Genome Publications
https://doi.org/10.61096/978-81-990998-7-6_2

Chapter 2

Data Foundations in CADD: Structural Databases, Descriptors and Force Fields

Udaya Kumari Tula
Research scientist, DSK Biopharma Inc, Morrisville, North Carolina, USA
G. Sumithira
Professor, Department of Pharmacology,
The Erode College of Pharmacy and Research Institute, Erode – 638112
Dr. VS. Thiruvengadarajan
Professor, KMCH College of Pharmacy, Coimbatore, Tamilnadu, India

Abstract: Computer-aided drug design (CADD) fundamentally relies on the systematic representation, management, and interpretation of molecular data. The field's evolution from empirical chemistry to predictive modeling is anchored in data-centric foundations comprehensive structural databases, numerical descriptors that encode chemical and biological information, and force fields that approximate molecular energetics. This chapter explores these pillars in depth, outlining how curated repositories such as PDB, ChEMBL, PubChem, and ZINC enable reproducible discovery and model training. It examines molecular descriptors from simple constitutional counts to advanced quantum-derived and hybrid fingerprints emphasizing their critical role in quantitative structure–activity relationship (QSAR) modeling and virtual screening. Force fields, representing the physical basis of molecular mechanics, are discussed as engines that convert chemical structures into energetically meaningful configurations. Collectively, these elements form the data–model continuum that sustains modern in silico drug discovery. The chapter concludes with an integrated discussion of database–descriptor–force field interoperability, data quality issues, and emerging trends toward AI-driven, interoperable, and FAIR-compliant CADD ecosystems.

Keywords: Structural Databases, Molecular Descriptors, Force Fields, Chemoinformatics, Computer-Aided Drug Design.

Citation: Udaya Kumari Tula, G. Sumithira, VS. Thiruvengadarajan. Data Foundations in CADD: Structural Databases, Descriptors and Force Fields. *Comprehensive Approaches in Computer-Aided Drug Design: QSAR, Docking, Screening, Homology, Pharmacophore and AI-Driven Insights.* Genome Publication. 2025; Pp11-22. https://doi.org/10.61096/978-81-990998-7-6_2

2.0 INTRODUCTION

The progress of computer-aided drug design (CADD) depends on the availability, quality, and interpretability of molecular and biological data. Each computational workflow from ligand-based screening to structure-based modeling begins with data that describe the chemical space, biological targets, and physicochemical interactions underlying drug–target binding. In the modern pharmaceutical landscape, the predictive power of any CADD model is determined less by algorithmic sophistication than by the integrity and diversity of its input data. This principle underlies the concept of data-driven discovery, in which curated repositories, standardized descriptors, and physically consistent force fields together form the triad of CADD foundations [1]. Historically, molecular modeling in the 1970s relied on manually drawn structures and energy calculations using simplified empirical equations. The introduction of the Protein Data Bank (PDB) in 1971 provided the first standardized format for storing biomolecular structures, while the 1990s witnessed the emergence of large-scale chemical libraries such as PubChem and ZINC [2]. These repositories enabled reproducibility, data mining, and the training of statistical and machine learning models for property and activity prediction. Today, the scale of CADD data is unprecedented: millions of experimentally validated compounds, thousands of resolved protein–ligand complexes, and petabytes of molecular dynamics (MD) simulation data are openly available [3].

The conceptual foundation of CADD data can be viewed as three interlinked layers. The first layer, structural databases, stores the atomic coordinates and physicochemical annotations of small molecules and macromolecules. The second layer, molecular descriptors, translates structures into numerical representations amenable to statistical learning and QSAR modeling. The third layer, force fields, encapsulates the physicochemical interactions between atoms, serving as the computational analog of potential energy surfaces. Together, these layers convert chemical intuition into computational knowledge, allowing predictions of binding affinities, conformational dynamics, and drug-likeness to be made with remarkable precision [4]. As the pharmaceutical industry increasingly embraces data-centric discovery, these foundations are being reshaped by artificial intelligence, graph-based molecular encodings, and cloud-integrated repositories that support real-time curation and cross-database interoperability. The following sections detail each component structural databases, molecular descriptors, and force fields demonstrating how they underpin predictive modeling and rational drug discovery.

2.1 Structural Databases: Chemical, Biological and Hybrid Repositories

Structural databases are the backbone of CADD, providing the standardized and validated data required for model building, benchmarking, and reproducibility. They are broadly classified into chemical structure databases, which archive small-molecule compounds, and biological structure databases, which store macromolecular targets such as proteins, nucleic acids, and complexes. Hybrid repositories integrate both, enabling structure–activity mapping across molecular hierarchies [5]. Chemical Databases such as PubChem, ChEMBL, ZINC, and DrugBank represent distinct yet complementary paradigms. PubChem, maintained by the National Center for Biotechnology Information (NCBI), houses over 110 million compounds with biological assay results, making it an indispensable source for activity data [6]. ChEMBL, curated by the European Bioinformatics Institute (EBI), provides manually verified compound–target–activity relationships, particularly valuable for QSAR model training [7]. ZINC, developed at UCSF, serves as a repository of commercially available compounds formatted for virtual screening, providing three-dimensional (3D) structures in multiple

protonation and tautomeric states [8]. DrugBank bridges experimental and clinical data, linking molecular structures with pharmacokinetic, pharmacodynamic, and regulatory information.

Biological Databases provide atomic-level insights into target macromolecules. The Protein Data Bank (PDB) remains the primary repository, containing over 220,000 experimentally determined protein, nucleic acid, and complex structures [9]. Advances in cryo-electron microscopy (cryo-EM) have expanded this dataset beyond crystallographic constraints, enabling near-atomic resolution for flexible and membrane-bound proteins. Complementary resources such as UniProtKB, which provides protein sequence and functional annotation, and BindingDB, which aggregates experimentally determined binding affinities, create the essential link between structural and biochemical data [10]. Hybrid and Derived Databases such as PDBbind, Binding MOAD, and BioLip extract protein–ligand complexes and their binding energies, enabling the benchmarking of docking and scoring algorithms [11]. These datasets serve as gold standards for validating CADD workflows, facilitating reproducible comparisons across different force fields and scoring functions. Other integrative repositories, including ChemBL–PDB crosslinks and AlphaFold Protein Structure Database, provide computationally predicted protein structures for targets lacking experimental data, greatly expanding the accessible structural space [12].

Recent developments emphasize FAIR data principles (Findable, Accessible, Interoperable, Reusable), ensuring that structural data can be effectively shared and reused across platforms. Metadata standards (e.g., SDF, MOL2, PDBx/mmCIF formats) and RESTful APIs have facilitated automated workflows where molecular structures are directly retrieved and analyzed within CADD software. As open data ecosystems evolve, the challenge shifts from data scarcity to data quality and standardization, making curation and validation critical aspects of any computational pipeline [13].

2.2 Molecular Descriptors: Quantitative Encodings of Chemical Structure

Molecular descriptors translate complex chemical structures into mathematical forms that capture their physicochemical essence. They serve as the bridge between raw chemical data and predictive models, enabling algorithms to infer structure–activity relationships. A molecular descriptor can be defined as a numerical value derived from a chemical structure that quantitatively represents one or more of its properties such as size, shape, hydrophobicity, or electronic distribution [14]. Descriptors are essential in QSAR, QSPR (Quantitative Structure–Property Relationship), and machine learning applications across drug design, toxicology, and material science. They provide a means to compare compounds, measure similarity, and construct models correlating structure with biological activity. The explosion of chemoinformatics software (e.g., RDKit, Dragon, PaDEL, CDK) has enabled the calculation of thousands of descriptors from a single molecule, spanning from simple counts (atoms, bonds) to complex quantum-chemical parameters [15].

The generation of descriptors typically follows a multi-step process: (i) Structure Standardization, where tautomers, stereochemistry, and protonation states are normalized; (ii) Feature Extraction, calculating descriptors based on molecular graph theory, 3D geometry, or quantum mechanics; and (iii) Feature Selection, where redundant or non-informative descriptors are removed to prevent overfitting in predictive models. Statistical and machine learning methods such as principal component analysis (PCA), recursive feature elimination (RFE), or mutual information are frequently applied to optimize descriptor sets [16]. The interpretability of descriptors is equally critical. While deep-learning-based encodings (e.g., molecular fingerprints, graph embeddings) have gained prominence, classical descriptors remain indispensable for mechanistic insight. For instance, hydrophobic descriptors (e.g., logP) explain membrane permeability, while electronic descriptors (e.g.,

HOMO–LUMO gap) elucidate reactivity trends [17]. Thus, descriptor design must balance interpretability and predictive performance.

Beyond individual molecules, global descriptors can capture dataset-level characteristics such as chemical diversity, scaffold complexity, and physicochemical coverage. These metrics guide library design and virtual screening campaigns. Moreover, descriptor computation forms the foundation of automated pipelines integrating with structural databases retrieving compounds, computing features, and feeding them into QSAR or docking workflows seamlessly [18].

2.3 Descriptor Categories: Constitutional, Topological, Geometrical, Electronic and Hybrid

Descriptors are systematically classified based on the nature of the information they encode and the level of structural detail they require. The five principal categories constitutional, topological, geometrical, electronic, and hybrid together provide a multiscale representation of molecular properties suitable for diverse CADD applications [19]. Constitutional descriptors are the simplest, derived directly from molecular formulae or connectivity tables without considering geometry. Examples include molecular weight, atom count, number of hydrogen bond donors or acceptors, and rotatable bonds. They are fast to compute and useful for rule-based filters such as Lipinski's "rule of five" for drug-likeness evaluation [20]. However, they fail to capture 3D conformational or electronic nuances.

Topological descriptors encode molecular connectivity through graph-theoretical indices. Notable examples are the Wiener index, Balaban index, and Kier–Hall electrotopological states. These descriptors quantify branching, cyclicity, and electronic influence propagation through the molecular graph. They are particularly useful in similarity searching and 2D-QSAR modeling, offering a balance between interpretability and computational simplicity [21]. Geometrical descriptors incorporate 3D information derived from spatial coordinates. Parameters such as molecular volume, surface area, dipole moment, and shape indices belong to this group. They are essential for modeling steric interactions, receptor–ligand complementarity, and binding affinity estimation in 3D-QSAR and docking studies [22].

Electronic descriptors capture charge distribution and reactivity-related properties. Quantum-chemical calculations provide quantities such as HOMO/LUMO energies, Mulliken charges, polarizability, and electrostatic potential surfaces. Although computationally intensive, these descriptors correlate strongly with molecular recognition and chemical reactivity patterns, making them indispensable in mechanistic drug design [23]. Finally, hybrid descriptors combine multiple categories or integrate experimental data with computational parameters. For example, 4D-fingerprints encode atomic interactions across conformations, while pharmacophore-based fingerprints integrate steric and electronic features relevant to bioactivity [24]. Recent AI-driven representations, such as graph neural network embeddings and message-passing fingerprints, extend this hybridization further, generating latent descriptors directly from molecular graphs that can be fine-tuned for specific predictive tasks [25].

Collectively, these categories form the quantitative backbone of CADD. Selecting the appropriate descriptor type depends on the target property, computational budget, and interpretability requirements. In advanced workflows, multiple descriptor classes are fused to form multimodal feature spaces, improving the generalizability and robustness of predictive models.

2.4 Force Fields and Molecular Mechanics: Foundations of Energetic Modelling

Force fields constitute the physical core of molecular mechanics, describing how atoms and molecules interact through potential energy functions derived from both empirical and theoretical foundations. In computer-aided drug design (CADD), force fields allow the conversion of static molecular structures into dynamic, energetically consistent systems that approximate real-world behavior. They underpin molecular docking, molecular dynamics (MD) simulations, and free energy calculations providing the mechanistic link between structure and function [26]. Bonded terms model the stretching of bonds, bending of angles, and torsional rotations, typically represented by harmonic or cosine functions. Non-bonded terms include van der Waals interactions modeled using Lennard–Jones potentials and electrostatic interactions derived from Coulomb's law. Together, they define the potential energy surface (PES) governing molecular stability and motion [27].

Classical force fields such as AMBER, CHARMM, OPLS-AA, and GROMOS have become standards in biomolecular simulation. Each is defined by unique parameter sets for bond lengths, force constants, and partial atomic charges optimized to reproduce experimental and quantum-mechanical data [28]. For example, AMBER (Assisted Model Building with Energy Refinement) emphasizes biomolecules like proteins and nucleic acids, while OPLS-AA (Optimized Potentials for Liquid Simulations) is widely applied to small organic molecules. CHARMM (Chemistry at HARvard Macromolecular Mechanics) offers a flexible framework with a broad range of lipid and carbohydrate parameters, and GROMOS (GROningen MOlecular Simulation) is known for its efficiency in aqueous systems [29]. Modern developments have extended these classical formulations into polarizable force fields, which dynamically adjust atomic charges in response to changing electrostatic environments, capturing effects like induction and polarization more accurately. Examples include AMOEBA and Drude Oscillator models, which have shown improved agreement with experimental binding energies [30].

The selection of an appropriate force field depends on the molecular system, target property, and computational resources. For instance, coarse-grained force fields like MARTINI simplify atomistic details to accelerate simulations of large biomolecular assemblies, while quantum mechanics/molecular mechanics (QM/MM) hybrid methods couple quantum accuracy with classical efficiency for active site modeling. In CADD, these formulations collectively enable virtual experiments such as ligand binding, conformational sampling, and energy minimization under realistic physical conditions [31].

2.5 Parameterization and Validation of Force Fields

Force field parameterization is a critical process ensuring that calculated energies, geometries, and dynamic behaviors align with experimental or high-level quantum-mechanical results. Parameters are derived through fitting procedures that minimize the difference between computed and reference data for small representative molecules. These reference datasets include vibrational spectra, lattice energies, hydration free energies, and conformational preferences [32]. Parameter optimization typically follows a hierarchical approach. Initially, bonded parameters (bonds, angles, torsions) are fitted to quantum-mechanical potential energy scans, while non-bonded parameters (Lennard–Jones coefficients, partial charges) are adjusted to reproduce macroscopic observables such as densities, heats of vaporization, and solvation energies. Tools such as Antechamber (for AMBER), CGenFF (for CHARMM), and LigParGen (for OPLS) automate this process, providing transferable parameters for small organic ligands in drug discovery contexts [33].

Validation is as crucial as parameterization. A well-parameterized force field must reproduce structural properties (bond lengths, RMSD distributions), thermodynamic properties (enthalpies, free energies), and dynamic behaviors (diffusion, conformational sampling) across diverse systems. Benchmarking against experimental datasets like PDBbind or thermodynamic databases (e.g., FreeSolv, ThermoML) ensures generalizability beyond the training set [34]. Challenges arise from the trade-off between transferability and accuracy. Force fields optimized for proteins may perform poorly for nucleic acids or small molecules, necessitating domain-specific variants. Additionally, fixed-charge models inherently neglect electronic polarization, leading to inaccuracies in highly charged or flexible systems. Emerging methodologies such as machine-learned force fields (MLFFs) address these limitations by training neural networks on quantum-mechanical data to reproduce potential energy surfaces with near-ab initio precision at classical computational cost [35].

Validation metrics such as root-mean-square deviation (RMSD), mean unsigned error (MUE), and correlation coefficients between experimental and computed energies quantify performance. In modern workflows, automated benchmarking pipelines like OpenFF Evaluator and ForceBalance allow reproducible, community-wide validation, ensuring that newly developed parameters meet rigorous accuracy standards [36]. The trend toward open, interoperable force fields e.g., OpenFF (Open Force Field Initiative) exemplifies the convergence of data science, physics, and community-driven reproducibility. These collaborative frameworks use machine learning, Bayesian inference, and quantum mechanical data to continually refine force field parameters for small molecules, marking a paradigm shift toward adaptive, data-centric molecular mechanics [37].

2.6 Interfacing Databases, Descriptors and Force Fields in CADD Workflows

In a modern CADD pipeline, structural databases, molecular descriptors, and force fields interact as interconnected modules within an integrated computational ecosystem. The workflow typically begins with data acquisition from chemical or biological databases, proceeds to feature extraction through descriptor computation, and culminates in energetic modeling using molecular mechanics or docking algorithms guided by force fields [38]. For instance, in a structure-based drug design (SBDD) workflow, protein structures are retrieved from PDB or AlphaFold databases, and ligands are selected from ChEMBL or ZINC. These structures are standardized (e.g., protonation, tautomerization), and descriptors such as molecular weight, hydrophobicity, or 3D pharmacophoric patterns are computed. Subsequently, molecular docking simulations apply force field–derived potentials (e.g., AMBER or OPLS) to predict binding poses and estimate interaction energies [39].

In ligand-based workflows, descriptors derived from chemical databases inform QSAR or machine learning models, predicting activity or ADMET properties. The integration of these models with molecular mechanics simulations refines predictions by accounting for conformational dynamics and energetics. The ability to seamlessly connect structural and numerical representations ensures predictive continuity across scales from atomistic to statistical modeling [40]. Data interoperability is achieved through standardized file formats and APIs. Structural data are typically stored in SDF, MOL2, or PDB formats; descriptors in CSV or JSON; and force field parameters in XML or topology files (e.g., PRMTOP, PSF, TOP). Software frameworks such as KNIME, Pipeline Pilot, and OpenMM allow visual or script-based integration, while scripting languages like Python facilitate automation via RDKit, MDAnalysis, and ParmEd libraries [41]. Recent advances emphasize cloud-based CADD ecosystems, where all three layers databases, descriptors, and force fields are orchestrated in real time. Examples include Schrödinger's LiveDesign, DeepChem, and BioSimSpace, enabling dynamic feedback between data sources and simulations. This integrated approach not only improves reproducibility but also

allows active learning, where machine learning models iteratively refine descriptors or force field parameters based on simulation outcomes, creating a closed-loop optimization cycle [42].

2.7 Software Platforms and Computational Pipelines

The efficient handling of vast chemical and biological data necessitates specialized software ecosystems capable of integrating database querying, descriptor computation, and molecular mechanics simulation. Prominent database and descriptor platforms include RDKit, Open Babel, ChemAxon's JChem, PaDEL-Descriptor, and Dragon, each providing thousands of descriptor calculations encompassing 1D–6D representations [43]. These tools facilitate high-throughput feature extraction directly from SMILES or 3D coordinate files, often coupled with data-cleaning modules to handle large compound libraries. For force field–based simulations, GROMACS, AMBER, CHARMM, and OpenMM dominate academic and industrial use. These packages offer comprehensive workflows from structure preparation and energy minimization to long-timescale molecular dynamics and free energy perturbation (FEP) analyses. Interoperability tools like MDAnalysis, ParmEd, and PLUMED enhance cross-platform compatibility, allowing users to transfer systems and parameters between simulation engines [44].

Workflow management systems such as KNIME Analytics Platform, Pipeline Pilot, and Galaxy enable drag-and-drop integration of data retrieval, descriptor generation, docking, and simulation tasks. They are particularly valuable in automated virtual screening campaigns where thousands of compounds are processed through identical pipelines for consistency and reproducibility [45]. Cloud-based AI-integrated platforms including DeepChem, Autodock-GPU, BioSimSpace, and RosettaScripts represent the current frontier, merging deep learning with physics-based modeling. These environments support massive parallelism, distributed data management, and model retraining, thereby reducing computational bottlenecks and enhancing scalability [46]. Visualization and analysis are facilitated through tools like PyMOL, VMD, and UCSF ChimeraX, which bridge the interpretability gap between raw data and molecular insight. Collectively, this ecosystem exemplifies how CADD has evolved into a data- and computation-driven discipline, sustained by modular interoperability and algorithmic transparency [47].

Table 2.1. Overview of Core Data Foundations in Computer-Aided Drug Design (CADD)

Category	Representative Examples	Primary Role in CADD	Key Features and Notes
Structural Databases	**Protein Data Bank (PDB)**	3D macromolecular structures	Repository for protein, nucleic acid, and complex structures; essential for docking and molecular dynamics.
	ChEMBL	Bioactivity data and QSAR model development	Curated compound–target–activity relationships; supports machine learning and QSAR pipelines.
	PubChem	Chemical structure and bioassay data	Extensive open-access repository of over 110 million compounds; integration with assay results.
	ZINC	Virtual screening compound library	3D-ready small molecules with multiple protonation states; used for hit discovery.

	DrugBank	Drug–target and pharmacokinetic data	Integrates approved and investigational drugs with mechanism and ADMET data.
Descriptor Categories	**Constitutional Descriptors**	Basic molecular composition	Atom counts, bond types, molecular weight; used in drug-likeness filters.
	Topological Descriptors	2D molecular connectivity	Graph-theoretical indices (Wiener, Balaban, Kier–Hall) capturing branching and cyclicity.
	Geometrical Descriptors	3D shape and size	Volume, surface area, dipole moment; useful for docking and QSAR alignment.
	Electronic Descriptors	Quantum-chemical properties	HOMO–LUMO gap, charge distribution, polarizability; vital for reactivity modeling.
	Hybrid Descriptors	Multimodal representations	Combine 3D, electronic, and pharmacophoric features; used in AI-enhanced QSAR.
Force Fields	**AMBER**	Biomolecular simulation	Suitable for proteins, nucleic acids, and ligands; integrates with Antechamber for small molecules.
	CHARMM	Macromolecular modeling	Comprehensive parameter sets for proteins, lipids, and carbohydrates.
	OPLS-AA	Organic and drug-like molecules	Balanced force field for liquids and small-molecule dynamics.
	GROMOS	Biomolecular dynamics	Emphasizes water and solvation effects; efficient for long MD runs.
	GAFF/MMFF94	Small-molecule parameterization	Transferable force fields for ligand docking and mixed protein–ligand systems.

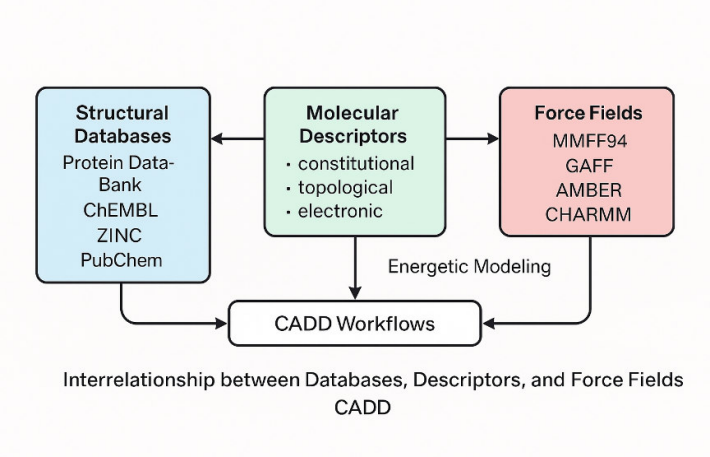

Figure 2.1. Interrelationship between Databases, Descriptors, and Force Fields in CADD

2.8 Challenges, Data Biases and Future Directions

Despite remarkable progress, the data foundations of CADD face persistent challenges related to data quality, representativeness, and interpretability. Many structural databases contain errors misannotated binding sites, incomplete protonation states, or missing residues that propagate into predictive models. Similarly, descriptor redundancy and overfitting remain major pitfalls in QSAR and machine learning pipelines, leading to inflated performance metrics on training data but poor generalization to new chemical spaces [48]. Another challenge is **bias** systematic overrepresentation of certain molecular scaffolds, assay types, or protein families which skews model learning. For instance, kinase inhibitors dominate ChEMBL datasets, biasing activity prediction models toward ATP-competitive mechanisms. Addressing such imbalance requires rigorous dataset curation, diversity analysis, and bias correction strategies [49].

In the realm of force fields, parameter transferability and polarization limitations continue to restrict accuracy, especially for flexible or charged systems. Emerging machine-learned force fields (MLFFs) trained on quantum data (e.g., ANI, DeePMD, NequIP) promise ab initio accuracy across chemical space, but their integration into large-scale workflows remains computationally demanding [50]. Future directions point toward AI-augmented, interoperable CADD ecosystems. Integration of graph-based molecular representations with dynamic simulations will yield more physically grounded predictions, while federated learning frameworks will enable collaborative model training across proprietary datasets without compromising data privacy [51]. The implementation of FAIR data standards ensuring findability, accessibility, interoperability, and reusability will remain central to sustainable innovation. Moreover, quantum computing and hybrid physics–AI modeling are expected to redefine force field development, allowing electronic correlation effects to be captured at near real-time computational speeds. As data generation continues to accelerate, the future of CADD will depend on curating high-quality, interpretable, and ethically shared datasets that sustain predictive accuracy and scientific reproducibility [52].

CONCLUSION

The success of computer-aided drug design (CADD) rests on its capacity to translate raw molecular information into actionable chemical and biological insights. Structural databases, molecular descriptors, and force fields together form the triad that supports this transformation—from molecular representation to energetic prediction and biological interpretation. Over the past five decades, these foundational pillars have evolved from isolated data sources and static equations into interconnected, dynamic systems powered by artificial intelligence, automation, and open data initiatives.

Structural databases now span millions of compounds and hundreds of thousands of biomolecular structures, allowing researchers to navigate an unprecedented breadth of chemical and biological space. When integrated with curated bioactivity data and molecular annotation, these repositories enable model training, validation, and benchmarking with a rigor that was once impossible. Molecular descriptors, in turn, convert this wealth of structural data into quantifiable features that bridge chemistry, physics, and biology. From classical 1D–3D metrics to graph-based and learned embeddings, descriptors have become both interpretable and computationally adaptable, facilitating QSAR modeling, virtual screening, and multi-parameter optimization.

Force fields complement these layers by providing a physically grounded means of exploring the conformational and energetic landscapes of molecules. As parameterization methods improve through quantum-mechanical calibration and machine learning, molecular mechanics simulations

increasingly approximate experimental precision, allowing for more reliable prediction of binding affinities and dynamic behavior. Together, databases, descriptors, and force fields create a closed feedback system where data drive hypotheses, simulations validate predictions, and new insights refine models in an iterative cycle of discovery.

The future of CADD lies in integrative, FAIR-compliant, and AI-augmented frameworks. Cloud-connected repositories, open-source software ecosystems, and adaptive force fields will converge to support reproducible, interpretable, and scalable drug discovery. Challenges such as data bias, interoperability, and interpretability must continue to be addressed through global collaboration and ethical governance. As the boundaries between computational and experimental drug design blur, the strength of CADD will increasingly depend on the robustness, accessibility, and integration of its data foundations.

In essence, data are not mere inputs but the intellectual infrastructure of modern drug design. The continued refinement and integration of structural databases, descriptors, and force fields will determine how effectively future scientists can explore the vast landscape of chemical space, accelerate therapeutic innovation, and uphold the principles of transparency and reproducibility that define modern pharmaceutical science.

REFERENCES

[1] Walters WP, Barzilay R. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of Chemical Research*. 2021;54(2):263–270.

[2] Berman HM et al. The Protein Data Bank: a historical perspective. *Acta Crystallographica Section D*. 2019;75(1):14–28.

[3] Kim S et al. PubChem 2023 update: improved data content and user interfaces. *Nucleic Acids Research*. 2023;51(D1):D1373–D1381.

[4] Schaduangrat N et al. Data-driven drug discovery: From big data to predictive models. *Pharmaceuticals*. 2022;15(5):556.

[5] Gaulton A et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*. 2017;45(D1):D945–D954.

[6] Irwin JJ, Shoichet BK. ZINC20—A free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*. 2020;60(1):6065–6073.

[7] Gilson MK et al. BindingDB in 2024: expanding biomolecular interaction data for structure-based drug design. *Nucleic Acids Research*. 2024;52(D1):D339–D348.

[8] Jamasb AR et al. FAIR data practices in cheminformatics and CADD. *Drug Discovery Today*. 2023;28(10):103842.

[9] Yap CW. PaDEL-Descriptor: An open-source software for calculating molecular descriptors and fingerprints. *Journal of Computational Chemistry*. 2011;32(7):1466–1474.

[10] Vanommeslaeghe K, MacKerell AD. CHARMM general force field (CGenFF): a force field for drug-like molecules. *Journal of Computational Chemistry*. 2012;33(31):2492–2511.

[11] Wang J et al. Development and testing of a general AMBER force field. *Journal of Computational Chemistry*. 2004;25(9):1157–1174.

[12] Halgren TA. The OPLS all-atom force field: parameter sets for organic liquids. *Journal of Computational Chemistry*. 1996;17(5-6):490–519.

[13] Ponder JW, Case DA. Force fields for protein simulations. *Advances in Protein Chemistry*. 2003;66:27–85.

[14] Boothroyd S et al. Open Force Field Evaluator: automated, reproducible evaluation of molecular

force fields. *Journal of Chemical Theory and Computation*. 2021;17(9):5868–5882.
[15] Rufa D et al. Toward chemical accuracy for protein–ligand binding free energies with polarizable force fields. *Journal of Chemical Theory and Computation*. 2020;16(3):1689–1702.
[16] Chan HCS et al. Integration of AI, QSAR, and molecular mechanics in drug discovery. *Frontiers in Drug Discovery*. 2024;4:1284751.
[17] Wang Y et al. Recent advances in machine learning-based force fields for drug discovery. *Frontiers in Chemistry*. 2023;11:1174902.
[18] Pires DE, Blundell TL, Ascher DB. pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *Journal of Medicinal Chemistry*. 2015;58(9):4066–4072.
[19] Sztanke K, Pospieszny T. Future perspectives of CADD in light of AI and FAIR data. *Computational and Structural Biotechnology Journal*. 2025;23:1789–1801.
[20] Lipinski CA. Rule of five in 2020 and beyond: Targeting new chemical space. Advanced Drug Delivery Reviews. 2021;173:1–15.
[21] Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. 2nd ed. Weinheim: Wiley-VCH; 2022.
[22] Doweyko AM. The nature of QSAR models: Insights from the application of geometrical and shape descriptors. Journal of Computer-Aided Molecular Design. 2018;32(6):1055–1071.
[23] Lu T, Chen F. Multiwfn: A multifunctional wavefunction analyzer for electronic descriptor analysis. Journal of Computational Chemistry. 2019;40(16):1250–1258.
[24] Zaliani A, Gohlke H. Pharmacophore fingerprinting and hybrid descriptors in modern QSAR and virtual screening. Drug Discovery Today: Technologies. 2020;37:27–38.
[25] Stokes JM et al. A deep learning approach to antibiotic discovery. Cell. 2020;181(2):475–483.
[26] Leach AR, Gillet VJ. An Introduction to Chemoinformatics. 3rd ed. Dordrecht: Springer; 2023.
[27] Ponder JW, Case DA. Force fields for protein simulations: Recent advances and future directions. Annual Review of Biophysics. 2020;49:275–300.
[28] Maier JA et al. ff14SB: Improving the accuracy of protein side chain and backbone parameters in the AMBER force field. Journal of Chemical Theory and Computation. 2015;11(8):3696–3713.
[29] Huang J, MacKerell AD Jr. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. Biophysical Journal. 2018;114(8):1609–1622.
[30] Lemkul JA, Huang J, Roux B, MacKerell AD. An empirical polarizable force field based on the classical Drude oscillator model. Chemical Reviews. 2016;116(9):4983–5013.
[31] Marrink SJ, Souza PCT, Tieleman DP. Perspective on the MARTINI model. Chemical Reviews. 2023;123(4):2405–2441.
[32] Harder E et al. OPLS3: A force field providing broad coverage of drug-like small molecules and proteins. Journal of Chemical Theory and Computation. 2016;12(1):281–296.
[33] Vanommeslaeghe K, Raman EP, MacKerell AD Jr. Automation of the CHARMM general force field (CGenFF) I: Bond perception and atom typing. Journal of Chemical Information and Modeling. 2012;52(12):3144–3154.
[34] Mobley DL, Guthrie JP. FreeSolv: A database of experimental and calculated hydration free energies, with input files. Journal of Computer-Aided Molecular Design. 2014;28(7):711–720.
[35] Smith JS et al. ANI-1: An extensible neural network potential with DFT accuracy at force-field computational cost. Chemical Science. 2017;8(4):3192–3203.
[36] Qiu Y et al. The Open Force Field (OpenFF) 2.0.0 release: Improved accuracy for small molecules. Journal of Chemical Theory and Computation. 2022;18(4):2901–2919.

[37] Horton JT, Boothroyd S, Wagner J, Mobley DL. Toward automated and data-driven force field parameterization. Journal of Physical Chemistry B. 2023;127(12):2723–2734.
[38] Sousa SF, Ribeiro AJM, Coimbra JTS. Bridging structural databases and force fields in CADD. Frontiers in Molecular Biosciences. 2021;8:653455.
[39] Meng XY, Zhang HX, Mezei M, Cui M. Molecular docking: A powerful approach for structure-based drug discovery. Current Computer-Aided Drug Design. 2021;17(4):451–461.
[40] Tropsha A, Golbraikh A. Predictive QSAR modeling workflow and challenges. Journal of Chemical Information and Modeling. 2019;59(7):1919–1934.
[41] Berthold MR et al. KNIME: The Konstanz Information Miner—Version 5.0. Data Mining and Knowledge Discovery. 2024;38(2):427–435.
[42] Eastman P et al. OpenMM 8: Accelerating molecular dynamics simulation on GPUs and cloud platforms. PLoS Computational Biology. 2024;20(1):e1010853.
[43] Landrum G. RDKit: Open-source cheminformatics. Software Release. 2023; https://www.rdkit.org
[44] Abraham MJ et al. GROMACS 2024: Fast, flexible, and free. SoftwareX. 2024;21:101540.
[45] Cerqueira NMFSA et al. Automated workflows in KNIME and Pipeline Pilot for high-throughput CADD. Molecules. 2022;27(3):752.
[46] Ramsundar B et al. Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More. Sebastopol: O'Reilly Media; 2019.
[47] Goddard TD et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Science. 2018;27(1):14–25.
[48] Bjerrum EJ, Sattarov B. Bias in chemical datasets: Causes, impact, and mitigation. Journal of Cheminformatics. 2024;16(1):39.
[49] Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity prediction using deep learning. Frontiers in Environmental Science. 2016;3:80.
[50] Schütt KT et al. SchNetPack 2.0: A neural network toolbox for atomistic modeling. Journal of Chemical Theory and Computation. 2023;19(10):2999–3013.
[51] Chen Y, Coley CW. Federated learning in molecular design. Nature Machine Intelligence. 2024;6(2):153–165.
[52] Lavecchia A, Cerchia C. Artificial intelligence in drug discovery: From concept to clinical reality. Drug Discovery Today. 2025;30(1):103745.