Molecular Docking and Bioinformatics Analysis of Drug Molecule–Target Protein Interactions

Dongxing Liu

Qingdao Jiaozhou Yingzi Private School, Qingdao, China

15866751837@163.com

Keywords: drug molecule; target protein; molecular docking; bioinformatics; interaction

Abstract: This paper focuses on the interaction between drug molecules and target proteins, elaborating on the core contents of molecular docking and bioinformatics analysis. Molecular docking utilizes computer simulations to predict the binding conformation and strength between drugs and target proteins, relying on conformation searching and energy evaluation. Bioinformatics integrates multidimensional data to interpret the biological significance of these interactions. Together, they form a "computational prediction–mechanistic interpretation" chain that addresses the questions of "how binding occurs" and "why it is effective," thus promoting innovation in drug development models. This approach offers theoretical support for shortening development cycles and improving the efficiency of targeted drug design, bearing both practical and theoretical significance.

1. Introduction

1.1. Research Background

Traditional drug development faces challenges such as long cycles, high costs, and low efficiency. Many candidate molecules fail in clinical trials due to insufficient understanding of the "drug–target protein" interactions. Although the rise of targeted therapies focuses on the specific binding between drugs and disease-related target proteins to improve efficacy and reduce side effects, accurately predicting binding patterns and strength remains a major challenge.

Molecular docking technology rapidly predicts binding conformations and strengths between drugs and target proteins through computer simulations, thus reducing the cost of experimental screening. Bioinformatics, from perspectives such as gene sequences, protein structures, and disease networks, uncovers biological patterns of interactions. The integration of these two methods shifts research from "experiment-led" to "computational prediction–experimental validation," providing essential tools for analyzing binding mechanisms, shortening development time, and enhancing screening efficiency, thereby becoming a key support in modern drug development.

1.2. Research Significance

The study of molecular docking and bioinformatics analysis of drug molecule–target protein interactions have significant theoretical and practical value.

From a practical perspective, it helps overcome the bottlenecks of traditional drug development: molecular docking enables rapid screening of potentially effective drug molecules, reducing the blindness of trial-and-error experiments and significantly lowering R&D costs and cycle time. Bioinformatics helps extract biological rules from interactions, allowing precise identification of highly specific targets and providing a basis for designing efficient and low-toxicity targeted drugs, thereby directly improving the accuracy and safety of clinical treatment.

From a theoretical standpoint, this research deepens the understanding of the relationship between "molecule and function": it reveals the structural basis of drug–target protein binding (such as key binding sites and types of intermolecular forces), improves the theoretical system of molecular recognition mechanisms, and promotes interdisciplinary integration among computational biology, structural biology, and others. It also offers methodological reference for exploring more complex biomolecular interactions and has a profound impact on technological innovation and disciplinary

DOI: 10.25236/iwmecs.2025.028

development in drug research.

2. Fundamentals of Drug Molecule–Target Protein Interactions

2.1. Basic Concepts of Drug Molecules and Target Proteins

Drug molecules and target proteins are core elements in drug action, and understanding their basic concepts is crucial for interpreting the mechanisms of interaction [1].

Drug molecules generally refer to small-molecule compounds with therapeutic activity, usually with a molecular weight between 100–1000. Their structures contain specific functional groups (e.g., hydroxyl, carboxyl), and they regulate physiological functions by binding with biomolecules in the body. Their main role is to intervene in disease progression, such as inhibiting abnormal enzyme activity or blocking receptor signals, and they must possess specificity—binding only to target molecules to reduce side effects.

Target proteins are the action sites of drugs in the body, typically disease-related macromolecules (e.g., receptors, enzymes, ion channels) with specific three-dimensional structures and binding sites (e.g., enzyme active centers) [2]. Structural and functional abnormalities in these proteins are key causes of disease—for instance, overexpressed receptors on the surface of cancer cells may promote cell proliferation, becoming targets for anticancer drugs.

The specific binding between these two is the basis for drug efficacy, and understanding their concepts is a prerequisite for analyzing interaction mechanisms and conducting targeted drug development.

2.2. The Nature of Intermolecular Interactions

The specific binding between drug molecules and target proteins essentially depends on the synergistic effect of non-covalent intermolecular forces—although weaker than covalent bonds, these forces enable reversible binding, achieving the dynamic balance of "activation–dissociation" and forming the core mechanism of drug action.

Hydrogen bonds are critical specific forces: formed between hydrogen atoms (bonded to electronegative atoms) in the drug or protein and electronegative atoms like oxygen or nitrogen on the counterpart, functioning like "molecular glue."[3] They precisely match structural sites and determine binding specificity.

Hydrophobic interactions occur when hydrophobic groups (e.g., alkyls, aromatic rings) aggregate to reduce contact with water molecules, lowering system energy and enhancing binding stability. These are especially significant within hydrophobic pockets of target proteins.

Van der Waals forces are weak but ubiquitous attractive forces arising from transient dipoles between molecules. While individual interactions are weak, the cumulative effect of many atoms can substantially increase binding strength.

Electrostatic interactions arise from charge differences on molecular surfaces, such as the attraction between positively charged drug groups and negatively charged regions of the target protein, directly influencing the initial recognition of binding.

The coordinated match of these forces determines both the binding strength (affecting drug efficacy) and specificity (avoiding off-target effects), serving as a foundation for understanding drug action mechanisms, as shown in Table 1:

Table 1 Major Non-Covalent Intermolecular Forces between Drug Molecules and Target Proteins

Type of Interaction	Nature	Primary Role
Hydrogen Bond	Interaction formed between a hydrogen atom (bonded to an electronegative atom) in the drug or target protein and an electronegative atom such as oxygen or nitrogen on the counterpart	Precisely matches structural sites and determines binding specificity
Hydrophobic Interaction	Aggregation of hydrophobic groups (e.g., alkyl, aromatic rings) to reduce contact with water molecules and lower system energy	Enhances binding stability, particularly significant in hydrophobic pockets of target proteins
Van der Waals Force	Weak attractive force caused by transient dipoles between molecules	Weak individually, but collectively enhances binding strength when many atoms are involved
Electrostatic Interaction	Attraction (or repulsion) arising from differences in surface charge distribution on molecules	Affects initial recognition of binding (e.g., attraction between positively and negatively charged regions)

3. Molecular Docking Technology

3.1. Basic Principles

The basic principle of molecular docking is to simulate, through computer modeling, the binding process between a drug molecule (ligand) and a target protein (receptor) under physiological conditions, thereby predicting the most probable binding conformation and binding affinity between the two [4]. The core concept is derived from the theory of "molecular recognition": the binding between ligand and receptor must meet both structural complementarity (such as shape and charge distribution matching) and energy compatibility (the system reaches its lowest energy upon binding).

In the simulation process, the receptor's binding pocket (e.g., the active site of an enzyme or the ligand-binding domain of a receptor) is regarded as the "interaction region" for the ligand. The program uses algorithms to search all possible spatial conformations (poses) of the ligand within this region, including flexible changes such as molecular rotation and folding. Meanwhile, based on molecular mechanics and energy calculations, each conformation's stability is evaluated—with particular attention to the synergistic contributions of non-covalent bonds such as hydrogen bonds and hydrophobic interactions. The conformation with the lowest total binding energy is selected (the lower the energy, the more stable the binding).

In summary, molecular docking uses a two-step approach of "conformational search + energy evaluation" to simulate the spontaneous binding of ligands and receptors under natural conditions. It ultimately outputs the most probable binding mode and predicted affinity, providing theoretical support for assessing the potential activity of drug molecules and reducing the blind spots in experimental screening.

3.2. Main Methods and Classifications

The main methods and classifications of molecular docking are usually categorized based on how molecular flexibility is handled. The core difference lies in whether the receptor or ligand is allowed to undergo conformational changes during binding. Accordingly, molecular docking is divided into three types:

Rigid docking is the most basic method. It assumes both the receptor and ligand structures are rigid (i.e., no bond rotation or conformational changes occur) and adjusts the ligand's position only by translation and rotation within the receptor's binding pocket [5]. This method is fast and computationally efficient but neglects the natural flexibility of molecules. It is suitable for preliminary screening of large numbers of candidate molecules or scenarios where the receptor structure is stable (e.g., the active center of a rigid enzyme).

Semi-flexible docking allows the ligand to undergo limited flexible changes (e.g., partial side-chain rotations or small-molecule backbone torsions), while the receptor remains rigid. This method balances computational accuracy and efficiency—it accounts for the adaptive adjustment of the ligand while avoiding the computational burden caused by receptor flexibility. It is commonly used for medium-scale molecule screening.

Flexible docking allows both the receptor (e.g., side chains of amino acids near the binding pocket) and ligand to undergo flexible changes, more closely resembling the actual binding process under physiological conditions. It offers higher predictive accuracy but involves handling a large number of conformation combinations, leading to significantly higher computational costs. It is often used for detailed analysis of the binding modes of a small number of candidate molecules.

The three types of methods have their own focuses. They need to be selected based on the research objectives (such as rapid screening or precise prediction), and together they form the technical system of molecular docking, as presented in Table 2:

Table 2 Main Methods and Classifications of Molecular Docking Technology

Method Type	Treatment of Molecular Flexibility	Core Features	Applicable Scenarios
Rigid Docking	Both receptor and ligand are rigid (no conformational changes)	Fast computation, high efficiency, ignores natural flexibility	Preliminary screening of large numbers of candidate molecules; stable receptor structures
Semi-Flexible Docking	Ligand undergoes limited flexible changes (e.g., side-chain rotation); receptor remains rigid	Balances accuracy and efficiency, accounts for ligand adaptability	Medium-scale molecule screening
Flexible Docking	Both receptor (near binding pocket) and ligand undergo flexible changes	High predictive accuracy, significantly increased computational cost	Detailed binding mode analysis for a small number of candidate molecules

3.3. Key Steps

The key steps of molecular docking technology can be divided into four interlinked parts to ensure the accuracy of predictions.

Step one is Molecular Preparation. Both the receptor (target protein) and the ligand (drug molecule) must undergo preprocessing: The receptor's three-dimensional structure is obtained from databases (e.g., PDB), and redundant elements such as crystallographic water molecules and heteroatoms are removed. Missing hydrogen atoms are added, and charges are calculated to reflect physiological conditions. For the ligand, a three-dimensional structure is constructed, the initial conformation is optimized, and the protonation state is clarified (e.g., the dissociation state of acidic groups at physiological pH), ensuring a "clean" molecular model for subsequent docking.

Step two is Binding Pocket Definition. By analyzing the receptor's structure—such as known active sites or amino acid residue distributions—the potential binding region for the ligand is defined. This avoids meaningless global searches and improves computational efficiency [6].

Step three is Conformational Search. Algorithms such as genetic algorithms and Monte Carlo methods are used to explore all possible poses of the ligand within the binding pocket, including bond rotations and flexible changes. A large number of candidate conformations are generated during this process.

Step four is Result Evaluation: Scoring functions are applied to calculate the binding energy of each conformation, taking into account contributions from non-covalent interactions such as hydrogen bonds and hydrophobic forces. The conformation with the lowest energy and highest structural compatibility is selected as the most probable binding mode.

3.4. Evaluation Metrics

1) Binding Energy Calculation (Molecular Mechanics Force Field Equation):

$$\mathrm{EBinding\ Energy} = \mathrm{Ecomplex} - (\mathrm{EReceptor} + \mathrm{ELigand})$$

(Where E represents the molecular mechanical energy, which includes bond energy, angle energy, non-bonded interaction energy, etc.)

2) Scoring Function:

$$\mathrm{S} = \mathrm{aEvdW} + \mathrm{bEele} + \mathrm{cEsol} + \mathrm{d}$$

(Where a,b,c are weighting coefficients and d is a constant)

4. Bioinformatics Analysis

4.1. Basic Principle

The basic principle of bioinformatics analysis is to integrate and mine massive biological data in order to interpret the biological significance of interactions between drug molecules and target proteins, thereby building a bridge between "molecular binding patterns" and "biological functional effects." [7] Its core logic lies in the understanding that the interaction between a drug and its target protein is not an isolated event—it is closely related to the target protein's sequence features, structural functions, involvement in physiological pathways, and disease associations. These relationships can be revealed through data patterns.

This analysis relies on multidimensional biological data, including gene sequences, protein structures, interaction networks, and disease databases. Computational algorithms are used to integrate these datasets and conduct analysis on three levels: At the sequence level, homologous alignment is used to identify conserved binding sites on the target protein, revealing the evolutionary conservation of the interaction. At the structural level, combining data such as protein domains and surface charge distribution helps to explain why the binding conformation predicted by molecular docking is stable. At the network level, a "drug–target protein–disease" interaction network is constructed to analyze the signaling pathways the target protein is involved in, clarifying the effect of the interaction on disease progression.

In short, bioinformatics analysis follows a process of "data association → pattern extraction → functional interpretation," transforming the physicochemical binding information obtained from molecular docking into interpretable biological mechanisms. This provides theoretical support for evaluating the rationality of drug action and predicting potential efficacy and side effects.

4.2. Main Methods and Tools

The main methods and tools of bioinformatics analysis revolve around data mining and functional interpretation, with a focus on multidimensional analysis of the biological significance of drug–target protein interactions.

Database retrieval is a foundational method, relying on authoritative databases to obtain core data: Protein structure databases (e.g., PDB) provide 3D structures of target proteins. Drug databases (e.g., DrugBank) contain detailed information on drug molecules [8]. Interaction databases (e.g., STRING) compile known protein–ligand interaction data, offering primary materials for analysis.

Sequence analysis commonly uses homology alignment methods. Tools such as BLAST compare amino acid sequences of target proteins with homologous proteins to identify conserved binding sites (e.g., key residues in active centers) and assess the evolutionary stability of binding sites.

Structure analysis utilizes tools like PyMOL (for visualizing protein structures) and Swiss-Model (for homology modeling to complete missing structures). These tools help analyze domain distribution, surface charge, and hydrophobic regions of target proteins to explain the rationality of predicted binding conformations.

Network and functional analysis is carried out with tools such as Cytoscape to construct drug–target–disease association networks, visually displaying molecular relationships. Tools like DAVID or Metascape are used for functional enrichment analysis to identify KEGG pathways (e.g., cancer signaling pathways) involving the target proteins and to define the biological functions of the interactions.

Together, these methods and tools convert scattered data into interpretable biological patterns, supporting a deeper understanding of the mechanisms underlying molecular interactions, as summarized in Table 3:

Table 3 Summary of Main Methods, Tools, and Functions in Bioinformatics Analysis

Method Type	Main Tools	Core Function
Database Retrieval	PDB, DrugBank, STRING	Obtain target protein structures, drug information, and protein–ligand interaction data
Sequence Analysis	BLAST	Align homologous sequences, identify conserved binding sites, assess evolutionary stability
Structural Analysis	PyMOL, Swiss-Model	Analyze structural domains, surface charge, etc., to explain the rationality of binding conformations
Network & Functional Analysis	Cytoscape, DAVID/Metascape	Construct interaction networks, perform functional enrichment, clarify biological functions

4.3. Key Steps

The key steps of bioinformatics analysis can be divided into four progressive stages, forming a logical pathway from raw data to mechanistic interpretation.

The first step is Data Collection and Preprocessing. Multidimensional data must be retrieved from authoritative databases: for example, the PDB database provides 3D structures of target proteins; DrugBank offers physicochemical properties of drug molecules; STRING or PharmGKB contain known "drug–target–disease" association data. Simultaneously, data cleaning is performed to ensure reliability—this includes removing duplicates or low-quality entries (e.g., protein structures with poor resolution) and standardizing data formats (e.g., using consistent molecular identifiers). This step lays the foundation for all subsequent analyses.

The second step is Target Feature Analysis. This step focuses on characterizing the target protein from both sequence and structural perspectives: Sequence analysis with BLAST helps identify conserved amino acid residues in binding sites, enabling evaluation of the evolutionary stability of predicted docking sites.

Structural analysis with tools like PyMOL investigates domain architecture, surface charge distribution, and hydrophobic pockets of the protein to verify structural complementarity with the drug molecule (e.g., whether predicted hydrogen bonds correspond to conserved polar residues on the target) [9]. The third step is Functional Association Mining. This step links the drug–target interaction to biological function: Use DAVID or Metascape for GO annotation and KEGG pathway enrichment of the target protein to identify its involvement in biological processes (e.g., cell proliferation, signal transduction).

Integrate data from disease databases such as OMIM to associate the target protein with specific diseases (e.g., determine whether it is a cancer driver gene), thereby revealing the potential therapeutic significance of the interaction [10]. The fourth step is Network Integration and Interpretation. This step transforms fragmented data into a coherent, visual network: Cytoscape is used to construct a "drug–target–pathway–disease" network that clearly presents molecular regulatory relationships. Network topology analysis (e.g., node degree, betweenness centrality) helps identify core nodes such as key target proteins or critical pathways. Finally, these results are interpreted in conjunction with existing studies or experimental evidence to refine the biological mechanism of interaction, providing a theoretically grounded direction for drug development.

4.4. Result Evaluation

1) Sequence Similarity Assessment (BLAST Score):

$$\text{Similarity score S} = \frac{\text{Number of matched residues}}{\text{Total alignment length}} \times 100\%$$

2) Structural Consistency Assessment (RMSD Value):

$$\text{RMSD} = \sqrt{\frac{1}{\text{N}}\sum^{\text{N}}\text{i} = 1(\text{xi} - \text{x}'\text{i})^2 + (\text{yi} - \text{y}'\text{i})^2 + (\text{zi} - \text{z}'\text{i})^2}$$

(Where N is the number of atoms, $(\text{xi}, \text{yi}, \text{zi})$ are the atomic coordinates after docking.$(\text{x}'\text{i}, \text{y}'\text{i}, \text{z}'\text{i})$ are the reference coordinates)

5. Summary and Outlook

5.1. Main Conclusions

The integration of molecular docking and bioinformatics analysis provides a systematic approach to decipher drug–target interactions. Molecular docking simulates the binding process to accurately predict the optimal binding conformation and affinity between a drug and its target protein. Its core value lies in efficiently and cost-effectively screening potential active compounds, thereby reducing experimental trial-and-error. Bioinformatics, on the other hand, interprets the physical-chemical binding data through multiple dimensions—such as sequence conservation, structural features, and functional pathways—to reveal the biological significance and disease relevance of the interactions.

The synergy between the two establishes a complete workflow of "computational prediction – mechanistic interpretation": molecular docking addresses how binding occurs, while bioinformatics explains why the interaction is functionally effective. Together, they promote a shift in drug discovery from traditional "experiment-driven" to "computation-guided experimentation." This model not only shortens the screening cycle for candidate compounds but also deepens the mechanistic understanding of drug actions, offering a full-spectrum theoretical framework for rational drug design.

5.2. Limitations and Challenges

Despite their value, current technologies face several limitations. The major bottleneck in molecular docking lies in insufficient handling of molecular flexibility: most methods struggle to simulate large-scale conformational changes in both receptor and ligand (e.g., domain motions in proteins), leading to deviations in predicting binding modes in complex systems. While scoring functions estimate binding energies, they still inadequately capture hydrogen bond directionality and solvent effects, potentially misjudging the contribution of weak interactions. Bioinformatics analysis is constrained by data quality and integration complexity. Many databases contain low-resolution or redundant structures that compromise analytical reliability. Multi-omics datasets (e.g., gene expression, protein interactions) are highly heterogeneous, making deep integration difficult. Moreover, functional enrichment analyses rely heavily on existing knowledge bases, which may introduce bias when interpreting novel targets or pathways. In addition, a gap persists between computational predictions and experimental outcomes—some molecules predicted to have high affinity show limited activity in vitro or in vivo, underscoring the need for extensive validation and limiting the efficiency of clinical translation.

5.3. Future Directions

Future developments will emphasize "precision, intelligence, and multidimensional integration." In molecular docking, artificial intelligence (e.g., deep learning) will enhance conformational search algorithms, enabling full-flexibility simulations of receptor–ligand systems. Coupled with quantum mechanical calculations, scoring functions will improve their accuracy in capturing weak interactions such as π–π stacking, narrowing the gap between prediction and experimental results.

Bioinformatics will advance toward deeper integration of multi-omics data. By employing knowledge graph technologies to connect genes, proteins, and metabolites, dynamic "drug–target–disease" regulatory networks will be constructed. The incorporation of single-cell sequencing data will enable the elucidation of target protein functions in specific cell subtypes, enhancing the cell-specific interpretation of mechanisms.

Interdisciplinary integration will be another key direction: embedding molecular docking and bioinformatics into a closed-loop system of "computational prediction – organoid-based validation – clinical verification." Organoid models will allow rapid validation of predicted interactions, while high-resolution structural data from cryo-EM will refine computational models. Ultimately, these advances will move the field from "possibility prediction" toward "precise design," providing more efficient tools for personalized targeted drug development.

References

[1] Yan C, Liu D, Li L, Wempe MF, Guin S, Khanna M, et al. Discovery and characterization of small molecules that target the GTPase Ral[J]. Nature, 2014, 515(7527):443–447. DOI:10.1038/nature13713.

[2] Gomes I , Fujita W , Chandrakala M V ,et al. Disease-specific heteromerization of G-protein-coupled receptors that target drugs of abuse.[J].Progress in Molecular Biology & Translational Science, 2013, 117:207-265.DOI:10.1016/B978-0-12-386931-9.00009-X.

[3] Derewenda Z S , Lee L , Derewenda U .The occurrence of C-H. O hydrogen bonds in proteins.[J].Journal of Molecular Biology, 1995, 252(2):248-262.DOI:10.1006/jmbi.1995.0492.

[4] Zubair M S , Anam S , Al-Footy K O ,et al. Cembranoid Diterpenes as Antitumor: Molecular Docking Study to Several Protein Receptor Targets[C]//International Conference on Computation for Science & Technology.2015.DOI:10.2991/iccst-15.2015.23.

[5] Pan H, Agarwalla S, Moustakas DT, Finer-Moore J, Stroud RM. Structure of tRNA pseudouridine synthase TruB and its RNA complex: RNA recognition through a combination of rigid docking and induced fit[J]. Proceedings of the National Academy of Sciences of the United States of America, 2003, 100(22):12648–12653. DOI:10.1073/pnas.2135585100.

[6] Gloriam D E , Foord S M , Blaney F E ,et al.Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design.[J].Journal of Medicinal Chemistry, 2009, 52(14):4429-4442.DOI:10.1021/jm900319e.

[7] Gomes L C , Simoes M .13C Metabolic Flux Analysis: From the Principle to Recent Applications[J]. Current Bioinformatics, 2012, 7(1).DOI:10.2174/157489312799304404.

[8] Wishart D S , Craig K , Chi G A ,et al.DrugBank: a comprehensive resource for in silico drug discovery and exploration[J].Nucleic Acids Research, 2006, 34(Database issue):D668-D672.DOI:10.1093/nar/gkj067.

[9] Dominique H , Van B I A E M ,Morán Luengo Tania,et al.A script to highlight hydrophobicity and charge on protein surfaces[J].Frontiers in Molecular Biosciences, 2015, 2:56. DOI:10.3389/fmolb.2015.00056.

[10] Mckusick V A .Mendelian Inheritance in Man and its online version, OMIM.[J].American Journal of Human Genetics, 2007, 80(4):588-604.DOI:10.1086/514346.