TYPE Review
PUBLISHED 17 March 2025
DOI 10.3389/fmolb.2025.1553667

Check for updates

OPEN ACCESS

EDITED BY
Yonglan Liu,
National Cancer Institute at Frederick (NIH),
United States

REVIEWED BY
Padhmanand Sudhakar,
Kumaraguru College of Technology, India
Yanxing Wang,
The University of Texas at Austin, United States

*CORRESPONDENCE
Raviprasad Aduri,
✉ aduri@goa.bits-pilani.ac.in
B. Jayaram,
✉ B.Jayaram@chemistry.iitd.ac.in

RECEIVED 31 December 2024
ACCEPTED 27 February 2025
PUBLISHED 17 March 2025

CITATION
Kattuparambil AA, Chaurasia DK, Shekhar S, Srinivasan A, Mondal S, Aduri R and Jayaram B (2025) Exploring chemical space for "druglike" small molecules in the age of AI.
Front. Mol. Biosci. 12:1553667.
doi: 10.3389/fmolb.2025.1553667

Exploring chemical space for "druglike" small molecules in the age of AI

Aman Achuthan Kattuparambil[1], Dheeraj Kumar Chaurasia[2,3], Shashank Shekhar[3], Ashwin Srinivasan[4], Sukanta Mondal[1], Raviprasad Aduri[1]* and B. Jayaram[3,5]*

[1]Department of Biological Sciences, BITS Pilani K K Birla Goa Campus, Zuarinagar, Goa, India, [2]School of Interdisciplinary Research, Indian Institute of Technology Delhi, New Delhi, India, [3]Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology Delhi, New Delhi, India, [4]Department of Computer Science & Information Systems, BITS Pilani K K Birla Goa Campus, Zuarinagar, Goa, India, [5]Department of Chemistry, Indian Institute of Technology Delhi, New Delhi, India

The announcement of 2024 Nobel Prize in Chemistry to Alphafold has reiterated the role of AI in biology and mainly in the domain of "drug discovery". Till few years ago, structure-based drug design (SBDD) has been the preferred experimental design in many academic and pharmaceutical R and D divisions for developing novel therapeutics. However, with the advent of AI, the drug design field especially has seen a paradigm shift in its R&D across platforms. If "drug design" is a game, there are two main players, the small molecule drug and its target biomolecule, and the rules governing the game are mainly based on the interactions between these two players. In this brief review, we will be discussing our efforts in improving the state-of-the-art technology with respect to small molecules as well as in understanding the rules of the game. The review is broadly divided into five sections with the first section introducing the field and the challenges faced and the role of AI in this domain. In the second section, we describe some of the existing small molecule libraries developed in our labs and follow-up this section with a more recent knowledge-based resource available for public use. In section four, we describe some of the screening tools developed in our laboratories and are available for public use. Finally, section five delves into how domain knowledge is improving the utilization of AI in drug design. We provide three case studies from our work to illustrate this work. Finally, we conclude with our thoughts on the future scope of AI in drug design.

KEYWORDS

machine learning (ML), artificial intelligence, computer aided drug design (CADD), small molecules, BIMP

1 Introduction

Small molecule libraries play a pivotal role in modern drug discovery, serving as essential collections of chemical compounds for identifying molecules with desired biological activity (Dandapani et al., 2012; Saldívar-González et al., 2020). These libraries can be broadly categorized into diverse libraries, which offer broad structural variety, and focused libraries that target specific protein

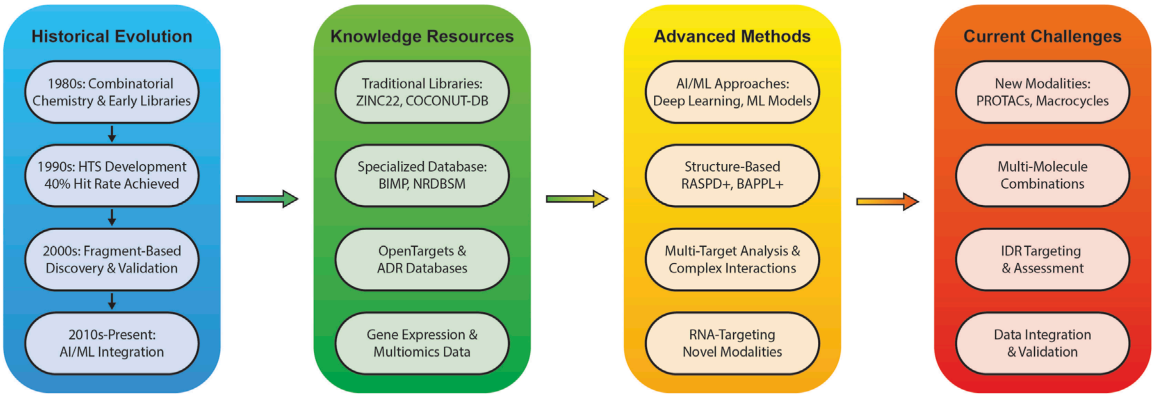
Historical Evolution
1980s: Combinatorial Chemistry & Early Libraries
1990s: HTS Development 40% Hit Rate Achieved
2000s: Fragment-Based Discovery & Validation
2010s-Present: AI/ML Integration
Knowledge Resources
Traditional Libraries: ZINC22, COCONUT-DB
Specialized Database: BIMP, NRDBSM
OpenTargets & ADR Databases
Gene Expression & Multiomics Data
Advanced Methods
AI/ML Approaches: Deep Learning, ML Models
Structure-Based RASPD+, BAPPL+
Multi-Target Analysis & Complex Interactions
RNA-Targeting Novel Modalities
Current Challenges
New Modalities: PROTACs, Macrocycles
Multi-Molecule Combinations
IDR Targeting & Assessment
Data Integration & Validation

FIGURE 1
An overview of the evolution, current state, and future directions of small molecule drug discovery in the age of AI is presented in this Figure, highlighting the key developments from historical advances to emerging challenges.

families or biological pathways, such as GPCR kinases (Dandapani et al., 2012; Harris et al., 2011). The generation of these libraries employs various methodologies, including combinatorial chemistry, diversity-oriented synthesis, fragment-based approaches, natural product extraction, and computational generation of virtual libraries (Dandapani et al., 2012; Saldívar-González et al., 2020; Sadybekov and Katritch, 2023).

The success of *in silico* drug design is significantly influenced by the selection of appropriate small molecule libraries through multiple factors (Dandapani et al., 2012). While diverse libraries enable broad exploration of chemical space, focused libraries can enhance hit rates for specific targets (Dandapani et al., 2012; Harris et al., 2011). The assessment of physicochemical properties and drug-likeness, particularly through established filters like Lipinski's RO5 (Lipinski et al., 1997) ensures appropriate absorption, distribution, metabolism, and excretion characteristics (Dandapani et al., 2012; Sadybekov and Katritch, 2023). Aqueous solubility remains a critical factor, as poorly soluble molecules can lead to false positives and limited optimization potential.

These libraries find application across various drug discovery approaches. In virtual screening, libraries undergo computational assessment for target binding potential, often in conjunction with experimental screening to enrich compound collections (Dandapani et al., 2012; Sadybekov and Katritch, 2023). *De novo* drug design utilizes these libraries as foundations for generating novel molecules, particularly when existing libraries have been exhausted, incorporating target constraints and leveraging machine learning approaches (Sadybekov and Katritch, 2023; Chang et al., 2023). Fragment-based drug design employs libraries of small fragments to identify weak-binding molecules that can be elaborated into more potent compounds (Dandapani et al., 2012; Sadybekov and Katritch, 2023), while lead optimization uses libraries to enhance existing compounds' properties through quantitative structure-activity relationship (QSAR) models (Sadybekov and Katritch, 2023; Chang et al., 2023). Figure 1 provides a comprehensive overview of how these approaches have evolved from historical developments to current AI-integrated methodologies, highlighting the interconnected nature of various tools and resources in modern drug discovery.

1.1 Historical context of small molecule libraries

The evolution of small molecule drug discovery has been marked by transformative technological advances since the 1980s. The field was revolutionized by combinatorial chemistry, progressing from Geysen's multi-pin technology to the first small-molecule combinatorial library by Bunin and Ellman in 1992 (Liu et al., 2017). This advancement, integrated with high-throughput screening (HTS) and computational methods, became fundamental to pharmaceutical lead discovery by the late 1990s (Appell et al., 2001).

Screening methodologies evolved in parallel, with laboratory robotics enabling automated biological assays that could generate up to 100,000 data points daily through ultrahigh-throughput screening platforms (Appell et al., 2001). The field progressed from random to focused libraries, with discovery libraries decreasing from 57% (1992–1997) to 21% (1999) (Appell et al., 2001), in contrast to targeted and optimization libraries. Fragment-Based Drug Discovery (FBDD) emerged as a complementary approach, leading to FDA-approved drugs like Vemurafenib (2011) and Venetoclax (Mureddu and Vuister, 2022; Bon et al., 2022).

The success of this evolution is exemplified by landmark drugs such as Imatinib (Gleevec), which revolutionized chronic myeloid leukemia treatment in 2001 (Müller, 2009; Druker and Lydon, 2000). Venetoclax demonstrated the feasibility of targeting protein-protein interactions, representing one of the first non-natural product clinical agents in this space (Congreve et al., 2008; Murray and Rees, 2009). While recent trends show a shift towards biologics due to lower clinical trial attrition rates (Sun et al., 2011), small molecules continue to comprise approximately 40% of FDA approvals annually (Mureddu and Vuister, 2022). However, challenges persist, with only 1% of compounds progressing from discovery to approved New Drug Application (NDA), and a 50% failure rate in clinical trials due to ADME issues (Appell et al., 2001), emphasizing the ongoing need for innovative approaches in small molecule drug discovery.

1.2 Types of small molecule libraries

A fundamental distinction exists between physically synthesized and virtually synthesizable libraries. Synthesized libraries represent physical collections created through chemical synthesis techniques, available through in-house programs, vendors, or contract research organizations (Dandapani et al., 2012). Conversely, synthesizable libraries exist as digital collections of compounds designed *in silico*, considered feasible to synthesize using known chemical reactions and commercially available reagents (Saldívar-González et al., 2020; Sadybekov and Katritch, 2023).

Several specialized categories have emerged to address specific drug discovery needs. Fragment libraries consist of low molecular weight compounds (typically <300 Da) with minimal hydrogen bond donors/acceptors, low lipophilicity, and few rotatable bonds (Dandapani et al., 2012). Lead-like libraries contain compounds with properties desirable for drug candidates, designed with a balance between structural diversity and drug-like properties (Dandapani et al., 2012; Saldívar-González et al., 2020). Natural product libraries comprise compounds derived from natural sources, providing valuable structural diversity and novel scaffolds for targeting macromolecule interactions (Dandapani et al., 2012; Heinzke et al., 2024).

Computationally generated libraries, exemplified by the GDB-17 library (160 billion molecules) and CHIPMUNK library (95 million compounds), enable cost-effective exploration of vast chemical spaces (Saldívar-González et al., 2020; Korablyov et al., 2024). While offering flexibility in design and novel structures, these face challenges including uncertainty in predicted properties and potential synthetic inaccessibility (Dandapani et al., 2012; Sadybekov and Katritch, 2023; Popova et al., 2018).

1.3 Filters and assessment criteria

Drug-likeness assessment primarily relies on established parameters, with RO5 setting fundamental criteria for oral bioavailability, including molecular weight under 500 Daltons, CLogP less than 5, and specific limits on hydrogen bond donors and acceptors (Dandapani et al., 2012; Ress et al., 2004). Additional guidelines have emerged for specialized applications, such as the “rule of 3” for fragment-based design and “rule of 2” for reagents, providing more targeted parameters for different molecular categories (Saldívar-González et al., 2020).

ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties form a crucial component of molecular assessment (Dandapani et al., 2012; Saldívar-González et al., 2020; Chang et al., 2023). Optimal passive membrane absorption correlates with logP values between 0.5 and 3, while metabolism considerations focus particularly on cytochrome P450 interactions. Toxicity evaluation encompasses various factors, including cardiac risks through hERG channel binding, with specific attention paid to identifying pan-assay interference compounds (PAINS) to avoid false positives in biological assays.

Synthetic feasibility evaluation employs metrics such as the synthetic accessibility score (SAS), where scores above 6 indicate potentially challenging synthesis (Popova et al., 2018). Structural properties assessment includes molecular complexity measures, examining features such as chiral centers and sp2:sp3 hybridization ratios, alongside diversity analysis using molecular fingerprints and clustering algorithms (Dandapani et al., 2012).

The integration of adverse drug reaction (ADR) databases has emerged as an additional filtering criterion, enhancing toxicity predictions. Databases such as DrugCentral provide comprehensive structural and pharmacological details for early safety assessment (Halip et al., 2023). Deep learning models trained on data from Open TG-GATEs and FAERS enable ADR likelihood prediction (Mohsen et al., 2020), while the SIDER database offers drug-ADR pairs from FDA drug labels for validation (Ietswaart et al., 2020). These resources, combined with machine learning approaches, facilitate more accurate correlation of structural features with specific adverse effects, particularly through analysis of properties aligned with traditional drug-likeness criteria (Halip et al., 2023). Feature selection methods and random forest models have achieved significant improvements in ADR detection, with some studies reaching 100% accuracy for specific compounds (Liu and Aickelin, 2021).

1.4 Limitations of current approaches and emerging solutions

Traditional molecular filters, while valuable, often oversimplify molecular behavior in biological systems (Dandapani et al., 2012). ADMET prediction models frequently demonstrate reduced accuracy when based on computational rather than experimental data (Dandapani et al., 2012; Patel et al., 2020). These models particularly struggle with non-traditional molecules like macrocycles and PROTACs, partly due to insufficient high-quality training data. The PDBbind (Wang et al., 2005) database, for instance, inadequately represents “negative space” or suboptimal interactions, limiting its utility in predicting general binding behaviour (Sadybekov and Katritch, 2023; Vamathevan et al., 2019).

Rigid adherence to conventional filters can exclude promising compounds with unique properties that might prove effective, particularly for non-oral drugs or specific targets (Dandapani et al., 2012). This narrow focus on conventional “drug-like” space reduces the possibility of discovering novel scaffolds or chemotypes (Saldívar-González et al., 2020; Ress et al., 2004). Current approaches often overlook crucial molecular complexity factors such as three-dimensionality, chirality, and sp2:sp3 hybridization ratios (Dandapani et al., 2012).

A significant limitation emerges with newer drug modalities like macrocycles and PROTACs. Traditional molecular descriptors, developed for conventional small molecules, fail to capture features of macrocycle chemotypes relevant to their pharmacological behavior (Viarengo-Baker et al., 2021). For macrocycles, computed parameters like clogP often do not reflect true conformation-dependent lipophilicity, and traditional rules about rotatable bonds become questionable when applied to macrocyclic structures (Viarengo-Baker et al., 2021). Successful macrocycles often achieve oral bioavailability through “chameleonicity,” adapting their conformations to different environments (Garcia Jimenez et al., 2023), as exemplified by cyclosporin A (Kingwell, 2023).

PROTACs present additional challenges due to their heterobifunctional nature and large molecular weight (700–1,100 Da) (An and Fu, 2018). Their size provides more opportunities for metabolic attack (An and Fu, 2018), and their optimization requires focus on the whole molecule rather than individual components (Weng et al., 2021). Critical considerations include linker design, which affects entropy, selectivity, activity, and permeability (Weng et al., 2021). These compounds may bury up to 800–900 Å2 of ligand surface area when binding to their target, approaching protein-protein interface areas (Doak and Kihlberg, 2017).

The field is adopting various strategies to address these limitations. Chemical space exploration is expanding through combinatorial chemistry, DNA-encoded libraries, and virtual libraries of on-demand compounds (Sadybekov and Katritch, 2023; Korablyov et al., 2024; Gottipati et al., 2020). Machine Learning (ML) and Artificial Intelligence (AI) enable *de novo* design of novel scaffolds (Chang et al., 2023; Popova et al., 2018; Gottipati et al., 2020; McNaughton et al., 2022), while fragment-based approaches provide systematic methods for developing new molecules (Dandapani et al., 2012; Ress et al., 2004). Target-focused libraries leverage structural data and ligand knowledge to improve hit rates for specific protein families (Dandapani et al., 2012; Harris et al., 2011). Performance diversity strategies, selecting compounds based on assay results rather than chemical diversity alone, are showing promise (Wawer et al., 2014). Advanced computational methods, including molecular dynamics, DFT, and MMPBSA/MMGBSA/MMBAPPL, provide more sophisticated analysis capabilities (Chang et al., 2023). New approaches emphasize synthetic accessibility assessment (Popova et al., 2018; Ivanenkov et al., 2023) and improved scoring functions for virtual screening (Sadybekov and Katritch, 2023), while hybrid strategies combine computational methods with experimental validation to enhance library design effectiveness.

New assessment criteria are emerging to address these challenges. For macrocycles, modified rules suggest maintaining HBD ≤7 combined with either MW < 1,000 Da, cLogP >2.5, or TPSA <300 Å2 (Garcia Jimenez et al., 2023). PROTAC assessment requires new metrics, with fa × fg = 0.25 suggested as a minimum threshold for drug-likeness (Hornberger and Araujo, 2023). Success stories like ARV-110 and ARV-471, which entered phase I clinical trials in 2019, demonstrate the potential of these approaches despite breaking traditional rules (Békés et al., 2022; Blanco and Gardinier, 2020).

Recent advances in ADR prediction models offer potential solutions to these limitations. Random Forest models can now predict drug-ADR and target-ADR associations using *in vitro* secondary pharmacology data (Ietswaart et al., 2020), while deep learning frameworks like DeepSide utilize gene expression profiling experiments and chemical structures to predict ADRs (Uner et al., 2019). These approaches enable early identification of potential safety issues, allowing for structural modifications to reduce interactions with targets linked to severe ADRs (Ietswaart et al., 2020). The integration of multiple data types, from chemical structures to literature mining, has enhanced the predictive power of these models (Mohsen et al., 2020).

1.5 Role of artificial intelligence

AI and ML are revolutionizing library design and selection through multiple avenues (Patel et al., 2020). At the core of these advances lies the crucial aspect of molecular representation, where deep learning algorithms perform feature learning or representation learning, contrasting with traditional feature engineering approaches (Chuang et al., 2020). The effectiveness of these representations depends on key considerations: expressiveness to capture chemical space diversity, parsimony to maintain compactness without losing critical information, and invariance to ensure consistent representation regardless of atom numbering (Chuang et al., 2020).

When dealing with high-dimensional chemical descriptor spaces, several challenges emerge. The empty space phenomenon results in sparse dataset coverage, while the vanishing sphere volumes and distance concentration effects can complicate meaningful molecular comparisons (Reutlinger and Schneider, 2012). To address these challenges, various dimensionality reduction and feature extraction methods are employed. These include Principal Component Analysis (PCA) for uncorrelated variable transformation, Kernel PCA for nonlinear relationship analysis, and advanced techniques like symmetric encoder networks, self-organizing maps (SOM), and stochastic proximity embedding (SPE) (Reutlinger and Schneider, 2012; Sarkar et al., 2023).

Enhanced virtual screening utilizing deep learning models enables efficient analysis of large chemical spaces and improved prediction of ligand properties. Machine learning facilitates *de novo* design through generative models and reinforcement learning, creating novel molecules with desired properties and overcoming existing library limitations (Gottipati et al., 2020; McNaughton et al., 2022).

Recent developments in AI have expanded to include sophisticated ADR prediction models. Deep learning architectures trained on drug chemical structures and gene expression profiles can now predict adverse reactions with unprecedented accuracy (Uner et al., 2019). These models, integrated with databases like FAERS and SIDER, provide comprehensive safety assessments early in the drug development process (Mohsen et al., 2020; Ietswaart et al., 2020). The success of these approaches is evidenced by models achieving high accuracy in detecting major ADRs, particularly when combining multiple data sources and advanced feature selection methods (Liu and Aickelin, 2021).

Generative models, such as REINVENT (Loeffler et al., 2024), have become particularly instrumental in creating novel, synthesizable compounds by exploring vast chemical spaces beyond traditional limitations (Chang et al., 2023; Patel et al., 2020; Loeffler et al., 2024). Based on recurrent neural networks or transformers, these models can perform multi-objective optimization, simultaneously considering factors like potency, selectivity, solubility, and ADMET properties (Popova et al., 2018). They enable scaffold hopping and linker design while incorporating synthetic feasibility predictions through reinforcement learning algorithms that navigate synthetically accessible chemical space (Sadybekov and Katritch, 2023; Gottipati et al., 2020).

These advances aid in predicting ADMET and pharmacokinetic properties, guiding hit-to-lead optimization through QSAR models, and supporting target identification through omics data analysis (Sadybekov and Katritch, 2023; Patel et al., 2020). However, careful consideration must be given to the application of dimensionality reduction methods, as their misuse can lead to erroneous results and misinterpretation, particularly in hit finding and hit-to-lead optimization stages of early drug discovery (Reutlinger and Schneider, 2012). When properly implemented and combined with fragment-based drug discovery approaches and ADR prediction models, these systems provide a comprehensive framework for developing safer and more effective drugs (Korablyov et al., 2024; Ietswaart et al., 2020).

2 Traditional approaches to molecular library development

The efficient exploration of chemical space for drug discovery necessitates robust approaches for generating and organizing molecular libraries. Here, we present two complementary methodologies developed by our group: a chemical template-based generation system and a curated molecular database, each addressing different aspects of the drug discovery pipeline.

2.1 Chemical template-based generation system

We developed a comprehensive chemical template library comprising 160 distinct chemical moieties, categorized into rings, sidechains, and linkers. This modular system enables the sequential construction of both known and novel molecular structures through systematic combination and arrangement of these template elements (Latha et al., 2004). The methodology incorporates a structured workflow for molecule generation, optimization, and evaluation against target proteins.

The system's implementation involves several key steps: initial molecule generation through template combinations, structural optimization of the generated molecules, molecular docking against target proteins, and subsequent scoring and ranking of potential candidates. This approach has been successfully implemented in the Sanjeevini software platform, facilitating active-site directed lead design (Jayaram et al., 2006; Jayaram et al., 2012).

However, during implementation, we identified a significant limitation: the disparity between computational feasibility and synthetic accessibility. Specifically, molecules that can be readily generated *in silico* may present substantial challenges for practical synthesis *in vitro*. This observation prompted the development of complementary approaches focused on curated molecular databases (Latha and Jayaram, 2005).

2.2 NRDBSM: a curated database for virtual screening

To address the limitations of template-based generation, we developed the Non-Redundant Database of Small Molecules (NRDBSM), specifically designed to facilitate virtual high-throughput screening (vHTS). This database represents a carefully curated collection of approximately 17,000 compounds, each selected based on stringent physicochemical criteria and optimized for lead-like characteristics (Shaikh et al., 2007; Shaikh et al., 2012).

The database construction prioritizes compliance with established drug-likeness parameters, including Lipinski's Rule of Five and additional criteria crucial for evaluating solubility, membrane permeability, and transport characteristics. Key molecular descriptors used in the curation process include molecular weight, hydrogen bond donor and acceptor counts, partition coefficient (logP), and molar refractivity.

A distinctive feature of NRDBSM is its uniform distribution of physicochemical parameters, deliberately deviating from the typical normal distribution observed in conventional databases. The parameters are distributed across carefully selected ranges: logP values from −1.0 to 6.0, molar refractivity spanning 40 to 130, molecular weights between 150 and 480, hydrogen bond donors from 0 to 3, and hydrogen bond acceptors from 2 to 9. This distribution strategy optimizes the coverage of chemical space while maintaining drug-like characteristics.

The compounds in NRDBSM are characterized by simplified molecular structures, conservative molecular weights, minimal ring systems, controlled numbers of rotatable bonds, and moderate hydrophobicity. This intentional simplicity facilitates their prospective evolution into drug-like compounds post-vHTS, allowing for systematic structural refinement and controlled complexity augmentation (Shaikh et al., 2007; Shaikh et al., 2012).

The database incorporates a comprehensive search engine enabling users to query and filter molecules based on multiple physicochemical parameters. This functionality supports both independent virtual screening campaigns and targeted searches within larger molecular datasets, effectively streamlining the early stages of drug discovery by identifying promising candidates while minimizing subsequent optimization challenges.

These complementary approaches - template-based generation and curated database development - provide researchers with versatile tools for exploring chemical space in drug discovery. While the template-based system offers flexibility in molecular design, NRDBSM ensures practical applicability through careful curation and optimization of physicochemical properties.

2.3 IDRs as targets and their limitations

Intrinsically Disordered Regions (IDRs) represent an emerging class of drug targets that challenge traditional small molecule screening approaches. These regions, characterized by their structural flexibility, play crucial roles in protein-protein interactions and are frequently associated with disease states, making them attractive therapeutic targets (Han et al., 2023; Wang et al., 2023). The structural plasticity of IDRs enables them to interact with multiple partners through Short Linear Motifs (SLiMs), promoting various biological processes including cell signaling and protein modification (Han et al., 2023).

In small molecule screening, IDRs present unique opportunities and challenges. The dynamic nature of IDP-ligand interactions, where small molecules can interact with multiple sites simultaneously, necessitates modified screening approaches (Wang et al., 2023). IDP drug virtual screening (IDPDVS) has emerged as an efficient strategy, employing conformation sampling, clustering, and selection of druggable conformations to identify potential binding molecules (Ruan et al., 2021). This approach has proven particularly valuable for IDPs without known active small-molecule ligands.

Several computational methods enhance IDP-targeted drug discovery. Ensemble-based drug discovery (EBDD) employs many-to-many scoring, evaluating multiple protein conformations against numerous ligands (Wang et al., 2023). The integration of multiple experimental techniques, including NMR, SAXS, and smFRET, with computational simulations has improved IDP model accuracy (Wang et al., 2023). Recent advances in deep learning and molecular dynamics have accelerated this field, with enhanced sampling methods enabling direct generation of IDP conformations (Wang et al., 2023).

Despite these advances, the structural flexibility of IDRs complicates traditional binding site prediction and docking approaches. However, successful examples of IDP-targeting drugs advancing to clinical trials demonstrate the feasibility of this approach (Wang et al., 2023), suggesting that integrating IDR-specific considerations into screening workflows could significantly expand the druggable target space.

3 Specialized knowledge-based resources

The evolution of drug discovery has been significantly enhanced by specialized databases that integrate diverse data types and provide comprehensive insights into molecular interactions. These knowledge bases serve as crucial resources for improving prediction accuracy and streamlining the drug development process through the integration of traditional knowledge, experimental data, and computational approaches.

3.1 BIMP database

The Bioactivity of Phytochemicals of Indian Medicinal Plants (BIMP) Database (https://scfbio.iitd.ac.in/bimp/) is a comprehensive and meticulously curated resource developed to assist researchers, scientists, and professionals in exploring the therapeutic potential of India's extensive medicinal flora. By bridging the gap between traditional knowledge and modern scientific research, the BIMP Database facilitates the discovery of bioactive compounds and therapeutic properties rooted in India's rich botanical heritage. This database is a crucial tool in advancing the understanding of medicinal plants and their role in drug discovery and development.

The BIMP Database encompasses an extensive inventory of 6,209 unique plant species and 105,909 phytochemicals. Each entry is annotated with detailed physicochemical properties and categorized into relevant compound classifications. This exhaustive resource allows researchers to systematically explore the therapeutic applications of Indian medicinal plants, providing valuable insights for both experimental and computational studies.

One of the key features of the database is the availability of molecular data in multiple formats, including SDF, PDB, XYZ, and MOL2. These formats provide both 2D and 3D representations of molecular structures, enabling detailed visualization and analysis.

Each phytochemical entry is supplemented with an extensive profile of physicochemical properties such as solubility, polarity, and molecular weight. Additionally, the inclusion of molecular descriptors provides further structural insights, allowing researchers to better understand compound behavior and bioactivity. These detailed annotations equip users with the tools needed to evaluate the potential of compounds in therapeutic contexts.

The BIMP Database also evaluates phytochemicals against widely accepted druglikeness rules, including Lipinski's Rule of Five, Egan's Rule, Muegge's Rule, Ghose's Rule, and Veber's Rules. Compounds that violate any of these rules are flagged, offering researchers critical insights into their suitability as viable drug candidates. This feature ensures that users can efficiently screen compounds for drug development potential.

Another significant feature of the BIMP Database is its integration of both predicted and experimentally validated pharmacological targets for phytochemicals. This dual approach provides comprehensive insights into the bioactivity of compounds and aids in identifying specific therapeutic applications. By offering predicted and experimental data, the database enables researchers to make more informed decisions in their investigations of pharmacological properties.

To further support drug discovery efforts, the database includes robust tools for virtual screening, scaffold identification, and similarity searches. These tools allow researchers to evaluate compounds efficiently based on specific criteria, streamlining the identification of potential drug candidates. Moreover, the database's search functionality supports diverse parameters, enabling users to search for compounds by ID, name, plant species, plant family, or links to external databases such as PubChem, DrugBank, FooDB, KnapSack, ChemSpider, and CAS.

The BIMP Database serves as a valuable resource for multiple sectors, including academia, the pharmaceutical industry, and healthcare. Its applications extend to facilitating novel therapeutic discoveries, supporting evidence-based medical research, informing sustainable policymaking regarding medicinal plant usage, and promoting biodiversity conservation (Chaurasia et al., 2024). By seamlessly integrating traditional knowledge with advanced scientific methodologies, the BIMP Database fosters significant advancements in natural product research, drug development, and sustainable healthcare solutions.

3.2 Comparative analysis of chemical libraries

The landscape of chemical libraries encompasses various specialized databases, each offering unique features and complementary strengths. While BIMP focuses on Indian medicinal flora with 105,909 phytochemicals from 6,209 plant species, other major databases like ZINC22 provide broader coverage with over 37 billion commercially available compounds (Tingle et al., 2023). This diversity in scope and focus enables researchers to access different segments of chemical space for drug discovery.

ZINC22's strength lies in its extensive coverage of commercially available compounds, offering advanced search capabilities and pre-calculated 3D conformers for virtual screening. The database's CartBlanche GUI facilitates analog searching, and its tranche browser allows tailored subsetting for specific project requirements (Tingle et al., 2023; Irwin et al., 2020). In contrast, BIMP's specialization in traditional medicine-derived compounds, complete with experimentally validated targets and comprehensive physicochemical annotations, provides a unique resource for natural product-based drug discovery.

Natural product databases like COCONUT complement these resources by aggregating information from multiple sources, improving annotations, and offering specialized focus areas (Chandrasekhar et al., 2025; Sorokina et al., 2021). While COCONUT combines data from 53 openly accessible natural product databases, specialized databases like NPAtlas focus on microbial natural products, and others like NuBBEDB and KNap-Sack concentrate on phytochemicals (Sorokina et al., 2021).

Each database offers distinct advantages in data organization and accessibility. ZINC22 provides rapid lookup of molecular properties and regular updates, with 90% of catalogs refreshed every 90 days (Tingle et al., 2023; Irwin et al., 2020). BIMP's strength lies in its detailed physicochemical profiling, multiple molecular format availability (SDF, PDB, XYZ, MOL2), and integration of both predicted and experimental target data.

Notably, analysis of druggability across these databases reveals interesting patterns. In BIMP's collection of over 100,000 phytochemicals, 33% conform to all major druggability rules (Lipinski, Ghose, Veber, Egan, Muegge's), while 72% satisfy at least one rule. These proportions are relatively high compared to databases of chemically synthesized compounds, suggesting that natural product libraries might offer a richer source of drug-like molecules. This observation aligns with the historical success of natural products in drug discovery and their evolutionary optimization for biological interactions.

Together, these resources create a complementary ecosystem for drug discovery, combining commercial availability, natural product diversity, and traditional medicine knowledge. The higher druggability ratio in natural product databases like BIMP provides an additional strategic advantage for drug discovery efforts, particularly when seeking novel scaffolds with inherent biological relevance.

3.3 Integration of openTargets for enhanced prediction

The Open Targets Platform, an open-source knowledge base integrating data from 23 independent public sources, offers valuable insights for drug target identification and prioritization (Buniello et al., 2025). This resource uniquely combines multiple data types: genetic associations, somatic mutations, transcriptomics, pathway biology, and critically, information about approved drugs and their targets (Han et al., 2022; Koscielny et al., 2017). For approved pharmaceuticals, the platform provides extensive molecular attributes and target information, enabling more accurate prediction models through validated drug-target pairs (Koscielny et al., 2017).

The platform's comprehensive architecture supports multiple prediction enhancement strategies. At the molecular level, it enables the creation of three-dimensional data tensors comprising gene targets, diseases, and evidence attributes (Ye et al., 2024). This integration has demonstrated significant improvements in prediction accuracy, particularly when combining target tissue specificity with functional interactions (Buniello et al., 2025). The ML-GPS (machine learning-assisted genetic priority score) framework exemplifies this approach, utilizing predicted phenotypes to enhance target identification for chronic diseases. This method has substantially expanded our understanding of drug-target relationships, supporting over 15,000 previously unvalidated gene-disease associations and identifying promising targets such as LRRK2 inhibitors for Parkinson's disease (Chen et al., 2024).

Gene expression data within OpenTargets provides an additional layer for screening refinement. By incorporating expression profiles with molecular attributes of successful drugs, prediction models can better account for both tissue-specific and cell-type specific effects. This granular understanding of cellular responses enables more precise predictions of drug effects across different cellular contexts and tissues. The integration has proven particularly valuable for target validation and novel indication discovery, although challenges remain in normalizing heterogeneous data sources and managing computational resources for large-scale expression analysis. Despite these limitations, the combined use of validated drug-target pairs and multi-level expression data has demonstrably improved prediction accuracy, with some studies reporting significant increases in both AUROC and AUPRC metrics (Ye et al., 2024).

4 Advanced computational methods for screening

Virtual screening of extensive chemical libraries targeting protein binding sites is a pivotal stage in modern drug discovery. This involves computational docking of ligands into protein binding sites to estimate their binding affinities. Traditional docking methods often generate multiple poses for ligands, leading to significant computational costs and challenges in accurately predicting protein-ligand binding affinities. To address these issues, advanced computational methods like RASPD+ (Holderbach et al., 2020) and BAPPL+ (Soni et al., 2020) have been developed, building on earlier versions of our CADD/Sanjeevini Pipeline (Figures 2, 3), which established foundational approaches for bracketing drug-like compounds from templates or databases (Jayaram et al., 2006; Jayaram et al., 2012).

4.1 RASPD+

RASPD+ represents a pre-filtering approach designed to prioritize ligands efficiently in drug discovery workflows. By leveraging machine learning (ML) models and physicochemical descriptors that are independent of ligand conformation, RASPD+ overcomes the limitations of traditional docking methods. Unlike conventional approaches, RASPD+ does not require the generation of ligand poses, focusing instead on pose-invariant descriptors of ligands and protein binding pockets. This pose-independent methodology reduces computational costs significantly while maintaining strong predictive performance.

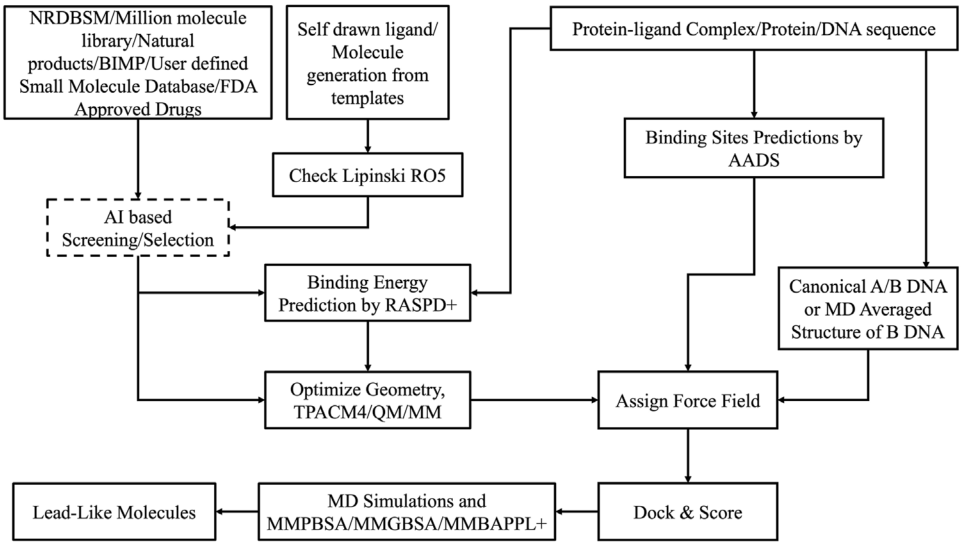
NRDBSM/Million molecule library/Natural products/BIMP/User defined Small Molecule Database/FDA Approved Drugs
Self drawn ligand/ Molecule generation from templates
Protein-ligand Complex/Protein/DNA sequence
Check Lipinski RO5
Binding Sites Predictions by AADS
AI based Screening/Selection
Binding Energy Prediction by RASPD+
Canonical A/B DNA or MD Averaged Structure of B DNA
Optimize Geometry, TPACM4/QM/MM
Assign Force Field
Lead-Like Molecules
MD Simulations and MMPBSA/MMGBSA/MMBAPPL+
Dock & Score

FIGURE 2
Workflow and architecture of *Sanjeevini*.

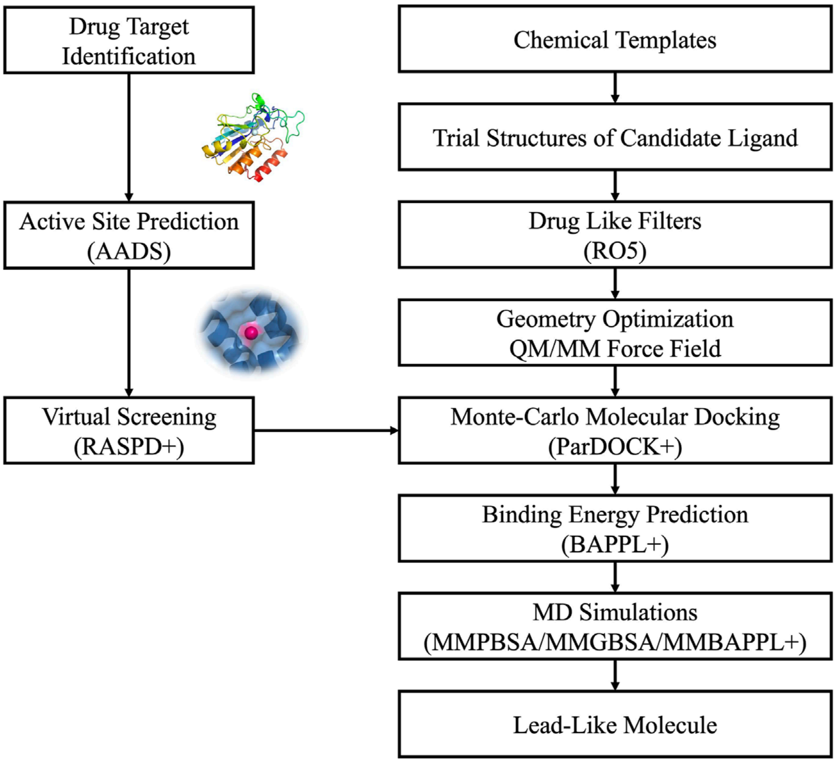
Drug Target Identification
Chemical Templates
Trial Structures of Candidate Ligand
Active Site Prediction (AADS)
Drug Like Filters (RO5)
Geometry Optimization QM/MM Force Field
Virtual Screening (RASPD+)
Monte-Carlo Molecular Docking (ParDOCK+)
Binding Energy Prediction (BAPPL+)
MD Simulations (MMPBSA/MMGBSA/MMBAPPL+)
Lead-Like Molecule

FIGURE 3
The *Sanjeevini* pathway for active site directed lead compound design *in silico*.

The ML models employed in RASPD+ are trained on the PDBbind dataset, enabling accurate prediction of protein-ligand binding affinities. When benchmarked against its predecessor and traditional scoring functions, RASPD+ demonstrates superior regression performance across multiple test datasets. These advancements make RASPD+ an ideal tool for pre-screening compound libraries in pharmaceutical research. Its ability to prioritize compounds rapidly without compromising accuracy expedites the identification of promising drug candidates, offering a highly efficient solution for early-stage drug discovery.

Performance evaluations of RASPD+ reveal its consistent and reliable regression performance, underscoring its potential to streamline the identification of prospective leads from extensive chemical libraries (Holderbach et al., 2020). This innovative method represents a significant step forward in computational drug discovery, combining computational efficiency with robust predictive capabilities.

Performance evaluations of RASPD+ reveal its consistent and reliable regression performance, underscoring its potential to streamline the identification of prospective leads from extensive chemical libraries (Holderbach et al., 2020). This innovative method represents a significant step forward in computational drug discovery, combining computational efficiency with robust predictive capabilities. The practical utility of RASPD+ is demonstrated through its successful implementation in various drug discovery projects. For instance, in the Dhanvantari platform, RASPD+ enables rapid screening of small molecule libraries against target protein active sites (Bhat et al., 2020). Its efficiency was particularly evident in a large-scale screening effort, where it successfully processed a million-molecule library against an identified site in HBsAg (Kiruthika et al., 2021), highlighting its capability to handle extensive chemical libraries while maintaining computational efficiency.

4.2 BAPPL+

BAPPL+ is an advanced scoring function designed to predict the binding affinities of protein-ligand (PL) complexes with enhanced accuracy. Evolved from earlier scoring functions such as BAPPL and BAPPLZ, BAPPL+ incorporates machine learning to improve prediction reliability. This new scoring function is particularly versatile, accommodating both metallo and non-metallo PL complexes, thus expanding its applicability in structure-based drug design.

The performance of BAPPL+ is underpinned by an enlarged and diverse training dataset, contributing to its enhanced predictive capabilities. It achieves a high Pearson correlation coefficient of approximately 0.76 with low standard deviations, demonstrating its reliability and precision in predicting binding affinities. These results surpass traditional scoring methods, positioning BAPPL+ as a robust tool for ranking drug candidates effectively.

BAPPL+ has been rigorously evaluated against state-of-the-art scoring systems, consistently exhibiting superior efficacy in predicting binding affinities. While its overall performance is robust, evaluations of target-specific proteins reveal certain limitations that provide opportunities for further refinement. These insights pave the way for iterative improvements, ensuring that BAPPL+ remains a dependable and precise framework for evaluating candidate compounds. The versatility of BAPPL+ is exemplified through its integration with various computational methods. It effectively calculates overall binding free energies of protein-inhibitor complexes throughout MD simulations, and can be seamlessly combined with molecular docking, quantum mechanical calculations, and molecular dynamics simulations to provide comprehensive understanding of inhibitor binding mechanisms (Kiruthika et al., 2021).

By accurately predicting binding affinities, BAPPL+ facilitates the ranking of drug candidates, streamlining the drug discovery process for both metallo and non-metallo protein targets (Soni et al., 2020; Jain and Jayaram, 2005; Jain and Jayaram, 2007). Its integration of machine learning and comprehensive dataset training, coupled with its proven applications in complex computational workflows, underscores its potential as a transformative tool in computational drug discovery, driving innovation in the identification and optimization of therapeutic compounds.

4.3 Molecular property predictor

MolPropPrep (MP2) takes advantage of a novel "bond order" matrix representation of SMILES notation and utilizes the message passing neural networks (MPNNs) with a built-in semi master node to predict 15 different physico-chemical properties such as HOMO-LUMO energy gaps, dipole moments, zero-point vibrational energies (Brahmavar et al., 2024). With the introduction of semi master node in the MPNN network, one can reverse engineer the possible contributions of various functional groups to the druglikeness of small molecules. With the current implementation of this architecture and "Bond order" matrix (Figure 4), MP2 could achieve an average error ratio of 0.61, across all the predicted properties, which is an order of magnitude better than the state-of-the-art tools.

4.4 AI driven methods for RNA-Small molecule interactions

While traditional computational drug design has primarily focused on proteins, recent advances in AI have enabled effective screening of small molecules targeting RNA structures. Structure-based drug design (SBDD) targeting RNA presents unique challenges due to RNA's conformational plasticity and dynamic nature, making sequence information alone insufficient for accurate predictions (Kozlovskii and Popov, 2021).

Several AI-driven approaches have emerged to address these challenges. BiteNetN, a pioneering structure-based deep learning method, effectively detects binding sites in nucleic acid structures, working with arbitrary nucleic acid complexes to demonstrate state-of-the-art performance (Kozlovskii and Popov, 2021). CplxCavity employs a two-step process, first determining surface cavities using atomic coordinates, then utilizing machine learning to predict binding sites (Pan, 2023). Additionally, geometric deep learning methods using RNA-ligand Surface Interaction Fingerprinting (RLASIF) have shown excellent performance in characterizing binding affinity through molecular surface features (Xia et al., 2025).

RNA-specific considerations have led to specialized prediction tools. RSAPred offers quantitative models for predicting RNA-small molecule binding affinity across six RNA subtypes, incorporating RNA sequence-based and small molecule structure-based features (Krishnan et al., 2024). DrugPred_RNA, though trained on protein pockets, successfully identifies druggable RNA binding sites using descriptors applicable to both RNA and protein binding sites (Rekand and Brenk, 2021).

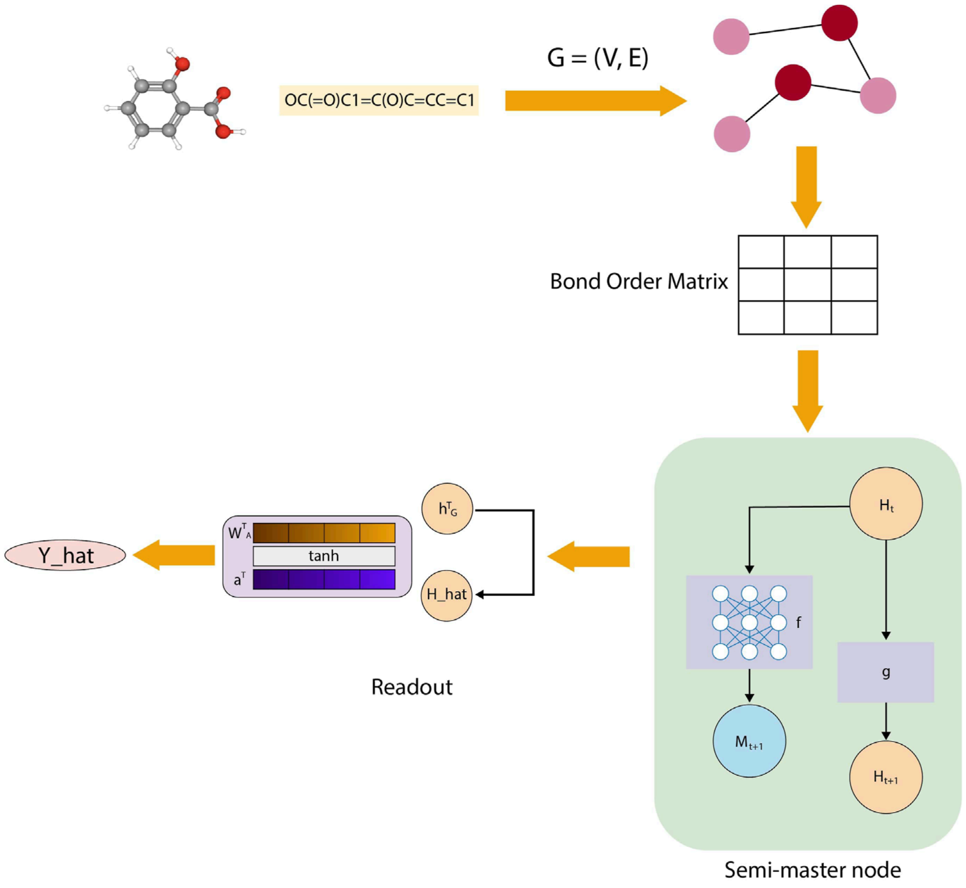

FIGURE 4
Workflow schematic of the molecular modeling pipeline. The process begins with molecular graph construction, followed by generation of a bond order matrix. A semi-master node is then incorporated to enhance information flow across the molecular structure. Finally, the Message Passing Neural Network (MPNN) processes this enhanced representation to generate predictions. This stepwise approach enables comprehensive molecular analysis while maintaining computational efficiency.

The development of these tools account for unique characteristics of RNA-binding compounds, which typically exhibit lower octanol-water partition coefficients, greater topological polar surface areas, and more hydrogen bond donors and acceptors compared to protein-binding compounds (Childs-Disney et al., 2022). Despite these advances, the field faces limitations due to the relatively small number of available RNA structures for training deep learning models (Kozlovskii and Popov, 2021). However, continued development of AI methods, combined with experimental techniques like molecular dynamics simulations, promises to enhance our ability to predict and optimize RNA-small molecule interactions.

4.5 Screening approaches for multi-molecule and complex systems

The complexity of biological systems often necessitates considering multiple molecules and protein complexes in screening approaches. Recent advances in computational methods have made it feasible to predict drug combination effects and protein complex interactions efficiently.

Drug combination prediction has evolved into both classification and regression tasks, with deep learning models demonstrating superior performance in handling large High-Throughput Screening (HTS) datasets (Liu et al., 2023). Sequential Model Optimization (SMO) methods iteratively adapt to new observations, identifying highly synergistic combinations while reducing experimental burden compared to exhaustive searches (Bertin et al., 2022). The RECOVER platform exemplifies this approach, utilizing deep neural networks to predict synergy scores based on molecular fingerprints and structural features (Bertin et al., 2022).

In the target space, protein-protein interactions (PPIs) present unique challenges due to their typically large, flat interfaces (Voet and Zhang, 2012; Koes et al., 2018). AnchorQuery, a specialized web application, enables rational structure-based design of PPI inhibitors through rapid screening of synthesizable compounds. This approach particularly focuses on anchor side chains, which form energetic hot spots at binding interfaces (Koes et al., 2018).

Virtual Screening methodologies have been adapted for Small Molecule Protein-Protein Interaction Inhibitors (SMPPII) discovery, incorporating molecular docking simulations and pharmacophore modeling (Voet and Zhang, 2012). Pharmacophore models provide abstract 3D representations of essential chemical functionalities, guiding the docking of compounds to ensure desired conformations and interactions. The comparison of PPI complexes with receptor-SMPPII structures enables visual observation of interaction mimicry by small molecule ligands (Voet and Zhang, 2012).

Challenges in these complex systems include data discrepancies from inconsistent generation processes, systematic biases limiting model generalizability, and protein mutations driving resistance (Bertin et al., 2022). Advanced models like ComboKR address these challenges by predicting drug combination response surfaces using normalized data schemes (Huusari et al., 2025). The integration of cheminformatics techniques, including structure-based virtual screening and molecular dynamics, with experimental validation has proven effective in identifying dual inhibitors (Melagraki et al., 2017), demonstrating the power of combining multiple computational approaches for complex system analysis.

5 Integration of domain knowledge with deep learning

The progression of scientific knowledge is typically characterized by the gradual refinement of existing theories, interspersed with revolutionary breakthroughs. Recent advancements in artificial intelligence (AI) have made the prospect of AI-based scientific assistants increasingly viable, offering the potential to accelerate routine reasoning and even generate transformative ideas. For such systems to be effective, they must incorporate concepts, relations, and hypotheses familiar to human scientists. While symbolic techniques have long been employed for hypothesis generation and testing due to their ability to reuse knowledge, modern neural-based deep learning approaches provide distinct advantages. These include significantly higher predictive performance, the ability to directly process diverse observational data, and the development of interactive systems through advancements in neural language models.

However, neural methods face challenges in leveraging formalized scientific knowledge to improve predictions, offer meaningful explanations, or ensure model correctness when generating new concepts or relationships. This section explores the feasibility of embedding formal domain knowledge into deep neural networks, demonstrating its utility through case studies focused on toxicity prediction, explanation, and molecular generation in drug discovery. By integrating symbolic knowledge with graph neural networks (GNNs), these studies highlight how hybrid approaches can enhance data representation, predictive accuracy, and overall system effectiveness. Figure 5 represents a black-box model with the following inputs and outputs, as employed in these AI-driven methodologies.

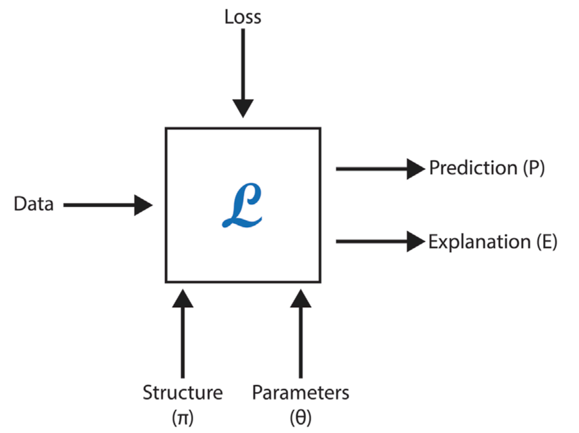

FIGURE 5
Summary of module L, which incorporates a deep network. The structure, parameters, and loss function correspond to the inputs of the deep network.

5.1 Case study 1: inclusion of domain knowledge to improve prediction

Understanding how domain knowledge can enhance deep learning models' data representation is crucial in the field of machine learning. Recent research has demonstrated an innovative approach combining Graph Neural Networks (GNN) as the machine learning engine (L) with a logical inference engine (⊢) for integrating domain-specific knowledge. In a comprehensive investigation of toxicity prediction, researchers analyzed a substantial dataset comprising approximately 225,000 molecules distributed across 73 individual datasets, with each dataset containing around 3,000 molecules classified as either "toxic" or "non-toxic" based on IC50 values. The methodology incorporated domain knowledge through formal symbolic definitions of chemical concepts, including functional groups, rings, and connected structures, encompassing roughly 100 relations expressed in formalized logical notation. By employing a logical inference engine to apply these definitions, the researchers enriched molecular graphs with detailed domain-specific information. These enhanced graphs were then processed using a specialized GNN model called BotGNN to distinguish between toxic and non-toxic molecules (Dash et al., 2022).

Figure 6 illustrates the setup of this methodology, showcasing the flow from domain knowledge integration to GNN processing. The results demonstrate a significant improvement in predictive accuracy across most datasets. The study compares BotGNN models built using five different GNN architectures. In the comparison, baseline GNN models and state-of-the-art models using approximate background knowledge (referred to as VEGNN) are outperformed by BotGNN. This highlights the value of incorporating comprehensive background knowledge in enhancing model performance.

5.2 Case study 2: inclusion of domain knowledge to improve explanations

The intersection of domain knowledge and machine learning interpretability presents compelling opportunities for research advancement. A key development in this area, as demonstrated by Srinivasan et al. (2024), employs a feedforward

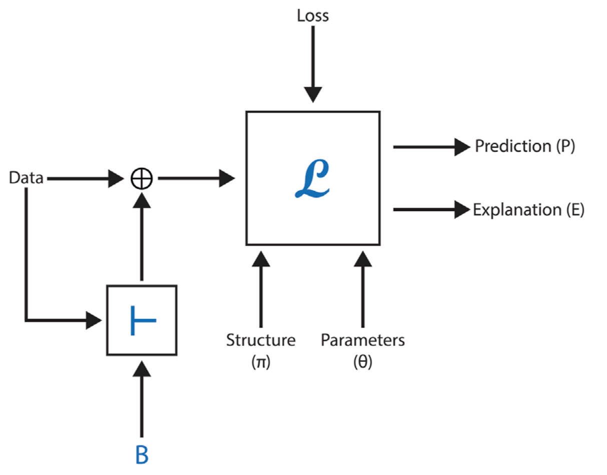

FIGURE 6
This figure shows the process of integrating domain knowledge into molecular graphs using a logical inference engine, followed by toxicity prediction with BotGNN.

neural network featuring a Compositional Relational Machine (CRM), which reformulates how domain knowledge structures data provided to deep learning models. This innovative framework was tested on a focused subset of data, specifically utilizing 10 datasets from a larger collection of 73 toxicity datasets, along with synthetic data containing known correct explanations as benchmarks. The research extends beyond basic chemical definitions to incorporate meta-information about chemical concepts and relations, establishing important constraints such as how rings and groups consist of sets of atoms, and how fused or connected structures require at least two structures of potentially different types.

Figure 7 illustrates the CRM architecture and its process of iterative feature construction. The framework utilizes provided meta-information to automatically generate a unique set of “simple features,” from which all other complex molecular features can be provably obtained through logical inference. These features are combined iteratively, with each step incorporating at least one simple feature with either another simple feature or a complex feature. The resulting feedforward network structure positions each non-input node as a complex feature, and the CRM is trained using stochastic gradient descent (SGD). The CRM serves as a proxy explainer for the BotGNN model, providing explanations by examining activations within the CRM when both models predict the same label. This approach enables a tree-like explanation structure, as the most relevant nodes can be backtracked to reveal the features involved in the CRM’s prediction. An example of this tree-like explanation structure is depicted in Figure 8.

The study effectively demonstrates how CRMs can function as interpretable proxies for more complex models like BotGNN, offering a structured approach to understanding model predictions through the lens of domain-specific features and relationships.

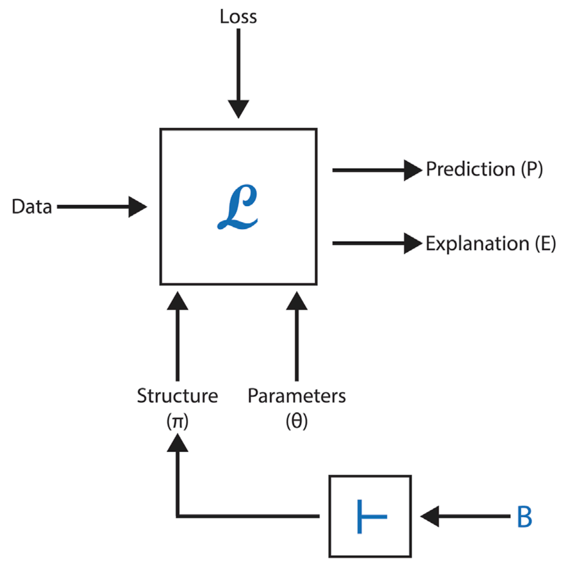

FIGURE 7
Architectural overview of the Compositional Relational Machine (CRM). The diagram illustrates how the model iteratively constructs complex features from simple features, resulting in a feedforward network structure where each non-input node represents a complex feature that can be used for toxicity prediction explanations.

5.3 Case study 3: inclusion of domain knowledge to improve generation

Recent advances in molecular design have opened new possibilities for generating molecules with specific binding properties and physicochemical characteristics. One such approach, developed by Bhat et al. (2024), leverages domain knowledge to refine the loss function of deep learning models, specifically focusing on generating molecules capable of binding to known target macromolecules while satisfying various constraints. As illustrated in Figure 9, this method incorporates a unique feedback mechanism to enhance molecular generation.

The framework’s core component employs a Large Language Model (LLM) augmented with a novel feedback loop, known as “language models with logical feedback” (LMLF). This system iteratively identifies and reinforces constraints that guide the model toward generating “good molecules” - those meeting specified logical criteria - effectively modifying the model’s loss function through indirect means. When tested on established benchmarks for Janus kinase inhibition, the LMLF approach demonstrated superior performance, generating molecules with higher estimated binding affinity compared to both state-of-the-art methods and conventional LLMs without logical feedback. Notably, computational chemists provided favourable evaluations of the LMLF-generated molecules, particularly highlighting their novelty and potential efficacy.

6 Conclusion

The evolution of molecular libraries continues to play a pivotal role in the drug discovery process, bridging traditional

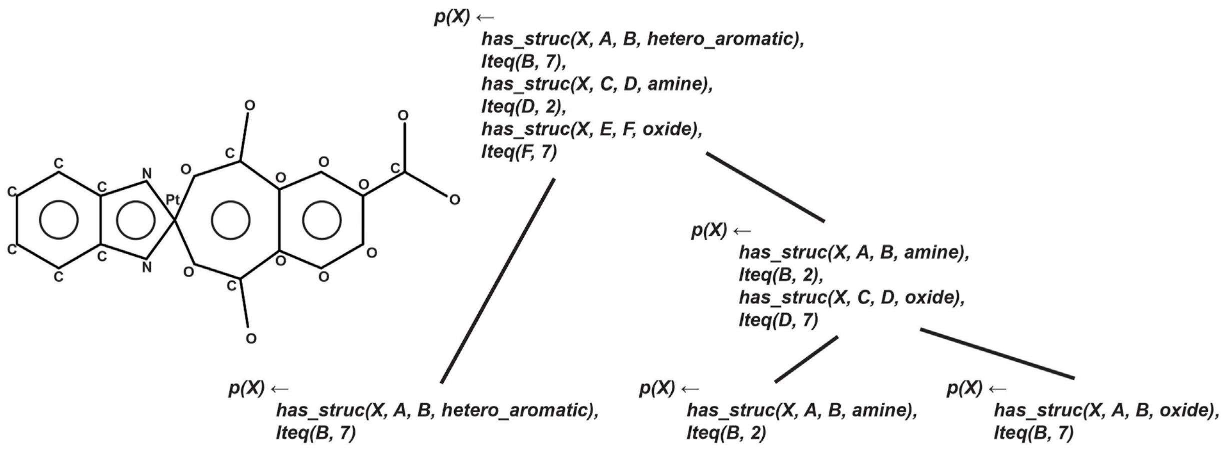
p(X) ←
has_struc(X, A, B, hetero_aromatic),
lteq(B, 7),
has_struc(X, C, D, amine),
lteq(D, 2),
has_struc(X, E, F, oxide),
lteq(F, 7)
p(X) ←
has_struc(X, A, B, amine),
lteq(B, 2),
has_struc(X, C, D, oxide),
lteq(D, 7)
p(X) ←
has_struc(X, A, B, hetero_aromatic),
lteq(B, 7)
p(X) ←
has_struc(X, A, B, amine),
lteq(B, 2)
p(X) ←
has_struc(X, A, B, oxide),
lteq(B, 7)

FIGURE 8
Example of a tree-like explanation structure generated by the CRM. The visualization demonstrates how activations are backtraced through the network to provide interpretable explanations of toxicity predictions, showing the hierarchical relationship between simple and complex features that contributed to the model's decision.

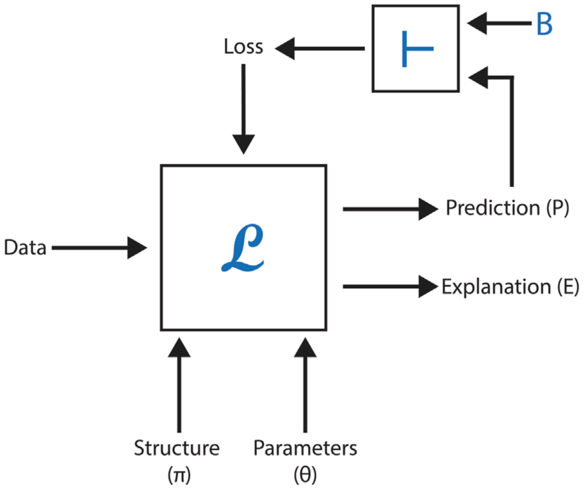

FIGURE 9
Schematic representation of the Language Models with Logical Feedback (LMLF) framework. The diagram shows the integration of domain knowledge into the model's loss function, highlighting the feedback loop mechanism that iteratively refines the generation of molecules meeting specified logical constraints for target macromolecule binding.

methodologies with new advancements in computational and AI technologies. From the historical progression of combinatorial chemistry in the 1980s to modern AI-driven approaches, the field has demonstrated remarkable adaptability in addressing emerging challenges. Traditional approaches, such as template-based generation and curated libraries, have laid a strong foundation for molecular exploration. For instance, curated databases like NRDBSM and BIMP streamline the discovery process by offering pre-screened collections of drug-like molecules and phytochemicals, each designed to address specific research needs. These resources emphasize physicochemical properties, drug-likeness criteria, and accessibility for high-throughput virtual screening, enhancing their utility for early-stage discovery.

The integration of ML and AI into modern approaches marks a transformative step in molecular library development. These technologies not only expand the exploration of chemical space but also enable the generation of novel, synthesizable compounds with tailored properties. Tools such as RASPD+ and BAPPL+ exemplify how computational methods are advancing ligand screening and binding affinity predictions, reducing computational costs while maintaining robust accuracy. The incorporation of adverse drug reaction databases and OpenTargets data has further enhanced prediction efficacy, while new approaches for screening RNA targets and multi-molecule combinations demonstrate the field's expanding scope.

Looking ahead, emerging trends signal a shift toward more diverse and complex molecular libraries, incorporating hybrid approaches that blend computational predictions with experimental validation. The rise of ultra-large virtual libraries, target-focused collections, and AI-driven generative models underscores the growing emphasis on innovation and efficiency in drug discovery. Additionally, the development of specialized assessment criteria for newer modalities like macrocycles and PROTACs reflects the field's adaptability to emerging therapeutic approaches. By leveraging these advancements while acknowledging the limitations of traditional filtering methods, researchers can identify and optimize promising candidates more effectively, accelerating the path from molecular design to therapeutic application.

Ultimately, the convergence of traditional expertise, modern computational tools, and specialized knowledge bases promises to reshape the landscape of drug discovery, unlocking new opportunities for addressing complex biological challenges and improving human health.

Author contributions

AK: Writing–original draft, Writing–review and editing, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization. DC: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. SS: Writing–review and editing. AS: Writing–original draft, Writing–review and editing. SM: Writing–original draft, Writing–review and editing. RA: Writing–original draft, Writing–review and editing. BJ: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The authors would like to thank the Department of Biotechnology, Govt. of India, under the DBT NNP scheme (BT/PR40236/BTIS/137/51/2022) for their generous funding to carry out this work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

An, S., and Fu, L. (2018). Small-molecule PROTACs: an emerging and promising approach for the development of targeted therapy drugs. *EBioMedicine* 36, 553–562. doi:10.1016/j.ebiom.2018.09.005

Appell, K., Baldwin, J. J., and Egan, J. W. (2001). "Combinatorial chemistry and high-throughput screening in drug discovery and development," in *Handbook of modern pharmaceutical analysis* (Academic Press), 24–56.

Békés, M., Langley, D. R., and Crews, C. M. (2022). PROTAC targeted protein degraders: the past is prologue. *Nat. Rev. Drug Discov.* 21, 181–200. doi:10.1038/s41573-021-00371-6

Bertin, P., Rector-Brooks, J., Sharma, D., Gaudelet, T., Anighoro, A., Gross, T., et al. (2022). RECOVER: sequential model optimization platform for combination drug repurposing identifies novel synergistic compounds *in vitro*. *ArXiv*. Available online at: http://arxiv.org/abs/2202.04202.

Bhat, B. S., Srinivasan, A., Dash, T., Krishnan, S. R., Vig, L., Roy, A., et al. (2024). "Generating novel leads for drug discovery using LLMs with logical feedback," in Proceedings of the AAAI conference on artificial intelligence. USA, February 20–27, 2024, Available online at: https://ojs.aaai.org/index.php/AAAI/article/view/27751.

Bhat, R., Kaushik, R., Singh, A., DasGupta, D., Jayaraj, A., Soni, A., et al. (2020). A comprehensive automated computer-aided discovery pipeline from genomes to hit molecules. *Chem. Eng. Sci.* 222, 115711. doi:10.1016/j.ces.2020.115711

Blanco, M. J., and Gardinier, K. M. (2020). New chemical modalities and strategic thinking in early drug discovery. *Am. Chem. Soc.* 11, 228–231. doi:10.1021/acsmedchemlett.9b00582

Bon, M., Bilsland, A., Bower, J., and McAulay, K. (2022). Fragment-based drug discovery—the importance of high-quality molecule libraries. *Mol. Oncol.* 16, 3761–3777. doi:10.1002/1878-0261.13277

Brahmavar, S. B., Shelar, M. M., Harinarthini, R., Sai Krishna, B. H., Kumta, N. H., Wadhwani, O., et al. (2024). Efficient integration of molecular representation and message-passing neural networks for predicting small molecule drug-like properties. *ChemRxiv*. Available online at: https://chemrxiv.org/engage/chemrxiv/article-details/65f321099138d23161741bae.

Buniello, A., Suveges, D., Cruz-Castillo, C., Llinares, M. B., Cornu, H., Lopez, I., et al. (2025). Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Res.* 53 (D1), D1467–D1475. doi:10.1093/nar/gkae1128

Chandrasekhar, V., Rajan, K., Kanakam, S. R. S., Sharma, N., Weiβenborn, V., Schaub, J., et al. (2025). COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database. *Nucleic Acids Res.* 53 (D1), D634–D643. doi:10.1093/nar/gkae1063

Chang, Y., Hawkins, B. A., Du, J. J., Groundwater, P. W., Hibbs, D. E., and Lai, F. (2023). A guide to *in silico* drug design. *Pharmaceutics* 15 (1), 49. doi:10.3390/pharmaceutics15010049

Chaurasia, D. K., Anjum, R., Sharma, A., Mishra, M., Jayaram, B., and Patel, A. K. (2024). BIMP Database. Available online at: https://scfbio.iitd.ac.in/bimp/.

Chen, R., Duffy, Á., Petrazzini, B. O., Vy, H. M., Stein, D., Mort, M., et al. (2024). Expanding drug targets for 112 chronic diseases using a machine learning-assisted genetic priority score. *Nat. Commun.* 15 (1), 8891. doi:10.1038/s41467-024-53333-y

Childs-Disney, J. L., Yang, X., Gibaut, Q. M. R., Tong, Y., Batey, R. T., and Disney, M. D. (2022). Targeting RNA structures with small molecules. *Nat. Rev. Drug Discov.* 21 (10), 736–762. doi:10.1038/s41573-022-00521-4

Chuang, K. V., Gunsalus, L. M., and Keiser, M. J. (2020). Learning molecular representations for medicinal chemistry. *J. Med. Chem.* 63 (16), 8705–8722. doi:10.1021/acs.jmedchem.0c00385

Congreve, M., Chessari, G., Tisi, D., and Woodhead, A. J. (2008). Recent developments in fragment-based drug discovery. *J. Med. Chem.* 51, 3661–3680. doi:10.1021/jm8000373

Dandapani, S., Rosse, G., Southall, N., Salvino, J. M., and Thomas, C. J. (2012). Selecting, acquiring, and using small molecule libraries for high-throughput screening. *Curr. Protoc. Chem. Biol.* 4 (3), 177–191. doi:10.1002/9780470559277.ch110252

Dash, T., Srinivasan, A., and Baskar, A. (2022). Inclusion of domain-knowledge into GNNs using mode-directed inverse entailment. *Mach. Learn* 111 (2), 575–623. doi:10.1007/s10994-021-06090-8

Doak, B. C., and Kihlberg, J. (2017). Drug discovery beyond the rule of 5 - opportunities and challenges. *Expert Opin. Drug Discov.* 12, 115–119. doi:10.1080/17460441.2017.1264385

Druker, B. J., and Lydon, N. B. (2000). Lessons learned from the development of an Abl tyrosine kinase inhibitor for chronic myelogenous leukemia. *J. Clin. Investigation* 105, 3–7. doi:10.1172/JCI9083

Garcia Jimenez, D., Poongavanam, V., and Kihlberg, J. (2023). Macrocycles in drug Discovery—Learning from the past for the future. *J. Med. Chem.* 66 (8), 5377–5396. doi:10.1021/acs.jmedchem.3c00134

Gottipati, S. K., Sattarov, B., Niu, S., Pathak, Y., Wei, H., Liu, S., et al. (2020). Learning to navigate the synthetically accessible chemical space using reinforcement learning. Available online at: http://arxiv.org/abs/2004.12485.

Halip, L., Avram, S., Curpan, R., Borota, A., Bora, A., Bologa, C., et al. (2023). Exploring DrugCentral: from molecular structures to clinical effects. *J. Computer-Aided Mol. Des.* 37, 681–694. doi:10.1007/s10822-023-00529-x

Han, B., Ren, C., Wang, W., Li, J., and Gong, X. (2023). Computational prediction of protein intrinsically disordered region related interactions and functions. *Genes (Basel)* 14 (2), 432. doi:10.3390/genes14020432

Han, Y., Klinger, K., Rajpal, D. K., Zhu, C., and Teeple, E. (2022). Empowering the discovery of novel target-disease associations via machine learning approaches in the open targets platform. *BMC Bioinforma.* 23 (1), 232. doi:10.1186/s12859-022-04753-4

Harris, C. J., Hill, R. D., Sheppard, D. W., Slater, M. J., and Stouten, P. F. W. (2011). The design and application of target-focused compound libraries. *Comb. Chem. High. Throughput Screen* 14 (6), 521–531. doi:10.2174/138620711795767802

Heinzke, A. L., Pahl, A., Zdrazil, B., Leach, A. R., Waldmann, H., Young, R. J., et al. (2024). Occurrence of "natural selection" in successful small molecule drug discovery. *J. Med. Chem.* 67 (13), 11226–11241. doi:10.1021/acs.jmedchem.4c00811

Holderbach, S., Adam, L., Jayaram, B., Wade, R. C., and Mukherjee, G. (2020). RASPD+: fast protein-ligand binding free energy prediction using simplified physicochemical features. *Front. Mol. Biosci.* 7, 601065. doi:10.3389/fmolb.2020.601065

Hornberger, K. R., and Araujo, E. M. V. (2023). Physicochemical property determinants of oral absorption for PROTAC protein degraders. *J. Med. Chem.* 66 (12), 8281–8287. doi:10.1021/acs.jmedchem.3c00740

Huusari, R., Wang, T., Szedmak, S., Aittokallio, T., and Rousu, J. (2025). Predicting drug combination response surfaces. *npj Drug Discov.* 2 (1), 2. doi:10.1038/s44386-024-00004-z

Ietswaart, R., Arat, S., Chen, A. X., Farahmand, S., Kim, B., DuMouchel, W., et al. (2020). Machine learning guided association of adverse drug reactions with *in vitro* target-based pharmacology. *EBioMedicine* 57, 102837. doi:10.1016/j.ebiom.2020.102837

Irwin, J. J., Tang, K. G., Young, J., Dandarchuluun, C., Wong, B. R., Khurelbaatar, M., et al. (2020). ZINC20 - a free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model* 60 (12), 6065–6073. doi:10.1021/acs.jcim.0c00675

Ivanenkov, Y. A., Polykovskiy, D., Bezrukov, D., Zagribelnyy, B., Aladinskiy, V., Kamya, P., et al. (2023). Chemistry42: an AI-driven platform for molecular design and optimization. *J. Chem. Inf. Model* 63 (3), 695–701. doi:10.1021/acs.jcim.2c01191

Jain, T., and Jayaram, B. (2005). An all atom energy based computational protocol for predicting binding affinities of protein-ligand complexes. *FEBS Lett.* 579 (29), 6659–6666. doi:10.1016/j.febslet.2005.10.031

Jain, T., and Jayaram, B. (2007). Computational protocol for predicting the binding affinities of zinc containing metalloprotein-ligand complexes. *Proteins Struct. Funct. Genet.* 67 (4), 1167–1178. doi:10.1002/prot.21332

Jayaram, B., Latha, N., In, J., Sharma, P., Gandhimathi, A., and Pandey, V. S. (2006). Sanjeevini: a comprehensive active site directed lead design software. *Indian J. Chem.* 45A, 1834–1837. Available online at: http://www.scfbio-iitd.res.in/drugdes/sanjeevini.html.

Jayaram, B., Singh, T., Mukherjee, G., Mathur, A., Shekhar, S., and Shekhar, V. (2012). Sanjeevini: a freely accessible web-server for target directed lead molecule discovery. *BMC Bioinforma.* 13 (Suppl. 17), S7. doi:10.1186/1471-2105-13-S17-S7

Kingwell, K. (2023). Macrocycle drugs serve up new opportunities. *Nat. Rev. Drug Discov.* 22, 771–773. doi:10.1038/d41573-023-00152-3

Kiruthika, S., Bhat, R., Dash, R., Rathore, A. S., Vivekanandan, P., and Jayaram, B. (2021). A novel piperazine derivative that targets hepatitis B surface antigen effectively inhibits tenofovir resistant hepatitis B virus. *Sci. Rep.* 11 (1), 11723. doi:10.1038/s41598-021-91196-1

Koes, D. R., Dömling, A., and Camacho, C. J. (2018). AnchorQuery: rapid online virtual screening for small-molecule protein–protein interaction inhibitors. *Protein Sci.* 27 (1), 229–232. doi:10.1002/pro.3303

Korablyov, M., Liu, C. H., Jain, M., van der Sloot, A. M., Jolicoeur, E., Ruediger, E., et al. (2024). Generative active learning for the search of small-molecule protein binders. Available online at: http://arxiv.org/abs/2405.01616.

Koscielny, G., An, P., Carvalho-Silva, D., Cham, J. A., Fumis, L., Gasparyan, R., et al. (2017). Open Targets: a platform for therapeutic target identification and Validation. *Nucleic Acids Res.* 45 (D1), D985-D994–94. doi:10.1093/nar/gkw1055

Kozlovskii, I., and Popov, P. (2021). Structure-based deep learning for binding site detection in nucleic acid macromolecules. *Nar. Genom Bioinform* 3 (4), lqab111. doi:10.1093/nargab/lqab111

Krishnan, S. R., Roy, A., and Michael Gromiha, M. (2024). Reliable method for predicting the binding affinity of RNA-small molecule interactions using machine learning. *Brief. Bioinform* 25 (2), bbae002. doi:10.1093/bib/bbae002

Latha, N., Jain, T., Sharma, P., and Jayaram, B. (2004). A free energy based computational pathway from chemical templates to lead compounds: a case study of cox-2 inhibitors. *J. Biomol. Struct. Dyn.* 21 (6), 791–804. doi:10.1080/07391102.2004.10506969

Latha, N., and Jayaram, B. (2005). A binding affinity based computational pathway for active-site directed lead molecule design: some promises and perspectives. *Drug Des. Reviews-Online* 2, 145–165. doi:10.2174/1567269053202688

Lipinski, C. A., Franco, L., Dominy, B. W., and Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25. doi:10.1016/s0169-409x(96)00423-1

Liu, H., Fan, Z., Lin, J., Yang, Y., Ran, T., and Chen, H. (2023). The recent progress of deep-learning-based *in silico* prediction of drug combination. *Drug Discov. Today* 28 (7), 103625. doi:10.1016/j.drudis.2023.103625

Liu, R., Li, X., and Lam, K. S. (2017). "Combinatorial chemistry in drug discovery," in *Current opinion in chemical biology*. Elsevier Ltd, 38, 117–126.

Liu, Y., and Aickelin, U. (2021). Feature selection in detection of adverse drug reactions from the health improvement network (THIN) database.

Loeffler, H. H., He, J., Tibo, A., Janet, J. P., Voronov, A., Mervin, L. H., et al. (2024). Reinvent 4: modern AI–driven generative molecule design. *J. Cheminform* 16 (1), 20. doi:10.1186/s13321-024-00812-5

McNaughton, A. D., Bontha, M. S., Knutson, C. R., Pope, J. A., and Kumar, N. (2022). *De novo* design of protein target specific scaffold-based Inhibitors via Reinforcement Learning. Available online at: http://arxiv.org/abs/2205.10473.20.

Melagraki, G., Ntougkos, E., Rinotas, V., Papaneophytou, C., Leonis, G., Mavromoustakos, T., et al. (2017). Cheminformatics-aided discovery of small-molecule protein-protein interaction (PPI) dual inhibitors of tumor necrosis factor (TNF) and receptor activator of NF-κB ligand (RANKL). *PLoS Comput. Biol.* 13 (4), e1005372. doi:10.1371/journal.pcbi.1005372

Mohsen, A., Tripathi, L. P., and Mizuguchi, K. (2020). Deep learning prediction of adverse drug reactions using open TG-GATEs and FAERS databases. Available online at: http://arxiv.org/abs/2010.05411.

Müller, B. A. (2009). Imatinib and its successors-how modern chemistry has changed drug development. *Curr. Pharm. Des.* 15, 120–133. doi:10.2174/138161209787002933

Mureddu, L. G., and Vuister, G. W. (2022). Fragment-based drug discovery by NMR. Where are the successes and where can it Be improved? *Front. Mol. Biosci.* 9, 9. doi:10.3389/fmolb.2022.834453

Murray, C. W., and Rees, D. C. (2009). The rise of fragment-based drug discovery. *Nat. Chem.* 1 (3), 187–192. doi:10.1038/nchem.217

Pan, N. (2023). Predicting RNA-small molecule binding sites by 3D structure. Available online at: http://arxiv.org/abs/2310.18985.

Patel, L., Shukla, T., Huang, X., Ussery, D. W., and Wang, S. (2020). Machine learning methods in drug discovery. *Molecules* 25 (22), 5277. doi:10.3390/molecules25225277

Popova, M., Isayev, O., and Tropsha, A. (2018). Deep reinforcement learning for de-novo drug design. *Sci. Adv.* 4 (7), eaap7885. doi:10.1126/sciadv.aap7885

Rekand, I. H., and Brenk, R. (2021). DrugPred_RNA - a tool for structure-based druggability predictions for RNA binding sites. *J. Chem. Inf. Model* 61 (8), 4068–4081. doi:10.1021/acs.jcim.1c00155

Ress, D. C., Congreve, M., Murray, C. W., and Carr, R. (2004). Fragment-based lead discovery. *Nat. Rev. Drug Discov.* 3, 660–672. doi:10.1038/nrd1467

Reutlinger, M., and Schneider, G. (2012). Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J. Mol. Graph Model* 34, 108–117. doi:10.1016/j.jmgm.2011.12.006

Ruan, H., Yu, C., Niu, X., Zhang, W., Liu, H., Chen, L., et al. (2021). Computational strategy for intrinsically disordered protein ligand design leads to the discovery of p53 transactivation domain I binding compounds that activate the p53 pathway. *Chem. Sci.* 12 (8), 3004–3016. doi:10.1039/d0sc04670a

Sadybekov, A. V., and Katritch, V. (2023). Computational approaches streamlining drug discovery. *Nature* 616, 673–685. doi:10.1038/s41586-023-05905-z

Saldívar-González, F. I., Huerta-García, C. S., and Medina-Franco, J. L. (2020). Chemoinformatics-based enumeration of chemical libraries: a tutorial. *J. Cheminform* 12 (1), 64. doi:10.1186/s13321-020-00466-z

Sarkar, C., Das, B., Rawat, V. S., Wahlang, J. B., Nongpiur, A., Tiewsoh, I., et al. (2023). Artificial intelligence and machine learning technology driven modern drug discovery and development. *Int. J. Mol. Sci.* 24 (3), 2026. doi:10.3390/ijms24032026

Shaikh, S. A., Jain, T., Sandhu, G., Latha, N., and Jayaram, B. (2007). From drug target to leads-sketching A physicochemical pathway for lead molecule design *in silico*. *Curr. Pharm. Des.* 13, 3454–3470. doi:10.2174/138161207782794220

Shaikh, S. A., Jain, T., Sandhu, G., Soni, A., and Jayaram, B. (2012). From drug target to leads-sketching A physico-chemical pathway for lead molecule design *in silico*. *Front. Med. Chem.* 6, 324–360. doi:10.2174/9781608054640113060015

Soni, A., Bhat, R., and Jayaram, B. (2020). Improving the binding affinity estimations of protein–ligand complexes using machine-learning facilitated force field method. *J. Comput. Aided Mol. Des.* 34 (8), 817–830. doi:10.1007/s10822-020-00305-1

Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A., and Steinbeck, C. (2021). COCONUT online: collection of open natural products database. *J. Cheminform* 13 (1), 2. doi:10.1186/s13321-020-00478-9

Srinivasan, A., Baskar, A., Dash, T., and Shah, D. (2024). Composition of relational features with an application to explaining black-box predictors. *Mach. Learn* 113 (3), 1091–1132. doi:10.1007/s10994-023-06399-6

Sun, C., Petros, A. M., and Hajduk, P. J. (2011). Fragment-based lead discovery: challenges and opportunities. *J. Comput. Aided Mol. Des.* 25 (7), 607–610. doi:10.1007/s10822-011-9451-z

Tingle, B. I., Tang, K. G., Castanon, M., Gutierrez, J. J., Khurelbaatar, M., Dandarchuluun, C., et al. (2023). ZINC-22—A free multi-billion-scale database of tangible compounds for ligand discovery. *J. Chem. Inf. Model* 63 (4), 1166–1176. doi:10.1021/acs.jcim.2c01253

Uner, O. C., Gokberk Cinbis, R., Tastan, O., and Cicek, A. E. (2019). DeepSide: a deep learning framework for drug side effect prediction. Available online at: http://biorxiv.org/lookup/doi/10.1101/843029.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18 (6), 463–477. doi:10.1038/s41573-019-0024-5

Viarengo-Baker, L. A., Brown, L. E., Rzepiela, A. A., and Whitty, A. (2021). Defining and navigating macrocycle chemical space. *Chem. Sci.* 12 (12), 4309–4328. doi:10.1039/d0sc05788f

Voet, A., and Zhang, K. Y. J. (2012). Pharmacophore modelling as a virtual screening tool for the discovery of small molecule protein-protein interaction inhibitors. *Curr. Pharm. Des.* 18, 4586–4598. doi:10.2174/138161212802651616

Wang, H., Xiong, R., and Lai, L. (2023). Rational drug design targeting intrinsically disordered proteins. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 13 (6). doi:10.1002/wcms.1685

Wang, R., Fang, X., Lu, Y., Yang, C. Y., and Wang, S. (2005). The PDBbind database: methodologies and updates. *J. Med. Chem.* 48 (12), 4111–4119. doi:10.1021/jm048957q

Wawer, M. J., Li, K., Gustafsdottir, S. M., Ljosa, V., Bodycombe, N. E., Marton, M. A., et al. (2014). Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proc. Natl. Acad. Sci. U. S. A.* 111 (30), 10911–10916. doi:10.1073/pnas.1410933111

Weng, G., Shen, C., Cao, D., Gao, J., Dong, X., He, Q., et al. (2021). PROTAC-DB: an online database of PROTACs. *Nucleic Acids Res.* 49 (D1), D1381–D1387. doi:10.1093/nar/gkaa807

Xia, W., Shu, J., Sang, C., Wang, K., Wang, Y., Sun, T., et al. (2025). The prediction of RNA-small-molecule ligand binding affinity based on geometric deep learning. *Comput. Biol. Chem.* 115, 108367. doi:10.1016/j.compbiolchem.2025.108367

Ye, C., Swiers, R., Bonner, S., and Barrett, I. P. (2024). A knowledge graph-enhanced tensor factorisation model for discovering. *Drug Targets*. Available online at: https://github.com/AstraZeneca/kg-enhanced-tf-for-target-discovery.