, an antiviral drug used to treat hepatitis B, is primarily eliminated through renal excretion, where its interaction with the organic anion transporter 1 (OAT1) plays a crucial role in its clearance. Molecular dynamics (MD) simulations have been employed to study how adefovir binds to OAT1 and the conformational changes the transporter undergoes during substrate recognition and transport. A recent study by Janaszkiewicz et al., 2023 utilized MD simulations to investigate the structural basis of adefovir's interaction with human OAT1. The simulations identified two potential binding sites on the transporter, revealing key molecular interactions responsible for substrate affinity and transport efficiency. These insights provided a mechanistic understanding of how OAT1 facilitates adefovir uptake into renal cells, helping predict its renal clearance.

MD simulations have emerged as a transformative approach in preclinical drug development, offering an unprecedented level of detail in understanding drug behavior at the molecular level. By simulating interactions between drug candidates and biological macromolecules, MD simulations provide valuable insights into absorption, distribution, metabolism, and excretion (ADME) processes, which are critical for optimizing pharmacokinetic properties. These simulations allow researchers to explore how drugs cross biological membranes, bind to plasma proteins, interact with metabolic enzymes, and are ultimately cleared from the body.

One of the key advantages of MD simulations is their ability to bridge the gap between in vitro experiments and in vivo outcomes. Traditional experimental techniques, while essential, often face limitations in accurately predicting human physiological responses due to species-specific differences and experimental constraints. By integrating computational methods with experimental data, researchers can refine predictions, enhance drug design, and reduce reliance on costly and time-consuming animal studies. Furthermore, MD simulations enable the identification of potential drug-drug interactions, enzyme inhibition risks, and transporter-mediated clearance mechanisms, all of which are essential for ensuring drug safety and efficacy.

As computational power continues to advance and algorithms become more sophisticated, the application of MD simulations in drug development will only expand. Their ability to provide mechanistic insights into drug interactions at an atomic level allows for the rational design of safer and more effective therapeutics. Moving forward, the integration of MD simulations with machine learning, artificial intelligence, and multi-scale modeling approaches will further enhance their predictive capabilities, enabling researchers to optimize drug properties with greater accuracy. Ultimately, the growing role of MD simulations in

drug discovery and development has the potential to accelerate the translation of promising drug candidates from bench to bedside, reducing development costs, improving success rates, and contributing to more personalized and effective treatment strategies.

Integrating AI in Molecular (MD) Dynamic Simulation

- Machine learning in force field optimization.

Molecular dynamics (MD) simulations have revolutionized our understanding of molecular interactions, drug design, and material science. However, despite their power, traditional MD simulations face challenges in terms of computational cost, accuracy, and scalability. With the rise of artificial intelligence (AI) and machine learning (ML), these limitations are being mitigated. AI-assisted MD simulations, particularly through machine learning (ML) techniques, have become a powerful tool in optimizing force fields and analyzing molecular dynamics trajectories. This section explores the integration of AI with MD simulations, focusing on machine learning in force field optimization, AI-driven trajectory analysis, and how these advancements are reshaping molecular modeling.

The accuracy of MD simulations depends heavily on the quality of the force fields used to model molecular interactions. Traditional force fields, such as AMBER and CHARMM, rely on empirical parameters derived from quantum mechanical calculations and experimental data. However, these force fields often struggle to capture the complexity of molecular interactions, particularly in highly flexible or unconventional drug-like molecules. To address these limitations, researchers have turned to machine learning (ML) techniques to optimize force field parameters dynamically.

Machine learning has been employed to refine force fields by learning from large datasets of molecular systems. Instead of relying

on predefined parameters, AI algorithms can generate new force field parameters by identifying patterns in large datasets of experimental data and high-level quantum mechanical calculations. These learned parameters can be incorporated into the force fields, making them more accurate and adaptable to a broader range of molecular systems. For instance, Gromacs has incorporated AI techniques to generate adaptive force fields that better represent the dynamics of complex biological molecules like proteins and lipids. By using AI to update the parameters based on the molecular environment, simulations can more accurately reflect real-world behavior, especially for systems that are not well-represented in traditional force fields.

A notable example of AI-assisted force field optimization is DeepMind's application of deep learning techniques to improve protein folding simulations. In 2020, DeepMind introduced the AlphaFold algorithm, which accurately predicted protein structures based on their amino acid sequences. While AlphaFold was primarily focused on protein structure prediction, the underlying AI techniques were adapted to improve force fields used in MD simulations. AlphaFold's ability to predict protein folding has made a significant impact on drug discovery, as understanding protein structure is critical in designing effective therapeutics. Furthermore, machine learning can predict the effects of mutations on protein structure and stability. For example, GDF-15 protein structure changes due to mutations were modeled using machine learning, where traditional force fields had difficulty capturing such complexities.

While AI-driven optimization of force fields has shown great promise, it also comes with challenges. The main obstacle lies in the need for large and high-quality datasets to train AI models. Molecular systems are highly complex, and generating sufficient data to teach machine learning models can be time-consuming and

computationally expensive. Additionally, AI models tend to be black-box systems, making it difficult to interpret how the algorithms arrive at certain conclusions. This lack of interpretability can hinder the broader acceptance of AI in computational chemistry and drug design, where understanding the underlying mechanisms is crucial.

- AI-driven trajectory analysis for predicting molecular behavior.

One of the critical applications of AI in MD simulations is trajectory analysis. Traditional MD simulations generate vast amounts of data, such as atomic coordinates and velocities at each time step, which can be overwhelming to analyze. The manual extraction of meaningful insights from these trajectories often requires cumbersome and time-intensive post-processing. AI provides a more efficient way to analyze and extract relevant features from these large datasets.

1. Trajectory Clustering and Dimensionality Reduction

AI-driven techniques, such as unsupervised machine learning, are increasingly used to identify patterns in MD simulation trajectories. Clustering algorithms can group similar conformations together, allowing researchers to focus on the most relevant regions of phase space. Dimensionality reduction techniques, such as principal component analysis (PCA) and t-SNE, enable the simplification of high-dimensional data to identify key conformational changes in molecular systems. For example, AI can cluster the binding modes of a drug molecule to its target protein, reducing the complexity of the dataset while maintaining essential information about the system's behavior. By grouping similar conformations together, AI helps identify the most probable pathways of drug-receptor interactions.

2. Case Study: AI in Predicting Protein-Ligand Interactions

AI-driven trajectory analysis has been applied to predict protein-ligand interactions. In one case, DeepChem, an open-source library, used deep learning algorithms to predict how small molecules bind to proteins. By analyzing the binding trajectories of different ligands, AI models can predict the binding affinity and optimize drug candidates for better efficacy. MD simulations combined with AI allow researchers to simulate millions of ligand-receptor interactions, significantly improving the speed and accuracy of drug discovery.

3. Predicting Conformational Transitions

MD simulations often explore conformational transitions, such as the folding of proteins or the changes in shape that occur during molecular recognition. These transitions can be rare events, making them difficult to capture with traditional methods. AI algorithms, especially reinforcement learning, can help predict rare conformational transitions by learning from previous simulations. By continuously updating models based on the feedback from simulations, reinforcement learning enables a more comprehensive understanding of the molecular mechanisms governing conformational changes.

Example: The use of reinforcement learning to predict protein conformational changes was demonstrated in the simulation of G-protein-coupled receptors (GPCRs), which undergo complex conformational changes when activated by ligands. Traditional MD simulations require extensive time scales to observe these transitions, but AI models can accelerate this process, providing insights that would be difficult to capture otherwise.

4. Challenges and Limitations in Trajectory Analysis

Despite its promise, AI-driven trajectory analysis faces several challenges. The quality of predictions depends heavily on the data used to train AI models, and in some cases, the molecular systems being studied may not have sufficient high-quality data. In addition, the complexity of AI models can sometimes lead to overfitting, where the model becomes too specialized to a specific dataset and performs poorly on others.

Another issue is the interpretability of AI-driven analysis. While AI models can identify patterns in data, understanding the underlying reasons for these patterns is more challenging. This lack of transparency can make it difficult to draw conclusions about the molecular mechanisms involved in a particular process.

The Future of AI-Assisted MD Simulations

The future of AI-assisted MD simulations is incredibly promising. As computational power continues to grow and more sophisticated AI algorithms are developed, the integration of AI with MD simulations will lead to faster, more accurate, and more comprehensive predictions of molecular behavior.

1. **Integration with Quantum Mechanics** One of the major frontiers in computational chemistry is the integration of MD simulations with quantum mechanics (QM). While MD simulations typically rely on classical mechanics, quantum mechanics provides a more accurate description of molecular interactions, especially for reactions involving electrons. AI can play a crucial role in bridging the gap between these two methods, allowing for more precise predictions of molecular behavior in complex systems.

2. **Real-Time Drug Discovery** In the future, AI-assisted MD simulations may enable real-time drug discovery by combining trajectory analysis with high-throughput screening methods. This would allow researchers to rapidly simulate and analyze the effects of thousands of potential drug candidates, significantly accelerating the drug discovery process.
3. **Personalized Medicine** AI-driven MD simulations can also play a key role in personalized medicine. By simulating drug interactions in patient-specific models, researchers can predict how a particular individual will respond to a drug, taking into account genetic variations and other factors. This approach could lead to more effective and tailored treatments, minimizing side effects and improving therapeutic outcomes.

We will explore these promising advancements and their potential impact in greater detail in **Chapter 3**, which delves deeper into the future applications of AI-assisted MD simulations in both drug discovery and personalized medicine.

AI-assisted MD simulations represent a transformative approach in molecular modeling and drug discovery. By leveraging machine learning for force field optimization and trajectory analysis, researchers can gain deeper insights into molecular interactions, optimize drug candidates, and accelerate the drug development process. While there are challenges to overcome, particularly in terms of data quality, model interpretability, and computational demands, the integration of AI with MD simulations is poised to revolutionize how we approach drug design, personalized medicine, and molecular dynamics as a whole.

As computational tools evolve, the synergy between AI and MD simulations will continue to improve the accuracy and efficiency of drug discovery, leading to more effective treatments and a better understanding of the molecular mechanisms underlying diseases.

From QSAR Applications to In Vitro Experimental Validation

Quantitative Structure-Activity Relationship (QSAR) modeling is a crucial tool in modern drug discovery. It helps researchers predict the biological activity of chemical compounds based on their molecular structure, providing valuable insights long before experimental (in vitro) testing is performed. This predictive modeling approach can significantly reduce the time and resources spent on synthesizing and testing compounds in the lab. By understanding how the chemical structure of a drug influences its biological activity, researchers can design more efficient drug candidates and optimize molecular properties in silico.

QSAR modeling operates on the principle that the relationship between a molecule's chemical structure and its biological activity can be quantitatively described. This relationship is captured in a mathematical model that can predict the activity of a compound based on its molecular features. These models rely on the correlation between the molecular descriptors, such as physicochemical properties, molecular topology, and electronic characteristics, and the observed biological activity. Before any compound reaches the laboratory for testing, QSAR can be employed to predict how a set of candidate molecules will perform. This allows researchers to select compounds that are most likely to exhibit the desired biological activity, thus prioritizing them for synthesis and in vitro screening.

Molecular Descriptors: The Key to Predicting Activity

The foundation of QSAR modeling is the use of molecular descriptors, quantitative representations of a molecule's structural and chemical features. These descriptors can be divided into several categories:

1. **Physicochemical Descriptors**: These include molecular weight, lipophilicity (logP), and hydrogen bond donors/acceptors. These properties are crucial for understanding how a compound will interact with biological targets, such as receptors or enzymes.
2. **Topological Descriptors**: These describe the connectivity and spatial arrangement of atoms within a molecule, including the number of rings, branching, and bonds. Topological descriptors are important for understanding how the molecule's shape influences its ability to bind to its target.
3. **Electrostatic Descriptors**: These represent the distribution of electronic charge across the molecule. Electrostatic interactions play a vital role in molecular recognition and binding to biological targets, influencing the potency and selectivity of a drug.
4. **Geometrical Descriptors**: These describe the three-dimensional shape of a molecule, including its size, flexibility, and the orientation of functional groups. The geometric features of a molecule can determine how well it fits into a biological receptor site.

By compiling these descriptors, QSAR models can generate equations that relate the chemical structure of a compound to its

biological activity, allowing researchers to predict how new molecules will perform.

QSAR Models for Activity Prediction

Once the relevant molecular descriptors are extracted, they are used to build a QSAR model. The model is trained using a set of compounds with known biological activities. These compounds are often classified based on their ability to bind to a specific target, such as a receptor or enzyme, or their ability to produce a desired therapeutic effect.

The QSAR model is constructed by applying statistical or machine learning techniques to correlate the molecular descriptors with biological activity. Common methods for building QSAR models include:

1. **Linear Regression**: A simple statistical technique that establishes a linear relationship between the descriptors and the biological activity.
2. **Partial Least Squares (PLS)**: A multivariate statistical method used to handle data with high dimensionality, commonly employed when dealing with a large number of molecular descriptors.
3. **Support Vector Machines (SVM) and Random Forests**: More advanced machine learning algorithms used to improve the predictive power of QSAR models, especially when the dataset is complex or contains non-linear relationships.
4. **Neural Networks**: A computational technique inspired by the human brain that can model highly non-linear relationships between molecular structure and biological activity.

Once the model is trained, it can predict the biological activity of new, untested molecules based on their descriptors. This predictive ability allows researchers to evaluate a vast number of compounds before they enter the lab, significantly narrowing down the field of candidates for further testing.

The application of QSAR models in drug design offers numerous advantages, particularly in the early stages of drug discovery:

1. **Reduced Time and Cost**: Traditional drug discovery involves the synthesis and testing of many compounds, which can be both time-consuming and expensive. By predicting the activity of compounds before they are synthesized, QSAR models help prioritize the most promising candidates for further development. This significantly reduces both the cost and time involved in drug development.
2. **Optimization of Drug Candidates**: QSAR models can be used to optimize molecular properties by predicting how small changes in structure will affect biological activity. For example, if a compound has low potency, QSAR can suggest modifications to improve binding affinity, solubility, or bioavailability. This process, known as "structure-activity optimization," allows for the fine-tuning of drug candidates.
3. **Incorporation of Multiple Target Interactions**: QSAR models are not limited to single-target drugs. They can also account for compounds that interact with multiple targets, allowing for the design of polypharmacological agents. This is particularly useful for complex diseases like cancer or

neurological disorders, where multiple pathways may need to be modulated simultaneously.

4. **Identification of Toxicity Risks**: QSAR models can also predict the potential toxicity of a drug candidate by evaluating its interactions with biological targets that are associated with adverse effects. For example, QSAR models can be used to predict hepatotoxicity, cardiotoxicity, or other harmful effects before in vitro testing, reducing the risk of failure in later stages of development.
5. **Virtual Screening of Chemical Libraries**: QSAR can be integrated into virtual screening workflows to analyze large chemical libraries in silico. By ranking compounds based on predicted biological activity, QSAR models can help identify lead compounds for experimental validation.

Examples of QSAR in Drug Design

Several successful applications of QSAR in drug design highlight its power and versatility:

- **HIV Protease Inhibitors**: QSAR modeling has been extensively used to design inhibitors of HIV protease, a key enzyme involved in the replication of the HIV virus. By analyzing the molecular structure of known inhibitors, researchers have been able to develop new compounds with improved potency and selectivity against the HIV protease (Darnag et al., 2017).
- **Anticancer Agents**: QSAR has played a significant role in the development of anticancer drugs, including those targeting the epidermal growth factor receptor (EGFR). By correlating the structure of

small molecules with their ability to inhibit EGFR, QSAR models have helped identify lead compounds with enhanced antitumor activity (Zhao et al., 2017).

- **Anti-inflammatory Drugs**: In the design of non-steroidal anti-inflammatory drugs (NSAIDs), QSAR models have been used to predict the ability of molecules to inhibit cyclooxygenase enzymes (COX-1 and COX-2). This has led to the identification of drugs with improved selectivity for COX-2, reducing side effects like gastric irritation (Asirvatham et al., 2019).
- **Antibiotics**: QSAR has also been used to design novel antibiotics by predicting the activity of compounds against bacterial targets such as DNA gyrase or beta-lactamase. Through QSAR, researchers have been able to develop antibiotics with enhanced efficacy and reduced resistance potential (Jakhar et al., 2022).

Challenges and Limitations of QSAR Models

While QSAR modeling is a powerful tool in drug design, it does come with some challenges and limitations:

1. **Data Quality and Quantity**: The accuracy of a QSAR model depends on the quality and size of the dataset used to train it. Insufficient or poorly curated data can lead to inaccurate predictions. Moreover, models may not generalize well to new, unseen compounds, particularly if the dataset is too narrow or unrepresentative of chemical diversity.
2. **Model Interpretability**: QSAR models can sometimes be difficult to interpret, especially when

advanced machine learning techniques are used. Understanding why a particular structure leads to high or low activity is often crucial for drug optimization, but some methods, like neural networks, can act as black boxes with little insight into the underlying mechanisms.

3. **Complexity of Biological Systems**: Biological systems are inherently complex, and QSAR models often rely on simplifications that may not capture all aspects of molecular interactions. For example, models may not fully account for protein flexibility, receptor conformational changes, or the dynamic nature of drug-target interactions.

4. **Limited Applicability to Novel Chemistries**: QSAR models are typically built on historical data, meaning they may not perform well when applied to novel chemistries or compounds that differ significantly from the training set. New classes of drug candidates, such as biologics or synthetic peptides, may require different predictive approaches.

QSAR models have become an indispensable tool in drug design, guiding the development of compounds with the desired biological activity before in vitro testing. By providing insights into the relationship between molecular structure and biological function, QSAR enables researchers to prioritize compounds for synthesis, optimize drug properties, and reduce the overall cost and time required for drug development. Despite challenges such as data quality and model interpretability, the continuous advancement of QSAR modeling techniques holds great promise for accelerating the discovery of novel and effective therapeutics.

In Vivo Applications:

As drug discovery progresses from in silico and in vitro stages toward clinical development, understanding how a compound behaves in vivo becomes increasingly important. The ability to predict how a drug will interact with the body, its absorption, distribution, metabolism, excretion (ADME) properties, and potential toxicity, is crucial for deciding whether a compound should move forward in development. While traditional experimental testing provides essential information about pharmacokinetics and toxicity, these processes can be time-consuming and expensive. This is where Quantitative Structure-Activity Relationship (QSAR) models offer significant advantages by allowing researchers to predict these critical in vivo properties before experimental testing.

QSAR models have evolved beyond predicting simple biological activity to provide valuable insights into how a drug candidate will behave in the body. By analyzing the chemical structure of a compound and correlating it with known pharmacokinetic and toxicity profiles, QSAR can guide the selection of compounds with favorable properties for in vivo testing. This predictive modeling approach helps streamline the drug development process, reducing the number of candidates that need to undergo expensive and lengthy in vivo testing.

Pharmacokinetics (PK) refers to how the body absorbs, distributes, metabolizes, and excretes a drug. Each of these processes significantly influences the efficacy and safety of a drug, and predicting these properties early in the drug discovery process can help identify compounds with optimal PK profiles.

1. Absorption

The first step in determining pharmacokinetics is predicting how well a compound will be absorbed into the bloodstream after

administration. Several factors influence drug absorption, such as molecular size, lipophilicity (fat solubility), hydrogen bonding potential, and the ability of a molecule to pass through biological membranes. QSAR models use molecular descriptors related to these properties to predict a compound's absorption profile.

For example, lipophilicity is a key predictor of absorption. Drugs with moderate lipophilicity (neither too hydrophilic nor too lipophilic) tend to pass through cell membranes more efficiently. QSAR models that include descriptors like partition coefficient (logP), which measures the relative solubility of a compound in water and fat, are commonly used to predict a drug's absorption potential. A logP value in a certain range indicates a good balance between solubility and permeability across biological barriers, such as the intestinal wall, which is crucial for oral absorption.

Furthermore, molecular weight is another important factor. Compounds with higher molecular weight may face challenges in crossing cell membranes or being absorbed efficiently. QSAR models often use a combination of molecular weight, hydrogen bond donors and acceptors, and the surface area of the molecule to predict absorption efficiency.

2. Distribution

After a drug enters the bloodstream, it must be distributed to the target tissues where it will exert its therapeutic effect. Predicting drug distribution involves understanding the molecule's ability to bind to plasma proteins, such as albumin, and its tendency to accumulate in various tissues, including the liver, kidneys, and fat.

Plasma protein binding is an important factor in predicting how a compound will distribute throughout the body. QSAR models often use molecular descriptors related to charge distribution, dipole moment, and lipophilicity to predict the likelihood of a compound

binding to plasma proteins. High protein binding can reduce the free drug concentration available for therapeutic action, affecting both efficacy and safety.

Moreover, volume of distribution (Vd) is a critical pharmacokinetic parameter that QSAR models can predict. This parameter refers to the extent to which a drug is distributed into body tissues. Compounds with high lipophilicity are more likely to accumulate in fatty tissues, whereas polar or hydrophilic compounds tend to stay in the bloodstream or accumulate in organs like the liver or kidneys.

3. Metabolism

Once absorbed and distributed, drugs undergo metabolism in the liver, primarily by cytochrome P450 enzymes. The metabolic stability of a drug plays a vital role in determining its half-life in the body and whether it will accumulate to toxic levels. Predicting a drug's metabolic profile early in the development process helps researchers avoid compounds that may have a high risk of rapid metabolism, leading to poor efficacy or the formation of toxic metabolites.

Cytochrome P450 inhibition is a common focus in QSAR models aimed at predicting metabolic interactions. By analyzing the molecular structure and predicting whether a compound is likely to inhibit or be metabolized by specific P450 enzymes, QSAR models can flag potential drug-drug interactions or toxic metabolites. Moreover, metabolic stability can be predicted using descriptors such as polar surface area (PSA) and topological polar surface area (TPSA), which correlate with a drug's ability to interact with metabolic enzymes. Compounds with a large polar surface area may be metabolized more easily, whereas those with smaller polar regions may be more stable and resistant to metabolism.

4. Excretion

Finally, the excretion of a drug is another critical factor in its pharmacokinetic profile. The kidneys are primarily responsible for eliminating water-soluble drugs, whereas lipophilic compounds are excreted in bile. Predicting a drug's excretion is essential for understanding its clearance rate and potential accumulation in the body.

Toxicity Predictions Using QSAR

The ability to predict toxicity is one of the most valuable aspects of QSAR models in drug development. Toxicity can manifest in many forms, including liver toxicity, cardiac toxicity, and genotoxicity, and it is a major reason why drug candidates fail during clinical development. Early prediction of potential toxicities can help eliminate problematic candidates before expensive animal studies or clinical trials.

1. Hepatotoxicity

Liver toxicity is a common and serious adverse effect associated with many drugs. Compounds that are metabolized by the liver can form reactive metabolites that interact with cellular components, leading to liver damage. QSAR models can be used to predict the likelihood of hepatotoxicity based on molecular structure. Descriptors such as lipophilicity, electrophilicity, and aromaticity are often included in models aimed at predicting liver toxicity.

For instance, compounds with highly electrophilic centers or specific aromatic structures may be more likely to undergo biotransformation into reactive metabolites that can bind to liver proteins, causing damage. Predicting these reactivity patterns early in development can help researchers avoid compounds with high hepatotoxicity potential.

2. Cardiotoxicity

Drugs that affect the heart can lead to arrhythmias, heart failure, or other severe cardiovascular effects. QSAR models for cardiotoxicity often focus on predicting the ability of a compound to interact with ion channels in the heart, such as the hERG (human Ether-à-go-go-Related Gene) channel, which is crucial for regulating the heart's electrical activity.

By modeling molecular interactions with hERG channels, QSAR can predict the likelihood that a compound will cause QT prolongation, a condition that can lead to fatal arrhythmias. Researchers can use QSAR models to filter out compounds with a high risk of cardiac toxicity before they proceed to in vivo testing.

3. Genotoxicity

Genotoxicity refers to the potential of a drug to cause genetic mutations or chromosomal damage, leading to cancer or birth defects. QSAR models for genotoxicity often focus on predicting whether a compound can cause DNA damage or induce mutations in cells. These models use molecular descriptors that indicate whether a compound has electrophilic centers or other reactive features capable of interacting with DNA.

By identifying compounds with a high likelihood of causing genotoxicity, QSAR models help prioritize safer drug candidates and reduce the risk of harmful effects in humans.

In vivo predictions of pharmacokinetics and toxicity are crucial for guiding drug development and minimizing the risks associated with new drug candidates. QSAR models have become an essential tool in this regard, allowing researchers to predict how a compound will behave in the body long before it undergoes costly and time-consuming in vivo testing. By predicting absorption, distribution, metabolism, excretion, and potential toxicity, QSAR

models help prioritize compounds with favorable pharmacokinetic profiles and reduced risks of harmful side effects. As computational tools continue to advance, QSAR modeling will play an even more significant role in optimizing drug development, reducing attrition rates, and accelerating the time it takes to bring safe and effective drugs to market.

AI-Powered QSAR Models

Quantitative Structure-Activity Relationship (QSAR) modeling has long been a fundamental tool in drug discovery, enabling the prediction of biological activity based on the chemical structure of a molecule. However, traditional QSAR models often face limitations due to the complexity of molecular systems and the sheer volume of data required to accurately model molecular interactions. With the rise of artificial intelligence (AI) and machine learning (ML), these traditional QSAR approaches are being enhanced, leading to more accurate predictions and a deeper understanding of how molecular structures influence biological activity.

AI-powered QSAR models leverage advanced machine learning techniques to optimize prediction accuracy, providing a powerful tool for drug discovery. These models not only improve the predictive power of QSAR but also handle more complex and larger datasets, uncovering hidden relationships that were previously difficult to identify using traditional methods. Two prominent machine learning techniques that have been successfully applied to QSAR models are random forests and deep learning models. These methods are particularly useful in overcoming the limitations of conventional QSAR approaches, such as the need for manual feature selection and the challenges of predicting activity for new compounds with high accuracy.

Machine learning models, especially random forests and deep learning, have revolutionized the way QSAR predictions are made. By using these methods, researchers can automate the process of identifying important molecular features and relationships, significantly improving the prediction accuracy of drug candidates.

1. Random Forests in QSAR

Random forests (RF) are a type of ensemble learning algorithm that builds a collection of decision trees to solve a problem. Each tree is constructed by randomly selecting a subset of features and data points, which helps improve the model's robustness and reduces overfitting. When applied to QSAR, random forests can handle complex, non-linear relationships between molecular descriptors and biological activity, which may be challenging for traditional linear regression models.

In the context of QSAR, random forests offer several advantages:

- **Feature Importance**: One of the key strengths of random forests is their ability to determine the importance of various molecular descriptors (such as molecular weight, hydrophobicity, and electrostatic potential) in predicting the activity of a drug candidate. This helps in understanding which features drive biological activity and provides insights for designing better drug candidates.
- **Handling Missing Data**: Random forests can handle missing or incomplete data, which is often a challenge in large datasets. By using bootstrapping and sampling techniques, random forests can make accurate predictions even when certain molecular descriptors are missing.

- **Non-Linear Relationships**: Unlike traditional linear models, random forests can capture complex, non-linear relationships between molecular descriptors and biological activity. This is important because many biological processes are non-linear, and random forests can provide a more accurate representation of these complexities.

Example:

In a study by Lind & Anderson, 2019, random forests were used to predict the anti-cancer activity of a series of small molecules. The random forest model identified key molecular features such as the size and shape of the compounds, as well as specific functional groups, which contributed to their cytotoxic effects. The model's ability to predict activity with high accuracy allowed the researchers to prioritize drug candidates for further experimental validation.

2. Deep Learning in QSAR

Deep learning models, particularly neural networks, have become increasingly popular in QSAR due to their ability to learn from large datasets and uncover hidden patterns. Deep learning models consist of multiple layers of interconnected nodes (neurons), and they excel at automatically extracting features from raw data without the need for manual feature engineering. In QSAR, deep learning methods can analyze molecular descriptors, raw chemical representations, and even 3D structures of compounds to predict biological activity.

Deep learning approaches, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been applied to QSAR for several reasons:

- **Automatic Feature Learning**: Unlike traditional QSAR methods, deep learning can automatically learn the most

relevant features from raw chemical data, including molecular fingerprints, 2D and 3D molecular representations, and sequence-based information (for proteins or nucleic acids).

- **Handling Complex Data Types**: Deep learning models can integrate diverse types of data, such as molecular structure, biological data, and even textual descriptions of chemical reactions, into a unified predictive framework. This ability makes deep learning particularly useful in predicting complex biological phenomena like protein-ligand binding.
- **Non-Linear Modeling**: Deep neural networks can model complex, highly non-linear relationships between molecular features and biological activity, which is difficult for traditional QSAR models to capture effectively.

Example:

In a study by Falls et al., 2021 deep learning models were applied to predict the binding affinity of drug candidates to the human immunodeficiency virus (HIV) protease. The model utilized molecular fingerprints and deep learning techniques to predict binding affinity with high accuracy, outperforming traditional QSAR models and highlighting the potential of deep learning in drug discovery.

3. Enhancing QSAR Prediction Accuracy with AI

AI-powered QSAR models have the potential to significantly enhance prediction accuracy by addressing some of the key limitations of traditional QSAR approaches. These limitations include overfitting, difficulty in modeling non-linear relationships, and the challenges of feature selection. By using machine learning

algorithms like random forests and deep learning, researchers can build more robust models that generalize better to new compounds and reduce the risk of false positives or negatives.

a. Model Generalization and Overfitting

Overfitting occurs when a model becomes too tailored to the training data, making it less effective when applied to new data. One of the key benefits of machine learning techniques, such as random forests and deep learning, is their ability to reduce overfitting. Random forests, for example, use ensemble methods to combine the predictions of multiple decision trees, which reduces the impact of individual trees overfitting the training data. Similarly, deep learning models, especially when properly tuned, can generalize better to new compounds by learning hierarchical features that are more representative of biological activity.

b. Feature Selection and Dimensionality Reduction

Traditional QSAR models often require manual selection of molecular descriptors, which can be time-consuming and may result in overlooking important features. Machine learning algorithms like random forests and deep learning can automatically identify the most relevant features for prediction, eliminating the need for manual feature selection. This not only saves time but also leads to models that are more efficient and accurate.

Deep learning models, in particular, are effective at handling high-dimensional data, such as 3D molecular structures or molecular dynamics simulations. They can reduce the dimensionality of the input data while preserving important information, leading to more efficient models that still capture the essential aspects of molecular activity.

c. Handling Large Datasets

One of the most significant advantages of machine learning-based QSAR models is their ability to handle large, complex datasets. With the increasing availability of large chemical and biological databases, deep learning and random forests can process vast amounts of data that traditional QSAR models would struggle to handle. This allows researchers to train models on more diverse datasets, which improves their ability to predict the activity of novel compounds and identify potential drug candidates with high accuracy.

d. Transfer Learning and Data Augmentation

Transfer learning is another powerful technique in AI that has been applied to QSAR modeling. This approach allows a model trained on one dataset to be fine-tuned and adapted to a new dataset, even with limited data. Transfer learning is particularly useful in drug discovery when only a small dataset of relevant compounds is available for a particular target or disease.

Data augmentation, a technique commonly used in deep learning, is another method that improves the generalization of QSAR models. By generating synthetic data through perturbations of existing data, researchers can expand their training datasets, making models more robust and less likely to overfit.

AI-powered QSAR models, incorporating machine learning techniques such as random forests and deep learning, are transforming drug discovery by improving prediction accuracy and enhancing the process of identifying promising drug candidates. These models overcome many of the challenges faced by traditional QSAR methods, including overfitting, manual feature selection, and the difficulty of modeling complex, non-linear relationships. With the ability to handle large datasets, automatically select features, and

generalize to novel compounds, machine learning-based QSAR models offer a more powerful and efficient approach to predicting biological activity and guiding the drug discovery process. As AI technologies continue to evolve, their integration with QSAR will continue to enhance the ability to design better drugs more quickly and accurately.

In wrapping up this chapter, it's clear that computational methods have reshaped the landscape of drug discovery, turning what was once a trial-and-error process into a precise science. Molecular simulations and other computational tools have given us the ability to peer into the molecular machinery of diseases, offering insights that were unimaginable just a few decades ago. From predicting how a drug will bind to a target protein, to simulating the dynamic environment of biological systems, these technologies have empowered researchers to design and optimize drugs with unprecedented accuracy.

But as we look ahead, the horizon of drug discovery is becoming even more exciting. The next frontier, driven by the fusion of CADD and artificial intelligence, is set to revolutionize the field further. Imagine a future where AI doesn't just analyze data, but actively predicts the next breakthrough treatment, continuously learning from vast datasets to optimize drug design in real-time. The potential is limitless. As we delve into the future of CADD and AI in the next chapter, one thing is certain: we're only scratching the surface of what's possible in the race to find new, more effective therapies.

References

Shamim, S., Munawar, R., Rashid, Y., Muhammad Zesshan Qadar, S., Bushra, R., Begum, I., … Quds, T. (2024). Molecular Docking: An Insight from Drug Discovery to Drug Repurposing Approach. IntechOpen. doi: 10.5772/intechopen.1005526

Aguiar, C., & Camps, I. (2024). Molecular docking in drug discovery: Techniques, applications, and advancements [Preprint]. *ChemRxiv*. https://doi.org/10.26434/chemrxiv-2024-gmhtx

Ferreira, L. G., Dos Santos, R. N., Oliva, G., & Andricopulo, A. D. (2015). Molecular Docking and Structure-Based Drug Design Strategies. *Molecules*, *20*(7), 13384-13421. https://doi.org/10.3390/molecules200713384

Türkoğlu, E. A., Taştekil, I., & Özbek Sarica, P. (2024). Surface plasmon resonance, molecular docking, and molecular dynamics simulation studies of lysozyme interaction with tannic acid. *Food Science & Nutrition, 12*(7), Article e4315. https://doi.org/10.1002/fsn3.4315

Alali, H., Bloch, I., Rapaport, I., Rodrigues, L., Sher, I., Ansbacher, T., & Gal, M. (2022). Application of In Silico Filtering and Isothermal Titration Calorimetry for the Discovery of Small Molecule Inhibitors of MDM2. *Pharmaceuticals (Basel, Switzerland)*, *15*(6), 752. https://doi.org/10.3390/ph15060752

Raghu, M. S., Kumar, K. Y., Prashanth, M. K., Anusuya Devi, V. S., Alharethy, F., & Jeon, B.-H. (2023). Synthesis, enzyme inhibition and molecular docking studies of pyrazolo[1,5-a][1,3,5] triazine derivatives as potential antioxidant agents. *Journal of Molecular Structure, 1294*, 136332. https://doi.org/10.1016/j.molstruc.2023.136332

Fradera, X., Reutershan, M. H., Machacek, M. R., Trotter, B. W., & McCoy, M. A. (2022). NMR data-driven docking of HDM2-

inhibitor complexes. *ChemBioChem, 23*(3), e202100570. https://doi.org/10.1002/cbic.202100570

Wang, H., Lin, Q., Liu, M., Ding, W., Weng, N., Ni, H., Lu, J., Lyu, M., & Wang, S. (2023). Molecular Docking and Site-Directed Mutagenesis of GH49 Family Dextranase for the Preparation of High-Degree Polymerization Isomaltooligosaccharide. *Biomolecules*, *13*(2), 300. https://doi.org/10.3390/biom13020300

Schindler, T., Bornmann, W., Pellicena, P., Miller, W. T., Clarkson, B., & Kuriyan, J. (2000). Structural mechanism for STI-571 inhibition of Abelson tyrosine kinase. *Science, 289*(5486), 1938–1942. https://doi.org/10.1126/science.289.5486.1938

Buchdunger, E., Zimmermann, J., Mett, H., Meyer, T., Müller, M., Druker, B. J., & Lydon, N. B. (1996). Inhibition of the Abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative. *Cancer research*, *56*(1), 100–104.

Gupta, S. K., Gupta, S. K., Smita, S., Srivastava, M., Lai, X., Schmitz, U., Rahman, Q., Wolkenhauer, O., & Vera, J. (2011). Computational analysis and modeling the effectiveness of 'Zanamivir' targeting neuraminidase protein in pandemic H1N1 strains. *Infection, Genetics and Evolution, 11*(5), 1072–1082. https://doi.org/10.1016/j.meegid.2011.03.018

Smith, B. J., Colman, P. M., Von Itzstein, M., Danylec, B., & Varghese, J. N. (2001). Analysis of inhibitor binding in influenza virus neuraminidase. *Protein science : a publication of the Protein Society*, *10*(4), 689–696. https://doi.org/10.1110/ps.41801

Cheer, S. M., & Wagstaff, A. J. (2002). Zanamivir: An update of its use in influenza. *Drugs, 62*, 71–106. https://doi.org/10.2165/00003495-200262010-00007

Alugubelli, Y. R., Geng, Z. Z., Yang, K. S., Shaabani, N., Khatua, K., Ma, X. R., Vatansever, E. C., Cho, C.-C., Ma, Y., Xiao, J., Blankenship, L. R., Yu, G., Sankaran, B., Li, P., Allen, R., Ji, H., Xu, S., & Liu, W. R. (2022). A systematic exploration of boceprevir-based main protease inhibitors as SARS-CoV-2 antivirals. *European Journal of Medicinal Chemistry, 240*, 114596. https://doi.org/10.1016/j.ejmech.2022.114596

Krismer, L., Schöppe, H., Rauch, S., Bante, D., Sprenger, B., Naschberger, A., Costacurta, F., Fürst, A., Sauerwein, A., Rupp, B., Kaserer, T., von Laer, D., & Heilmann, E. (2024). Study of key residues in MERS-CoV and SARS-CoV-2 main proteases for resistance against clinically applied inhibitors nirmatrelvir and ensitrelvir. *npj Viruses, 2*, Article 23. https://doi.org/10.1038/s44238-024-00023-x

Gong, Q., Song, J., Song, Y., Tang, K., Yang, P., Wang, X., Zhao, M., Ouyang, L., Rao, L., Yu, B., Zhan, P., Zhang, S., & Zhang, X. (2025). New techniques and strategies in drug discovery (2020–2024 update). *Chinese Chemical Letters, 36*(3), 110456. https://doi.org/10.1016/j.cclet.2024.110456

Mo, X., Rao, D. P., Kaur, K., Hassan, R., Abdel-Samea, A. S., Farhan, S. M., Bräse, S., & Hashem, H. (2024). Indole Derivatives: A Versatile Scaffold in Modern Drug Discovery—An Updated Review on Their Multifaceted Therapeutic Applications (2020–2024). *Molecules*, *29*(19), 4770. https://doi.org/10.3390/molecules29194770

Li, Y., Guo, B., Xu, Z., Li, B., Cai, T., Zhang, X., Yu, Y., Wang, H., Shi, J., & Zhu, W. (2016). Repositioning organohalogen drugs: A case study for identification of potent B-Raf V600E inhibitors via docking and bioassay. *Scientific Reports, 6*, Article 31074. https://doi.org/10.1038/srep31074

Talele, T. T., Khedkar, S. A., & Rigby, A. C. (2010). Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Current topics in medicinal chemistry*, *10*(1), 127–141. https://doi.org/10.2174/156802610790232251

Turk B. (2006). Targeting proteases: successes, failures and future prospects. *Nature reviews. Drug discovery*, *5*(9), 785–799. https://doi.org/10.1038/nrd2092

Yan, Y., Li, Y., Munshi, S., Sardana, V., Cole, J. L., Sardana, M., Steinkuehler, C., Tomei, L., De Francesco, R., Kuo, L. C., & Chen, Z. (1998). Complex of NS3 protease and NS4A peptide of BK strain hepatitis C virus: a 2.2 A resolution structure in a hexagonal crystal form. *Protein science : a publication of the Protein Society*, *7*(4), 837–847. https://doi.org/10.1002/pro.5560070402

Buchdunger, E., Cioffi, C. L., Law, N., Stover, D., Ohno-Jones, S., Druker, B. J., & Lydon, N. B. (2000). Abl protein-tyrosine kinase inhibitor STI571 inhibits in vitro signal transduction mediated by c-kit and platelet-derived growth factor receptors. *The Journal of pharmacology and experimental therapeutics*, *295*(1), 139–145.

von Itzstein, M., Wu, W.-Y., Kok, G. B., Pegg, M. S., Dyason, J. C., Jin, B., Phan, T. V., Smythe, M. L., White, H. F., Oliver, S. W., Colman, P. M., Varghese, J. N., Ryan, D. M., Woods, J. M., Bethell, R. C., Hotham, V. J., Cameron, J. M., & Penn, C. R. (1993). Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature, 363*, 418–423. https://doi.org/10.1038/363418a0

Ryan, D. M., Ticehurst, J., Dempsey, M. H., & Penn, C. R. (1994). Inhibition of influenza virus replication in mice by GG167 (4-guanidino-2,4-dideoxy-2,3-dehydro-N-acetylneuraminic acid) is consistent with extracellular activity of viral neuraminidase (sialidase). *Antimicrobial agents and chemotherapy*, *38*(10), 2270–2275.

Dunn, C. J., & Goa, K. L. (1999). Zanamivir: a review of its use in influenza. *Drugs*, *58*(4), 761–784. https://doi.org/10.2165/00003495-199958040-00016

Ryan, D. M., Ticehurst, J., Dempsey, M. H., & Penn, C. R. (1994). Inhibition of influenza virus replication in mice by GG167 (4-guanidino-2,4-dideoxy-2,3-dehydro-N-acetylneuraminic acid) is consistent with extracellular activity of viral neuraminidase (sialidase). *Antimicrobial agents and chemotherapy*, *38*(10), 2270–2275. https://doi.org/10.1128/AAC.38.10.2270

Moscona, A. (2005). Neuraminidase inhibitors for influenza. *The New England Journal of Medicine, 353*(13), 1363–1373. https://doi.org/10.1056/NEJMra050740

Bollag, G., Tsai, J., Zhang, J., Zhang, C., Ibrahim, P., Nolop, K., & Hirth, P. (2012). Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nature reviews. Drug discovery*, *11*(11), 873–886. https://doi.org/10.1038/nrd3847

Kim, J. H., & Scialli, A. R. (2011). Thalidomide: the tragedy of birth defects and the effective treatment of disease. *Toxicological sciences : an official journal of the Society of Toxicology*, *122*(1), 1–6. https://doi.org/10.1093/toxsci/kfr088

Li, Y., Guo, B., Xu, Z., Li, B., Cai, T., Zhang, X., Yu, Y., Wang, H., Shi, J., & Zhu, W. (2016). Repositioning organohalogen drugs: A case study for identification of potent B-Raf V600E inhibitors via docking and bioassay. *Scientific Reports, 6*, 31074. https://doi.org/10.1038/srep31074

Torng, W., & Altman, R. B. (2019). Graph convolutional neural networks for predicting drug-target interactions. *Journal of Chemical Information and Modeling, 59*(10), 4131-4149. https://doi.org/10.1021/acs.jcim.9b00411

Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R. P., & Pande, V. (2017). Is multitask deep learning practical for pharma? *Journal of Chemical Information and Modeling,* *57*(8), 2068-2076. https://doi.org/10.1021/acs.jcim.7b00335

Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., & Siedlecki, P. (2018). Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics, 34*(21), 3666–3674. https://doi.org/10.1093/bioinformatics/bty374

Yan, J., Zhang, Z., Zhu, J., Zhang, K., Pei, J., & Liu, Q. (2024). DeltaDock: A unified framework for accurate, efficient, and physically reliable molecular docking. *arXiv:2410.11224v2 [q-bio.BM]*. https://doi.org/10.48550/arXiv.2410.11224

AlRawashdeh, S., & Barakat, K. H. (2023). Applications of molecular dynamics simulations in drug discovery. In *Methods in Molecular Biology* (Vol. 2714, pp. 127–141). Springer Nature. https://doi.org/10.1007/978-1-0716-2956-3_9

Liu, X., Shi, D., Zhou, S., Liu, H., Liu, H., & Yao, X. (2018). Molecular dynamics simulations and novel drug discovery. *Expert opinion on drug discovery*, *13*(1), 23–37. https://doi.org/10.1080/17460441.2018.1403419

Loschwitz, J., Olubiyi, O. O., Hub, J. S., Strodel, B., & Poojari, C. S. (2020). Computer simulations of protein–membrane systems. In *Progress in Molecular Biology and Translational Science* (Vol. 170, pp. 273–403). https://doi.org/10.1016/bs.pmbts.2020.01.001

Rudling, A., Orro, A., & Carlsson, J. (2018). Prediction of ordered water molecules in protein binding sites from molecular dynamics simulations: The impact of ligand binding on hydration

networks. *Journal of Chemical Information and Modeling, 58*(2). https://doi.org/10.1021/acs.jcim.7b00662

Badaya, A., & Sasidhar, Y. U. (2020). Inhibition of the activity of HIV-1 protease through antibody binding and mutations probed by molecular dynamics simulations. *Scientific Reports, 10*, Article 5501. https://doi.org/10.1038/s41598-020-62283-7

Kutzner, C., Grubmüller, H., de Groot, B. L., & Zachariae, U. (2011). Computational electrophysiology: The molecular dynamics of ion channel permeation and selectivity in atomistic detail. *Biophysical Journal, 101*(4), 809-817. https://doi.org/10.1016/j.bpj.2011.06.010

Martinotti, C., Ruiz-Perez, L., Deplazes, E., & Mancera, R. L. (2020). Molecular dynamics simulation of small molecules interacting with biological membranes. *ChemPhysChem, 21*(10), 1160–1173. https://doi.org/10.1002/cphc.202000219

Matamoros-Recio, A., Franco-Gonzalez, J. F., Forgione, R. E., Torres-Mozas, A., Silipo, A., & Martín-Santamaría, S. (2021). Understanding the Antibacterial Resistance: Computational Explorations in Bacterial Membranes. *ACS omega*, *6*(9), 6041–6054. https://doi.org/10.1021/acsomega.0c05590

Badaya, A., & Sasidhar, Y. U. (2020). Inhibition of the activity of HIV-1 protease through antibody binding and mutations probed by molecular dynamics simulations. *Scientific Reports, 10*, 5501. https://doi.org/10.1038/s41598-020-62352-7

Beer, M., Oliveira, A. S. F., Tooke, C. L., Hinchliffe, P., Li, A. T. Y., Balega, B., Spencer, J., & Mulholland, A. J. (2024). Dynamical responses predict a distal site that modulates activity in an antibiotic resistance enzyme. *Chemical Science, 15*(41), 17232-17244. https://doi.org/10.1039/d4sc03295k

Latallo, M. J., Cortina, G. A., Faham, S., Nakamoto, R. K., & Kasson, P. M. (2017). Predicting allosteric mutants that increase activity of a major antibiotic resistance enzyme. *Chemical Science, 8*(9), 6484-6492. https://doi.org/10.1039/c7sc02676e

Róg, T., Girych, M., & Bunker, A. (2021). Mechanistic Understanding from Molecular Dynamics in Pharmaceutical Research 2: Lipid Membrane in Drug Design. *Pharmaceuticals*, *14*(10), 1062. https://doi.org/10.3390/ph14101062

Zapata-Morin, P. A., Sierra-Valdez, F. J., & Ruiz-Suárez, J. C. (2014). The interaction of local anesthetics with lipid membranes. *Journal of molecular graphics & modelling*, *53*, 200–205. https://doi.org/10.1016/j.jmgm.2014.08.001

Saeedi, M., Lyubartsev, A. P., & Jalili, S. (2017). Anesthetics mechanism on a DMPC lipid membrane model: Insights from molecular dynamics simulations. *Biophysical Chemistry, 226*, 1-13. https://doi.org/10.1016/j.bpc.2017.03.006

Evoli, S., Mobley, D. L., Guzzi, R., & Rizzuti, B. (2016). Multiple binding modes of ibuprofen in human serum albumin identified by absolute binding free energy calculations. *Physical chemistry chemical physics : PCCP*, *18*(47), 32358–32368. https://doi.org/10.1039/c6cp05680f

Mundada, A. B., Pradhan, P., Raju, R., Sujitha, Y. S., Kulkarni, P. A., & Mundada, P. A. (2025). Molecular dynamics in pharmaceutical nanotechnology: Simulating interactions and advancing applications. *Journal of Biomaterials Science, Polymer Edition*. https://doi.org/10.1080/09205063.2025.2450150

Salahshoori, I., Golriz, M., Nobre, M. A. L., Mahdavi, S., Malekshah, R. E., Javdani-Mallak, A., Jorabchi, M. N., Khonakdar, H. A., Wang, Q., Mohammadi, A. H., Mirnezami, S. M. S., & Kargaran, F. (2024). Simulation-based approaches for drug delivery

systems: Navigating advancements, opportunities, and challenges. *Journal of Molecular Liquids, 395*, 123888. https://doi.org/10.1016/j.molliq.2023.123888

Jiang, T., Sun, G., He, T., & Yang, P. (2018). Comparative analysis of bind energies and sites between simvastatin and atorvastatin into CYP3A4. *JACC: Journal of the American College of Cardiology, 72*(16_Supplement), C174–C175. https://doi.org/10.1016/S0735-1097(18)32209-9

Jiang, T.-C., Kong, N., Zhao, S.-S., & Yang, P. (2013). Comparative Analysis of Bind Energies and Sites Between Simvastatin and Atorvastatin into CYP3A4. *Asian Journal of Chemistry*, *25*(17), 9909–9912. https://doi.org/10.14233/ajchem.2013.15588

Denisov, I. G., Grinkova, Y. V., Camp, T., McLean, M. A., & Sligar, S. G. (2021). Midazolam as a Probe for Drug-Drug Interactions Mediated by CYP3A4: Homotropic Allosteric Mechanism of Site-Specific Hydroxylation. *Biochemistry*, *60*(21), 1670–1681. https://doi.org/10.1021/acs.biochem.1c00161

Janaszkiewicz, A., Tóth, Á., Faucher, Q., Arnion, H., Védrenne, N., Barin-Le Guellec, C., Marquet, P., & Di Meo, F. (2023). Substrate binding and lipid-mediated allostery in the human organic anion transporter 1 at the atomic-scale. *Biomedicine & Pharmacotherapy, 160*, 114342. https://doi.org/10.1016/j.biopha.2023.114342

Zhao, M., Wang, L., Zheng, L., Zhang, M., Qiu, C., Zhang, Y., Du, D., & Niu, B. (2017). 2D-QSAR and 3D-QSAR Analyses for EGFR Inhibitors. *BioMed research international*, *2017*, 4649191. https://doi.org/10.1155/2017/4649191

Asirvatham, S., Dhokchawle, B. V., & Tauro, S. J. (2019). Quantitative structure activity relationships studies of non-steroidal

anti-inflammatory drugs: A review. *Arabian Journal of Chemistry, 12*(8), 3948–3962. https://doi.org/10.1016/j.arabjc.2016.03.002

Jakhar, R., Khichi, A., Kumar, D., Dangi, M., & Chhillar, A. K. (2022). Discovery of novel inhibitors of bacterial DNA gyrase using a QSAR-based approach. *ACS Omega, 7*(36). https://doi.org/10.1021/acsomega.2c04272

Lind, A. P., & Anderson, P. C. (2019). Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PloS one*, *14*(7), e0219774. https://doi.org/10.1371/journal.pone.0219774

Falls, Z., Fine, J., Chopra, G., & Samudrala, R. (2022). Accurate prediction of inhibitor binding to HIV-1 protease using CANDOCK. *Frontiers in Chemistry, 9*, 775513. https://doi.org/10.3389/fchem.2021.775513

THE FUTURE OF COMPUTER-AIDED DRUG DESIGN (CADD)

ELIFSU PERSILIOGLU[1]

The field of Computer-Aided Drug Design (CADD) has evolved significantly over the past few decades, transforming from a computational supplement to an integral component of modern drug discovery. Driven by the convergence of biotechnology, artificial intelligence (AI), and computational sciences, CADD is now an essential tool in optimizing drug discovery pipelines, improving efficiency, and reducing failure rates in clinical trials. However, despite the progress made, current methodologies still face fundamental limitations in handling vast biomedical datasets, accurately modeling complex molecular interactions, and predicting patient-specific responses. As the industry moves toward data-driven and precision-driven drug development, the future of CADD will be defined by AI, quantum computing, high-performance computing (HPC), federated learning, and CRISPR-driven genomic editing—all of which promise to unlock unprecedented capabilities in drug design and development.

[1] School of Medicine, Bahcesehir University, Istanbul 34734, Turkey, ORCHID: 0009-0009-0971-816X

The first chapter of this book established the fundamental concepts of CADD, providing a historical perspective on molecular docking, molecular dynamics simulations, and QSAR (Quantitative Structure-Activity Relationship) models. These computational techniques have played a vital role in reducing the time and cost associated with early-stage drug discovery, allowing researchers to screen vast chemical libraries, predict molecular interactions, and optimize drug-like properties before entering experimental phases. Additionally, Chapter 1 highlighted the initial steps toward integrating AI and quantum computing into CADD, emphasizing how computational power and intelligent algorithms are beginning to reshape the field.

Building on these foundational principles, Chapter 2 focused on the real-world applications of CADD across different domains of drug discovery. By exploring AI-driven enhancements in molecular docking, MD simulations, and QSAR modeling, it showcased how machine learning and deep learning techniques are improving the accuracy, efficiency, and scalability of computational drug design. The discussion extended to case studies demonstrating the successful application of AI-based predictive models in identifying and optimizing drug candidates, further solidifying CADD's role in modern pharmaceutical research.

While these advancements have significantly improved the efficiency and reliability of drug discovery, several critical challenges remain. The sheer volume of biological, chemical, and clinical data generated today presents a data-processing bottleneck, requiring innovative approaches to extract meaningful insights. Furthermore, traditional computational models often struggle to accurately simulate complex molecular interactions, leading to gaps between computational predictions and experimental validation. Additionally, as drug development becomes more personalized, there is an urgent need to integrate multi-omics data, patient-specific

biomarkers, and real-world evidence into computational models. CADD's progression embodies a cumulative evolution, with each innovation building upon prior methodologies to further refine computational drug discovery.

This chapter explores the cutting-edge advancements and future prospects in Computer-Aided Drug Design (CADD), concentrating on two central domains: Target Prediction and Drug-Target Interaction and Optimization.

In the first section, the discussion begins with recent progress in Target Prediction, highlighting the transformative impacts of artificial intelligence (AI) and quantum computing. AI-driven methods integrate diverse biological datasets—genomics, proteomics, transcriptomics, and metabolomics—to uncover novel mechanisms of disease and therapeutic targets. Key developments in this area include:

- The integration of multi-omics data through advanced AI algorithms to identify and validate potential therapeutic targets.
- Utilization of knowledge graphs and network-based machine learning to systematically organize and interpret complex biological relationships, significantly enhancing predictive accuracy.
- Application of large language models (LLMs) and sophisticated literature mining techniques to rapidly extract valuable insights from extensive biomedical literature.
- Incorporation of CRISPR-based functional genomics into AI frameworks for high-throughput validation of gene-disease associations.

- Adoption of federated learning models to facilitate secure, collaborative data analysis across multiple institutions, overcoming data scarcity while ensuring privacy and compliance.
- Employing AI-driven structural biology and druggability assessments to evaluate therapeutic viability of targets.
- Exploration of quantum computing as a powerful tool to surpass traditional computational limitations in analyzing intricate biological systems, thereby broadening the scope of target discovery.

The second section delves into computational prediction of Drug–Target Interactions and Lead Optimization, showcasing how AI significantly enhances predictions and optimization strategies. This section identifies the limitations inherent in traditional computational approaches, such as accuracy issues and scalability constraints, and examines transformative advancements enabled by AI:

- Development of advanced AI-driven predictive models to accurately forecast drug-target interactions.
- Application of generative AI techniques for the creation of novel, optimized ligands.
- Improvements in molecular docking methods and binding affinity predictions using AI.
- Exploration of quantum computing and quantum-inspired machine learning algorithms to refine interaction predictions further.
- Breakthrough advancements in protein structure prediction technologies, exemplified by AlphaFold,

enhancing precision in structural assessments of molecular targets.

- Implementation of systems biology and network-centric approaches to provide holistic analyses of complex biological interactions.
- Integration with synthetic biology and patient-derived models to advance personalized and precision medicine strategies.

The chapter also critically addresses current challenges, including data quality, interpretability of AI models, ethical considerations, and regulatory compliance, emphasizing the necessity for responsible AI deployment.

Looking forward, the chapter underscores the importance of explainable and causative AI to ensure transparency and trustworthiness in drug discovery. It concludes by outlining the promising trajectory of CADD and its potential to fundamentally transform pharmaceutical research and therapeutic development.

1. AI-Driven Target Prediction in Drug Discovery

Artificial intelligence (AI) is increasingly transforming the early stages of drug discovery by revolutionizing how new therapeutic targets are identified from vast biological data. Target identification – finding a disease-associated biomolecule whose modulation could yield a therapeutic benefit – has traditionally been an expensive, labor-intensive, and high-risk endeavor. Many drug projects fail because the chosen target ultimately lacks efficacy or safety, underscoring the need for better initial target selection. In recent years, the explosion of omics data (genomics, transcriptomics, proteomics, etc.) and patient information has far outpaced human ability to interpret it. This deluge of complex data, often termed “big data,” now provides an opportunity for AI and machine learning

(ML) to uncover hidden patterns linking genes or proteins to diseases. By integrating heterogeneous datasets and learning from subtle multidimensional correlations, modern AI systems can propose novel disease targets that would likely be missed by conventional approaches. Indeed, AI-driven analyses have already identified previously unrecognized targets, and even AI-designed drugs against those targets are entering clinical trials.

Traditional target discovery relied on experimental techniques like biochemical assays, affinity probes, genetic knockdowns, and phenotypic screens to pinpoint disease-relevant proteins. While powerful, these methods are slow and often "failure-prone" when applied at scale. The rise of high-throughput genomics and systems biology in the past decade has introduced multi-omics strategies, which combine data from multiple biological layers (DNA, RNA, protein, metabolites, epigenome, etc.) to infer disease-associated genes. Multi-omics integration can generate richer hypotheses than single-data-source studies, but making sense of such complex data requires computational intelligence beyond manual analysis. AI has emerged as the enabling technology to handle this complexity. By training on large datasets, AI models can detect weak but consistent signals of disease association, distinguish true causal drivers from spurious correlations, and prioritize targets with a higher likelihood of success. Moreover, AI methods can integrate not only experimental data but also knowledge mined from literature and clinical records, providing a more comprehensive picture of each candidate target. The convergence of machine learning with biomedical data is thus reshaping target discovery into a more systematic, data-driven discipline, in contrast to the serendipitous or single-hypothesis approaches of the past. Researchers now speak of an emerging paradigm in which AI acts as a "target finder" – scouring databases, publications, and even real-world patient data to

propose the most promising intervention points for a given disease (Liu et al., 2024).

In this chapter, we survey the state-of-the-art computational methods for target prediction, with an emphasis on AI and machine learning techniques. We focus on how these methods leverage diverse data sources to identify new targets and how they complement traditional experimental discovery. Key areas include multi-omics data integration, knowledge graph-based inference, natural language processing for literature mining, federated learning for collaborative modeling, quantum computing approaches, and the integration of high-dimensional data such as 3D genomics. Throughout, we highlight recent advances from top-tier research (Q1 and Q2 journals in the last five years) to provide a technically rich, up-to-date perspective. We avoid detailed experimental protocols except where needed to illustrate how AI interfaces with lab techniques (for example, AI analysis of CRISPR screens). The content is intended for an expert audience in computational drug discovery, so we delve into the algorithms and data structures that underpin these AI systems. By the end, it should be clear how modern AI-driven target prediction is accelerating the discovery of novel therapeutic targets and ushering in a new era for computer-aided drug design.

AI-Powered Multi-Omics Integration for Target Discovery

One of the most transformative applications of AI in target discovery is the integration of multi-omics data – genomics, transcriptomics, proteomics, metabolomics, epigenomics, and beyond – to pinpoint disease drivers. Complex diseases often involve alterations at multiple molecular levels, and single-omic studies can miss the bigger picture. AI algorithms excel at synthesizing such high-dimensional datasets and finding convergent signals that implicate certain genes or proteins as causal players. By examining

a gene from many angles – DNA variants, RNA expression, protein activity, epigenetic marks, etc. – an AI model can determine if that gene consistently appears dysfunctional in the disease context, boosting confidence that it is a valid target.

Multi-omics AI platforms typically ingest data from large patient cohorts and public databases. For example, genome-wide association studies (GWAS) may reveal DNA variants linked to a disease, but interpreting which genes those variants affect is challenging. Transcriptomic data (e.g. gene expression from RNA sequencing) and proteomic data can provide functional context. An AI model can learn to connect the dots: if a particular gene harbors risky variants (from GWAS) and is abnormally expressed in patients (from transcriptomics) and its protein shows altered levels or activity (from proteomics), then that gene is likely involved in disease pathology. Studies have shown that such integrative approaches yield more robust target hypotheses than any single data source alone . For instance, in inflammatory bowel disease, a multi-omics analysis using AI was able to highlight the gene IL23R (interleukin-23 receptor) as a key driver by combining genetic association signals with gene expression profiles . IL23R's involvement was not obvious from genomics alone (amid hundreds of GWAS hits), but the multi-omics convergence singled it out, and indeed therapies targeting the IL-23 pathway have since proven effective (Pun et al., 2023).

Behind the scenes, various machine learning techniques enable multi-omics data fusion. Kernel-based methods and ensemble learning can combine different feature sets (variants, expression levels, etc.) while weighting their relative contributions. More recently, deep learning architectures with multiple input branches have been used, where each branch processes one omic modality and internal layers learn a joint representation (Bhat et al., 2022). This allows the model to capture interactions between omics – for

example, a specific DNA mutation's effect on RNA and protein. Such models can be trained to predict a phenotype (like disease vs. healthy) from all omics at once, thereby identifying which features (genes) across modalities are most predictive of the disease. Those features often make plausible targets. Another approach is unsupervised integration, where AI looks for clusters or networks of patients that consistently show multi-omic perturbations in certain pathways (Liu et al., 2024). If patients cluster based on a dysregulated gene network, the hubs of that network become attractive targets.

A concrete example of AI-driven multi-omics target discovery is the use of transcriptome-wide association studies (TWAS) augmented with 3D genomics. TWAS integrates GWAS (genetic variant) data with gene expression data to find genes likely affected by disease-associated variants. However, many regulatory variants influence genes over long genomic distances, sometimes on different chromosomes, via 3D DNA folding. Researchers recently enhanced TWAS with chromatin contact maps (3D genomics) and epigenomic marks to better link non-coding variants to their target genes . Using an AI model to weigh genomic proximity, chromatin loops, and expression correlation, Khunsriraksakul et al. (2022) identified novel target genes for diseases that were missed by 2D analysis . The 3D genome data provided missing links between GWAS hits and gene expression changes, yielding a more complete network of gene–disease relationships. This illustrates how high-dimensional data integration, guided by AI, can reveal key regulatory genes as drug targets that traditional analyses overlook.

Multi-omics AI is also proving valuable in cancer target discovery. Cancer is driven by genomic mutations, but the consequences of those mutations manifest in transcriptomic and proteomic changes. AI models have been developed to predict synthetic lethal gene pairs – combinations of genes where

simultaneous disruption is lethal to cancer cells – by integrating somatic mutation profiles with functional genomics and expression data (Feng et al., 2024). In one benchmarking study, dozens of machine learning methods (random forests, support vector machines, neural networks, etc.) were evaluated on their ability to predict synthetic lethal partners in cancer . The top performers were AI models that leveraged multi-omic features: they considered not only mutation co-occurrence patterns but also whether the two genes operate in compensatory pathways (inferred from expression and protein networks) . Such predictions suggest targeting gene B if gene A is mutated, and AI helps prioritize which gene pairs to experimentally validate. As another example, integrating CRISPR screen data (which genes are essential to cancer cells) with transcriptomes of tumors can highlight context-specific essential genes – potential cancer targets that are critical only in tumors with certain expression signatures (Bhat et al., 2022).

It is important to note that simply throwing data into an algorithm is not a magic bullet. Careful curation and normalization of each omics dataset are needed so that the AI model isn't misled by technical biases. Issues like batch effects, differences in sample sources, and noise must be addressed, often by preprocessing or by the model architecture itself. Some advanced AI frameworks use autoencoders or other dimensionality reduction techniques to compress each omic dataset into a latent representation, filtering out noise, and then align these representations across modalities. The result is a unified view of the biological system where each sample (patient) is represented by a compact vector capturing multi-omic state. Clustering or classification on these vectors can then identify which genes are consistently influential. This strategy was employed in an AI platform called PandaOmics, which integrates multiple omics and human-curated databases to rank disease targets (Kamya et al., 2024). PandaOmics uses a combination of statistical scoring

and ML to evaluate thousands of genes against multiple evidentiary criteria (genetic associations, expression changes, druggability, etc.), outputting a prioritized target list for a given disease (PandaOmics: An AI-Driven Platform for Therapeutic Target and Biomarker Discovery | Journal of Chemical Information and Modeling) (PandaOmics: An AI-Driven Platform for Therapeutic Target and Biomarker Discovery | Journal of Chemical Information and Modeling). Notably, it also incorporates text-mined knowledge (discussed later) and expert feedback, exemplifying how multi-omics AI platforms are becoming comprehensive decision-support systems for target discovery.

In summary, AI-enabled multi-omics analysis provides a powerful, unbiased approach to target identification. By examining the disease from multiple molecular angles simultaneously, these methods increase confidence in targets that show consistent perturbations. Multi-omics AI has yielded successes like discovering new immune regulators in inflammatory diseases and new vulnerabilities in cancer subtypes (Pun et al., 2023). As data grows, these models continue to refine their predictions, especially when coupled with experimental feedback. Multi-omics integration has thus moved from a data challenge to a critical asset in target discovery, with AI as the key that unlocks its potential.

Knowledge Graphs and Network Learning for Target Prediction

Complementing multi-omics data, another major AI approach to target prediction involves knowledge graphs – comprehensive networks that map relationships between biomedical entities (genes, diseases, drugs, pathways, etc.). In a knowledge graph, nodes represent entities and edges represent known relationships (e.g. a gene is involved in a pathway, a drug treats a disease, a protein interacts with another protein). The premise is that

new therapeutic targets can be discovered by analyzing the topology of this biomedical network. AI algorithms, particularly graph neural networks (GNNs) and link prediction models, can learn from the structure of these graphs to suggest new connections that imply a gene is associated with a disease (Liu et al., 2024).

A prototypical example is the recently developed Progeni platform, which built a probabilistic knowledge graph (prob-KG) for target identification . In Progeni's graph, nodes included diseases, known targets, "unknown" candidate genes, drugs, and side effects, all integrated from multiple databases (DrugBank, DisGeNET, etc.) (Liu et al., 2024). Each edge was annotated with a probability score reflecting the strength of association between two nodes, often derived from literature co-occurrence or database evidence. The task of target identification was then framed as a link prediction problem on this graph: essentially, find which disease–gene pairs (edges between disease nodes and "unknown target" nodes) are likely missing links that should exist . By learning the patterns of existing connections, the AI model predicts new connections that have not been recorded yet but are plausible given the network structure.

In training, Progeni's GNN-based model "learns" the network by iteratively trying to reconstruct the known edges. It uses multiple specialized GNNs, one per relation type, so that it processes e.g. drug–disease links separately from protein–protein interaction links . The node embedding produced by the GNN captures a summary of everything connected to that node. If two nodes (say a disease and a gene) have embeddings that can be projected to match an edge with high probability, the model is essentially saying "this disease and this gene are likely linked." By optimizing the model to minimize the error in reconstructing known links (with higher weight on high-probability known links), the model becomes proficient at emphasizing biologically meaningful connections and downplaying random ones. After training, it then outputs a list of

disease–gene pairs ranked by predicted probability of association. Intriguingly, Progeni's top predictions for diseases like melanoma and colorectal cancer included genes not previously seen as targets. Subsequent laboratory experiments confirmed several of these predictions, validating that the graph-based AI had indeed discovered novel, biologically relevant targets . This success demonstrates the power of combining diverse data sources into a unified graph and applying AI to navigate it.

Knowledge graph approaches benefit from encoding a wide breadth of knowledge: genetic interactions, metabolic pathways, clinical observations, even text-mined relationships can all be thrown into the graph. For example, a system might include an edge between a gene and a side effect if mutations in that gene cause a similar symptom, or between a drug and a protein if the drug is known to bind that protein. By leveraging such connections, AI can sometimes make imaginative leaps that humans might miss. An often-cited use case is repurposing: a graph model might notice that a drug for Disease X targets a protein that is closely connected in the network to a protein implicated in Disease Y, suggesting the drug could modulate Disease Y as well by network proximity. In the target discovery realm, if a protein has a similar neighborhood in the graph as known targets of a disease (e.g., it interacts with many of the same partners), the AI might flag it as a potential target too.

Graph neural networks are especially well-suited to these problems because they can handle the graph's complexity natively. A GNN will take each node, look at its neighbors and edges, and update the node's representation. Stacking multiple GNN layers means the representation of a node can incorporate information from far-reaching parts of the network (the neighbors of neighbors, etc.). For target prediction, this means, for example, a disease node's embedding encodes not just the few genes directly linked to it, but also related diseases, drugs that treat it, pathways it involves, etc.

Likewise a gene's embedding encodes its interactome and known phenotypic effects. If a gene embedding and disease embedding end up similar in the model's latent space, it suggests a possible association. Techniques like GraphSAGE, R-GCN (relational GNN), and knowledge graph embeddings (TransE, DistMult) have all been applied to biomedical graphs to score candidate links.

A notable example of knowledge-graph-driven discovery is a probabilistic graph approach applied to identify cancer drivers. In a 2023 study, researchers constructed a heterogeneous network including gene–gene interactions, gene–disease links, drug–target links, and tissue-specific information (Liu et al., 2024). By running link prediction, they successfully pinpointed a gene (CDK20) as a novel hepatocellular carcinoma target, which was experimentally validated (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences). Interestingly, the model had learned from the network that CDK20 was functionally analogous to other known cancer drivers and was connected to several cancer pathways, even though CDK20 itself had limited prior evidence in liver cancer. This "guilt-by-association" logic, formalized through AI, is a powerful motif of knowledge graph analysis.

Another emerging approach is to incorporate literature knowledge into graphs. Systems like the one by Zhang et al. (2022) use natural language processing to extract mentions of gene–disease associations from millions of papers, adding edges for those mentions in a knowledge graph with a weight corresponding to mention frequency or sentiment. An AI model can then integrate this literature-driven subgraph with experimental data-driven subgraphs. In fact, the Progeni prob-KG included a literature co-occurrence score for every edge . This helps capture information like "Gene A is often mentioned alongside Disease B in publications," which might indicate a causal relationship worth exploring. Knowledge

graphs can thus serve as a scaffold to combine empirical data with collective scientific knowledge.

While knowledge graph AI has great promise, it also faces challenges. A major issue is incompleteness – the graph can only include known entities and relationships. If a disease's true causal gene is entirely unstudied, it may be absent from the graph or too isolated, making discovery difficult. There is also bias toward well-studied genes (they will have many edges and thus be easier for algorithms to rediscover). Efforts like open-source project OpenBioLink are trying to systematically compile balanced biomedical graphs for benchmarking algorithms. Moreover, interpreting why an AI predicted a certain link can be non-trivial; researchers often must trace through the graph to find the supporting paths or node similarities that led to a prediction.

Nonetheless, knowledge graph modeling has become a staple in AI-driven target discovery. Pharmaceutical companies have begun building enterprise knowledge graphs that integrate their proprietary data (omics, screening results, clinical trial data) with public knowledge, and then applying GNNs to prioritize targets. These methods scale well with data – as new studies add edges or nodes, the graph only becomes richer. In the future, combining knowledge graphs with causal inference techniques might even enable automated hypothesis generation, where an AI not only proposes a target but also suggests the biological mechanism connecting target and disease (in terms of network pathways). This could greatly aid researchers in evaluating and triaging AI-proposed targets.

Language Models and Literature Mining for Target Identification

A vast amount of biomedical knowledge relevant to target discovery is embedded in text: journal articles, patents, clinical trial reports, and databases. Artificial intelligence in the form of natural

language processing (NLP) and large language models (LLMs) is increasingly used to unlock this "dark data" and support target prediction. Modern language models can read and comprehend millions of publications to find latent connections between genes and diseases that might not be obvious from raw data alone (Lee et al., 2020). By mining the literature, AI can identify hints that a particular protein is involved in a disease – for example, if many papers independently report that the protein's levels are altered in patients, or that it interacts with known disease factors.

Early approaches to text mining for target discovery relied on keyword searches and manual curation. Those methods could retrieve direct statements (e.g. "Gene X is associated with Disease Y") but struggled with nuance and context. Today's transformer-based language models (like BERT, GPT, etc.) understand biomedical text with much greater sophistication. Models such as BioBERT (a BERT model pre-trained on biomedical corpora) have been fine-tuned to extract specific relationships from text (Lee et al., 2020). For instance, given a sentence from a paper, BioBERT can classify whether it describes a gene–disease association, a drug–target interaction, or no relationship. By applying such models across the entire PubMed database, one can build a massive knowledge repository of potential target relationships annotated with literature evidence. Lee et al. (2020) demonstrated that BioBERT significantly outperformed earlier text-mining tools in recognizing gene–disease and gene–drug interactions, due to its deeper contextual understanding (Lee et al., 2020).

Beyond extracting explicit relationships, large language models can perform literature-based discovery, which is more exploratory. These models might be asked: "What do we know about protein X in the context of Alzheimer's disease?" and generate a summary that highlights relevant findings (Pun et al., 2023). They might note, for example, that "Protein X is a stress-response enzyme

found upregulated in Alzheimer's patient brains and has been shown to interact with tau protein," which could suggest it as a target (with supporting citations). A notable advancement is the use of LLMs to identify implicit links. For example, separate papers might report that gene A is linked to process P, and process P is linked to disease B, but no paper directly links gene A to disease B. An NLP system can synthesize these and suggest gene A as a novel disease B candidate. This paradigm was famously used in discovering a connection between magnesium and migraines decades ago using literature indices; today's AI can do this at scale across thousands of topics.

Recently, specialized large language models for biomedicine have emerged, such as BioGPT (Luo et al., 2022) and PubMedBERT, which are trained on millions of biomedical abstracts and full-text articles. These models not only retrieve facts but can also reason and generate hypotheses. For target discovery, a model like BioGPT can be prompted with, say, "Gene A and Disease B" and asked to produce any known connections. If trained well, it might output something like: "Gene A encodes a kinase that regulates immune response; aberrant activation of Gene A has been observed in patients with Disease B ." This helps researchers quickly gather evidence. Notably, BioGPT was reported to achieve high performance on question-answering benchmarks about drug–gene interactions, showing its potential in querying for target relationships (Pun et al., 2023).

A cutting-edge development is the integration of LLMs with dedicated target discovery platforms. ChatGPT-like models have been tailored for biomedical use – for example, Insilico Medicine's ChatPM (nicknamed "ChatPandaGPT") was integrated into their PandaOmics platform to assist in target rationalization. After PandaOmics' multi-omics AI prioritizes a gene, researchers can query the integrated LLM: "Why might this gene be relevant to the

disease?" The LLM will generate an explanation drawing on known biology: e.g. "This gene encodes a cytokine receptor subunit involved in inflammatory signaling; multiple studies report its pathway is overactive in patients, supporting it as a potential target" (Ozerov et al., 2024). Impressively, the system can also provide references for its claims . By doing so, the LLM adds a layer of interpretability and validation to AI-generated target lists – scientists get not just a ranked gene name, but a narrative of evidence.

Another use of NLP is to mine clinical texts and real-world data for target clues. Electronic health records (EHRs) and pathology reports contain phenotypic descriptions and co-morbidities that can highlight disease mechanisms. For example, an NLP analysis of EHR notes might reveal that patients with Disease X often have high levels of Biomarker Y (mentioned in lab test sections), suggesting Y's pathway is involved. If Biomarker Y is regulated by Gene Z, Gene Z could be a target. Privacy-preserving NLP on anonymized patient data has already identified such correlations that were later confirmed biologically (Liu & Panagiotakos, 2022). Furthermore, text mining of patents and trial reports can uncover targets that companies are investigating but haven't published in journals, providing intelligence on novel target ideas in the industry.

It's worth noting that language models can suffer from hallucinations or errors, especially if asked open-ended questions. They might incorrectly connect a gene to a disease if the training data had confounding information. To mitigate this, LLM-based systems for target discovery often operate in a constrained setting: either extracting facts (where the model's outputs are grounded in specific source texts) or requiring the model to provide citations for every claim. For instance, an AI might output: "Gene X is associated with neurodegeneration ," pointing to a specific source, rather than just stating it without proof. This ensures that any hypothesis generated can be traced back to supporting evidence.

In summary, AI language models serve as powerful research assistants sifting through the world's biomedical literature to support target discovery. They can rapidly produce summaries of known information on a candidate target, highlight connections across studies, and even suggest new targets based on indirect evidence aggregation. In practical use, language model outputs are being combined with data-driven evidence. An AI pipeline might first use a multi-omics algorithm to propose a list of candidate genes, and then use NLP to prioritize among them by checking what's known: a gene with rich literature backing (e.g. multiple papers hinting it's important in the disease) might be ranked higher than one with no information (or alternatively, a completely novel gene might be more risky but potentially more groundbreaking). The synergy of data mining and text mining AI thus helps ensure that promising targets with existing support are recognized, while also flagging intriguing out-of-the-box targets.

The rapid progress in NLP suggests that future target discovery could involve conversational AI systems where researchers literally "ask" an AI about a disease and get target suggestions with rationale. Foundation models with billions of parameters (like GPT-4) are already being tested on biomedical reasoning tasks. As these models become integrated with structured databases and ontologies, we can expect them to perform sophisticated reasoning – for example, suggesting a target because "it lies upstream of two other proteins known to drive pathology," even if that specific chain of reasoning was never published explicitly. Such capabilities will make AI an even more invaluable partner in the ideation phase of drug discovery, scouring both data and knowledge to propose the next therapeutic breakthroughs.

AI in Functional Genomics: CRISPR Screens and Beyond

While AI can generate many hypotheses, experimental validation is the ultimate test for potential targets. In recent years, functional genomics techniques like CRISPR-Cas9 screening have provided powerful experimental methods to probe gene function on a large scale. These methods systematically knock out or modify genes in cell models to see which genes affect disease-relevant phenotypes (such as cell survival, proliferation, or specific biomarkers). However, interpreting the results of large-scale screens and prioritizing hits as drug targets is a complex task – one increasingly aided by AI (Bhat et al., 2022).

A typical genome-wide CRISPR knockout screen might target 18,000+ genes in parallel to find which ones a cancer cell needs to survive. The raw output is a list of genes that, when knocked out, cause the cells to die or grow slower (so-called "negatively selected" genes), as well as genes that cause faster growth or drug resistance when knocked out ("positively selected"). This list can be long and context-specific. AI comes into play by analyzing the list in the broader biological context: identifying pathways that are enriched, interactions between hits, and cross-referencing multi-omics data to distinguish the true driver targets from secondary effects.

In the analysis stage, machine learning can perform tasks such as hit scoring, noise filtering, and multi-screen integration. Bayesian models have been used to differentiate true signals from experimental noise in CRISPR screens, accounting for variable sgRNA cutting efficiencies and random dropout events (Bhat et al., 2022). These models output a probability that each gene is a real hit. Additionally, AI can integrate results from multiple screens – for instance, the same cell line treated with different drugs – to find hits that consistently appear, indicating a robust target. Unsupervised

clustering of screen results can also identify genetic interactions: if knocking out gene A alone has little effect, but knocking out A and B together is lethal, AI might detect that pattern across conditions (a form of synthetic lethality detection).

Crucially, AI helps in functional interpretation of screen hits. A list of 200 essential genes from a CRISPR screen is not immediately actionable. AI tools map those genes onto known pathways and protein–protein interaction networks (often using knowledge graphs as described earlier). If the hits significantly overlap a certain pathway, that pathway becomes a prime target (the individual genes in it might all point to the same drug target). AI-based network analysis can identify "communities" of hits within the larger interactome . For example, a set of co-essential genes might all be part of the spliceosome complex – indicating the spliceosome is a vulnerability. Perhaps a drug can target a core spliceosome component. Conversely, if an AI finds that the top hits are scattered but their only common link is gene X (they all interact with X), gene X could be the master regulator causing the phenotype.

Another powerful use of AI in functional genomics is predicting synthetic lethal partners and genetic interactions from partial data. As mentioned, Feng et al. (2024) benchmarked ML methods for this purpose. One approach is to train a classifier on known synthetic lethal pairs (from previous screens or databases like DepMap) using features like co-expression, pathway overlap, and mutual exclusivity of mutations in tumors. The classifier then predicts new gene pairs that likely have a lethal interaction (Feng et al., 2024). These predictions guide combinatorial screens or suggest combination therapies. For instance, if AI predicts that gene A and gene B are synthetic lethal in cancer, and gene A is frequently mutated in tumors, then inhibiting gene B might selectively kill those tumor cells – making gene B a promising target for patients with gene A mutations.

AI is also closing the loop by guiding experiments in real-time. Some groups have developed active learning frameworks where the AI analyzes interim results from a functional screen and decides which perturbations to test next. For example, in an iterative CRISPR screening approach, an AI might notice partway that knocking out a certain family of genes yields strong effects, and then advise testing additional genes in that family or related pathways in the next round. This kind of adaptive experiment, sometimes called an "AI-guided laboratory," was demonstrated in a different context of enzyme engineering . Researchers created an autonomous lab where an AI model chose which genetic mutations to introduce in an enzyme to improve its activity, the robotic system made those mutations and tested them, and the data was fed back to the AI to choose the next mutations (ending up with dramatically improved enzyme function in a short time) . One can imagine a similar setup for target discovery: an AI directs a CRISPR or compound screen towards the most informative experiments. AstraZeneca's prototype iLab is one such effort, where AI "scientists" control lab automation to optimize experiments in drug discovery.

Incorporating functional genomics data also helps validate AI predictions from other sources. Suppose a knowledge graph algorithm suggests Gene X is a disease target. If CRISPR screening in a disease model shows that knocking out Gene X indeed alters the disease phenotype (e.g. kills disease cells or reverses a pathogenic process), that is strong supporting evidence. AI can assist in this validation by quantifying how "on-target" the effects are – e.g., checking that knockout of X affects disease outcomes but does not severely affect normal cells (perhaps using transcriptomic readouts to ensure specificity). Such analyses might involve training a classifier to distinguish on-target vs. off-target effects of perturbations, or to predict cell viability outcomes from gene expression changes.

Finally, AI contributes to optimizing the experimental tools themselves. In CRISPR screens, not all sgRNA guides are equal; some cut efficiently, others don't, and some have off-target effects. Machine learning models (often deep learning) have been trained on large datasets of CRISPR guide experiments to predict guide efficiency and specificity (Bhat et al., 2022). By using these models to design better sgRNA libraries (picking guides that are predicted to be highly efficient and low off-target), researchers obtain cleaner screening data. This creates a virtuous cycle: AI improves the experiment, the experiment yields better data, which feeds back into AI models that interpret the results and propose targets. For instance, the deep learning model DeepCRISPR can predict on-target knockout potency and was used to filter out poor guides, resulting in more reliable identification of essential genes in a screen. In summary, AI ensures that the outputs of functional screens are as informative as possible for target discovery.

In conclusion, AI and functional genomics now go hand in hand. Large-scale perturbation screens provide causal evidence of gene function, and AI provides the analytical muscle to extract targets from those screens. This synergy has led to discoveries of new cancer vulnerabilities, identification of genes that cause drug resistance (pointing to combination therapy targets), and overall more confidence in AI-suggested targets (since they've been functionally tested). As screening technologies expand (CRISPRa/CRISPRi for activation/repression, base editing, pooled drug synergy screens), AI's role will only grow in designing, analyzing, and learning from these experiments to pinpoint the best targets for therapeutic development.

Federated Learning and Collaborative Data Models

Discovering drug targets often requires learning from large datasets that are distributed across different institutions – for

example, patient genomic data from multiple hospitals or assay results held by different pharmaceutical companies. Federated learning (FL) is an emerging AI approach that enables collaborative modeling on such distributed data, without the data ever leaving its source site (Rieke et al., 2020). This is particularly important in biomedical research, where patient privacy and proprietary data concerns traditionally create "data silos." Federated learning allows AI models to be trained on a union of datasets to improve power and generalizability, which can directly benefit target prediction by incorporating diverse populations and experimental contexts.

In a federated learning setup, a common machine learning model (for instance, a neural network) is initialized and sent to each data-holding site (hospital or company). Each site trains the model locally on its own data (for example, genotype-phenotype datasets linking patient DNA to disease status) and computes only the model parameter updates, not exposing any raw data. The updates or gradients from all sites are then securely aggregated by a central server to yield an improved global model . This global model now has learned patterns from all sites combined, and is sent back out for another round of local training – iterating until convergence (Rieke et al., 2020). Throughout this process, no site ever shares individual records; only abstracted model parameters are exchanged, often with encryption or differential privacy to prevent reconstruction of the data. The end result is a machine learning model that has essentially been trained on the union of all datasets, without those datasets ever being pooled in one place.

Federated learning can be applied to target discovery in several ways. One example is federated genomic analysis: suppose multiple research centers have patient genomic data with disease phenotypes. A federated model can be trained to predict disease risk from genomic features across all centers. The internal weights of that model can highlight genomic regions or genes that are important

predictors – effectively identifying candidate disease-related genes (potential targets) in a way that leverages the full sample size without compromising privacy. This was demonstrated in a proof-of-concept FL study predicting neurological disease risk from patient MRI and genetic data spread across institutions, which achieved similar accuracy to a pooled-data model (Rieke et al., 2020).

A real-world landmark of federated learning in drug discovery was the MELLODDY consortium, a collaboration of ten major pharma companies who jointly trained AI models on each other's chemical-screening data without sharing the data directly. Each company had large datasets of compounds tested in various biological assays (target binding, toxicity, etc.). Using FL, they trained a multitask neural network on a combined >20 million compounds and ~40,000 assays across companies . The resulting global model – essentially an AI that had "seen" the collective experience of the industry – could predict drug–target activities and off-target effects much better than models any single company could train alone (Heyndrickx et al., 2024). Importantly, this was done without any company divulging its proprietary compounds or assay results to others . The MELLODDY model yielded stronger predictive performance for target activity (QSAR models) across the board, meaning it could, for example, predict if a new molecule will hit a particular target with higher accuracy by drawing on subtle patterns learned from billions of data points spread across companies (Heyndrickx et al., 2024). This illustrates how federated AI can enhance target discovery – in this case by improving virtual screening and target activity prediction – through cross-organization collaboration.

Another scenario is federated learning across hospitals for genomic and clinical data integration. Imagine multiple hospitals each have cohorts of patients with rare diseases, including genome sequences and clinical phenotypes. Individually, each cohort might

be too small to confidently pinpoint genetic targets. FL can train a joint model to predict disease severity or outcomes from genome variants across all hospitals' data. That model might identify, say, that mutations in a certain gene consistently predict severe disease across populations – highlighting that gene as a key disease driver (and thus a therapeutic target). Because the data never left the hospitals, issues of patient consent and data locality (e.g. regulations like GDPR) are mitigated, allowing inclusion of data from regions that otherwise couldn't share. Rieke et al. (2020) note that federated models in healthcare achieved performance close to centrally trained models in tasks like medical image analysis, underscoring its viability.

A significant advantage of federated learning for target discovery is diversity. Models trained on a single data source can be biased or not generalize well. By training on multi-center data, an FL model learns patterns that are consistent across different demographics, sample prep methods, etc. – arguably making the identified targets more robust. For example, a federated genomic study that includes data from Europe, Asia, and Africa will be able to find disease genes that are important across ancestries, not just in one lineage (Liu & Panagiotakos, 2022). This is crucial in avoiding the historical bias where many drug targets were discovered in European-ancestry cohorts and sometimes failed to translate universally. FL opens the door to more inclusive discovery efforts.

Federated learning does introduce technical challenges. One is that data at different sites may be heterogeneous (batch effects, different distributions). Models like neural networks can sometimes struggle if one site's data dominates the gradients. Advanced FL algorithms include techniques to balance contributions from sites and mitigate site-specific biases . For instance, methods exist to scale gradients from smaller sites so they aren't drowned out by larger sites, or to do a few personalized training steps for each site

after the global model is learned (to adapt to local quirks). There are also robust aggregation methods that resist a single bad actor site (in case data from one site is noisy or even maliciously modified). In MELLODDY, a secure aggregation protocol was used so that even the central server could not see individual company updates – it only saw an encrypted combined update. This way, no participant could infer another's proprietary info from the model. Such cryptographic safeguards are essential in sensitive collaborations.

From a target discovery perspective, one can envision federated networks focusing on specific data modalities: e.g., a federated network of proteomics labs pooling their phosphoproteomics datasets to identify consistently dysregulated kinases in a disease; or academic groups federating single-cell RNA-seq data to find cell-type-specific targets in a rare disease. By keeping data local, FL also allows use of datasets that otherwise couldn't be shared due to size (shipping terabytes of raw data) or bureaucracy. Instead of months negotiating data transfer agreements, collaborators can agree to train an FL model in weeks.

The outputs of federated learning relevant to target prediction might be an improved prediction model (like MELLODDY's QSAR model) that directly aids target selection or a set of discovered features (genes, pathways) that are globally predictive of disease. In either case, FL broadens the evidence base. For example, if a federated model identifies Gene X as a top predictor of patient survival in three independent cohorts, that strengthens the rationale to explore Gene X as a therapeutic target to improve survival.

It's worth noting that federated learning is still an emerging technology in biomedicine, and not a panacea. If data is too heterogeneous or if certain sites have small datasets, the combined model might not help much over local ones. Also, debugging and interpreting federated models can be trickier – if a model weights a

certain gene feature highly, is that driven by all sites or just one? Researchers have developed methods to examine contribution per site (e.g., leave-one-site-out experiments) to address this.

In summary, federated learning extends AI-driven target discovery into a collaborative, privacy-preserving domain. It enables consortiums to jointly train AI models on previously siloed data, yielding insights that no party could obtain alone. By unlocking more data for analysis, FL can reveal new targets (especially for diseases where data is scarce per institute) and build more powerful predictive models for target validation and drug response. Given the success of projects like MELLODDY, it's likely we will see federated approaches applied to areas like in silico target validation (e.g., predicting clinical trial outcomes based on shared data from multiple pharma companies, which could highlight which targets are likely to succeed) and to global biobank analyses for target gene discovery. As the infrastructure and trust in federated systems grow, AI models will routinely learn from data across continents, accelerating the identification of drug targets that are truly globally relevant (Heyndrickx et al., 2024).

AI for Structural Biology and Druggability Assessment

Understanding the three-dimensional structure of proteins and their interactions is vital for target validation and subsequent drug design. AI has made remarkable strides in structural biology – most famously with DeepMind's AlphaFold2, which achieved near-experimental accuracy in predicting protein 3D structures from sequence (Jumper et al., 2021). While structure prediction might seem more relevant to drug design (e.g. docking small molecules), it also plays a role in target identification. Knowing a protein's structure can reveal whether and how it is druggable (i.e., has pockets for small molecules or epitopes for biologics) and can even

hint at function if the structure resembles known proteins (Jumper et al., 2021).

AlphaFold2's breakthrough of providing reliable models for most human proteins was a watershed moment. Overnight, researchers gained structural information on thousands of previously uncharacterized proteins. For target discovery, this means that when an AI nominates a novel gene, we can often immediately examine its predicted structure. If the structure reveals, for example, an active site pocket or a kinase domain fold, that increases interest – it suggests existing chemistry might modulate it. Alternatively, the structure might show that the protein is an intrinsically disordered protein (harder to drug with small molecules) or lacks any obvious binding pocket, which might deprioritize it or steer the strategy towards biologics (like therapeutic peptides or protein degraders). In this way, structural AI helps filter and triage targets by druggability.

Moreover, structure can inform function: AlphaFold has predicted structures for many proteins of unknown function, and by comparing those structures to known ones, scientists can infer what the protein might do . If a novel disease-associated protein is predicted to have, say, a kinase-like domain, one can hypothesize it's an enzyme involved in signaling, making it appealing as a target (kinases are a well-exploited class in drug development). Indeed, researchers reported cases where AlphaFold structures suggested biochemical functions that were later confirmed in the lab . Thus, AI-based structure prediction turns sequence data (like a gene identified in a GWAS) into actionable knowledge (e.g. "this gene's product is a membrane receptor with a ligand-binding pocket"), aiding target discovery.

Beyond static structures, AI is now being applied to model protein interactions and dynamics. Protein-protein interaction prediction with AI can identify if a potential target interacts with

known disease proteins. For example, if AI predicts that a candidate target protein binds to the transcription factor that drives pathological gene expression, that provides a mechanistic link. Structural models of complexes (like an antibody bound to a receptor) can guide therapeutic development directly. The integration of structural AI into target ID is exemplified by efforts to map the human protein interactome with predicted structures, discovering new interactions that could be disrupted or mimicked by drugs.

Another frontier is AI-driven generative design of proteins and biologics to modulate targets. If a target is deemed "undruggable" by small molecules (perhaps it has no pockets or is not an enzyme), one approach is to design a therapeutic protein that can bind and inhibit or modulate it. Here, AI generative models come into play. In 2023, an AI method called RFdiffusion (a diffusion-based generative model for protein structures) was introduced that can design novel protein shapes with specified functions (Watson et al., 2023). In one demonstration, RFdiffusion was used to create de novo proteins that bind tightly to the human insulin receptor and activate it . The AI was given the task of designing a protein that fits into the receptor's binding site, and it hallucinated entirely new protein sequences that fold into the desired shape. These AI-designed proteins acted as receptor agonists – essentially artificial mimics of insulin – and had sub-nanomolar binding affinity . This example is striking: it shows AI can invent a biologic drug from scratch to engage a target (in this case, a well-known target, the insulin receptor). For novel targets, generative AI could similarly propose therapeutic modalities. For instance, if an AI finds target protein X is implicated in a disease but small molecules cannot easily inhibit X, an AI like RFdiffusion might design a protein that binds X and blocks its function, giving a starting point for a biologic therapy (Watson et al., 2023).

Generative AI is not limited to proteins – it also creates new antibodies, peptides, and even gene therapies (by designing guide RNAs or gene editors). The common theme is that once a target is identified, AI can rapidly generate candidate therapeutic molecules tailored to that target, shortening the drug design cycle. While this strays into drug design from target discovery, the linkage is important: it means "difficult" targets (like protein-protein interactions) are less likely to be discarded now, because AI may find a way to drug them via novel modalities. This feedback loop encourages casting a wider net in target discovery – including those that earlier might have been labeled undruggable.

On the flip side, structural AI helps avoid dead-ends by revealing truly intractable targets. For example, if a target is an ultra large scaffolding protein with no pockets and primarily intracellular (not accessible to biologics), one might deprioritize it. Or structural analysis might show a target is nearly identical to another essential protein (raising specificity issues), steering attention to alternatives.

In silico structural analysis has also been integrated with virtual screening early in target selection. If multiple candidate targets emerge from an AI analysis, researchers can quickly perform virtual screening for each: dock a library of drug-like compounds to the AlphaFold model of each target to gauge druggability. AI is used here too – modern docking can be aided by ML scoring functions that better predict binding. If one target yields many high-scoring hits in silico and another yields none, that could influence which target to pursue first.

Furthermore, molecular dynamics simulations enhanced by AI (e.g. using deep learning potentials) can examine the conformational flexibility of a target and identify transient pockets or allosteric sites. A target might appear "smooth" and binding-pocket-free in a static structure, but simulations reveal that it

occasionally opens a pocket that a drug could exploit. AI models like DeepDriveMD are accelerating such simulations. This level of insight helps in designing allosteric inhibitors or stabilizers for targets that were previously thought unligandable.

In summary, AI in structural biology provides critical support to the target discovery pipeline by evaluating target druggability, suggesting functions, and even designing molecules to probe or modulate the target. Tools like AlphaFold2 have essentially solved one piece of the puzzle (protein structure prediction), enabling downstream AI and computational chemistry to flourish. Combined with experimental structural techniques (cryo-EM, X-ray crystallography), which AI also aids in analysis, the structural lens on potential targets is sharper than ever. As a result, target discovery no longer treats structure and function as afterthoughts; they are integral from the beginning, with AI ensuring that for any given target hypothesis, we rapidly know what the target likely looks like, how it might be engaged, and whether it is worth the effort. This tight coupling of AI-driven target identification and AI-driven target validation/design will be a hallmark of in silico drug discovery in the coming years.

Quantum Computing for Drug Target Discovery

Quantum computing, though still nascent, offers the promise of handling certain computational problems exponentially faster than classical computers – a potential boon for the complex optimization and simulation tasks in drug discovery. While practical quantum algorithms are in early stages, researchers have started exploring how quantum computing and quantum machine learning could impact target prediction and related areas (Cao et al., 2019). One of the most direct contributions of quantum computing would be in quantum chemistry simulations: accurately calculating molecular energies and interactions. This can enhance our

understanding of how a small molecule (or potential drug) binds to a protein target, and could even allow simulation of protein dynamics and ligand effects at a scale impossible for classical computers (Cao et al., 2019).

For target discovery specifically, quantum computing might not directly “find a target” out of data, but it can greatly aid in evaluating and prioritizing targets by improving the modeling of molecular systems. For instance, once a candidate target is identified, a quantum algorithm could rapidly evaluate how druggable it is by simulating a panel of drug-like fragments binding to it with high accuracy. Current classical methods use approximations (force fields, etc.) which quantum computing could overcome by solving the Schrödinger equation for the system more exactly (Cao et al., 2019). If target A shows many stable binding configurations with various fragments in quantum simulations, whereas target B shows none, one might favor A as more druggable.

Another area is quantum-assisted machine learning on high-dimensional biological data. Quantum machine learning algorithms (e.g., quantum kernel machines, variational quantum circuits) have been proposed to potentially detect patterns in data that classical algorithms might miss when the data is highly entangled or complex. In target discovery, an example could be a quantum kernel method that maps multi-omics patient data to a high-dimensional quantum Hilbert space and finds subtle separations corresponding to certain gene influences on disease. If such a method pinpointed gene X as a separator of disease vs. healthy in that space, it suggests gene X’s combination of effects is unique to disease, nominating it as a target. While this is theoretical at this stage, researchers are experimenting with small quantum models on genomics datasets to see if any advantage emerges.

Quantum computing could also enhance optimization problems that appear in drug discovery. One such problem is selecting an optimal subset of targets or genetic features that explain a disease – essentially feature selection, which can be combinatorially hard. Certain quantum algorithms (like QUBO solvers or Grover's search) might solve such combinatorial selection faster. Imagine trying to choose 5 key genes out of 1000 that, together, cover the disease pathways. A quantum annealer could, in principle, search this space more efficiently than brute force, potentially identifying sets of genes (targets) that maximize a certain objective (e.g., predictive power for disease phenotype). There has been a demonstration of using a D-Wave quantum annealer to do feature selection in a cancer dataset, which hinted at some speedups (Srivastava, 2023). As quantum hardware improves, these capabilities could become more practical.

Furthermore, quantum computing could simulate biophysical processes like protein folding or ligand binding in ways classical simulations struggle to. If one could simulate how a thousand different proteins misfold or aggregate, one might find common patterns or key intervention points (targets) for diseases like Alzheimer's where misfolding is central. Quantum simulations of small peptides have already been done, and scaling this up is an active area.

It is important, however, to temper expectations: current quantum computers are noisy and have limited qubits, so their use in target discovery today is mostly exploratory. That said, many pharmaceutical companies have started collaborations to be "quantum-ready." For example, some have used quantum algorithms to compute binding affinities of very small drug-target models as a proof of concept. These pilot studies haven't yet yielded a new target, but they are building the foundation.

One nearer-term quantum-inspired approach is using quantum-inspired algorithms on classical hardware, which mimic some quantum strategies. For instance, quantum-inspired tensor networks have been applied to genomics data analysis for dimensionality reduction, which could help in target discovery by finding better low-dimensional representations of patient data that highlight differentially activated pathways.

In summary, quantum computing holds potential to supercharge the computational analyses underlying drug target discovery by tackling the quantum mechanical nature of molecular interactions directly and exploring vast combinatorial spaces more efficiently. While it may not "replace" conventional AI methods that directly sift omics data for targets, it will complement them: AI might suggest a target, and quantum computing might then deeply evaluate that target's druggability or simulate intervention strategies. Or quantum-enhanced ML might sift through complex epistatic interactions in genetic data to propose novel synthetic lethal target pairs in cancer. The field is very much in flux, but given the rapid progress (Cao et al., 2019), it is plausible that within the next decade, quantum computing will become another tool in the computational drug discovery toolbox. The future might see hybrid workflows where classical AI proposes hypotheses and quantum computing refines and tests them in silico before any wet lab experiments, making the overall process of target discovery faster and more reliable.

Conclusion and Future Outlook

AI-driven target identification stands at the forefront of the future of computer-aided drug design, offering an expanded and accelerated toolkit for unraveling disease biology. We have seen how integrating multi-omics data can reveal hidden drivers of disease, how knowledge graphs and GNNs can infer new disease–gene links

from the tapestry of biomedical knowledge, and how language models can read and reason over decades of literature to support target hypotheses. Emerging approaches like federated learning enable these AI models to learn from previously inaccessible datasets (across institutions or companies), thereby broadening the scope of discovery. AI is also increasingly intertwined with experimental science: it enhances high-throughput screens (like CRISPR) by guiding experimental design and interpreting results, and it leverages breakthroughs in structural biology (like AlphaFold) to ensure potential targets are actionable. Even quantum computing, while early in its journey, hints at future paradigms for simulating and evaluating targets with unprecedented fidelity.

These approaches do not exist in isolation – importantly, they complement and reinforce each other. The multi-omics analysis might suggest a novel target gene, which a language model then contextualizes with mechanistic insights from literature and prior knowledge. A federated model might validate the target's association across diverse patient cohorts, increasing confidence that it's broadly relevant. A CRISPR or RNAi screen can experimentally confirm that modulating the target yields the desired phenotypic effect. An AlphaFold model provides a 3D structure indicating the target is druggable (or suggests how it might be tackled). Thus, AI helps take us from an initial big-data correlation to a concrete, validated target with a clear path for drug development, in a fraction of the time it used to require.

Of course, challenges remain. AI predictions are only as good as the data and assumptions that go into them. Interpretability of complex AI models is an ongoing concern – we need to understand why an AI chose target X, to trust it and to design appropriate experiments. Research into explainable AI for biology is progressing, such as methods that highlight which features (pathways, mutations, etc.) were most influential in a prediction.

Data biases are another issue: human genomic data, for instance, is still over-representative of certain ancestries; if not addressed, AI could miss targets relevant to understudied populations. Efforts to gather more diverse data and use techniques like re-weighting in training can mitigate this (as federated learning aims to do). Additionally, while AI can propose dozens of potential targets, experimental validation remains a bottleneck. Testing each hypothesis in vivo or even in vitro is time-consuming and resource-intensive. Automating and prioritizing validation is thus critical – this is where AI-driven lab automation and better predictive modeling of efficacy come in, to focus resources on the most promising leads.

Looking ahead, the ultimate vision is that future drug discovery begins with a comprehensive in silico analysis that maps the molecular landscape of a disease and immediately offers a shortlist of high-confidence, mechanism-backed targets – some conventional, some very unexpected – that human experts can then take forward. This kind of AI-generated "disease blueprint" could become the starting point for every new therapeutic program. We are moving toward an era where AI is an ever-present partner from hypothesis to cure: from suggesting the hypothesis (target) to designing the molecule and even monitoring clinical response. As AI models continue to improve by learning from more data (including negative results and trial outcomes), their target predictions will become increasingly reliable. In fact, AI might eventually predict not just what to target, but how to target it (for example, recommending that a given protein is best inhibited with a degrader rather than an enzyme inhibitor, based on holistic analysis).

In conclusion, AI-driven target prediction is revolutionizing the front end of drug discovery. It expands our reach to identify novel therapeutic strategies by decoding the complexity of human biology in ways that were not possible before. By embracing these

technologies and continuing to integrate them with expert knowledge and rigorous experimentation, we will significantly enhance our ability to understand diseases and develop effective medicines. The future of computer-aided drug design will be defined by this synergy between artificial intelligence and human ingenuity, accelerating the translation of big data into breakthrough cures.

Key Points

- AI Integration of Big Data: Artificial intelligence can analyze and integrate multi-omics datasets (genomic, transcriptomic, proteomic, etc.) to pinpoint genes that consistently show disease-associated perturbations across data types. This multi-omics approach yields more robust target candidates than single data sources, uncovering key drivers of disease that might elude human analysis.
- Knowledge Graphs and GNNs: AI models using knowledge graphs and graph neural networks learn from complex networks of genes, diseases, drugs, and other biomedical entities to predict new disease–gene links. For example, the Progeni system integrated diverse networks into a probabilistic graph and successfully identified novel cancer targets via link prediction (Liu et al., 2024).
- NLP and Literature Mining: Advanced language models (e.g. BioBERT, BioGPT) can “read” millions of publications to extract known and implicit relationships between genes and diseases. These models assist target discovery by summarizing evidence for candidate targets and even generating new hypotheses based on connections in literature (Pun et al., 2023).

- CRISPR Screens and AI: Genome-wide CRISPR-Cas9 knockout screens produce functional genomics data that AI algorithms analyze to prioritize targets. Machine learning filters out noise, finds pathway enrichment in gene hit lists, and identifies essential gene networks. AI-guided analysis of CRISPR screens helps validate computationally predicted targets and reveals novel vulnerabilities (Bhat et al., 2022).

- Federated Learning: Federated learning enables AI models to be trained on sensitive data from multiple institutions without pooling the data, preserving privacy (Rieke et al., 2020). By learning from diverse, siloed datasets (e.g. patient genomics across hospitals, or pharma screening data across companies), federated models improve target prediction and generalizability. A notable example is the MELLODDY project, where a federated model trained on 10 companies' assay data outperformed models trained on any single company's data (Heyndrickx et al., 2024).

- Structural AI and Druggability: AI breakthroughs in protein structure prediction (AlphaFold2) provide 3D models for most human proteins (Jumper et al., 2021). These structural insights help assess target druggability and function. Researchers can evaluate whether a candidate target has suitable binding pockets or resembles known protein families, informing target selection and strategy.

- Generative AI for Therapeutics: Generative models (e.g. protein diffusion models) can design novel molecules or biologics against chosen targets. AI-designed proteins have been created to bind and modulate traditionally

challenging targets (Watson et al., 2023, expanding the range of "druggable" targets by providing new therapeutic modalities (such as de novo agonists or inhibitors) for targets identified by computational methods.

- Quantum Computing Potential: Emerging quantum computing approaches aim to tackle complex simulations and optimizations in drug discovery (Cao et al., 2019). In the future, quantum algorithms could improve the accuracy of virtual screening and molecular simulations for target evaluation, and quantum machine learning might uncover patterns in high-dimensional biological data that aid target identification. While not yet mainstream, quantum advances are poised to complement classical AI in target discovery.
- Convergence of Techniques: Modern target discovery often combines multiple AI approaches in a pipeline. For instance, multi-omics AI might nominate a target, an LLM-based system finds supporting literature and mechanisms, a federated model confirms its relevance in diverse cohorts, and a structural AI model assesses its druggability. This convergence accelerates the progression from data to testable target hypothesis.
- Impact on Drug Discovery Pipeline: AI-driven target prediction significantly de-risks and speeds up the early phase of drug discovery. By providing a data-backed shortlist of candidate targets (with rationales), it allows researchers to focus experimental validation and development resources on the most promising biology. This increases the chance that subsequent drug development efforts hit the mark, potentially improving

success rates in clinical trials by starting with better target choices.

2. Computational Prediction of Drug–Target Interactions and Lead Optimization

To address these limitations, computational approaches have become increasingly essential, providing a cost-effective and scalable alternative to traditional target identification. By leveraging bioinformatics, structural biology, and network-based models, computational methods allow for the rapid screening and prioritization of potential drug targets. Over the past two decades, techniques such as molecular docking, molecular dynamics (MD) simulations, and network-based target prediction have emerged as powerful tools to analyze biological interactions, protein structures, and disease mechanisms. However, despite their advantages, these methods have inherent limitations, necessitating the development of more advanced, AI-driven approaches to improve accuracy, efficiency, and predictive reliability.

Why These Methods Alone Are Not Enough

Despite their success, traditional computational methods face several key challenges:

- Molecular docking struggles with protein flexibility and scoring errors, leading to false positives.
- MD simulations provide high-resolution insights but are computationally intensive.
- Network-based methods depend on incomplete or noisy biological datasets, making them prone to false discoveries.

As drug discovery becomes increasingly data-driven, these challenges highlight the need for AI-enhanced approaches that can integrate multiple levels of biological information and refine target predictions. Artificial intelligence (AI) and machine learning (ML) algorithms have emerged as powerful tools to overcome these obstacles by enhancing scoring functions, optimizing molecular simulations, and improving network-based disease modeling.

The landscape of CADD is rapidly evolving as artificial intelligence (AI) becomes deeply integrated into drug design and discovery. Traditional computer-aided drug design (CADD) techniques – from virtual screening and molecular docking to pharmacophore modeling – are being augmented and often reinvented by modern machine learning algorithms. AI-driven methods can dramatically accelerate early-stage discovery by predicting drug–target interactions (DTIs) and binding affinities in silico, screening vast chemical libraries, and even generating novel bioactive molecules de novo. In parallel, advances in genomics, proteomics, and structural biology are producing an explosion of data ripe for AI analysis, enabling a new era of data-driven target identification and validation. The convergence of these fields is yielding powerful workflows: for example, deep learning models now routinely predict which compounds are likely to bind a given protein, generative models propose optimized molecular structures, and high-resolution protein structure predictions (exemplified by AlphaFold) provide reliable starting points for structure-based design. This continuous, multi-disciplinary narrative explores the future of CADD, focusing on how AI is revolutionizing drug–target interaction prediction and binding affinity modeling in an end-to-end drug discovery pipeline.

AI in Drug–Target Interaction Prediction: One of the foundational applications of AI in CADD is predicting DTIs (Drug target interactions) – essentially determining whether a given drug molecule modulates a given protein target. Early machine learning approaches in this realm date back over a decade, using techniques like kernel methods and matrix factorization on chemical and genomic feature spaces. In recent years, however, deep learning has taken center stage. Models such as DeepDTA, DeepAffinity, and wide and deep neural networks have achieved success by learning complex non-linear mappings from molecular descriptors (or structures) of ligands and targets to binding affinities or interaction probabilities. A key advantage of deep learning is its ability to automatically extract features; for example, convolutional neural networks (CNNs) can treat protein sequences and ligand SMILES as "images" or sequences and learn predictive patterns, while transformer-based models leverage attention mechanisms to capture long-range dependencies in molecular sequences. In one recent study, a transformer-network incorporating graph representations (DeepMGT-DTI) significantly outperformed earlier methods on DTI prediction benchmarks. Likewise, graph neural networks (GNNs) have become invaluable for DTI prediction. Molecules are naturally represented as graphs (atoms as nodes, bonds as edges), and proteins can be represented by residue graphs or 3D contact networks; GNN-based models can learn joint representations of drug–target pairs that encode both chemical structure and protein context. Lim et al. (2019) introduced one such model combining 3D structural information with graph convolutions, improving DTI prediction accuracy by leveraging spatial features of molecular interactions. These deep models often surpass classic ligand-based QSAR approaches, especially in capturing cases where the binding determinants are subtle or context-dependent.

Despite their accuracy, deep learning DTI models face challenges. They are notoriously data-hungry, requiring large training sets of known drug–target pairs. Public databases like BindingDB, ChEMBL, and various bioassay repositories have enabled training of increasingly complex models, but biases in these datasets can limit model generalizability. For instance, models tend to perform well on protein families or chemotypes abundant in the training data but may struggle with truly novel scaffolds or targets (the classic applicability domain issue). Moreover, many deep models function as black boxes, providing predictions without mechanistic insight. This lack of interpretability has led to efforts to develop explainable AI techniques – for example, using attention weights to highlight which protein residues and ligand atoms contribute most to a predicted interaction, or applying methods like SHAP or integrated gradients to rationalize predictions. Interpretability is not just an academic concern; it is essential for trust and adoption in the pharma community. A clinician or medicinal chemist is more likely to act on a prediction if the model can suggest why a drug binds a target (e.g. identifying a key hydrogen bond or hydrophobic pocket) rather than treating the model as an oracle. Accordingly, current research in DTI prediction often couples high-performance architectures with interpretability modules, aiming to satisfy both accuracy and transparency.

Beyond binary interaction predictions, AI is also improving quantitative binding affinity prediction – effectively replacing or complementing traditional scoring functions. Whereas classical docking programs produce a single score or energy for a given pose, modern AI models can be trained on large affinity datasets (like IC50/Kd compilations) to predict binding strengths. Techniques range from graph-based deep learning models that predict continuous affinity values, to multitask networks that learn to predict binding to many targets simultaneously, thereby leveraging related

information to improve predictions for each (an approach that can mitigate data scarcity for low-data targets). There are also hybrid approaches where docking and AI are combined: for example, generating multiple docked poses and using a neural network to re-score or rank those poses (learning which docked pose features correlate with true binding. In one innovative approach, researchers have integrated language model techniques (like BERT or other transformers initially developed for text) to create embedding vectors for protein sequences, which can then be fused with ligand representations for DTI prediction. These language model embeddings of proteins (sometimes called "protein language models") capture subtle evolutionary and structural signals from sequences and have been shown to boost the accuracy of binding predictions, especially for targets lacking solved 3D structures.

Generative Models for Ligand Design: Perhaps the most exciting advance in AI-driven CADD is the rise of deep generative models that can create novel chemical structures with desired properties. Instead of searching existing libraries, generative models learn the rules of chemical structure from data and can propose new molecules, essentially functioning as AI "imaginations" of drug-like compounds. Early examples include variational autoencoders (VAEs) that learned to compress molecules into continuous latent vectors and sample new molecules from this latent space, and generative adversarial networks (GANs) that learned to produce realistic molecules by pitting a generator network against a discriminator (in an analogy to image generation). By around 2018, proof-of-concept studies had shown that such models could generate valid molecules that were structurally novel – but the field truly leaped forward when generative models were combined with reinforcement learning (RL) and property-conditioned design objectives. A landmark study by Zhavoronkov et al. in 2019 demonstrated the power of this approach. In a timed challenge, they

used a generative tensorial reinforcement learning system (GENTRL) to design small-molecule inhibitors targeting DDR1 kinase, an anti-fibrotic target, and managed to go from computer design to experimental validation in a matter of weeks. The AI model generated six novel chemical structures in only 21 days, four of which showed micromolar to nanomolar activity in enzyme assays, and two of which were active in cellular assays; one lead compound even demonstrated favorable pharmacokinetics in mice. This entire process – design, synthesis, in vitro tests, and in vivo mouse testing – was completed in 46 days, roughly 15 times faster than a conventional lead discovery pipeline. This milestone provided a striking proof-of-concept that AI could dramatically compress the drug design cycle.

Since then, generative models have proliferated with increasing sophistication. There are models based on recurrent neural networks (e.g. LSTMs treating SMILES strings as sentences to be "written"), graph-based generative models that assemble molecules atom-by-atom or fragment-by-fragment, and the latest breakthrough: diffusion models. Diffusion models, originally developed for image generation, have been adapted to molecular design by treating the generation process as a progressive denoising of a random graph or 3D point cloud. These models perturb molecules with noise and then learn to reverse the process, yielding novel molecules that follow a learned distribution (Diffusion Models in De Novo Drug Design - ACS Publications) (Diffusion Models in De Novo Drug Design - PubMed). Diffusion approaches are especially powerful for 3D structure-based design – for instance, the DiffDock model uses a diffusion generative process to place a ligand in a protein's binding site, treating the protein structure as a fixed context and diffusing the ligand's atomic coordinates until a plausible bound conformation emerges (Speeding up drug discovery with diffusion generative models). Such models have achieved state-

of-the-art performance in blind docking benchmarks, often surpassing classical physics-based docking in success rate (Speeding up drug discovery with diffusion generative models) . More broadly, diffusion models have emerged as among the most potent tools for de novo drug design, demonstrating the ability to generate candidate molecules with optimized properties (potency, selectivity, etc.) and valid synthetic routes (Diffusion Models in De Novo Drug Design - PubMed). Their strength lies in generating a diversity of high-quality structures while incorporating complex conditions (like docking constraints, pharmacophore features, or even synthetic accessibility filters) as part of the generative process.

Generative models are now being steered by multi-objective optimization through reinforcement learning. A generative model can propose a molecule, and an RL algorithm can treat a combined scoring function (incorporating predicted potency, ADMET properties, novelty, etc.) as a reward to bias generation towards optimal candidates. This approach, essentially an AI-driven closed-loop discovery, is increasingly used in adaptive discovery workflows. For example, one could generate a batch of molecules, have predictive models evaluate each for various criteria (on-target potency, off-target selectivity, toxicity, synthetic feasibility), then use those evaluations to update the generative model via RL to improve the next round of molecules. This iterative loop can be coupled with actual experimentation in what's known as Bayesian optimization or active learning cycles – wherein the AI not only designs molecules but also decides which ones should be synthesized and tested next, gradually refining its model of the target's requirements. Such self-driving laboratories, while still in their infancy, are a clear direction for the future of drug discovery, potentially enabling an autonomous cycle of hypothesis generation, testing, and learning.

Despite the promise, generative AI for chemistry has its own pitfalls. Models may generate molecules that are chemically unrealistic or infeasible to synthesize if not properly constrained. They might also exploit learned data biases in undesirable ways (for instance, overemphasizing particular scaffolds that appeared frequently in the training set, thereby producing analogs rather than truly novel chemotypes). There are also cases where a model optimizes the learned objective to an absurd degree – a classic example being when a generative model, instructed only to maximize predicted activity, produces structures that the predictive model scores highly but that violate known medicinal chemistry principles (often these turn out to be pan-assay interference compounds or other notorious false positives). Thus, in practice, generative outputs require filtration and human medicinal chemistry intuition. Researchers mitigate these issues by integrating additional filters (for toxicity, stability, drug-likeness) into the generation process and by using human-in-the-loop feedback at intermediate stages. Encouragingly, some AI-designed compounds have already entered clinical trials in recent years (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences) (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences), underscoring that with careful validation, AI-generated molecules can advance beyond just in silico curiosities.

Enhanced Docking and Binding Prediction with AI: Structure-based drug design traditionally relies on methods like molecular docking to predict how a ligand binds to a protein and to estimate binding affinity. While docking is fast and useful, it often assumes a relatively rigid protein structure and employs simplified scoring functions that cannot capture all aspects of molecular recognition (such as entropic effects, water-mediated interactions, or induced fit movements). AI is addressing these limitations on multiple fronts. One area is the incorporation of protein flexibility in

docking – historically approached via induced-fit docking protocols or ensemble docking of multiple protein conformations. The induced fit problem, where the binding of a ligand causes the protein to change shape for a better fit, has long been recognized as a cause of docking failure when only a single protein conformation is used . New AI-enhanced workflows generate and utilize many protein conformations to account for this. For example, a 2023 study presented a CHARMM-GUI based induced-fit docking pipeline that first refines the binding pocket (using molecular dynamics and minimization) to create an ensemble of receptor conformations, then docks ligands into each and evaluates the stability of each complex via short MD simulations . This approach achieved a remarkable ~80% success rate within 2.5 Å RMSD in cross-docking benchmarks (i.e. docking a ligand into a different conformation of the protein than the one it was co-crystallized with) (CHARMM-GUI-Based Induced Fit Docking Workflow to Generate Reliable Protein-Ligand Binding Modes - PubMed), far outperforming standard rigid docking on these challenging cases. The success underscores that a combination of physics-based simulation (to capture induced fit) and AI or efficient sampling (to explore multiple poses and conformations) can substantially improve docking outcomes. Similarly, ensemble docking – wherein one docks compounds against a panel of protein conformations (derived from MD simulations or multiple crystal structures) – has been shown to increase hit rates by accounting for protein flexibility, albeit with increased computational cost. AI can assist here by predicting which conformations are most relevant or by clustering MD trajectories to a manageable number of representative states. In essence, AI helps navigate the protein's conformational landscape more intelligently, choosing informative snapshots for docking rather than relying on brute force.

Machine learning is also being leveraged to re-score or refine docking results. Classical scoring functions often mis-rank poses or compounds, so researchers have trained ML models (random forests, neural networks, gradient boosting models, etc.) on large datasets of ligand poses with known binding outcomes to learn corrections to scoring functions. These ML scoring functions (sometimes called "data-driven scoring") learn from both successes and failures of classical scoring, and can incorporate descriptors beyond the physics-based energy terms (e.g., interaction fingerprints, chemical functionality patterns, etc.). Studies report that such models improve the enrichment of true hits in virtual screening campaigns (Application of Artificial Intelligence In Drug-target Interactions Prediction: A Review) (Application of Artificial Intelligence In Drug-target Interactions Prediction: A Review). Additionally, AI can analyze the enormous output of docking campaigns in ways humans cannot. For instance, given millions of docked poses from a virtual screen, a clustering algorithm or an autoencoder might detect common binding modes or novel chemotypes that bind in similar ways, guiding medicinal chemists to scaffold hop or merge features from different molecules.

Beyond docking, physics-based methods like free energy perturbation (FEP) and molecular dynamics (MD) simulations represent the gold standard for binding affinity prediction, accounting for conformational dynamics and thermodynamics. These are computationally intensive, but AI is making inroads here as well. One approach is using machine learning to accelerate molecular dynamics or augment it. Enhanced sampling techniques (such as metadynamics, accelerated MD, and replica exchange) have been complemented by AI methods that learn optimal collective variables or reaction coordinates to bias sampling toward rare events like ligand binding/unbinding. For example, variational autoencoders and other dimensionality-reduction techniques have

been applied to find low-dimensional representations of protein–ligand configurations that capture the progress of binding, which can then be used to run faster 1D or 2D free energy calculations. Reinforcement learning has also been explored to adaptively drive MD simulations – the simulation acts as an environment, and an RL agent can decide on-the-fly whether to push the ligand in a certain direction or apply a force to a protein loop to expedite an otherwise slow transition, with a reward for achieving unbinding or binding events. Though experimental, such AI-guided simulations show promise in capturing kinetic pathways that would normally require orders of magnitude more sampling.

Meanwhile, neural networks are being trained as fast approximate potentials or force fields. So-called neural force fields or physics-informed neural networks can ingest molecular coordinates and output energies and forces, essentially learning to emulate quantum mechanics or high-level molecular mechanics. DeepMind's recent work on AlphaFold's potential is one example in a different context (protein folding), but in the drug-binding context there are efforts like SchNet or ANI which learn potential energy surfaces. These can accelerate energy evaluations in FEP or molecular docking refinement. For instance, instead of a costly quantum mechanical calculation for every sampled configuration in a QM/MM simulation, a pre-trained neural network potential can provide near-instant estimates of energy, dramatically speeding up scoring of poses in a binding site (How exascale computing can shape drug design - ScienceDirect.com) . This marries well with the need for higher accuracy in cases such as metalloproteins or covalent inhibitors, where traditional force fields may falter – a trained network can capture subtle electronic effects if included in its training.

Importantly, AI is being used to interpret and learn from MD simulation data. A long unbiased MD simulation of a ligand binding

process (if one has the computing power of, say, special-purpose machines like Anton) yields a trajectory showing how the ligand diffuses, finds the pocket, and settles. AI pattern recognition can mine these trajectories to identify important transitional states or alternative binding poses. Clustering algorithms can identify metastable states along the pathway. More advanced, researchers have used long short-term memory (LSTM) networks to model the temporal sequence of protein–ligand contacts, essentially treating a binding or unbinding trajectory as a "language" to be learned, thereby identifying which early contacts are predictive of eventual successful binding. These analyses help elucidate mechanistic insights – for instance, an AI analysis might reveal that a ligand first binds shallowly to a surface patch and later migrates to the orthosteric site, suggesting opportunities for allosteric modulation. Moreover, by combining many short MD simulations (perhaps started from different initial orientations of a ligand around a protein) and analyzing them in aggregate, AI methods like Markov state models or kinetic network models can construct a comprehensive picture of the binding free energy landscape. The bottom line is that AI is not only accelerating predictions but also expanding what we can learn from simulations, thus bridging the traditional gap between modeling and experiment. Indeed, recent comments on MD and CADD highlight that machine learning can guide the selection of simulation snapshots for analysis, refine continuum solvent models like MM/GBSA, and even help decide how many simulations are needed to reach converged answers. By guiding simulation frame selection and energy evaluations, AI markedly improves the efficiency of physics-based binding predictions, bringing rigorous methods like free energy calculations closer to practical throughput.

Quantum Computing and Quantum-Inspired ML: An eye toward the future wouldn't be complete without noting the nascent contributions of quantum computing in drug design. While still in

very early stages, quantum machine learning (QML) is emerging as a potential game-changer for certain computational chemistry problems. Quantum computers have the theoretical ability to simulate quantum systems (like molecular electronic structures) exponentially faster than classical ones, which could revolutionize drug discovery by enabling exact (or highly accurate) calculations of binding energies for drug–target complexes, or rapid generation of novel compounds via quantum algorithms. In practice, current quantum hardware is limited in scale and prone to noise, but small-scale demonstrations have already been reported. Notably, in 2023 an experimental hybrid quantum-classical workflow was used to design inhibitors for the oncogenic protein KRAS, a target long deemed "undruggable." Researchers combined quantum computing with generative AI and classical screening: they trained a quantum circuit-based generative model alongside a classical LSTM generator to propose molecules, and used a massive classical virtual screen (100 million compounds) to guide the training dataset (Team uses AI and quantum computing to target 'undruggable' cancer protein) (Team uses AI and quantum computing to target 'undruggable' cancer protein). Out of this combined pipeline came promising KRAS binders, two of which showed low-micromolar efficacy in cell-based assays against multiple KRAS mutants (Team uses AI and quantum computing to target 'undruggable' cancer protein). This work, published in Nature Biotechnology in 2025, represents the first instance of quantum computers contributing directly to the discovery of new drug leads for a challenging protein target (Team uses AI and quantum computing to target 'undruggable' cancer protein) (Team uses AI and quantum computing to target 'undruggable' cancer protein). It suggests that even as quantum hardware improves, near-term value can be extracted by hybridizing quantum algorithms with AI-driven design – for instance, using a quantum computer to efficiently sample from a complex chemical

space distribution, and a classical AI to fine-tune or filter those samples.

Apart from direct quantum computing, quantum-inspired methods are enhancing classical simulations. Quantum mechanical/molecular mechanical (QM/MM) hybrid models, where the key part of the binding site is treated quantum-mechanically and the rest of the system with classical force fields, have long been used to improve accuracy of binding mode prediction and affinity calculation . AI plays a role here by helping decide which states or poses to subject to expensive QM refinement, and by learning to predict QM corrections. For example, if a docked ligand involves a metal coordination or proton transfer, a machine learning classifier might flag that pose as requiring a QM re-score. Researchers are also exploring surrogate models for QM calculations: training neural nets on thousands of QM calculations of protein–ligand complexes to predict, say, the polarization energy or charge redistribution upon binding, which can then be added to classical scores for better accuracy. This is another facet of the broader trend of replacing expensive computations with learned models – analogous to how AlphaFold bypassed the need for slow physics-based folding simulations by training on structures, one can imagine a future system that bypasses brute-force quantum chemistry in affinity prediction by relying on a learned quantum-savvy scoring function.

Protein Structure Prediction and Folding Powered by AI: A revolution in CADD came not only from better algorithms for ligands, but also from breakthroughs on the target side – specifically, predicting protein structures from sequence. The success of DeepMind's AlphaFold2 in 2020–2021 has been described as transformational for structural biology . AlphaFold's deep learning approach achieved, for the first time, accuracy comparable to experimental structures in many cases, delivering high-resolution predictions of protein 3D structure purely from amino acid

sequences. The implications for drug discovery are enormous. Many proteins of therapeutic interest that were previously intractable to crystallography or cryo-EM now have AlphaFold models available. In fact, AlphaFold and its collaborators have released predictions for over 200 million proteins (essentially every protein sequence in UniProt) into a public database , meaning that a vast majority of human proteins and pathogens have at least a theoretical model available. Medicinal chemists can now perform structure-based design or virtual screens on targets that lack an experimental structure, something that was impossible a few years ago (Application of Artificial Intelligence In Drug-target Interactions Prediction: A Review) (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences). For example, researchers recently reported the discovery of a novel CDK20 inhibitor by leveraging an AlphaFold model of CDK20 in the absence of any crystal structure; the AI-designed molecule was confirmed experimentally, illustrating how predicted structures can kick-start real drug discovery (Application of Artificial Intelligence In Drug-target Interactions Prediction: A Review) (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences). AlphaFold's impact also extends to polypharmacology – by predicting structures for off-targets and antitargets, it enables in silico selectivity profiling of new compounds by docking them across a panel of modelled proteins to foresee potential cross-reactivity.

However, it's important to use predicted structures with an understanding of their limitations . AlphaFold and similar algorithms provide a static snapshot (often corresponding to a dominant likely conformation), and they may not capture alternative conformations or induced-fit movements that occur upon ligand binding . Indeed, studies have found that while AlphaFold models are generally excellent in backbone placement, they can be less reliable for loop regions or domain orientations that might shift in

different functional states . Docking studies comparing performance on AlphaFold models vs. real crystal structures have shown mixed results: in many cases AlphaFold models suffice for docking, but in others subtle errors in the binding site (like a side chain rotamer or a slightly misplaced loop) can cause docking algorithms to mis-predict ligand poses . Encouragingly, AlphaFold provides a measure of its own confidence in different regions of the structure (the predicted aligned error, PAE) , which practitioners can use to gauge whether the binding site is likely predicted well or if caution is warranted. We often see a workflow where an AlphaFold model is taken as a starting point and then refined – either via energy minimization, brief MD simulations (to relax any steric clashes or unrealistic strain), or by using additional experimental data (e.g., cross-linking data or mutagenesis data that can be used to slightly tweak the model). In drug design projects, teams may treat AlphaFold models as another form of "homology model." Much like homology modeling has long been used when a template structure is available, now AlphaFold can provide a model even when no clear single template exists (in essence, it is like doing multiple-sequence homology modeling with a deep learning prior). These models can be plugged into all the usual CADD steps: virtual screening, pharmacophore mapping, pocket analysis for allosteric sites, etc. And when an experimental structure later becomes available, it often validates much of the AlphaFold model while highlighting which regions needed improvement.

AI has also improved protein–protein and protein–ligand structural prediction. Following AlphaFold's success in monomers, efforts like AlphaFold-Multimer and others have made progress in predicting the structures of protein complexes. We are approaching a point where given a protein target and a small molecule ligand, deep learning might directly predict the bound pose – essentially an AI-driven docking that places the ligand in the binding site without a traditional energy minimization, by learning pose preferences from

data. Preliminary models in this vein have appeared: for example, one approach encodes the protein structure in a grid or graph neural network and then uses an encoder–decoder model to place a ligand, yielding poses that can sometimes rival docking solutions. Diffusion models we mentioned (like DiffDock) also fall into this category, treating the protein atoms as an environment in which ligand atoms are grown in a physics-learned manner (Speeding up drug discovery with diffusion generative models). These approaches blur the line between what we consider "docking" and "structure prediction" – they are purely data-driven pose predictors. In benchmarks, DiffDock was able to predict correct binding poses for a substantial fraction of test cases where classical docking failed, especially for ligand classes underrepresented in training it could still generalize using learned chemical intuition (Speeding up drug discovery with diffusion generative models). Looking ahead, we can envision a fully AI-based binding mode prediction system that takes as input a protein sequence (predicts its structure and pocket via AlphaFold or a similar method) and a ligand structure (proposes plausible binding poses via a learned model), all in a matter of seconds and with quantified confidence.

Systems Biology, Networks, and Target Identification: While much of CADD focuses on the direct molecular interactions, AI is also transforming how we identify which targets to drug in the first place. Target identification and validation are crucial early steps – pick the wrong target and even the best-designed drug will fail in the clinic. Historically, target discovery often relied on academic research into disease biology, and on methods like genetic linkage studies, expression profiling, or serendipitous findings. Now, the availability of massive "omics" datasets (genomics, transcriptomics, proteomics, metabolomics) and functional genomics screens (e.g. CRISPR knockout screens across hundreds of cell lines) allows a more systematic approach. AI excels at finding patterns in these

high-dimensional datasets that might pinpoint disease vulnerabilities. For instance, integrating multi-omics data with phenotype data, machine learning models have been developed that prioritize certain genes as potential drug targets based on their network influence and disease-association patterns (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences) (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences). One notable example is the use of deep networks on heterogeneous networks of genes, compounds, and diseases: Zeng et al. (2020) constructed a deep learning model that ingested a network with nodes representing drugs, targets, and diseases and learned to predict new links (i.e., new target identifications) that were later experimentally confirmed (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences) (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences). Similarly, graph-based approaches have been applied to known drug–target–disease triplets to predict drug repurposing opportunities or polypharmacology, effectively treating target identification as a link prediction problem in a knowledge graph (Application of Artificial Intelligence In Drug-target Interactions Prediction: A Review) (Application of Artificial Intelligence In Drug-target Interactions Prediction: A Review).

Beyond correlation-based approaches, causal inference methods combined with AI are emerging. Large perturbational datasets like the Connectivity Map (which contains gene expression signatures of cells treated with thousands of compounds) and CRISPR knockout datasets (like the DepMap, which catalogues the effect of knocking out each gene on cell viability in many cancer cell lines) provide matrices of high-dimensional responses. AI models, including Bayesian machine learning methods, have been employed to integrate these diverse data types – chemical structures, gene expression changes, genetic dependency profiles – to propose which

protein a new compound might be hitting or which gene, if drugged, would most selectively kill cancer cells (Bayesian Model in Target Identification - DLeader) (A Bayesian machine learning approach for drug target identification ...). For example, Madhukar et al. (2019) developed a Bayesian integrative model called BANDIT that combined chemical proteomics data, gene expression perturbations, and known target information to predict novel targets for existing molecules; impressively, BANDIT achieved ~90% accuracy in recovering known drug–target pairs and suggested new ones that were experimentally validated (Bayesian Model in Target Identification - DLeader) (A Bayesian machine learning approach for drug target identification ...). These approaches underscore a paradigm shift: instead of starting with one target and searching for drugs, we can start with data from many drugs and many targets and let AI suggest new links, be it repurposing an old drug for a new target or identifying a previously unrecognized protein that, if drugged, modulates disease.

Functional genomics screenings are particularly powerful for target discovery, and AI helps make sense of them. In CRISPR-Cas9 knockout or interference screens, one might knock out each gene in a genome (one per cell line) and observe which genes are essential for cell survival or which modify some disease-relevant phenotype. These screens result in lists of genes with a score (e.g., depletion of cells when gene X is knocked out implies gene X is essential). AI can analyze such results across many contexts to pinpoint druggable targets. For example, AI analysis of a CRISPR screen identified BRD2 as a key regulator of the host inflammatory response to SARS-CoV-2, making it a candidate target for COVID-19 therapy (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences) (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences). In cancer, combining CRISPR screen data with gene expression and mutation

data, AI models have found synthetic lethal targets – genes that are not essential in a normal cell but become essential when a particular cancer mutation is present, thus providing a tumor-specific vulnerability. These synthetic lethal relationships (like the famous BRCA1–PARP example) can be discovered by pattern-finding: if only the cell lines with mutation Y die when gene X is knocked out, then X is likely a synthetic lethal partner of Y. Machine learning classifiers are adept at picking out such conditional dependencies from large, noisy screening datasets. What's more, by learning from multiple screening datasets, AI can suggest which hits are likely to be true positives versus context-specific or off-target effects.

AI can also propose targets by analyzing patient data and disease networks. In complex diseases, one often constructs networks of proteins (interaction networks, signaling pathways, gene regulatory networks) and looks for nodes that are dysregulated. AI-based network analysis (like network propagation algorithms or community detection augmented by learning) can highlight "hub" proteins or pathways that drive disease phenotypes. Increasingly, AI is used to combine heterogeneous evidence: for instance, in Alzheimer's disease, one might integrate genome-wide association study (GWAS) results (which implicate certain genes), gene expression changes in patients vs. controls, and known protein–protein interactions to identify a subnetwork of proteins that is both genetically and transcriptionally implicated. Such approaches have led to proposals of novel targets that might have been missed by any single modality. The trade-off often discussed is between target novelty and confidence (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences) (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences). AI can generate many hypotheses, including very novel ones (e.g., a protein never before linked to a disease); however, novel targets often come with low prior evidence and hence higher risk. Conversely, known targets

have high confidence but may be heavily trodden ground. The consensus emerging is that AI should be used to broaden the search space (to not overlook potential novel targets), but strong experimental validation is still required, especially for novel predictions. Indeed, some AI-identified targets have now progressed into validation: recent examples include a deep learning approach that predicted HDAC6 inhibition could be cardioprotective in a form of heart disease, which was subsequently confirmed in a mouse model (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences), and identification of CDK20 as a target for hepatocellular carcinoma via AI network analysis, leading to development of a new inhibitor that showed efficacy in preclinical models (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences).

Integration of Synthetic Biology and Patient-Derived Models: AI doesn't operate in isolation; its predictions need experimental testing and refinement. Cutting-edge experimental techniques are providing more accurate testbeds for AI hypotheses in the preclinical stage. For instance, patient-derived organoids (miniaturized organs grown from patient stem cells) and organ-on-a-chip systems can recapitulate human tissue architecture and function far better than traditional cell lines. Using these as avatars of disease, researchers can test whether modulating an AI-suggested target yields the expected therapeutic effect. These complex systems generate high-dimensional data (imaging, multi-omics readouts), which again loops back to AI for analysis. In a sense, AI helps choose what to test, and when new data comes from advanced models, AI helps interpret that data, creating a virtuous cycle. A recent review of human disease models in drug development notes that organoids and bioengineered tissues have improved translational success by providing more human-relevant insights early on (Human disease models in drug development) ((PDF) Human disease models in drug

development (2023) - SciSpace). AI can analyze organoid responses (for example, changes in high-content imaging or transcriptomic profiles after treatment) to quantitatively assess drug efficacy and off-target effects in a way that is more predictive of patient outcomes. This is particularly useful in diseases like cancer, where patient-derived tumor organoids can be used to test a panel of drugs and combinations. Machine learning can cluster response patterns and potentially predict which organoid (or patient) subtypes respond to which treatments, thus guiding precision medicine.

Synthetic biology also contributes to target validation: one can engineer cell systems or whole organisms to modulate targets in controlled ways (using inducible gene circuits, CRISPR activation/repression systems, etc.), essentially creating experimental models that simulate what a drug would do. AI is helpful in designing these genetic constructs (for example, designing guide RNA libraries, optimizing gene circuit designs for desired dynamical behavior, etc.) but also in analyzing the results. If an AI model predicts that co-inhibiting Pathway A and Pathway B will have a synergistic effect on killing cancer cells, scientists could engineer a synthetic circuit that knocks down a key gene in A and B simultaneously when triggered. The outcomes from such experiments, perhaps measured across many cell lines, produce a rich dataset of synergy measurements that AI can again mine to refine its understanding of the underlying network.

In the translational pipeline, omics-guided patient stratification is an area where AI shines. For example, given large cancer genomics datasets, AI can identify molecular subtypes of a disease that have distinct vulnerabilities. This can inform target selection – a target might not be universally relevant for all patients, but absolutely critical for one subtype. AI-driven analyses have led to the identification of such subtype-specific targets, e.g., a specific metabolic enzyme might emerge as a synthetic lethal target only in

tumors with a certain metabolic reprogramming signature. CRISPR screens can be segmented by these subtypes to find differential essential genes. In practice, this means AI might guide a pharma company to develop a drug not for "breast cancer" generally, but for the subset of breast cancer characterized by, say, a MYC gene amplification and a particular transcriptional program, where a certain chromatin regulator becomes essential. This approach aligns with the trend towards precision medicine and ensures that when a new drug enters clinical trials, it can be paired with biomarkers to select the patients most likely to benefit, thereby improving trial success rates.

Challenges and Considerations: Despite the impressive advancements, there remain significant challenges in AI-driven CADD. Data quality and bias are perennial issues. Models are only as good as the data they learn from; if there are systematic errors or biases in experimental assays, the model will inherit them. For instance, training a predictive model on binding affinity data that come mostly from a single assay format might inadvertently teach the model the quirks of that assay rather than true physical binding determinants. Similarly, chemical space is vast but our databases cover a thin slice of it, biased toward certain scaffolds (e.g., many drugs contain similar functional groups to each other, so models may get a skewed view of what chemical space looks like). An AI model might thus overestimate the drug-likeness or synthesizability of molecules that in reality are exotic. There is also the issue of publication bias and confirmation bias in target identification: AI might find what many papers report as important (since published data is used), potentially overlooking novel insights because they are, by definition, not prevalent in literature data (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences). Efforts are being made to mitigate bias – for example, through data augmentation (generating synthetic data points to balance classes),

transfer learning (where a model pretrained on a broad dataset is fine-tuned on a specific unbiased dataset), and federated learning (where models are trained on data from multiple organizations without pooling the data, thus leveraging diverse datasets that might reduce bias).

Model interpretability and trust present another challenge. While AI/ML models can achieve high predictive accuracy, their adoption in drug discovery teams (and especially by regulatory bodies) depends on how well their decisions can be understood and justified. A black-box model that predicts a certain compound will be toxic is less useful than one that can highlight a particular substructure likely responsible for toxicity, which a chemist can then modify. This has led to growing interest in explainable AI (XAI) in drug discovery. Techniques such as attention mapping in transformer models can point to which parts of a molecule align with which protein regions for a predicted strong binding (Application of Artificial Intelligence In Drug-target Interactions Prediction: A Review) (Application of Artificial Intelligence In Drug-target Interactions Prediction: A Review). Graph-based models can be probed by removing certain nodes or edges to see how it impacts prediction, effectively identifying important pharmacophores. And sometimes simpler surrogate models (like a decision tree) can be trained to mimic the behavior of a complex model within a local region of chemical space, yielding human-readable rules (for example, “presence of a positively charged moiety and a polyaromatic system is a strong indicator of activity for this target”) that align with the deep model’s reasoning. The field recognizes that improving interpretability is not just an academic exercise but critical for clinical translatability: when an AI-designed drug enters regulatory review, questions will be asked about why that molecule was chosen and whether any key liabilities were accounted for. Thus, the future likely holds AI models with interpretation layers explicitly

built in, or hybrid models that combine mechanistic components with learned components to retain some explainability.

Scalability is another practical concern. Some of the most powerful AI models (for instance, large transformers or deep GNNs on big graphs) require considerable computational resources, both for training and inference. While training can often be done offline by specialized teams, inference needs to be feasible in real-time for broad screening. There is ongoing work on model compression and optimization – e.g., distilling a large model into a smaller one, quantizing weights, or using cloud-based distributed inference – to ensure that even as models grow in complexity, their deployment remains scalable. In a real drug discovery project, one might need to screen millions (or even billions) of candidate compounds quickly. AI can narrow the funnel dramatically compared to brute force docking or experimental HTS, but if the model itself is huge, screening that many compounds may still be slow. Therefore, efficiency is being baked into new models. For example, graph neural nets have been optimized to run on GPUs in parallel for thousands of molecules at a time, and generative models can propose new compounds in a continuous manner rather than one-shot, allowing early termination if certain properties look poor. In terms of data, federated learning that we touched on is addressing scalability of data sharing. Multiple pharmaceutical companies each possess proprietary screening data. Due to competitive and privacy reasons, they cannot pool these into one giant database. Federated learning allows a shared model to be trained across companies without any data leaving individual premises – the model weights are updated iteratively by each company's data and only the weights (not the raw data) are shared (Application of Artificial Intelligence In Drug-target Interactions Prediction: A Review). This approach can effectively scale the training data size and diversity available to an AI model, making predictions more robust. A recent federated

learning benchmark for DTI prediction demonstrated that models trained in a federated manner across silos of data achieved performance close to a model trained on the combined data outright (Application of Artificial Intelligence In Drug-target Interactions Prediction: A Review) (Application of Artificial Intelligence In Drug-target Interactions Prediction: A Review), showing the feasibility of this approach for real-world drug discovery where data is fragmented across stakeholders.

When it comes to clinical **translatability**, we must acknowledge that despite all improvements in prediction, the true test of a drug is in human patients. The high failure rate in clinical trials (around 85% for new drug candidates) is often due to lack of efficacy or unexpected toxicity in humans (AI-powered therapeutic target discovery: Trends in Pharmacological Sciences). AI can help reduce this by better prediction of human-relevant effects: integrating human genetics (e.g., if human loss-of-function mutations in a target cause a disease-like or protective phenotype, indicating target validity), predicting off-target interactions that might cause side effects, and even modeling patient population variation (polymorphisms in drug metabolism genes, differing pathophysiology). However, the predictive models themselves need to be validated with rigorous experimental data – which is why AI-driven drug discovery has a strong emphasis on iterative validation. Early phases of projects now often include AI-predicted biomarkers and companion diagnostics. For example, if AI suggests a drug will only work in patients with a certain molecular signature, clinical trials can be designed to enroll based on that signature, and AI can analyze early trial data to adaptively update who is benefiting (adaptive trials). In a sense, AI continues to play a role even in clinical development, analyzing multi-modal patient data (e.g., imaging, genomics, lab tests) to identify responders vs. non-responders, perhaps identifying unforeseen secondary markers that

correlate with response, which can then refine the target or patient selection strategy mid-course.

Future Perspectives: Looking forward, the future of AI in CADD is likely to be characterized by greater integration – of data sources, of disciplines, and of AI with human expertise – rather than a wholesale replacement of one approach with another. One exciting concept is the development of digital twins for drug discovery. A digital twin is a virtual model of a complex system (in this case, a patient or a disease) that can be used to simulate interventions. With AI, one can conceive of a "digital twin" of a patient that incorporates that individual's genomic, proteomic, and clinical data to predict how they would respond to various drugs. This goes beyond traditional pharmacokinetic/pharmacodynamic models by using AI to handle the vast complexity of human physiology and disease heterogeneity . For example, an AI-driven digital twin of a cancer patient might simulate tumor evolution under different therapies, helping oncologists choose the optimal treatment sequence. In drug development, digital twins of diseases could be used to test a drug in a virtual population before actual trials – adjusting parameters to represent different patient subgroups, comorbidities, etc., to foresee challenges. Early efforts in this direction show that AI can integrate diverse clinical data to create patient-specific models and that these models can predict outcomes like trial survival or adverse effects with promising accuracy . As computing power grows and more real-world data becomes available (from electronic health records, wearable sensors, etc.), these digital twins will become more sophisticated. They also pose an interesting regulatory challenge – perhaps future clinical trials will include an arm that is not actual patients but a cohort of digital twins used as a control to augment statistical power. The FDA and other agencies are already exploring how such in silico evidence might be used in supporting drug approval.

Another important future trend is ethical AI and governance in drug discovery. The very power of AI to design potent bioactive molecules comes with a dual-use risk: the same algorithms that can create a life-saving drug could, in theory, be directed to design harmful compounds (e.g., chemical warfare agents or toxins). A startling proof-of-concept was demonstrated when researchers took a generative model trained to avoid toxicity and simply inverted its objective – in less than six hours, the AI generated tens of thousands of molecules predicted to be more toxic than VX (a deadly nerve agent) . Many known chemical warfare agents were rediscovered in the process, along with novel ones that were computationally predicted to be extremely lethal . This experiment, meant as a wake-up call, underscores the need for responsible AI use. The scientific community is now discussing guidelines and safeguards – for instance, controlling access to models or adding ethical "tripwires" that prevent certain objective functions from being pursued . On the regulatory side, agencies are grappling with how to evaluate AI-designed drugs. While a drug molecule itself can be assessed by the usual pharmacological and toxicological methods, regulators are interested in the provenance: was it designed using an AI model trained on data that might bias it toward a certain population? Does the AI design process raise any red flags that wouldn't be apparent from conventional analysis? The FDA has released discussion papers on the use of AI in drug development and is building internal expertise to review AI-related submissions (Artificial Intelligence for Drug Discovery: Are We There Yet?). We can expect future submissions of Investigational New Drugs (INDs) or New Drug Applications (NDAs) to include sections describing any AI methods used, their validation, and possibly even the model code or parameters as part of the documentation.

Explainable and Causative AI: The next generation of AI models will likely incorporate more mechanistic insight – blending

data-driven learning with causal reasoning. Instead of just correlating patterns, they will attempt to infer causal relationships (for example, that activating receptor X causes downstream pathway Y to induce disease Z). Efforts like causal inference algorithms and reinforcement learning environments that simulate biological processes are steps in this direction. This is crucial for drug discovery because we ultimately want to intervene in a system (human biology) in a causal way. A model that merely predicts "Compound A will lower biomarker B" is less useful than one that understands "Compound A binds receptor X which in turn modulates B via pathway C" because the latter allows extrapolation and anticipation of side effects (since receptor X might also affect other pathways). There is a push towards knowledge graphs and symbolic AI integration, where known biological knowledge (pathway maps, ontologies, etc.) is combined with statistical AI. This can make AI's suggestions more interpretable and grounded. For example, an AI might suggest a polypharmacology strategy: hit target M and target N together for synergistic effect. If this suggestion comes out of the blue, it's hard to trust. But if the AI can contextualize it: "Target M and N are both in a feedback loop controlling cell death; inhibiting M alone triggers compensation via N, so dual inhibition is needed" – which it might derive from incorporating pathway knowledge – then the suggestion carries weight and can be rationalized through existing biological frameworks.

Finally, we foresee an expansion of AI in clinical development and post-market monitoring. The role of AI won't stop when a drug is discovered; it will continue as drugs are tested and used. AI will help design smarter clinical trials (choosing optimal patient criteria, predicting outcomes to assist in trial go/no-go decisions), and once a drug is on the market, AI will sift through real-world data for patterns (pharmacovigilance, detecting rare adverse events via social media or electronic records). In essence, the drug

discovery pipeline may become an AI-guided continuum from target to clinic to market, with continuous learning. Each stage's data feeds back – for instance, post-market safety data could inform early discovery about what scaffolds to avoid. This closes the loop in a way that historically was difficult because of siloed stages. With modern data integration and machine learning, a kind of end-to-end learning system for drug discovery and development becomes conceivable.

In conclusion, the future of computer-aided drug design is being shaped by AI at every level: from pinpointing the right disease target, to designing and optimizing molecules, to predicting and testing how they will behave in biological systems. High-profile successes, like AI-discovered clinical candidates and the cracking of protein folding, have generated justified excitement, but also a healthy recognition of remaining challenges. We are moving toward AI as a collaborative partner in drug discovery – not replacing human insight but amplifying it, uncovering non-intuitive solutions, and handling complexity at scales beyond human cognitive capacity. As algorithms become more powerful and data more abundant, issues of trust, ethics, and validation come to the forefront, ensuring that this technology is used responsibly and effectively. The optimistic vision is that AI will help break the costly logjam of drug development, allowing more therapies to be discovered faster and at lower cost, including treatments for rare or currently "undruggable" diseases that human researchers alone might not crack. It is an exciting era where a deeply technical, highly interdisciplinary approach – combining cheminformatics, molecular biology, advanced computing, and now AI – is redefining what is possible in drug discovery. With careful integration of experimental feedback, continued innovation in algorithms, and a strong ethical compass, AI-driven CADD is poised to significantly accelerate the journey from bytes to bedside.

FURTHER READINGS

Ahmed, M., et al. (2023). From byte to bench to bedside: Molecular dynamics simulations and drug discovery. *BMC Biology, 21*(1), 66. https://doi.org/10.1186/s12915-023-01791-z

Bagherian, M., Saadati, F., Naghibzadeh, M., et al. (2021). Machine learning approaches and databases for prediction of drug–target interactions: A survey paper. *Briefings in Bioinformatics, 22*(1), 247–269. https://doi.org/10.1093/bib/bbz157

Bhat, A. A., Nisar, S., Maacha, S., & Haris, M. (2022). Integration of CRISPR/Cas9 with artificial intelligence for improved cancer therapeutics. *Journal of Translational Medicine, 20*(1), 1–14. https://doi.org/10.1186/s12967-022-03765-1

Cao, Y., Romero, J., Olson, J. P., Degroote, M., Johnson, P. D., Kieferová, M., … & Aspuru-Guzik, A. (2019). Quantum chemistry in the age of quantum computing. *Chemical Reviews, 119*(19), 10856–10915. https://doi.org/10.1021/acs.chemrev.8b00803

Feng, Q., Zhang, Q., Han, W., Li, D., Hu, D., Lu, C., … & Gao, Y. (2024). Machine learning for prediction of synthetic lethality in cancer therapeutics: A comprehensive benchmarking. *Bioinformatics Advances, 4*(1), vbad105. https://doi.org/10.1093/bioadv/vbad105

Guo, Z., Zhu, J., Zheng, S., et al. (2024). Diffusion models in bioinformatics and computational biology. *Nature Reviews Bioengineering, 2*(3), 136–154. https://doi.org/10.1038/s44222-023-00114-9

Guterres, H., & Im, W. (2023). CHARMM-GUI-based induced fit docking workflow to generate reliable protein–ligand binding modes. *Journal of Chemical Information and Modeling, 63*(15), 4772–4779. https://doi.org/10.1021/acs.jcim.3c00416

Heyndrickx, W., Simm, J., Cortes-Ciriano, I., & Bender, A. (2024). MELLODDY: Federated learning in a consortium setting for chemical compound activity prediction. *Journal of Chemical Information and Modeling, 64*(1), 56–72. https://doi.org/10.1021/acs.jcim.3c01654

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., … & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kamya, P., Wang, S., Zhao, Y., & Ozerov, I. (2024). PandaOmics: An AI-driven platform for therapeutic target and biomarker discovery. *Journal of Chemical Information and Modeling, 64*(3), 567–582. https://doi.org/10.1021/acs.jcim.3c01619

Khan, M. Z., et al. (2023). AI-powered therapeutic target discovery. *Trends in Pharmacological Sciences, 44*(11), 818–838. https://doi.org/10.1016/j.tips.2023.06.010

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Liao, Q., et al. (2025). Application of artificial intelligence in drug–target interactions prediction: A review. *NPJ Biomedical Innovations, 2*(1), 1. https://doi.org/10.1038/s44385-024-00003-9

Lim, J., Ryu, S., Park, K., & Kim, W. Y. (2019). Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation. *Journal of Chemical Information and Modeling, 59*(10), 3981–3988. https://doi.org/10.1021/acs.jcim.9b00387

Liu, F., & Panagiotakos, D. (2022). Mining real-world health data for causal insights: A review of AI methods. *Nature Reviews Drug Discovery, 21*(8), 567–584. https://doi.org/10.1038/s41573-022-00376-4

Liu, X., Zhang, W., Wang, L., Li, Z., & Wu, Y. (2024). A probabilistic knowledge graph for therapeutic target identification. *PLOS Computational Biology, 20*(2), e1011945. https://doi.org/10.1371/journal.pcbi.1011945

Loewa, A., Feng, J. J., & Hedtrich, S. (2023). Human disease models in drug development. *Nature Reviews Bioengineering, 1*(7), 545–559. https://doi.org/10.1038/s44222-023-00063-3

Luo, R., Sun, L., Xia, Y., Qin, Z., Zhu, H., Liu, J., & Li, Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *arXiv preprint*, arXiv:2204.06979. https://doi.org/10.48550/arXiv.2204.06979

Madhukar, N. S., Khade, P. K., Huang, L., et al. (2019). A Bayesian machine learning approach for drug target identification using diverse data types. *Nature Communications, 10*, 5221. https://doi.org/10.1038/s41467-019-12928-6

Ozerov, I. V., Mamoshina, P., & Zhavoronkov, A. (2024). AI-enabled hypothesis generation and target prioritization in the PandaOmics platform. *Trends in Pharmacological Sciences, 45*(1), 42–58. https://doi.org/10.1016/j.tips.2023.10.002

Pun, G., Zhang, L., Jin, Q., Yang, Y., Gao, X., & Wu, H. (2023). AI-powered therapeutic target discovery: From big data to precision medicine. *Trends in Pharmacological Sciences, 44*(9), 678–692. https://doi.org/10.1016/j.tips.2023.06.004

Ren, F., et al. (2023). AlphaFold accelerates artificial intelligence powered drug discovery: Efficient discovery of a novel

CDK20 small molecule inhibitor. *Chemical Science, 14*(5), 1443–1452. https://doi.org/10.1039/D2SC05709C

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., … & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine, 3*, 119. https://doi.org/10.1038/s41746-020-00323-1

Srivastava, A. (2023). Quantum annealing for feature selection in cancer classification datasets. *Computational and Structural Biotechnology Journal, 21*, 1231–1240. https://doi.org/10.1016/j.csbj.2023.01.016

Urbina, F., et al. (2022). Dual use of artificial intelligence–powered drug discovery. *Nature Machine Intelligence, 4*(3), 189–191. https://doi.org/10.1038/s42256-022-00465-9

Varadi, M., et al. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research, 50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Walters, W. P., & Murcko, M. (2020). Assessing the impact of generative AI on medicinal chemistry. *Nature Biotechnology, 38*(2), 143–145. https://doi.org/10.1038/s41587-020-0418-2

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., … & Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. *Nature, 620*(7975), 155–163. https://doi.org/10.1038/s41586-023-05934-5

Zeng, X., Zhu, S., Lu, W., et al. (2020). Target identification among known drugs by deep learning from heterogeneous networks. *Chemical Science, 11*(7), 1775–1797. https://doi.org/10.1039/C9SC03666D

Zhang, Y., Chen, X., Fan, J., & Wang, J. (2022). Integrative knowledge graph for biomedical literature mining and target identification. *Bioinformatics, 38*(17), 4089–4097. https://doi.org/10.1093/bioinformatics/btac396

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology, 37*(9), 1038–1040. https://doi.org/10.1038/s41587-019-0224-x