Genome Publications
https://doi.org/10.61096/978-81-990998-7-6_11

Chapter 11

Virtual Screening Paradigms: Structure-Based and Ligand-Based Approaches

Dr. Sethuramani A
Assistant professor, Department of Pharmacognosy, College of Pharmacy, Madurai Medical College, Madurai
P. Sivasubramaniyan
Assistant Professor, Department of Pharmaceutical Chemistry, Madurai medical college, Madurai.
A. Sebatini Sinsi
Department of Pharmacognosy, College of Pharmacy, Madurai Medical College, Madurai

Abstract: Virtual screening (VS) has become a central paradigm in computer-aided drug design (CADD), enabling the rapid identification of potential bioactive molecules from vast chemical libraries through computational filtering before experimental validation. This chapter explores the theoretical foundations, methodologies, and applications of both structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS), highlighting how these complementary strategies accelerate hit identification and lead optimization. SBVS utilizes the three-dimensional structure of target proteins to evaluate molecular binding affinities via docking and scoring, whereas LBVS relies on the similarity of known ligands to predict novel compounds with comparable biological activities. Recent advances in artificial intelligence, cheminformatics, and cloud-based screening have dramatically enhanced the scale and precision of virtual screening workflows. The chapter critically evaluates aspects such as library design, screening metrics (enrichment factor, ROC analysis), integration with experimental high-throughput screening (HTS), and hybrid strategies that unify structural and ligand-based information. Case studies illustrate the translational success of VS pipelines in discovering kinase inhibitors, GPCR modulators, and antiviral agents. Finally, emerging trends such as deep learning driven scoring functions, adaptive sampling, and ultra-large chemical space exploration are discussed in the context of future drug discovery ecosystems integrating automation and AI-guided decision support.

Keywords: Virtual screening, structure-based design, ligand-based design, docking, cheminformatics

Citation: Sethuramani A, P. Sivasubramaniyan, A. Sebatini Sinsi. Virtual Screening Paradigms: Structure-Based and Ligand-Based Approaches. *Comprehensive Approaches in Computer-Aided Drug Design: QSAR, Docking, Screening, Homology, Pharmacophore and AI-Driven Insights.* Genome Publication. Genome Publication. 2025; Pp133--144. https://doi.org/10.61096/978-81-990998-7-6_11

INTRODUCTION

1.0 Conceptual Overview of Virtual Screening

Virtual screening represents a computational strategy aimed at identifying bioactive compounds with a high likelihood of interacting favorably with a biological target before any experimental testing is performed. It is a critical component of modern drug discovery workflows, enabling researchers to prioritize chemical entities from vast molecular libraries containing millions to billions of compounds [1]. By mimicking the principles of physical high-throughput screening (HTS), virtual screening dramatically reduces experimental costs, accelerates the discovery cycle, and enhances the probability of finding viable leads with desired pharmacological properties. At its core, virtual screening encompasses two principal paradigms: Structure-Based Virtual Screening (SBVS) and Ligand-Based Virtual Screening (LBVS). SBVS leverages the three-dimensional (3D) structure of the target usually derived from X-ray crystallography, cryo-electron microscopy, or homology modeling to dock potential ligands into the binding site and estimate binding affinities through scoring functions [2]. Conversely, LBVS is utilized when the structure of the target protein is unavailable but a set of active ligands is known; it employs molecular descriptors, fingerprints, or pharmacophore models to identify compounds sharing similar physicochemical or topological features [3].

The workflow of virtual screening typically involves several stages: database preparation, filtering of compounds based on drug-likeness and ADMET properties, molecular docking or similarity searching, scoring and ranking, post-processing (rescoring or consensus scoring), and finally, visual or statistical analysis of top-ranked hits. The iterative refinement of this workflow, often in combination with experimental validation, forms the backbone of integrated computational-experimental discovery pipelines [4]. Virtual screening's impact is reflected in numerous successful case studies such as the discovery of HIV integrase inhibitors, selective kinase blockers, and GPCR-targeted small molecules where computationally predicted hits translated into clinically relevant drugs [5]. Moreover, the increasing integration of AI and machine learning has transformed traditional VS pipelines by enabling predictive scoring functions, intelligent filtering, and adaptive sampling of chemical space. Such systems, combined with cloud computing and distributed processing, now allow virtual screening of ultra-large compound libraries exceeding 10^9 molecules [6].

While virtual screening offers tremendous promise, its success depends heavily on data quality, algorithmic robustness, and the fidelity of the scoring functions used. Limitations such as inaccurate docking poses, scoring bias, and overfitting in ML-based models can lead to false positives or negatives. Thus, careful calibration, cross-validation, and benchmarking against experimental data are essential to maintain reliability and reproducibility [7]. Ultimately, the evolution of virtual screening from early rule-based approaches to modern AI-driven systems reflects a paradigm shift from heuristic selection to predictive molecular intelligence, where both structure and data-driven insights guide rational drug discovery.

1.1 Structure-Based Virtual Screening Workflow

Structure-Based Virtual Screening (SBVS) exploits the availability of a target's three-dimensional atomic coordinates to identify compounds likely to bind effectively within its active site. The fundamental principle rests on the assumption that molecular recognition is governed by complementarity in shape, electrostatics, and energetics between the ligand and the receptor binding pocket [8]. Modern SBVS workflows combine molecular docking algorithms, scoring functions, and post-docking analysis to simulate the binding process computationally. The SBVS workflow generally

begins with target structure acquisition, obtained through experimental methods (e.g., X-ray crystallography, NMR, cryo-EM) or predicted through computational modeling (homology modeling, AlphaFold) [9]. The protein preparation phase involves optimization of the 3D structure: adding missing residues, assigning protonation states, removing water molecules, defining binding sites, and optimizing hydrogen bonding networks. Software such as Schrödinger's *Protein Preparation Wizard*, *AutoDock Tools*, or *MGLTools* streamline this step [10].

Following protein preparation, ligand library preparation ensures that compounds are correctly protonated, energy-minimized, and conformationally diverse. Libraries may originate from publicly available databases like *ZINC*, *ChEMBL*, *PubChem*, or proprietary collections, often filtered for drug-likeness using criteria such as Lipinski's Rule of Five or Veber's rules [11]. Conformer generation tools like *OMEGA*, *RDKit*, and *Open Babel* generate multiple ligand geometries to account for molecular flexibility. The docking step constitutes the computational core of SBVS, where ligands are placed into the active site using search algorithms (systematic, stochastic, or hybrid) to explore possible binding poses. Popular docking engines *AutoDock Vina*, *Glide*, *GOLD*, *DOCK6*, *SwissDock*, and *Ledock* utilize diverse scoring functions (empirical, force-field–based, knowledge-based, or machine learning–derived) to estimate binding affinity [12]. Often, multiple docking runs with varied parameters are conducted to improve reliability.

Post-docking, scoring and ranking evaluate the predicted binding poses to prioritize compounds with the most favorable estimated free energies. However, single scoring functions may exhibit bias; therefore, consensus scoring combining multiple scoring functions or machine learning post-processors improves predictive accuracy [13]. Some workflows incorporate rescoring through MM–GBSA or MM–PBSA methods for higher precision. Finally, visual inspection and interaction analysis of top-ranked poses help verify key molecular interactions such as hydrogen bonds, π–π stacking, and hydrophobic contacts. Advanced visualization tools like *PyMOL*, *ChimeraX*, and *Biovia Discovery Studio* aid interpretation and validation [14].

Recent trends in SBVS emphasize automation and scalability through cloud platforms (e.g., AWS ParallelCluster, OpenEye Orion), GPU acceleration, and AI-enhanced scoring. Deep learning-based scoring models such as *DeepDock*, *GNINA*, and *DeltaVinaRF20* leverage convolutional neural networks to predict binding affinities with improved generalization [15]. Moreover, ensemble docking, where multiple receptor conformations are considered, enhances accuracy for flexible proteins. The reliability of SBVS depends on accurate protein modeling, correct identification of the binding site, and the robustness of scoring algorithms. While the computational cost is lower than experimental HTS, SBVS still faces limitations in capturing protein dynamics, solvation effects, and induced-fit phenomena. Integration with molecular dynamics (MD) simulations and free energy perturbation methods is increasingly used to overcome these shortcomings, offering a bridge between static docking and dynamic molecular recognition [16].

1.2 Ligand-Based Virtual Screening Workflow

Ligand-Based Virtual Screening (LBVS) operates on the principle that molecules exhibiting similar structural or physicochemical characteristics tend to elicit similar biological responses. Unlike SBVS, it does not require a known 3D structure of the target protein; instead, it leverages data from known active compounds to infer and predict new hits [17]. LBVS is particularly valuable in early discovery stages or for targets that are difficult to crystallize, such as GPCRs, ion channels, or membrane-associated proteins. The first stage in LBVS involves data collection and curation of known actives, inactives, and decoys from repositories such as *ChEMBL*, *BindingDB*, or *PubChem BioAssay*.

These datasets form the foundation for constructing predictive models or similarity metrics [18]. The quality of input data is critical, as experimental inconsistencies or annotation errors can propagate into misleading predictions.

Subsequently, feature extraction transforms molecular structures into computationally tractable representations. These features may include molecular descriptors (e.g., topological, physicochemical, geometrical, quantum chemical) or molecular fingerprints such as *MACCS keys*, *ECFP4/6*, or *pharmacophore fingerprints* [19]. Descriptor selection and normalization are essential to avoid redundancy and ensure model interpretability. Two major methodological categories define LBVS: similarity-based screening and pharmacophore-based screening.

Similarity-based approaches employ metrics such as the Tanimoto coefficient or cosine similarity to rank candidate molecules based on resemblance to known actives. Tools like *ROCS*, *RDKit*, and *OpenEye FastROCS* are widely used for shape and feature comparisons [20]. *Pharmacophore-based screening* abstracts essential molecular interaction features hydrogen bond donors/acceptors, aromatic rings, hydrophobic centers, and charged groups into a spatial 3D model that serves as a query to screen chemical libraries [21]. Software such as *LigandScout*, *Phase*, and *Discovery Studio* enable automatic pharmacophore hypothesis generation and virtual screening.

More advanced LBVS techniques involve machine learning and deep learning models, where molecular representations are mapped to biological activity predictions using algorithms such as random forests, support vector machines, or graph neural networks. These predictive frameworks allow generalization to new chemical spaces, enabling virtual screening even in the absence of close structural analogues [22]. Hybrid pipelines often integrate LBVS with SBVS to leverage both structural and ligand similarity insights, achieving higher hit enrichment and lower false positive rates. LBVS is computationally efficient and ideal for rapid hypothesis generation, though its accuracy depends heavily on the diversity and reliability of known actives. The absence of explicit receptor information may limit interpretability, and excessive similarity thresholds can lead to scaffold redundancy. Nonetheless, the integration of AI-based representation learning (e.g., SMILES transformers, graph embeddings) is revolutionizing LBVS, expanding its applicability across previously intractable target classes [23].

1.3 Library Design: Diversity, Focused Libraries and On-Demand Synthesis

The design and curation of compound libraries represent a pivotal determinant of success in virtual screening campaigns. Whether employing structure-based or ligand-based approaches, the coverage, diversity, and quality of the molecular library directly influence the likelihood of identifying novel, potent, and selective hits. Library design must therefore balance between chemical diversity to explore broad regions of chemical space and biological relevance, ensuring the compounds are synthetically accessible, stable, and drug-like [24]. Diversity-oriented libraries aim to represent a wide range of chemical scaffolds and functional groups. They are especially valuable in early discovery when limited prior information about the target or chemical class is available. Such libraries maximize the chance of discovering unanticipated scaffolds, which may exhibit novel mechanisms of action. Techniques for generating and assessing chemical diversity often rely on clustering algorithms based on molecular fingerprints (e.g., ECFP, MACCS) and principal component analysis (PCA) of physicochemical properties [25]. Computational platforms such as *KNIME*, *RDKit*, and *ChemAxon's Instant JChem* allow chemists to evaluate diversity using metrics like Tanimoto distances and Shannon entropy.

In contrast, focused libraries are constructed around a specific biological target class (e.g., kinases, proteases, GPCRs) or chemical scaffold family. The compounds are typically derived from known actives or pharmacophore models, emphasizing structure–activity relationship (SAR) expansion rather than novelty. Focused libraries dramatically improve screening enrichment when prior biological data are available, especially when integrated with ligand-based or docking-guided enrichment filters [26]. The development of fragment-based focused libraries, consisting of low-molecular-weight compounds with high binding efficiency, has become particularly popular in identifying weak binders that serve as starting points for fragment-based drug discovery (FBDD) [27]. Recent advances have introduced on-demand virtual libraries, which leverage automated synthesis prediction tools and reaction-based enumeration to explore theoretical chemical space far beyond commercially available compounds. Tools like *RECAP*, *ChemAxon Reactor*, and *Enamine REAL Space Navigator* allow the creation of billions of synthetically feasible virtual molecules from available building blocks [28]. Such libraries support make-on-demand synthesis, enabling researchers to rapidly transition from computational hits to experimental testing.

Library preparation also involves property-based filtering to exclude compounds likely to fail in downstream processes. Common filters include Lipinski's "Rule of Five," Veber's polar surface area criteria, PAINS (pan-assay interference compounds) filters, and reactive group eliminations [29]. Integration of ADMET filters using tools like *SwissADME*, *admetSAR*, or *pkCSM* ensures pharmacokinetic viability and toxicity risk assessment before screening. An emerging trend is the incorporation of AI-assisted library design, where machine learning models predict target affinity, diversity metrics, or synthetic accessibility scores (SAS) to generate adaptive libraries tailored to specific pharmacological profiles. Cloud-based platforms like *PostEra Manifold*, *MolSoft ICM-Pro*, and *DeepChem* provide end-to-end pipelines for library design, docking, and property optimization [30].

The future of library design lies in merging chemical space exploration with generative AI, creating adaptive, self-evolving compound repositories. Such systems will continuously update based on feedback from virtual and experimental results, establishing a dynamic loop between computational predictions and laboratory synthesis.

1.4 Screening Metrics: Enrichment Factor, ROC Curves and Hit Rates

Evaluating the effectiveness of a virtual screening campaign requires robust statistical metrics that quantify how well the computational model discriminates true actives from inactive or decoy compounds. Three primary categories of performance metrics are commonly used: enrichment metrics, classification metrics, and hit rate statistics [31]. The Enrichment Factor (EF) is one of the most widely employed measures for assessing screening performance. It quantifies how much better a virtual screening method performs in identifying actives compared to random selection. It is defined as:

$$\mathbf{EFx=(nx/Nx)(na/Na)}\ \boldsymbol{EFx=(na/Na)(nx/Nx)}$$

where n*x* is the number of actives found in the top x% of the ranked list, N*x* is the number of compounds in that subset, n*a* is the total number of actives in the dataset, and N*a* is the total number of compounds. An EF value greater than 1 indicates better-than-random performance. For example, an EF_{10} of 10 signifies that the top 10% of screened compounds are tenfold enriched in actives relative to a random selection [32]. Complementary to enrichment analysis, Receiver Operating Characteristic (ROC) curves plot the true positive rate (sensitivity) against the false positive rate (1-specificity) across all scoring thresholds. The Area Under the Curve (AUC) provides a global measure

of screening quality, with values approaching 1.0 indicating near-perfect discrimination. ROC-AUC analysis is particularly useful for comparing scoring functions or docking protocols across datasets [33]. However, as ROC-AUC can mask performance differences at early enrichment levels, researchers often prefer BEDROC (Boltzmann-enhanced discrimination of ROC), which weights early retrieval performance more heavily [34].

Hit rate is a straightforward metric representing the fraction of experimentally confirmed actives among the top-ranked predictions. While useful in practical terms, hit rates can be dataset-dependent and influenced by library bias. A typical acceptable hit rate for well-validated virtual screening pipelines ranges between 1–10%, although AI-driven hybrid methods have demonstrated higher success rates in focused campaigns [35]. To ensure meaningful performance evaluation, benchmarking against standard datasets such as *DUD-E (Directory of Useful Decoys – Enhanced)*, *ChEMBL Benchmark Suite*, or *DEKOIS 2.0* is essential [36]. These datasets provide balanced active and decoy sets with known physicochemical properties, allowing reproducible and comparable screening evaluations.

Modern workflows often employ consensus metrics, integrating enrichment factor, ROC-AUC, and precision–recall curves to achieve a more comprehensive evaluation. Furthermore, cross-validation and external validation ensure generalizability across chemical spaces and reduce overfitting of machine learning models [37]. AI-enhanced frameworks are increasingly employing active learning and uncertainty quantification to dynamically refine screening metrics, focusing computational effort on ambiguous or high-impact compounds. The transition from static metrics to adaptive performance feedback systems is expected to transform virtual screening into a more intelligent and data-driven discovery paradigm [38].

1.5 Integration with High-Throughput Screening: Opportunities and Limitations

The integration of virtual screening with experimental high-throughput screening (HTS) represents a powerful hybrid strategy that bridges computational prediction and empirical validation. Traditionally, HTS involves testing large compound libraries against biological targets using automated robotic systems, yielding millions of data points but also incurring substantial costs and time. Virtual screening serves as a computational pre-filter, drastically reducing the number of compounds subjected to physical testing and improving overall hit rates [39].

In a typical hybrid screening pipeline, virtual screening is first applied to filter large compound collections based on docking or similarity scores, ADMET criteria, or pharmacophore alignment. The top-ranked candidates are then prioritized for synthesis or experimental testing in HTS assays. This dual-layered strategy often termed *in silico–in vitro cascade screening* has proven effective in identifying high-quality leads while minimizing experimental redundancy [40]. For example, in kinase inhibitor discovery, combining structure-based virtual screening of 10^6 compounds with experimental testing of the top 1,000 candidates yielded multiple potent inhibitors with nanomolar activity, demonstrating a significant improvement over random HTS hit rates [41]. Similarly, hybrid screening has been successfully applied to GPCRs, proteases, and viral enzymes where structural data are partially available [42].

The benefits of integrating VS and HTS include

1. Cost and time efficiency – Computational filtering can reduce experimental workload by 90–95%.
2. Higher enrichment – Pre-selected compounds exhibit higher active-to-inactive ratios.

3. Diverse hit exploration – VS can identify non-obvious scaffolds missed by traditional chemical intuition.
4. Data-driven feedback – Experimental outcomes can retrain and refine computational models, creating iterative discovery cycles [43].

However, this integration also presents challenges. Virtual screening methods may produce false positives due to limitations in scoring accuracy or target flexibility. Conversely, false negatives may occur when potential actives are filtered out by overly strict thresholds. Moreover, HTS assays themselves may yield assay artifacts or false signals due to compound aggregation, fluorescence interference, or nonspecific binding [44]. Consequently, post-screening validation using orthogonal assays (e.g., SPR, ITC, NMR binding) remains crucial. The future of hybrid screening lies in AI-driven adaptive pipelines, where experimental results dynamically feed back into computational models. Emerging systems employ reinforcement learning to optimize compound selection after each experimental cycle, progressively improving predictive precision [45]. Moreover, cloud-based automation and robotics are enabling real-time VS–HTS integration, where computational scoring, synthesis, and bioassays operate in parallel under centralized data control.

Recent studies have demonstrated the efficacy of such closed-loop discovery systems, achieving iterative enrichment across multiple screening cycles for targets like SARS-CoV-2 main protease and Alzheimer's β-secretase [46]. These integrated frameworks are gradually redefining drug discovery from linear screening to autonomous, self-optimizing pipelines that blend computational intelligence with experimental throughput.

1.6 Case Studies Comparing SBVS and LBVS

To contextualize the theoretical frameworks of SBVS and LBVS, several case studies illustrate how each paradigm and their integration has led to tangible advances in hit identification and drug development. One notable example is the identification of novel acetylcholinesterase inhibitors for Alzheimer's disease. Researchers combined LBVS using pharmacophore modeling based on known inhibitors (e.g., donepezil, galantamine) with SBVS docking into the enzyme's active site. The combined approach yielded potent hits with submicromolar IC_{50} values, demonstrating that hybrid methodologies outperform either method alone in enriching biologically active scaffolds [47].

In kinase inhibitor discovery, SBVS was instrumental in identifying novel scaffolds targeting EGFR and VEGFR-2. Docking-based virtual screening of over one million compounds using *Glide* and *AutoDock Vina*, followed by MM–GBSA rescoring, led to several compounds with nanomolar inhibitory potency validated experimentally [48]. Complementary LBVS using ECFP6 fingerprint similarity to known kinase inhibitors helped identify additional compounds with distinct scaffolds, highlighting the power of ligand-based extrapolation to uncover scaffold diversity. For GPCR targets, where structural information is often limited or conformationally variable, LBVS and pharmacophore-based screening have proven particularly valuable. For instance, the discovery of novel dopamine D3 receptor antagonists relied heavily on ligand-based modeling guided by known dopaminergic ligands, later validated through docking once crystal structures became available [49]. Another illustrative case is the discovery of SARS-CoV-2 main protease inhibitors during the COVID-19 pandemic. SBVS enabled rapid repurposing of FDA-approved drugs and screening of millions of compounds using homology models and later cryo-EM structures. Concurrently, LBVS using molecular fingerprints and pharmacophore models trained on known viral protease inhibitors identified novel scaffolds, several of which advanced to experimental validation with promising antiviral activity [50].

Comparative analyses reveal that

- SBVS excels when high-quality structural information is available, allowing accurate pose prediction and interaction analysis.
- LBVS performs better for data-rich targets with multiple known actives but limited structural data.
- Hybrid SBVS–LBVS workflows yield the highest success rates, combining structural precision with chemical diversity exploration [51].

These case studies reinforce that the choice between SBVS and LBVS should be dictated by data availability, target class, and project stage. As AI-driven fusion models continue to evolve, future workflows will likely blur the distinction between these two paradigms, uniting them into adaptive, integrated screening systems.

Table 11.1 Comparative Overview of Structure-Based and Ligand-Based Virtual Screening

Parameter	Structure-Based Virtual Screening (SBVS)	Ligand-Based Virtual Screening (LBVS)
Primary Input	3D structure of target protein (from crystallography, cryo-EM, or modeling)	Known active ligands and their molecular descriptors/fingerprints
Key Principle	Predicts ligand binding by docking and scoring within the protein binding site	Predicts activity by similarity or pharmacophore relationships with known actives
Data Requirement	Requires accurate structural model of the target	Requires sufficient number of active ligands with confirmed bioactivity
Core Algorithms	Docking engines (AutoDock, Glide, GOLD, DOCK, SwissDock) and scoring functions (empirical, ML-based)	Similarity search, pharmacophore mapping, ML/QSAR models
Output	Predicted binding poses and affinity scores	Ranked list of compounds predicted to have similar activity
Advantages	Mechanistic insight into binding; applicable to novel scaffolds	Fast, computationally light; effective when structure unavailable
Limitations	Sensitive to protein flexibility, scoring inaccuracies	Limited by quality and diversity of known actives
Applications	Enzyme inhibitors, receptor-ligand complexes, fragment-based design	GPCRs, ion channels, metabolic enzymes, data-rich targets
Integration	Often combined with MD, MM–GBSA, or consensus scoring	Often integrated with pharmacophore modeling or ML-based similarity prediction
Representative Tools	AutoDock Vina, Glide, GOLD, GNINA, DeepDock	LigandScout, Phase, ROCS, RDKit, ChemAxon, DeepChem
Computational Cost	Moderate to high (depends on docking size)	Low to moderate (depends on dataset size)
Hit Enrichment	High for accurate structures (EF up to 10–20)	Moderate but scalable; efficient for large datasets
Recent AI Trends	Deep docking, CNN-based scoring, DiffDock	GNN-based representation learning, AI-driven similarity mapping

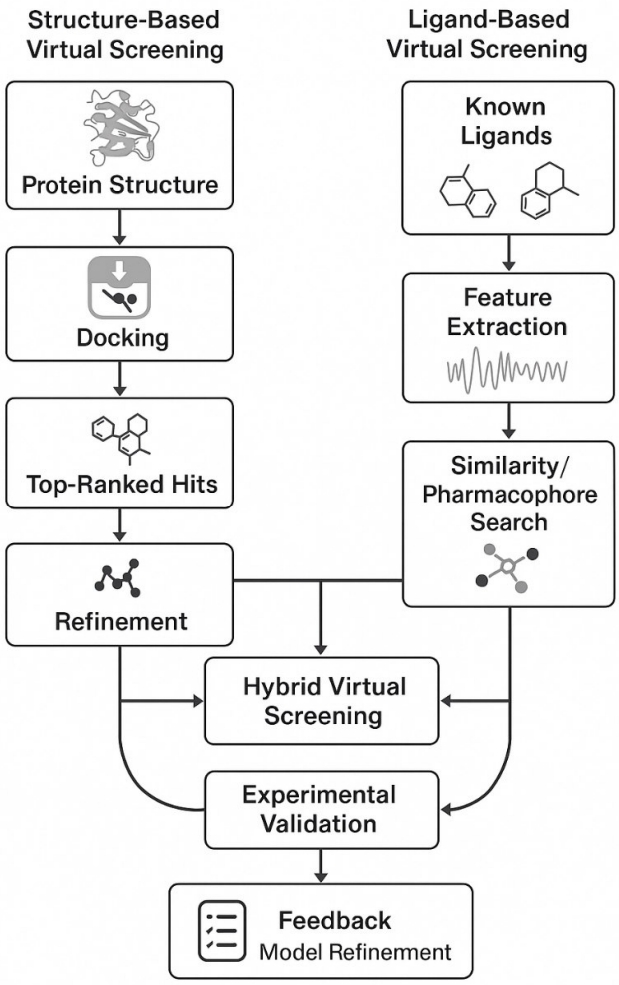

Figure 1: Workflow of Structure-Based and Ligand-Based Virtual Screening

1.7 Best Practices and Future Directions in Virtual Screening

The continuous evolution of virtual screening methodologies demands the establishment of standardized best practices to ensure reproducibility, reliability, and translatability of results. Effective virtual screening requires a harmonized framework that integrates computational accuracy, data quality, and experimental validation.

Best practices include

1. Comprehensive data preparation – Ensuring structural accuracy of proteins and ligands, proper protonation, tautomerization, and stereochemical validation.
2. Multi-level validation – Employing re-docking, cross-docking, enrichment benchmarking, and consensus scoring for unbiased assessment.
3. ADMET integration – Filtering screened compounds through physicochemical, pharmacokinetic, and toxicity models early in the workflow.
4. Consensus and hybrid strategies – Combining SBVS and LBVS, multiple scoring functions, or ensemble docking to mitigate algorithmic bias.

5. Transparent reporting – Adhering to reproducibility standards such as FAIR (Findable, Accessible, Interoperable, Reusable) data principles and documenting computational protocols [52].

Future trends in virtual screening are shaped by AI integration, quantum computing, and automation. Deep learning models now enable end-to-end molecular prediction pipelines, capable of learning implicit representations of protein–ligand interactions from raw structural data. Models such as *DeepDocking*, *DiffDock*, and *GraphScoreDTA* have demonstrated remarkable accuracy improvements in pose prediction and binding affinity estimation [53]. In parallel, cloud-based ultra-large screening platforms, such as *OpenEye Orion* and *Google DeepMind's AlphaFold–Docking integration*, are redefining scalability screening up to billions of compounds in days using distributed GPU resources [54]. These advances democratize access to large-scale screening, allowing academic and industry laboratories to participate in global collaborative drug discovery initiatives.

The convergence of virtual screening with automated synthesis and experimental robotics will usher in the era of autonomous drug design laboratories, where AI algorithms continuously refine models using real-time bioassay data. Integration with quantum mechanical scoring functions and physics-informed neural networks promises even higher accuracy in binding affinity prediction, bridging the current gap between docking scores and true thermodynamic free energies [55]. Ultimately, the next generation of virtual screening will not merely identify hits but intelligently design, synthesize, and test molecules through self-learning, closed-loop systems, transforming the pace, scale, and precision of modern drug discovery.

REFERENCES

1. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov*. 2004;3(11):935–949.
2. Ferreira LG, dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. *Molecules*. 2015;20(7):13384–13421.
3. Lavecchia A, Di Giovanni C. Virtual screening strategies in drug discovery: a critical review. *Curr Med Chem*. 2013;20(23):2839–2860.
4. Lionta E, Spyrou G, Vassilatis DK, Cournia Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr Top Med Chem*. 2014;14(16):1923–1938.
5. Walters WP, Barzilay R. Applications of deep learning in molecule generation and property prediction. *Acc Chem Res*. 2020;53(2):263–270.
6. Zhavoronkov A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*. 2019;37:1038–1040.
7. Ballester PJ, Mitchell JB. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*. 2010;26(9):1169–1175.
8. Meng XY, Zhang HX, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des*. 2011;7(2):146–157.
9. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583–589.
10. Schrödinger Release 2023-4: Protein Preparation Wizard, Schrödinger, LLC, New York, NY, 2023.
11. Irwin JJ, Shoichet BK. ZINC: a free database of commercially available compounds for virtual screening. *J Chem Inf Model*. 2005;45(1):177–182.
12. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function. *J Comput Chem*. 2010;31(2):455–461.

13. Charifson PS, Walters WP. Filtering databases and virtual screening for drug discovery. *J Comput Aided Mol Des*. 2002;16(5-6):311–323.
14. Pettersen EF, Goddard TD, Huang CC, et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci*. 2021;30(1):70–82.
15. McNutt AT, Francoeur PG, Aggarwal R, et al. GNINA 1.0: molecular docking with deep learning. *J Cheminform*. 2021;13(1):43.
16. Homeyer N, Gohlke H. Free energy calculations by the molecular mechanics Poisson–Boltzmann surface area method. *Mol Inform*. 2012;31(2):114–122.
17. Ripphausen P, Nisius B, Bajorath J. State-of-the-art in ligand-based virtual screening. *Drug Discov Today*. 2011;16(9-10):372–376.
18. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2023: extending experimental data coverage and data accessibility. *Nucleic Acids Res*. 2023;51(D1):D976–D987.
19. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010;50(5):742–754.
20. Hawkins PC, Skillman AG, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. *J Med Chem*. 2007;50(1):74–82.
21. Wolber G, Langer T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inf Model*. 2005;45(1):160–169.
22. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein–ligand scoring with convolutional neural networks. *J Chem Inf Model*. 2017
22. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discov Today*. 2018;23(6):1241–1250.
23. Reymond JL. The chemical space project. *Acc Chem Res*. 2015;48(3):722–730.
24. Warr WA. Scientific workflow systems for drug discovery. *J Comput Aided Mol Des*. 2012;26(5):479–496.
25. Shoichet BK. Screening in a spirit haunted world. *Drug Discov Today*. 2006;11(13–14):607–615.
26. Erlanson DA, Fesik SW, Hubbard RE, Jahnke W, Jhoti H. Twenty years on: the impact of fragments on drug discovery. *Nat Rev Drug Discov*. 2016;15(9):605–619.
27. Polishchuk PG, Madzhidov TI, Varnek A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J Comput Aided Mol Des*. 2013;27(8):675–679.
28. Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem*. 2010;53(7):2719–2740.
29. Imrie F, Bradley AR, van der Schaar M, Deane CM. Deep generative models for 3D linker design. *J Chem Inf Model*. 2020;60(4):1983–1995.
30. Truchon JF, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem. *J Chem Inf Model*. 2007;47(2):488–508.
31. Nicholls A. What do we know and when do we know it? *J Comput Aided Mol Des*. 2008;22(3–4):239–255.
32. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment factors, and decoy databases. *J Comput Aided Mol Des*. 2008;22(3–4):213–222.
33. Mysinger MM, Shoichet BK. Rapid context-dependent ligand desolvation in molecular docking. *J Chem Inf Model*. 2010;50(9):1561–1573.
34. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem*. 2012;55(14):6582–6594.

35. Gabel J, Desaphy J, Rognan D. Beware of machine learning-based scoring functions on the danger of developing black boxes. *J Chem Inf Model*. 2014;54(10):2807–2815.
36. Liu X, Jiang H, Li H. SHAFTS: a hybrid approach for 3D molecular similarity calculation. 1. Method and assessment of virtual screening performance. *J Chem Inf Model*. 2011;51(9):2372–2385.
37. Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. *Mol Inform*. 2016;35(1):3–14.
38. Macarron R, Banks MN, Bojanic D, et al. Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov*. 2011;10(3):188–195.
39. Sliwoski G, Kothiwale S, Meiler J, Lowe EW Jr. Computational methods in drug discovery. *Pharmacol Rev*. 2014;66(1):334–395.
40. Lionta E, Spyrou G, Kremidas A, et al. Virtual screening for kinase inhibitors: successes and challenges. *Curr Top Med Chem*. 2014;14(16):1923–1938.
41. Wang Z, Sun H, Yao X, et al. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys Chem Chem Phys*. 2016;18(18):12964–12975.
42. Schneider G. Automating drug discovery. *Nat Rev Drug Discov*. 2018;17(2):97–113.
43. Baell J, Walters MA. Chemistry: chemical con artists foil drug discovery. *Nature*. 2014;513(7519):481–483.
44. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*. 2019;37(9):1038–1040.
45. Ton A-T, Gentile F, Hsing M, Ban F, Cherkasov A. Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 1.3 billion compounds. *Mol Inform*. 2020;39(8):2000028.
46. Benmansour F, et al. Structure-based and ligand-based virtual screening for acetylcholinesterase inhibitors: discovery of novel scaffolds. *Eur J Med Chem*. 2021;225:113804.
47. Gupta M, Sharma R, Kumar A. Hybrid structure–ligand-based virtual screening for identification of novel VEGFR-2 inhibitors. *Sci Rep*. 2020;10(1):1–13.
48. Rodriguez D, Gutierrez-de-Teran H, Carlsson J. Advances in GPCR homology modeling driven by structural data. *Curr Opin Struct Biol*. 2020;63:120–127.
49. Douangamath A, et al. Crystallographic and virtual screening of the SARS-CoV-2 main protease reveals inhibitors and conserved binding sites. *Nat Commun*. 2020;11(1):5047.
50. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today*. 2015;20(3):318–331.
51. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
52. Corso G, Stärk H, Jing B, et al. DiffDock: Diffusion steps, twists, and turns for molecular docking. *Nat Methods*. 2023;20(10):1530–1538.
53. Irwin JJ, Tang KG, Young J, et al. ZINC20 A free ultralarge-scale chemical database for ligand discovery. *J Chem Inf Model*. 2020;60(12):6065–6073.
54. Noé F, Tkatchenko A, Müller KR, Clementi C. Machine learning for molecular simulation. *Annu Rev Phys Chem*. 2020;71:361–390.