Genome Publications
https://doi.org/10.61096/978-81-990998-7-6_1

Chapter 1

Historical Foundations of Computer-Aided Drug Design: From Trial-and-Error to Rational Discovery

Udaya Kumari Tula
Research scientist, DSK Biopharma Inc, Morrisville, North Carolina, USA
Chinmaya Rath
Lecturer, Usha college of Pharmacy, Dhadkidih,
Behind DC Office P.O Madhurpur, Seraikela - 831013, Jharkhand, India
Abhisek Pradhan
Lecturer, Usha college of Pharmacy, Dhadkidih,
Behind DC Office P.O Madhurpur, Seraikela - 831013, Jharkhand, India

Abstract: The evolution of computer-aided drug design (CADD) represents one of the most profound paradigms shifts in pharmaceutical research, transforming empirical, trial-and-error experimentation into a rational, hypothesis-driven scientific process. From the serendipitous discovery of penicillin in the early 20th century to the structure-based optimization of HIV protease inhibitors and the data-centric revolution driven by artificial intelligence, CADD has continually redefined how molecules are designed, analysed, and optimized for therapeutic efficacy. Pioneering developments such as the determination of myoglobin's crystal structure, the advent of molecular mechanics and quantum calculations, and the introduction of the first quantitative structure–activity relationship (QSAR) models laid the groundwork for modern rational drug discovery. Over subsequent decades, methodologies such as molecular docking, pharmacophore modelling, homology modelling, molecular dynamics simulations, and multi-dimensional QSAR expanded the predictive capacity of in silico research. Contemporary CADD integrates big data analytics, deep learning, and multi-omics data, bridging molecular insights with systems pharmacology. This chapter traces the historical foundations of CADD, outlines its multidisciplinary underpinnings, and contrasts structure-based and ligand-based paradigms. It also evaluates how computational approaches have reshaped drug discovery economics, accelerated innovation, and set the stage for future AI-integrated, ethically governed, and open-science-driven discovery frameworks.

Keywords: Computer-aided drug design, QSAR, molecular docking, structure-based design, artificial intelligence.

Citation: Udaya Kumari Tula, Chinmaya rath, Abhisek Pradhan. Historical Foundations of Computer-Aided Drug Design: From Trial-and-Error to Rational Discovery. *Comprehensive Approaches in Computer-Aided Drug Design: QSAR, Docking, Screening, Homology, Pharmacophore and AI-Driven Insights.* Genome Publication. 2025; Pp1-10. https://doi.org/10.61096/978-81-990998-7-6_1

1.0 INTRODUCTION

Computer-aided drug design (CADD) encapsulates the integration of computational methods with experimental pharmacology to rationalize and accelerate the discovery of therapeutically active compounds. The roots of CADD can be traced to the mid-20th century when the first structural data of biological macromolecules became available through X-ray crystallography. The determination of myoglobin and haemoglobin structures by Kendrew and Perutz in 1958–1960 represented a turning point that enabled scientists to visualize, at atomic resolution, how small molecules interact with protein active sites. These foundational insights laid the conceptual groundwork for structure-based drug design (SBDD) [1]. Before the computational era, drug discovery was predominantly empirical driven by natural product screening and serendipity. The identification of aspirin, sulphonamides, and penicillin relied more on observational pharmacology than molecular understanding. The emergence of theoretical models in the 1960s, particularly the Hansch–Fujita approach, provided the first quantitative framework to correlate chemical structure with biological activity through physicochemical parameters such as lipophilicity (log P), electronic properties (σ constants), and steric effects [2]. This development marked the origin of QSAR and the conceptual birth of rational drug design.

The 1970s and 1980s witnessed the parallel evolution of computational chemistry and molecular modelling. Advances in force field development (CHARMM, AMBER, GROMOS) and quantum mechanical methods enabled energy minimization and conformational analyses that could predict ligand–receptor interactions. The introduction of molecular docking algorithms, notably DOCK (1982) by Kuntz and colleagues, enabled in silico simulation of ligand fitting within protein active sites [3]. These innovations coincided with exponential increases in computational speed, allowing the simulation of increasingly complex biological systems. The completion of the Human Genome Project (2003) and the explosion of structural data within the Protein Data Bank (PDB) revolutionized target-based discovery, ushering in a new era of multi-target and systems pharmacology. CADD approaches became indispensable for hit identification, lead optimization, and virtual screening, significantly reducing the cost and time required for early-phase drug discovery [4]. The integration of machine learning and artificial intelligence since the mid-2010s exemplified by deep learning frameworks like Depeche and AlphaFold has further expanded the boundaries of predictive modelling, moving beyond static representations toward dynamic, data-driven inference [5].

In its modern form, CADD is no longer limited to small molecules. It encompasses peptides, nucleic acid therapeutics, biologics, and even protein–protein interaction inhibitors. The historical trajectory of CADD thus mirrors the evolution of the broader pharmaceutical sciences from serendipity and empiricism to rationality, automation, and artificial intelligence. This chapter aims to capture this continuum, contextualizing CADD as both a scientific discipline and a technological revolution that underpins 21st-century precision pharmacology.

1.1 Multidisciplinary Foundations: Chemical, Biological, and Computational Principles

CADD exists at the intersection of chemistry, biology, and computer science three disciplines whose integration has enabled the rationalization of molecular recognition processes. Its chemical foundation lies in medicinal chemistry and physical organic chemistry, where the principles of thermodynamics, molecular interactions, and structure–activity relationships provide the framework for understanding how ligands modulate biological function. The biological foundation is grounded in receptor theory, enzymology, and structural biology, which define the mechanistic basis of drug–target interactions. The computational foundation encompasses molecular mechanics, quantum chemistry,

data science, and algorithmic modelling that enable the simulation, prediction, and visualization of complex molecular systems [6]. Chemically, drug–receptor interactions are governed by non-covalent forces hydrogen bonding, electrostatic interactions, van der Waals forces, and hydrophobic contacts. Quantitative models derived from physical chemistry, such as the Gibbs free energy of binding ($\Delta G = \Delta H - T\Delta S$), describe the balance between enthalpic and entropic contributions that determine binding affinity. These principles form the foundation of scoring functions used in molecular docking and free energy calculations [7].

From a biological standpoint, the target structure dictates ligand complementarity. Advances in X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) have enabled the elucidation of biomolecular architectures at atomic or near-atomic resolution. These structural datasets are curated within repositories such as the PDB, which now contains over 220,000 entries, serving as the structural backbone for SBDD workflows [8]. Parallel developments in bioinformatics particularly sequence alignment algorithms and homology modelling have allowed the extrapolation of unknown protein structures from homologous templates, broadening the applicability of CADD beyond crystallographic ally resolved proteins. Computationally, the theoretical models underpinning CADD rely on both deterministic and statistical frameworks. Deterministic models, such as molecular mechanics and dynamics, solve the equations of motion for molecular systems based on classical physics, while quantum mechanical approaches (e.g., density functional theory, Hartree–Fock) provide electron-level precision for reaction mechanisms and energy states. Statistical and data-driven models QSAR, pharmacophore modelling, and machine learning capture complex non-linear relationships between molecular features and biological activities [9]. Together, these paradigms enable a unified continuum from atomistic simulations to predictive analytics.

The synthesis of these disciplines created a new scientific ecosystem one where computational predictions guide experimental synthesis and validation. This synergy not only enhances molecular understanding but also allows iterative feedback loops where experimental data refine computational models, leading to continuous improvement in predictive accuracy. Thus, the multidisciplinary foundation of CADD is not static but inherently adaptive integrating innovations from structural biology, theoretical chemistry, and data science to address emerging challenges in modern drug discovery.

1.2 Structure-Based vs Ligand-Based Paradigms

CADD methodologies are broadly categorized into structure-based and ligand-based approaches, distinguished by whether the 3D structure of the biological target is known. These paradigms, while distinct in theoretical foundation, are complementary in practice and often integrated in modern drug discovery pipelines. Structure-Based Drug Design (SBDD) relies on explicit knowledge of the target's three-dimensional structure to model ligand interactions within the binding pocket. The availability of high-resolution structural data allows computational tools to predict binding modes, affinities, and conformational changes upon ligand binding. Techniques under the SBDD umbrella include molecular docking, molecular dynamics simulations, free energy perturbation (FEP) calculations, and fragment-based design [10]. Notable historical milestones include the design of angiotensin-converting enzyme (ACE) inhibitors, such as captopril, which emerged from the structural understanding of the enzyme's zinc-binding site, and the rational development of HIV-1 protease inhibitors, marking one of the first triumphs of SBDD in antiviral therapy [11]. Advances in force fields

and scoring algorithms, along with GPU-accelerated simulations, have further refined the accuracy of binding predictions, enabling virtual screening of millions of compounds against structural targets.

Ligand-Based Drug Design (LBDD), in contrast, is employed when the 3D structure of the target is unknown or experimentally inaccessible. Instead, it leverages the known activities of a series of ligands to infer the molecular features necessary for biological activity. Techniques such as QSAR modelling, pharmacophore identification, and similarity searching form the backbone of LBDD [12]. The QSAR paradigm correlates physicochemical properties with biological effects, while pharmacophore modelling identifies the spatial arrangement of features hydrogen bond donors/acceptors, aromatic rings, hydrophobic centres that underpin receptor binding. The success of β-adrenergic antagonists (β-blockers) and benzodiazepine derivatives illustrates the practical power of ligand-based strategies before the advent of widespread structural data. While SBDD is grounded in physical models of molecular interactions, LBDD depends on statistical inference and pattern recognition. The boundary between the two has increasingly blurred, especially with the advent of hybrid and AI-driven methods. For instance, deep generative models can integrate structural data and ligand activity profiles to design novel compounds that satisfy both geometric and pharmacophoric constraints [13]. Furthermore, machine learning-assisted scoring functions now bridge the gap between physics-based and data-driven paradigms.

In practice, the choice between SBDD and LBDD depends on data availability, target type, and computational resources. Structure-based approaches are favoured for well-characterized protein targets with known binding sites, while ligand-based methods are indispensable for orphan receptors or targets lacking resolved structures. Integrating both strategies where ligand-derived pharmacophoric insights inform docking constraints or where docking results enhance QSAR datasets represents a hallmark of modern CADD workflows. This hybridization has proven particularly effective in identifying allosteric modulators, covalent inhibitors, and multitarget ligands in complex disease networks.

1.3 Roles of Molecular Modelling, QSAR, Docking, Virtual Screening and AI in CADD

The evolution of CADD is tightly linked to five core methodological pillars: molecular modelling, QSAR, molecular docking, virtual screening, and artificial intelligence. Each represents a progressive layer in the computational hierarchy, collectively driving the transition from descriptive chemistry to predictive pharmacology. Molecular modelling serves as the conceptual and practical foundation of CADD. It encompasses methods that visualize and simulate molecular structures to predict conformational behaviour, energetics, and interactions. Early molecular mechanics models such as MM2 and AMBER enabled energy minimization and conformational sampling of small molecules. Molecular dynamics extended this to simulate biomolecular motion in explicit solvent environments, allowing dynamic insight into ligand binding and receptor flexibility [14]. With increasing computational power, all-atom and coarse-grained simulations have become routine, revealing the dynamic plasticity of active sites critical for accurate docking and binding energy estimation.

QSAR represents the earliest form of computational predictive modelling in drug discovery. Since the seminal Hansch analysis, QSAR has evolved into multidimensional paradigms (2D–6D QSAR) that encode topological, spatial, electronic, and temporal molecular features. Methods such as Coma and CoMSIA introduced 3D fields to capture steric and electrostatic influences, while advanced machine learning models random forests, support vector machines, and deep neural networks have enhanced predictive accuracy [15]. QSAR remains integral for activity prediction, ADMET profiling, and

virtual screening filtering, particularly when experimental data are limited. Molecular docking bridges the structural and energetic dimensions of drug design. It predicts how ligands fit into target binding sites, estimating pose and binding affinity through scoring functions that combine empirical and physics-based terms. Docking software such as Auto Dock, Glide, GOLD, and DOCK have become industry standards, routinely screening millions of compounds in silico before synthesis. Docking outputs are often refined through molecular dynamics simulations and free energy calculations (MM–GBSA, FEP) for higher accuracy [16].

Virtual screening (VS) integrates QSAR, docking, and pharmacophore modelling into large-scale computational searches of compound libraries. VS allows prioritization of candidate molecules based on predicted activity, significantly reducing experimental screening costs. Structure-based virtual screening (SBVS) utilizes docking against known targets, whereas ligand-based virtual screening (LBVS) exploits molecular similarity and pharmacophore matching. The incorporation of cloud computing and distributed databases such as ZINC, Chambly, and Enamine REAL has expanded the accessible chemical space to billions of compounds, enabling global-scale hit discovery [17]. Finally, artificial intelligence (AI) represents the newest and most transformative layer in CADD. AI-driven systems employ deep learning, graph neural networks, and reinforcement learning to predict drug–target interactions, generate novel scaffolds, and optimize physicochemical properties. Platforms such as Depeche, Atom Net, and AlphaFold2 have demonstrated unprecedented accuracy in molecular property prediction and protein structure determination [18]. Beyond prediction, AI enables generative design creating entirely new molecules with desired activity and ADMET profiles. The integration of AI with molecular modelling and experimental feedback forms the foundation of next-generation autonomous drug discovery pipelines.

Together, these components illustrate CADD's evolution from descriptive modelling to predictive and generative intelligence. Each methodological advance has incrementally expanded the scope of what can be rationally designed, bringing the pharmaceutical sciences closer to the long-envisioned goal of fully in silico drug discovery.

1.4 Socio-Economic Impact: Cost, Speed and Success Rates in Drug Discovery

The incorporation of computer-aided drug design (CADD) into pharmaceutical pipelines has fundamentally reshaped the economic and temporal dimensions of drug discovery. Historically, the average cost of bringing a new drug to market has been estimated between USD 1.5 and 2.5 billion, with timelines extending up to 12–15 years [19]. This burden is largely attributed to high attrition rates where more than 90% of preclinical candidates fail during clinical development due to inadequate efficacy or unanticipated toxicity. CADD mitigates these inefficiencies by introducing predictive, hypothesis-driven workflows that minimize redundant synthesis and testing. One of the most significant socio-economic benefits of CADD is its role in **early-phase triage**. Virtual screening and QSAR modelling enable the pre-selection of candidates with desirable physicochemical and pharmacokinetic profiles before laboratory synthesis. This computational pre-filtering drastically reduces the number of compounds entering costly experimental pipelines, improving the hit-to-lead ratio [20]. For example, large-scale virtual screening campaigns against SARS-CoV-2 main protease during the COVID-19 pandemic demonstrated how in silico modelling can identify viable hit molecules within weeks an achievement that would have taken months using traditional screening approaches [21].

Furthermore, structure-based approaches enhance lead optimization by providing atomistic insight into binding interactions. This precision facilitates rational modification of chemical scaffolds to

improve potency, selectivity, and metabolic stability, thereby decreasing the number of synthesis iterations required. Pharmaceutical companies such as Pfizer, Merck, and Novartis have reported significant reductions in cycle times for target-to-lead programs using molecular docking and pharmacophore modelling tools integrated with AI-driven analytics [22]. CADD also contributes to cost reduction through repurposing and de-risking. Computational repurposing strategies exploit existing drugs' structural and pharmacological data to identify new therapeutic indications, offering a cost-effective alternative to de novo discovery. The repurposing of thalidomide for multiple myeloma and sildenafil for pulmonary hypertension exemplifies the clinical and economic viability of such strategies [23]. AI-enhanced network pharmacology now enables identification of multi-target interactions, further extending the value of approved drugs across new indications.

From a broader socio-economic perspective, the democratization of computational tools and open-access databases has made CADD accessible to academic institutions and startups, decentralizing innovation traditionally confined to major pharmaceutical companies. Platforms such as OpenMP, Depeche, and Swiss Dock empower resource-limited laboratories to conduct high-quality virtual experiments at minimal cost. Consequently, CADD not only accelerates innovation but also fosters inclusivity and global collaboration within the scientific community. In terms of measurable outcomes, analyses have shown that integrating CADD can reduce early discovery timelines by up to 30–50% and overall costs by approximately 20–40%, depending on target complexity and data availability [24]. Although these estimates vary across therapeutic domains, the consistent trend highlights computational drug design as a key enabler of efficiency, reproducibility, and sustainability in modern pharmaceutical R&D.

1.5 Current Challenges: Chemical Space Exploration, Data Integration and Biases

Despite its transformative potential, CADD faces persistent scientific and technical challenges. Chief among these are the vastness of chemical space, limitations in data quality and integration, and the propagation of algorithmic biases that influence model reliability. The chemical space problem arises from the nearly infinite number of theoretically possible small molecules estimated at over 10^{60} compounds [25]. Even the largest virtual libraries, such as Enamine REAL or GDB-17, cover only an infinitesimal fraction of this space. Sampling bias, restricted by existing chemistries and available descriptors, often leads to a narrow exploration of structural diversity. Although deep generative models and reinforcement learning have expanded the ability to propose novel scaffolds, ensuring synthetic feasibility and pharmacological relevance remains a major bottleneck [26]. Equally critical are data integration challenges. The explosion of omics data genomics, transcriptomics, proteomics, and metabolomics offers unprecedented insights into biological complexity. However, the lack of standardized data formats, inconsistent annotation, and incomplete metadata often hinder the creation of unified predictive models [27]. For example, integrating gene expression data with ligand-binding affinities demands sophisticated normalization and machine-learning techniques capable of handling high-dimensional, heterogeneous datasets. Efforts such as FAIR (Findable, Accessible, Interoperable, Reusable) data principles have attempted to establish common standards, yet adoption across the pharmaceutical ecosystem remains inconsistent.

Algorithmic and dataset biases further complicate the reproducibility of computational predictions. Machine learning models trained on skewed or under-representative datasets can overfit to specific chemotypes or physicochemical patterns, reducing their generalizability. A well-known example is the overrepresentation of hydrophobic ligands in public databases, which leads to an inflated prediction of lipophilic binding preferences [28]. Similarly, structural redundancy in training

sets can produce artificially high cross-validation metrics, masking poor real-world performance. Addressing these issues requires stringent curation protocols, external validation, and transparent reporting of dataset composition. Finally, the interpretability problem persists across deep learning applications in CADD. Although AI models demonstrate remarkable predictive accuracy, their decision-making processes often remain opaque a limitation for regulatory validation and scientific acceptance. Emerging frameworks for explainable AI (XAI), including attention visualization and feature attribution methods, are beginning to illuminate how molecular substructures influence predictions, thereby improving user trust [29]. Nonetheless, balancing model complexity with interpretability continues to be a delicate trade-off in next-generation drug design pipelines.

1.6 Ethical, Regulatory and Open-Science Considerations

As CADD systems increasingly influence real-world pharmaceutical decision-making, ethical and regulatory oversight becomes paramount. Computational predictions if improperly validated can lead to resource misallocation or, in extreme cases, unsafe clinical outcomes. Hence, responsible data stewardship, transparency, and regulatory harmonization are essential for sustaining public trust in AI-augmented discovery frameworks. Ethically, the use of proprietary patient-derived datasets and genomic information raises questions about data privacy, consent, and ownership. Adherence to international regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) is critical when integrating clinical or pharmacogenomic data into CADD workflows [30]. Federated learning and encrypted computation techniques have emerged as promising solutions, allowing collaborative model training without direct data exchange.

From a regulatory standpoint, agencies like the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have begun formalizing guidelines for in silico model validation. The FDA's "Model-Informed Drug Development" (MIDD) initiative emphasizes the use of computational models to support dose selection, risk assessment, and biomarker qualification. These frameworks demand transparency in algorithm design, version control, and documentation of training datasets [31]. Similarly, the Organization for Economic Co-operation and Development (OECD) has proposed validation principles for (Q)SAR models, emphasizing reproducibility, defined applicability domains, and mechanistic interpretability [32]. The open-science movement has further catalysed ethical and methodological progress in CADD. Open-access repositories such as Chambly, PubChem, and the PDB democratize access to structural and activity data, while collaborative projects like Folding home and COVID Moonshot illustrate the power of community-driven discovery. However, open models must balance transparency with data integrity ensuring that democratization does not compromise reliability or intellectual property rights.

An emerging dimension of ethical consideration involves the environmental sustainability of computation. High-performance simulations and deep learning models consume substantial energy, contributing to carbon emissions. Efforts toward green computing through energy-efficient hardware, optimized algorithms, and cloud-based workload distribution reflect an evolving awareness of sustainability in digital pharmaceutical science [33]. Ultimately, ethical governance in CADD extends beyond compliance to encompass fairness, inclusivity, and reproducibility. As AI and automation continue to dominate discovery pipelines, embedding ethical reflexivity within algorithm design, validation, and dissemination will determine whether CADD fulfils its promise of socially responsible scientific advancement.

Table 1. Representative Structural Databases and Their Key Features

Database	Type	Data Contents	Typical Applications	Access/Source
Protein Data Bank (PDB)	Protein structures	3D atomic coordinates, ligands	Structure-based drug design, docking	https://www.rcsb.org
ChEMBL	Bioactivity data	Ligand activity, target binding data	QSAR modeling, screening	https://www.ebi.ac.uk/chembl
PubChem	Small molecules	Chemical structures, properties, assays	Virtual screening, similarity search	https://pubchem.ncbi.nlm.nih.gov
BindingDB	Binding affinities	Kd, Ki, IC50 values	Affinity modeling, benchmarking	https://www.bindingdb.org
DrugBank	Approved and experimental drugs	Structures, pharmacology, ADMET	Drug repurposing, target analysis	https://go.drugbank.com
ZINC15	Virtual compound libraries	Ready-to-dock molecules	Structure-based virtual screening	https://zinc15.docking.org

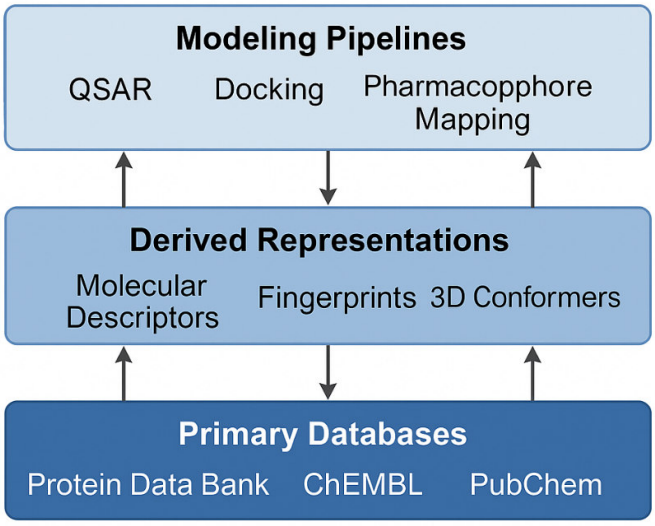

Figure 1. Hierarchical Framework of Data Foundations in Computer-Aided Drug Design

1.7 Chapter Summary and Structure of the Book

The historical trajectory of computer-aided drug design illustrates a transition from empiricism to rationality, from structure elucidation to predictive modelling, and from isolated experimentation to integrated computational ecosystems. Beginning with the early QSAR models of the 1960s, CADD evolved through decades of interdisciplinary convergence merging chemical intuition with structural biology and computational mathematics. This evolution enabled not only the visualization of molecular interactions but also their quantitative prediction across diverse biological contexts. Today, CADD represents a comprehensive framework encompassing molecular modelling, QSAR, docking, pharmacophore mapping, molecular dynamics, virtual screening, and AI-driven generative design. Each methodology contributes a unique layer of understanding, collectively enabling accelerated and cost-effective drug discovery. Yet, persistent challenges including the exploration of uncharted chemical space, biases in data-driven models, and ethical implications of AI decision-making highlight the need for continuous refinement and critical oversight.

1.8 CONCLUSION

The conclusion will summarize how data integrity, standardization, and interoperability underpin the entire field of computer-aided drug design (CADD). It will emphasize the continuum from raw structural data (e.g., Protein Data Bank, PubChem, ChEMBL) to derived representations (molecular descriptors, fingerprints, 3D conformers) and finally to computational models (QSAR, docking, molecular dynamics). The section will highlight that the accuracy of any in silico prediction is only as reliable as the quality of its underlying data. I t will also address emerging transformations AI-ready datasets, FAIR-compliant repositories, and automated data curation pipelines that now drive reproducibility and scalability in computational discovery. The conclusion will close by asserting that data-centric thinking is no longer peripheral but central to modern drug design, serving as the bridge that links computational innovation with experimental validation and translational pharmacology.

REFERENCES

1. Kendrew JC, Bodo G, Dentzes HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature.* 1958;181:662–666.
2. Hansch C, Fujita T. ρ–σ–π Analysis. A method for the correlation of biological activity and chemical structure. *J Am Chem Soc.* 1964;86(8):1616–1626.
3. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule–ligand interactions. *J Mol Biol.* 1982;161(2):269–288.
4. Berman HM, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000;28(1):235–242.
5. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596:583–589.
6. Leach AR. *Molecular Modelling: Principles and Applications.* 3rd ed. Pearson; 2010.
7. Gilson MK, Zhou HX. Calculation of protein–ligand binding affinities. *Annu Rev Biopsy's Bismol Struct.* 2007;36:21–42.
8. Burley SK, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures. *Nucleic Acids Res.* 2024;52(D1):D494–D503.
9. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatic.* 2nd ed. Wiley-VCH; 2020.
10. Jorgensen WL. The many roles of computation in drug discovery. *Science.* 2004;303(5665):1813–1818.
11. Woodware A, Vondrasek J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu Rev Biopsy's Bismol Struct.* 1998;27:249–284.

12. Gohlke H, Klebe G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Agnew Chem Int Ed.* 2002;41(15):2644–2676.
13. Zamorano A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol.* 2019;37(9):1038–1040.
14. Van Der Sopel D, et al. GROMACS: Fast, flexible, and free. *J Compute Chem.* 2005;26(16):1701–1718.
15. Cherkasov A, et al. QSAR modelling: where have you been? Where are you going to? *J Med Chem.* 2014;57(12):4977–5010.
16. Trott O, Olson AJ. Auto Dock Vina: improving the speed and accuracy of docking with a new scoring function. *J Compute Chem.* 2010;31(2):455–461.
17. Sterling T, Irwin JJ. ZINC 15 – Ligand discovery for everyone. *J Chem Inf Model.* 2015;55(11):2324–2337.
18. Walters WP, Barzilay R. Applications of deep learning in molecule generation and molecular property prediction. *Acs Chem Res.* 2021;54(2):263–270.
19. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J Health Econ.* 2016;47:20–33.
20. Schneider G. Automating drug discovery. *Nat Rev Drug Disco.* 2018;17(2):97–113.
21. Machiguenga M, Pagliai M, Procacci P. In silico study of COVID-19 main protease inhibitors. *Sci Rep.* 2020;10:20069.
22. Sliwoski G, Kothiwale S, Meiler J, Lowe EW Jr. Computational methods in drug discovery. *Pharmacal Rev.* 2014;66(1):334–395.
23. Pushpakar S, et al. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Disco.* 2019;18(1):41–58.
24. Paul SM, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Disco.* 2010;9(3):203–214.
25. Reymond JL. The chemical space project. *Acs Chem Res.* 2015;48(3):722–730.
26. Ståhl N, Falkman G, Karlsson A, Mathiason G, Boström J. Deep reinforcement learning for multiparameter optimization in de novo drug design. *J Chem Inf Model.* 2019;59(7):3166–3176.
27. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol.* 2017;18(1):83.
28. Xu Y, Dai Z, Chen F, Gao S, Pei J, Lai L. Deep learning for drug-induced liver injury. *J Chem Inf Model.* 2015;55(10):2085–2093.
29. Jiménez-Luna J, Grison F, Schneider G. Drug discovery with explainable artificial intelligence. *Nat Mach Intel.* 2020;2(10):573–584.
30. Mittelstadt BD, Florida L. The ethics of biomedical big data. *Synthase.* 2016;194(5):1741–1778.
31. FDA. *Model-Informed Drug Development Pilot Program.* Silver Spring: U.S. Food and Drug Administration; 2023.
32. OECD. *Principles for the Validation, for Regulatory Purposes, of (Q)SAR Models.* Paris: OECD Publishing; 2014.
33. Strobel E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. *ACL Proceedings.* 2019;3645–3650.