biomolecules

MDPI

Review

Fingerprinting Interactions between Proteins and Ligands for Facilitating Machine Learning in Drug Discovery

Zoe Li [1], Ruili Huang [2], Menghang Xia [2], Tucker A. Patterson [1] and Huixiao Hong [1,*]

[1] National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR 72079, USA; zoe.li@fda.hhs.gov (Z.L.); tucker.patterson@fda.hhs.gov (T.A.P.)
[2] National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, MD 20892, USA; ruili.huang@nih.gov (R.H.); mxia@mail.nih.gov (M.X.)
* Correspondence: huixiao.hong@fda.hhs.gov; Tel.: +1-870-543-7296

Abstract: Molecular recognition is fundamental in biology, underpinning intricate processes through specific protein–ligand interactions. This understanding is pivotal in drug discovery, yet traditional experimental methods face limitations in exploring the vast chemical space. Computational approaches, notably quantitative structure–activity/property relationship analysis, have gained prominence. Molecular fingerprints encode molecular structures and serve as property profiles, which are essential in drug discovery. While two-dimensional (2D) fingerprints are commonly used, three-dimensional (3D) structural interaction fingerprints offer enhanced structural features specific to target proteins. Machine learning models trained on interaction fingerprints enable precise binding prediction. Recent focus has shifted to structure-based predictive modeling, with machine-learning scoring functions excelling due to feature engineering guided by key interactions. Notably, 3D interaction fingerprints are gaining ground due to their robustness. Various structural interaction fingerprints have been developed and used in drug discovery, each with unique capabilities. This review recapitulates the developed structural interaction fingerprints and provides two case studies to illustrate the power of interaction fingerprint-driven machine learning. The first elucidates structure–activity relationships in β2 adrenoceptor ligands, demonstrating the ability to differentiate agonists and antagonists. The second employs a retrosynthesis-based pre-trained molecular representation to predict protein–ligand dissociation rates, offering insights into binding kinetics. Despite remarkable progress, challenges persist in interpreting complex machine learning models built on 3D fingerprints, emphasizing the need for strategies to make predictions interpretable. Binding site plasticity and induced fit effects pose additional complexities. Interaction fingerprints are promising but require continued research to harness their full potential.

Keywords: molecular fingerprints; 3D structural interaction fingerprints; machine learning; drug discovery; structure–activity relationships; protein–ligand interactions; predictive modeling

Citation: Li, Z.; Huang, R.; Xia, M.; Patterson, T.A.; Hong, H. Fingerprinting Interactions between Proteins and Ligands for Facilitating Machine Learning in Drug Discovery. *Biomolecules* **2024**, *14*, 72. https://doi.org/10.3390/biom14010072

Academic Editors: Monika Wasilewska, Aneta Michna and Maria Morga

Received: 2 November 2023
Revised: 26 December 2023
Accepted: 28 December 2023
Published: 5 January 2024

1. Introduction

Molecular recognition is a fundamental process in living organisms, involving specific and high-affinity interactions between biological macromolecules and various small molecules, leading to the formation of specific complexes [1,2]. Among these macromolecules, proteins play a vital role as they carry out their functions by binding to themselves or other molecules [2]. Consequently, a comprehensive understanding of protein–ligand interactions holds the key to unraveling the intricacies of molecular biology. Additionally, this knowledge about the mechanisms governing protein–ligand recognition and binding serves as a valuable resource in drug discovery, design, and development. By delving into the specifics of these interactions, researchers can better advance their quest for new therapeutic agents and foster scientific advancements in the field of drug development.

Traditional experimental methods have long been employed to predict the binding activity of small molecules [3]. These methods include isothermal titration calorimetry, fluorescence thermal shift assay, cellular thermal shift assay, and analytical ultracentrifugation, among others [3]. However, the vastness of the chemical space allows for an astounding number of approximately 10^{60} possible small molecules to be synthesized [4]. Despite this immense potential, only a small fraction of the potential protein–ligand interactions has yet to be explored [4]. Efficiently navigating through this vast search space poses challenges for traditional experimental methods due to their inherent drawbacks: high cost, time consumption, and labor intensiveness. Consequently, the increasing demand for more efficient approaches to predict the biological activities of small molecules has driven the development of computational methods. These computational approaches serve as invaluable tools to streamline the search process, narrowing down the possibilities and enabling researchers to focus on promising targets.

One of the most widely used computational approaches in drug discovery is quantitative structure–activity/property relationship (QSAR/QSPR) analysis [5]. This approach operates on the assumption that similar molecules exhibit similar bioactivities or physicochemical properties [5,6]. Leveraging this assumption, QSAR/QSPR analysis predicts the activities or properties of new molecules by establishing correlations between their chemical or structural features and their observed activities or properties [5,6]. This approach significantly reduces the need for time-consuming and costly experimental assays. Central to QSAR/QSPR analysis is the concept of molecular similarity, which is usually measured based on various molecular descriptors and fingerprints [7,8]. Molecular descriptors are numerical descriptions of the structural features of a chemical and are widely used in the development of predictive models of predicting biological activity and chemical properties [9–14]. Fingerprints encode the structural features of a molecule. These fingerprints serve as property profiles, typically presented in the form of vectors, where each vector element represents the existence, degree, or frequency of a specific structural feature [15–17]. Molecular fingerprints play a fundamental role in various drug discovery processes, including virtual screening, similarity-based compound searches, target molecule ranking, drug ADMET (absorption, distribution, metabolism, excretion, and toxicity) prediction, and more. Over the past few decades, different types of two-dimensional (2D) fingerprints have been developed for molecular feature encoding [18–20]. These fingerprints can be extracted from molecular connection tables without requiring three-dimensional (3D) structural information. The main categories of 2D fingerprints are as follows: substructure key-based fingerprints, topological or path-based fingerprints, circular fingerprints, and pharmacophore fingerprints [21–23]. Two-dimensional fingerprints are advantageous due to their ease, speed, and convenience of generation, as they solely rely on 2D structures [5]. Consequently, they are extensively utilized as input for machine learning algorithms in various drug discovery applications, such as binding affinity prediction, toxicity assessment, solubility analysis, and partition coefficient estimation [24]. A typical workflow for using machine learning to predict the properties of molecules is shown in Figure 1.

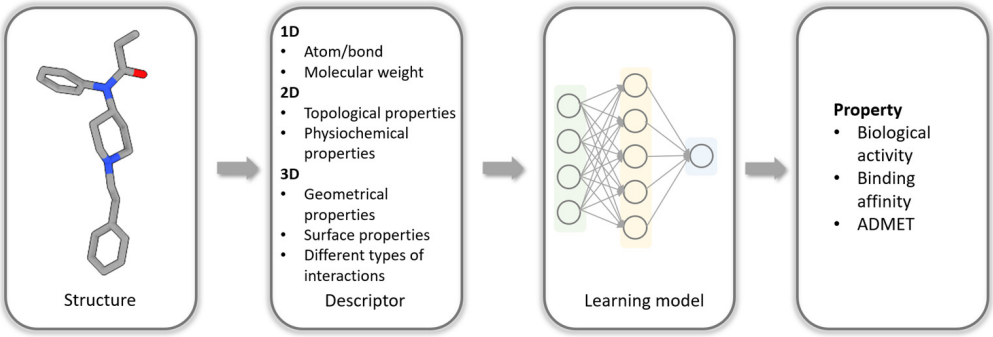

Figure 1. Typical workflow for using machine learning to predict properties of molecules.

In recent years, there has been a notable shift in the extensive use of machine learning from QSAR studies to focus on structure-based predictive modeling [25–28]. The availability of abundant structural and binding affinity data for protein–ligand complexes has enabled the training of binding affinity prediction models, leading to a surge in the development of machine-learning scoring functions [29]. These scoring functions exhibit exceptional performance in scoring works and have proven to outperform classical scoring functions, primarily due to their ability to handle large volumes of structural data effectively [29,30]. A critical aspect of constructing a machine-learning scoring function is feature engineering, which involves transforming complex structures into a series of descriptors. This process is guided by biologically-relevant interactions, such as hydrogen bonds, hydrophobic contacts, ionic interactions (salt bridges), π-stacking, and π-cation interactions [31].

Figure 2 illustrates a conventional fingerprint that is generated based only on the 2D structure of a small molecule and an emerging 3D interaction fingerprint that describes the interactions between a small molecule and its interacting macromolecule in a 3D structure. Recently, the focus of scoring function descriptors has shifted towards 3D interaction fingerprints (IFPs) because of their simplicity in representation and elaborate profiles of key interactions. IFPs are defined based on the interacting atoms between the protein and ligand within a protein–ligand complex structure. They are stored as one-dimensional (1D) vectors or matrices of Booleans, integers, or floating-point numbers, providing a concise and informative representation of the interaction patterns between the two entities [30,32]. The use of IFPs in machine-learning scoring functions holds significant promise in accurately characterizing and predicting protein–ligand interactions, thereby advancing the field of structure-based predictive modeling.

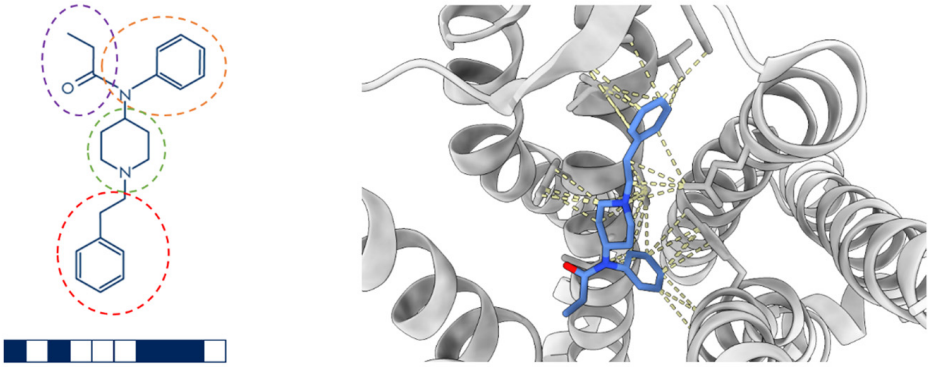

Figure 2. Illustration of a 2D molecular fingerprint (**left**) and a 3D molecular fingerprint (**right**). The dash circles in different colors indicate different structural features that are recorded in a bit string (under the 2D structure) as the fingerprint of the molecule. In the right sub-figure, the small molecule is represented by a stick model and the protein is drawn in a grey ribbon model. The interactions between the small molecule and the protein are indicated with yellow dashed lines and are recorded as the fingerprint of the small molecule in the protein.

2. Types of Structural Interaction Fingerprints

The development and application of various structural IFPs have been significant in advancing the field of protein–ligand interaction analysis. One of the pioneering structural IFP algorithms was introduced by Deng et al. in 2004, focusing on clustering kinase–inhibitor complexes [33]. Their fingerprint encompassed seven bits per interacting amino acid, representing predefined interaction types, including backbone, sidechain, polar, hydrophobic, and H-bond donor/acceptor interactions [33]. Mordalski et al. later extended this approach by adding two bits to encode aromatic and charged interactions, leading to improved technical implementation [34]. Notably, structural IFP was instrumental

in identifying the critical amino acids involved in interactions with antagonists within serotonin 5-HT7 receptor homology models [35].

Another widely used variant, developed by Marcou and Rognan in 2006, employs a seven-bit fingerprint encoding hydrophobic, aromatic face-to-face and edge-to-face, H-bond donor/acceptor, and cationic/anionic interactions [36]. Importantly, the geometric definitions in this variant can be customized, allowing for the inclusion of less common interaction types like weak H-bonds, cation-pi, and metal complexation [36]. This flexibility has enhanced the versatility of the fingerprinting approach. Later, the Rognan group devised a method to encode protein–ligand interactions into a 1D binary IFP string represented by an array of 11-bit substrings [37,38]. This novel approach effectively describes how each amino acid within the binding pocket interacts with the ligand. Specifically, every amino acid is encoded into one 11-bit substring, corresponding to 11 distinct types of interactions: hydrophobic interaction, aromatic interaction (face-to-face), aromatic interaction (edge-to-face), hydrogen bond interaction (protein atom as acceptor), hydrogen bond interaction (protein atom as donor), ionic interaction (protein atom with positive charge), ionic interaction (protein atom with negative charge), weak hydrogen bond interaction (protein atom as acceptor), weak hydrogen bond interaction (protein atom as donor), π-cation interaction, and metal ionic interaction with the ligand [37,38]. This encoding system provides a comprehensive representation of the intricate interactions between amino acids and the ligand, enabling a detailed analysis of their binding patterns.

The Rognan group also introduced triplet IFPs, where interaction points forming triangles are encoded into a fixed-length fingerprint of 210 bits [30]. The protein–ligand interaction is characterized by two interacting atoms and an interaction pseudoatom for ionic interaction, hydrogen bonding, and metal complexation. The interaction pseudoatom can be in three positions: the geometric center of the interacting atoms, near the interacting protein atom, and near the interacting ligand atom [30]. Interaction pseudoatoms can be computed using any of these three positions, allowing for mapping the interaction either on ligand atoms, protein atoms, or naturally at the mid-distance between the interacting atoms [30]. For hydrophobic interactions, when a ligand atom interacts with more than one protein atom, the interaction with the shortest distance is used to define the interaction pseudoatom. For aromatic interactions, an aromatic interaction pseudoatom is placed in the middle between the aromatic ring centroids. Although primarily designed for binding site comparison, triplet IFPs showed comparable performance to IFP in the post-processing of docking results [30].

Python-based protein–ligand interaction fingerprint (PyPLIF), an open-source Python tool developed by Radifar et al., aims at improving the accuracy of molecular docking results in virtual screening [39]. PyPLIF converts 3D interaction data from molecular docking into 1D bitstring representations, where each bit encodes the presence or absence of specific interaction types with binding site residues [39]. The similarity between these fingerprints and a reference ligand fingerprint is then evaluated using metrics like the Tanimoto coefficient [39]. Selecting top docking poses based on interaction fingerprint similarity, rather than relying solely on docking scores, significantly improves the identification of true binders [39].

Atomic pairwise interaction fingerprint (APIF) offers a binding site size-independent encoding of protein–ligand interactions. It achieves this by considering the relative position and interaction type of all pairs of interacting atoms between the ligand and protein [40]. Each interacting atom pair is categorized by its interaction type, such as the hydrophobic-acceptor, and sorted into discrete distance ranges between the ligand and protein atoms [40]. Consequently, a 294-bit fixed-length binary fingerprint is generated, encompassing various combinations of interaction pairs and distances. APIF's utilization of relative geometry rather than absolute positions allows for a comparison of binding modes across diverse binding sites [40]. This 1D fingerprint retains essential 3D information, making it valuable for virtual screening and docking pose selection. However, one limitation is the reduced precision in capturing geometric details, which may make interpreting interactions from

APIF challenging [40]. Despite this, APIF stands out for providing a concise representation of conserved interaction patterns, independent of the binding site size, although it may lack the intuitive interpretability found in residue-specific interaction fingerprints.

The simple ligand–receptor interaction descriptor (SILIRID) is an innovative fixed-length vector representation that derives from protein–ligand interaction fingerprints, serving to characterize binding sites. It condenses the interactions between ligand atoms and binding site residues into a concise 168-dimensional vector [41]. This is achieved by summing the binary fingerprint bits for identical amino acids and capturing their corresponding interaction types (such as hydrophobic, hydrogen bond donor/acceptor, etc.) [41]. SILIRID's distinct feature lies in its ability to merge residue-specific fingerprints into a binding site-independent summary, facilitating the comparison of interactions across binding sites of varying sizes [41]. As a result, SILIRID offers a compact representation of conserved interaction patterns that find applications in tasks like binding site comparison, virtual screening, and the visualization of chemogenomic space. One limitation to consider is the reduction in per-residue details, which may limit the granularity of interpretation [41]. Overall, SILIRID excels in encoding essential interaction features within a size-independent vector, although it may not possess the same level of interpretability found in residue-specific fingerprints.

Another unique approach to structural protein–ligand interaction fingerprints (SPLIF) was proposed by Da and Kireev [42]. It was designed to describe and compare protein–ligand interactions in a manner that is independent of the binding site. Unlike other approaches, SPLIFs explicitly encode the 3D structures of interacting ligand and protein fragments, capturing the nuances of the interaction modes and implicitly considers various contacts, such as π-π stacking [42]. The generation of SPLIF involves expanding contacting ligand and protein atoms to include neighboring atoms within a defined radius [42]. These circular fragments are assigned identifiers, and their 3D coordinates are retrieved [42]. The SPLIF then encodes the matching circular fragments between a docking pose and the reference complex, assessing similarity through a normalized score based on the fraction of matched fragments [42]. The evaluation involves both 2D fragment identity and 3D structural alignment, providing a comprehensive representation of the interaction patterns. A notable advantage of SPLIFs is their implicit inclusion of diverse interaction types in the 3D structure description [42]. However, the trade-off is the loss of precise geometric details. Overall, SPLIFs offer a robust platform for the quantitative comparison of conserved interaction patterns across binding sites of varying sizes.

Recently, Wojcikowski et al. introduced the protein–ligand extended connectivity fingerprint (PLECFP) [43], based on the atomic environment concept of the extended connectivity fingerprint initially proposed by Rogers and Hahn in 2010 [18]. PLECFP captures the local atomic environments between the interacting protein and ligand molecules. Its construction involves identifying contacting atom pairs and characterizing the neighborhood surrounding each atom within a specified bond depth. These ligand and receptor environments are paired, and their hashed bit positions create the final folded fingerprint. PLECFP's parameterization and evaluation on binding affinity prediction tasks using linear regression, random forest, and neural network models showcased its impressive descriptive capabilities. Surprisingly, the simple linear model performed similar with more complex methods, underscoring the richness of PLECFP's representation. Notably, PLECFP outperformed other interaction fingerprints like SILIRID and SPLIF, yielding Pearson correlation coefficients exceeding 0.8 on benchmark datasets [43]. Such exceptional performance suggests PLECFP's potential for diverse drug discovery tasks, including lead optimization and scaffold hopping, thanks to its implicit capacity to capture relevant interactions. A summary of different types of protein–ligand interaction fingerprints is listed in Table 1. A list of currently available software for calculating interaction fingerprints is shown in Table 2.

Table 1. Different types of protein–ligand interaction fingerprints and their characteristics.

Types of Protein–Ligand Interaction Fingerprints	Characteristics and Pattern Types	Length	Reference
Structural IFP	Uses well-defined interaction types such as hydrogen bond, halogen bonds, and π-π stacking	Each residue is represented by a seven-bit long bit string	[33,34]
Python-based protein–ligand interaction fingerprint (PyPLIF)	Uses well-defined interaction types such as hydrogen bond, halogen bonds, and π-π stacking	Seven bits represent seven different interactions for each residue	[39]
Triplet IFP	Uses two interacting atoms and an interaction pseudoatom positioned at three potential locations: the geometric center of the interacting atoms, the interacting protein atom, and the interacting ligand atom to encode different interaction types (7 types) at defined distance ranges (6 ranges)	210 integers	[30]
Atom-pairs-based interaction fingerprint (APIF)	Considers the relative positions of the atom pairs instead of the absolute locations of the individual interactions	294 bits	[40]
Simple ligand–receptor interaction descriptor (SILIRID)	Groups interactions by residue type, the interactions included are hydrophobic, aromatic face to face, aromatic edge to face, H-bond donated by the protein, H-bond donated by the ligand, ionic bond with protein cation and protein anion, and interaction with metal ion	168 integers (corresponds to the product of 20 amino acids and 1 co-factor and 8 interaction types per amino acid)	[41]
Structural protein–ligand interaction fingerprint (SPLIF)	Encodes interacting ligand and protein fragments by representing them as circular fingerprints using Extended Connectivity Fingerprints (ECFP2) and generates integer identifiers to represent each substructure fragment	Length depends on the number of interacting fragments identified	[42]
Protein–ligand extended connectivity fingerprint (PLECFP)	Pairs and hashes the ECFP environment from the interacting ligand and protein atoms to represent contacts and interactions between the molecules	The raw folded fingerprint consists of integers between 0 and 2^{32} (32 bits)	[43]

Table 2. Available software for calculating structural interaction fingerprints.

Software/Web Server	Types of Input Complex	Input Format	MD Trajectory Analysis	Reference
Arpeggio	All combinations between ligand, protein, DNA and RNA molecules	PDB	N/A	[44]

Table 2. *Cont.*

Software/Web Server	Types of Input Complex	Input Format	MD Trajectory Analysis	Reference
fingeRNAt	All combinations between ligand, protein, DNA and RNA molecules	PDB and SDF	N/A	[45]
getContacts	All combinations between ligand, protein, DNA and RNA molecules	VMD	N/A	getcontacts.github.io (accessed on 2 November 2023)
Ichem	Protein ligand complex only	Mol2	N/A	[37]
LUNA	Protein ligand and protein–protein complex	PDB, Mol, Mol2, and RDKit	N/A	[46]
MD-IFP	Ligand protein complex only	MDAnalysis	Yes	[47]
ODDT	Ligand protein complex only	OpenBabel and RDKit	N/A	[48]
PLIP	All combinations between ligand, protein, DNA and RNA molecules	PDB	N/A	[49]
ProLIF	All combinations between ligand, protein, DNA and RNA molecules	MDAnalysis and RDKit	Yes	[50]
PyPLIF HIPPOS	Ligand protein complex only	PDBQT and Mol2	N/A	[39]
Schrodinger	Ligand protein complex only	SDF, PDB, and MAE	N/A	[51,52]

3. Case Study of Structural Interaction Fingerprint Application

In this section, we highlight two case studies that incorporated structural interaction fingerprints into machine learning. The first case study demonstrated that molecular docking and machine learning can be combined to reveal key structure–activity relationships for drug targets [53]. The researchers compiled a dataset of approximately 2700 known ligands for the β2 adrenoceptor (β2AR). They computationally docked these ligands to β2AR structures to generate approximately 75,000 poses and calculated atomic interaction fingerprints describing receptor–ligand interactions. Machine learning models were trained on these fingerprints to predict whether ligands act as agonists or antagonists. Figure 3 shows the detailed workflow of this work. The models identified specific hydrophobic and polar contacts with receptor residues that differentiate agonists and antagonists. Agonists were found to preferentially interact with residues K97, F194, S203, S204, S207, H296, and K305 while antagonists were found to favor residues W286 and Y316. This structure–activity relationship modeling approach achieved high accuracy in predicting ligand pharmacological activity and provided molecular insights into β2AR activation and inhibition. This study demonstrates the power of interaction fingerprint-driven machine learning for elucidating ligand binding mechanisms and guiding rational drug design. The results from this case study revealed that structural interaction fingerprints derived from docking poses offer insights into the environment surrounding the ligand, which can be useful for differentiating the potential biological activities of ligands.

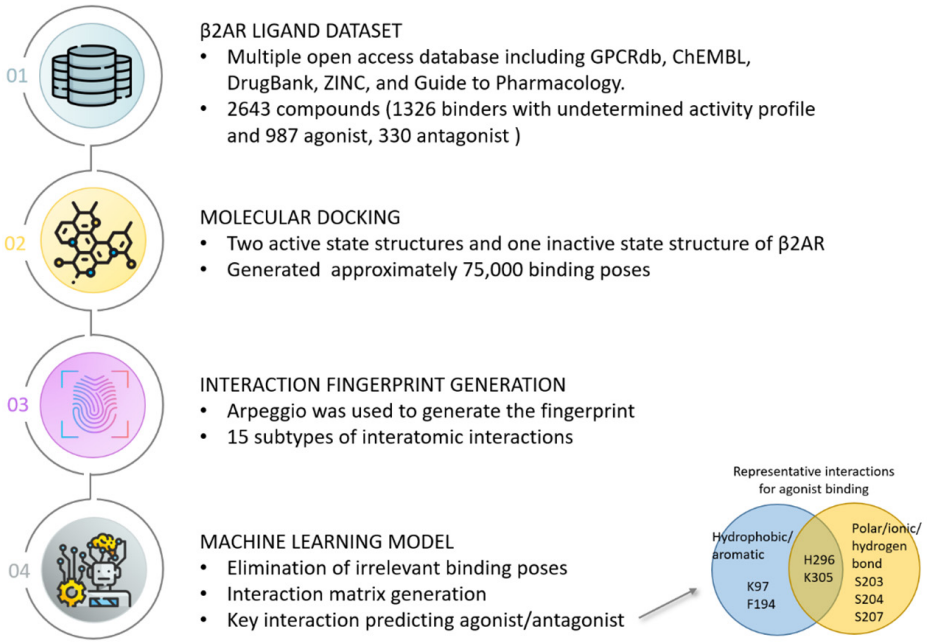

Figure 3. Case study of Jimenez-Roses et al. [53]. Workflow of utilizing interaction fingerprints extracted from docking poses as input for machine learning model to identify key residues for ligand pharmacological activity on β2 receptors.

The second case study introduced a machine learning strategy employing an innovative molecular representation termed RPM (retrosynthesis-based pre-trained molecular) representation to predict protein–ligand dissociation rates (k_{off}) [54]. The RPM representation was constructed through training on retrosynthesis reaction data, enabling the encapsulation of molecular reactivity and functional group information. Subsequently, these RPM features were fed into a partial least squares regression model to predict the k_{off} values for 501 inhibitors spanning 55 proteins. Impressively, the RPM-based model demonstrated superior performance compared to other pre-trained representations such as the molecular pre-training graph-based deep learning framework and geometry-enhanced molecular representation, achieving a noteworthy Pearson correlation coefficient of 0.76 on this specific dataset. To exemplify its application, the model was further evaluated using 38 novel inhibitors targeting the N-terminal domain of the heat shock protein 90α (HSP90), yielding a commendable correlation of 0.73 with experimental k_{off} values. In-depth mechanistic insights into the kinetics were sought through accelerated molecular dynamics simulations, which obtained data on relative retention times and protein–ligand IFPs along the dissociation trajectory. Figure 4 illustrates the detailed workflow of this case study. The simulated k_{off} values exhibited reasonable agreement with experimental results, with the IFPs elucidating important residues like N51, S52, and L107 that significantly influence the dissociation process. In an additional validation, the machine learning model coupling with molecular dynamics simulation was extended to two new HSP90 inhibitors absent from the training set. Encouragingly, the model accurately predicted their relative k_{off} values, which were aligned with experimental observations. Furthermore, the IFP analysis offered detailed insights into how substituents modulated binding kinetics. This case study combined different approaches and offered a comprehensive exploration of the molecular attributes and interactions that govern binding kinetics, thereby underlining its potential utility for kinetics-focused drug design endeavors.

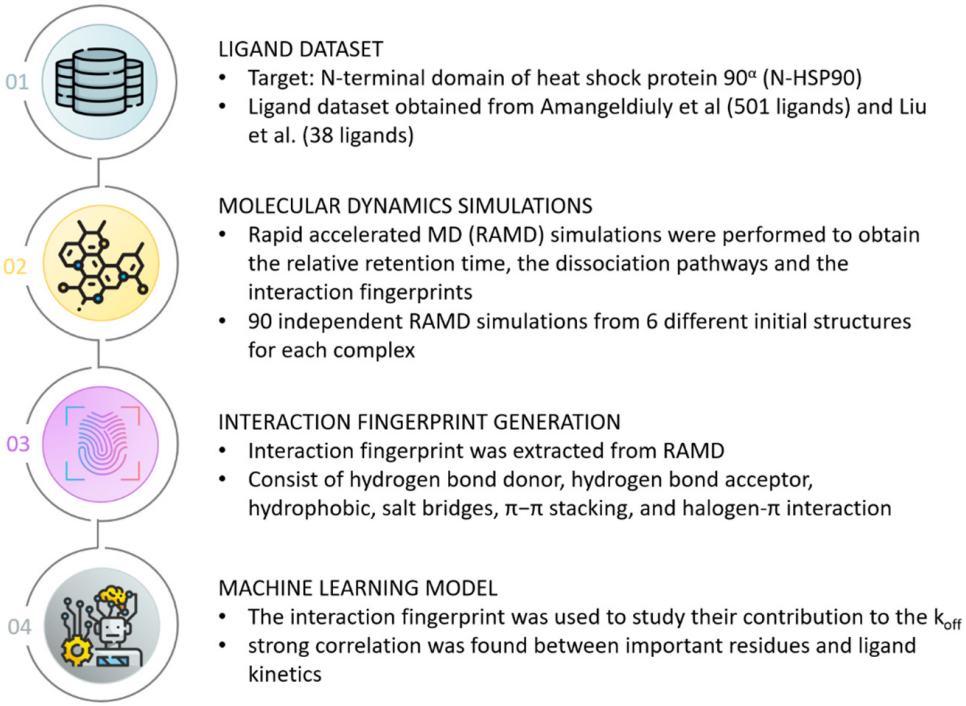

Figure 4. Case study of Zhou et al. [54]. Workflow of utilizing interaction fingerprints extracted from MD simulations as input for machine learning model to identify correlation between key residues and ligand kinetics. The ligand dataset in this case study was obtained from Amangeldiuly et al. [55] and Liu et al. [56].

4. Future Perspective

Molecular fingerprints have become indispensable cornerstones in the realm of computational drug discovery, offering informative representations of ligands for property prediction and activity modeling. In this landscape, the realm of molecular fingerprints stands at an exciting crossroads, with 2D fingerprints providing simplicity and ease of use, while 3D structural interaction fingerprints hold the tantalizing potential to intricately encapsulate the minutiae of interactions within protein–ligand complexes. The future trajectory of this field is poised for further advancement, driven by the synergy of hybrid fingerprint design and technological progress. The amalgamation of 3D structural interaction descriptors with other properties, such as physicochemical attributes, has the potential to elevate the accuracy of ligand bioactivity predictions. By encompassing both structural intricacies and physicochemical subtleties, hybrid fingerprints extend the horizons of molecular characterization, and the application of advanced machine learning techniques holds the key to their optimal integration. As computational methodologies advance and resources expand, the landscape for harnessing the potential of 3D fingerprints in drug discovery grows even more fertile. The interplay of refined machine learning algorithms, augmented structural datasets, and enhanced computational power opens new possibilities and opportunities in interaction fingerprint design, training, and prediction, with deep learning strategies poised to unveil profound insights from intricate 3D interaction patterns.

Yet, as the future of molecular fingerprints shines brightly, it is not without its challenges. One such limitation lies in the dependency of 3D fingerprints on the accessibility of protein–ligand complex structures. Nonetheless, the ongoing advancements in structural determination techniques contribute to an increasing abundance of structures, facilitating the progress of molecular fingerprint development. Another drawback is the insufficient incorporation of the energy terms necessary to comprehensively characterize the interactions occurring between proteins and ligands. Recent deep learning-based scoring functions may potentially solve this problem. Decoding complex machine learning models constructed on 3D fingerprints is another challenge. The process of unraveling the pivotal

interacting features driving a model's predictions remains an active area of exploration. Novel strategies are essential to deconstruct model outputs into interpretable interaction insights, which in turn can illuminate pathways for molecular optimization. Moreover, the intricacy of binding site plasticity and induced fit effects introduces complexities in accurately characterizing interactions solely from static structural data. Another limitation is the reliance on the availability of known ligand–protein interaction information. In both case studies, the target has a large number of known ligands that can be used for model training. However, for targets that have few or no known ligands, for which the discovery of new ligands is in higher demand, this method would not be as applicable.

Overall, interaction fingerprints hold immense promise but require continued research to fully harness their potential and overcome existing limitations, unlocking new vistas of discovery and application.

Author Contributions: Writing—original draft preparation, Z.L.; writing—review and editing, R.H., M.X., T.A.P. and H.H.; supervision, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by an appointment to the Research Participation Program at the National Center for Toxicological Research (Zoe Li), administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy and the US Food and Drug Administration and the Intramural/Extramural research program of the NCATS, NIH.

Conflicts of Interest: The authors declare no conflicts of interest.

Authors' Disclaimer: This article reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration.

References

1. Janin, J. Protein-Protein Recognition. *Prog. Biophys. Mol. Biol.* **1995**, *64*, 145–166. [CrossRef] [PubMed]
2. Du, X.; Li, Y.; Xia, Y.-L.; Ai, S.-M.; Liang, J.; Sang, P.; Ji, X.-L.; Liu, S.-Q. Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. *Int. J. Mol. Sci.* **2016**, *17*, 144. [CrossRef] [PubMed]
3. Kairys, V.; Baranauskiene, L.; Kazlauskiene, M.; Matulis, D.; Kazlauskas, E. Binding Affinity in Drug Design: Experimental and Computational Techniques. *Expert Opin. Drug Discov.* **2019**, *14*, 755–768. [CrossRef] [PubMed]
4. Colwell, L.J. Statistical and Machine Learning Approaches to Predicting Protein–Ligand Interactions. *Curr. Opin. Struct. Biol.* **2018**, *49*, 123–128. [CrossRef] [PubMed]
5. Gao, K.; Nguyen, D.D.; Sresht, V.; Mathiowetz, A.M.; Tu, M.; Wei, G.-W. Are 2D Fingerprints Still Valuable for Drug Discovery? *Phys. Chem. Chem. Phys.* **2020**, *22*, 8373–8390. [CrossRef] [PubMed]
6. Hansch, C.; Maloney, P.P.; Fujita, T.; Muir, R.M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178–180. [CrossRef]
7. Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold^2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48*, 1337–1344. [CrossRef]
8. Hong, H.; Liu, J.; Ge, W.; Sakkiah, S.; Guo, W.; Yavas, G.; Zhang, C.; Gong, P.; Tong, W.; Patterson, T.A. Mold^2 Descriptors Facilitate Development of Machine Learning and Deep Learning Models for Predicting Toxicity of Chemicals. In *Machine Learning and Deep Learning in Computational Toxicology*; Computational Methods in Engineering & the Sciences; Hong, H., Ed.; Springer International Publishing: Cham, Switzerland, 2023; pp. 297–321. ISBN 978-3-031-20729-7.
9. Guo, W.; Liu, J.; Dong, F.; Chen, R.; Das, J.; Ge, W.; Xu, X.; Hong, H. Deep Learning Models for Predicting Gas Adsorption Capacity of Nanomaterials. *Nanomaterials* **2022**, *12*, 3376. [CrossRef]
10. Liu, J.; Guo, W.; Dong, F.; Aungst, J.; Fitzpatrick, S.; Patterson, T.A.; Hong, H. Machine Learning Models for Rat Multigeneration Reproductive Toxicity Prediction. *Front. Pharmacol.* **2022**, *13*, 1018226. [CrossRef]
11. Liu, J.; Guo, W.; Sakkiah, S.; Ji, Z.; Yavas, G.; Zou, W.; Chen, M.; Tong, W.; Patterson, T.A.; Hong, H. Machine Learning Models for Predicting Liver Toxicity. In *In Silico Methods for Predicting Drug Toxicity*; Methods in Molecular Biology; Benfenati, E., Ed.; Springer US: New York, NY, USA, 2022; Volume 2425, pp. 393–415. ISBN 978-1-07-161959-9.
12. Huang, Y.; Li, X.; Xu, S.; Zheng, H.; Zhang, L.; Chen, J.; Hong, H.; Kusko, R.; Li, R. Quantitative Structure–Activity Relationship Models for Predicting Inflammatory Potential of Metal Oxide Nanoparticles. *Environ. Health Perspect.* **2020**, *128*, 067010. [CrossRef]
13. Idakwo, G.; Thangapandian, S.; Luttrell, J.; Li, Y.; Wang, N.; Zhou, Z.; Hong, H.; Yang, B.; Zhang, C.; Gong, P. Structure–Activity Relationship-Based Chemical Classification of Highly Imbalanced Tox21 Datasets. *J. Cheminform.* **2020**, *12*, 66. [CrossRef] [PubMed]

14. Wang, Z.; Chen, J.; Hong, H. Developing QSAR Models with Defined Applicability Domains on PPARγ Binding Affinity Using Large Data Sets and Machine Learning Algorithms. *Environ. Sci. Technol.* **2021**, *55*, 6857–6866. [CrossRef] [PubMed]
15. Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216. [CrossRef] [PubMed]
16. Khan, M.T.H. Predictions of the ADMET Properties of Candidate Drug Molecules Utilizing Different QSAR/QSPR Modelling Approaches. *Curr. Drug Metab.* **2010**, *11*, 285–295. [CrossRef] [PubMed]
17. Roy, K.; Mitra, I. Electrotopological State Atom (E-State) Index in Drug Design, QSAR, Property Prediction and Toxicity Assessment. *Curr. Comput. Aided-Drug Des.* **2012**, *8*, 135–158. [CrossRef] [PubMed]
18. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef] [PubMed]
19. Lo, Y.-C.; Rensi, S.E.; Torng, W.; Altman, R.B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546. [CrossRef]
20. Cereto-Massagué, A.; Ojeda, M.J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63. [CrossRef]
21. Hall, L.H.; Kier, L.B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045. [CrossRef]
22. Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280. [CrossRef]
23. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 33. [CrossRef] [PubMed]
24. Yang, J.; Cai, Y.; Zhao, K.; Xie, H.; Chen, X. Concepts and Applications of Chemical Fingerprint for Hit and Lead Screening. *Drug Discov. Today* **2022**, *27*, 103356. [CrossRef] [PubMed]
25. Dudek, A.; Arodz, T.; Galvez, J. Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. *Comb. Chem. High Throughput Screen.* **2006**, *9*, 213–228. [CrossRef]
26. Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. K_{DEEP}: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287–296. [CrossRef]
27. Feinberg, E.N.; Sur, D.; Wu, Z.; Husic, B.E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V.S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530. [CrossRef] [PubMed]
28. Nguyen, D.D.; Cang, Z.; Wu, K.; Wang, M.; Cao, Y.; Wei, G.-W. Mathematical Deep Learning for Pose and Binding Affinity Prediction and Ranking in D3R Grand Challenges. *J. Comput. Aided Mol. Des.* **2019**, *33*, 71–82. [CrossRef] [PubMed]
29. Ain, Q.U.; Aleksandrova, A.; Roessler, F.D.; Ballester, P.J. Machine-Learning Scoring Functions to Improve Structure-Based Binding Affinity Prediction and Virtual Screening: Machine-Learning SFs to Improve Structure-Based Binding Affinity Prediction and Virtual Screening. *WIREs Comput. Mol. Sci.* **2015**, *5*, 405–424. [CrossRef]
30. Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623–637. [CrossRef]
31. Salentin, S.; Haupt, V.J.; Daminelli, S.; Schroeder, M. Polypharmacology Rescored: Protein–Ligand Interaction Profiles for Remote Binding Site Similarity Assessment. *Prog. Biophys. Mol. Biol.* **2014**, *116*, 174–186. [CrossRef]
32. Crisman, T.J.; Sisay, M.T.; Bajorath, J. Ligand-Target Interaction-Based Weighting of Substructures for Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1955–1964. [CrossRef]
33. Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47*, 337–344. [CrossRef] [PubMed]
34. Mordalski, S.; Kosciolek, T.; Kristiansen, K.; Sylte, I.; Bojarski, A.J. Protein Binding Site Analysis by Means of Structural Interaction Fingerprint Patterns. *Bioorganic Med. Chem. Lett.* **2011**, *21*, 6816–6819. [CrossRef] [PubMed]
35. Vass, M.; Kooistra, A.J.; Ritschel, T.; Leurs, R.; de Esch, I.J.; de Graaf, C. Molecular Interaction Fingerprint Approaches for GPCR Drug Discovery. *Curr. Opin. Pharmacol.* **2016**, *30*, 59–68. [CrossRef] [PubMed]
36. Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207. [CrossRef] [PubMed]
37. Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem* **2018**, *13*, 507–510. [CrossRef]
38. Zhao, Z.; Bourne, P.E. Harnessing Systematic Protein–Ligand Interaction Fingerprints for Drug Discovery. *Drug Discov. Today* **2022**, *27*, 103319. [CrossRef] [PubMed]
39. Radifar, M.; Yuniarti, N.; Istyastono, E.P. PyPLIF: Python-Based Protein-Ligand Interaction Fingerprinting. *Bioinformation* **2013**, *9*, 325–328. [CrossRef] [PubMed]
40. Pérez-Nueno, V.I.; Rabal, O.; Borrell, J.I.; Teixidó, J. APIF: A New Interaction Fingerprint Based on Atom Pairs and Its Application to Virtual Screening. *J. Chem. Inf. Model.* **2009**, *49*, 1245–1260. [CrossRef]
41. Chupakhin, V.; Marcou, G.; Gaspar, H.; Varnek, A. Simple Ligand–Receptor Interaction Descriptor (SILIRID) for Alignment-Free Binding Site Comparison. *Comput. Struct. Biotechnol. J.* **2014**, *10*, 33–37. [CrossRef]
42. Da, C.; Kireev, D. Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. *J. Chem. Inf. Model.* **2014**, *54*, 2555–2561. [CrossRef]

43. Wójcikowski, M.; Kukiełka, M.; Stepniewska-Dziubinska, M.M.; Siedlecki, P. Development of a Protein–Ligand Extended Connectivity (PLEC) Fingerprint and Its Application for Binding Affinity Predictions. *Bioinformatics* **2019**, *35*, 1334–1341. [CrossRef] [PubMed]
44. Jubb, H.C.; Higueruelo, A.P.; Ochoa-Montaño, B.; Pitt, W.R.; Ascher, D.B.; Blundell, T.L. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J. Mol. Biol.* **2017**, *429*, 365–371. [CrossRef] [PubMed]
45. Szulc, N.A.; Mackiewicz, Z.; Bujnicki, J.M.; Stefaniak, F. FingeRNAt—A Novel Tool for High-Throughput Analysis of Nucleic Acid-Ligand Interactions. *PLoS Comput. Biol.* **2022**, *18*, e1009783. [CrossRef] [PubMed]
46. Fassio, A.V.; Shub, L.; Ponzoni, L.; McKinley, J.; O'Meara, M.J.; Ferreira, R.S.; Keiser, M.J.; de Melo Minardi, R.C. Prioritizing Virtual Screening with Interpretable Interaction Fingerprints. *J. Chem. Inf. Model.* **2022**, *62*, 4300–4318. [CrossRef] [PubMed]
47. Kokh, D.B.; Doser, B.; Richter, S.; Ormersbach, F.; Cheng, X.; Wade, R.C. A Workflow for Exploring Ligand Dissociation from a Macromolecule: Efficient Random Acceleration Molecular Dynamics Simulation and Interaction Fingerprint Analysis of Ligand Trajectories. *J. Chem. Phys.* **2020**, *153*, 125102. [CrossRef] [PubMed]
48. Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *J. Cheminform.* **2015**, *7*, 26. [CrossRef]
49. Salentin, S.; Schreiber, S.; Haupt, V.J.; Adasme, M.F.; Schroeder, M. PLIP: Fully Automated Protein–Ligand Interaction Profiler. *Nucleic Acids Res.* **2015**, *43*, W443–W447. [CrossRef]
50. Bouysset, C.; Fiorucci, S. ProLIF: A Library to Encode Molecular Interactions as Fingerprints. *J. Cheminform.* **2021**, *13*, 72. [CrossRef]
51. Sastry, M.; Lowrie, J.F.; Dixon, S.L.; Sherman, W. Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771–784. [CrossRef]
52. Duan, J.; Dixon, S.L.; Lowrie, J.F.; Sherman, W. Analysis and Comparison of 2D Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods. *J. Mol. Graph. Model.* **2010**, *29*, 157–170. [CrossRef]
53. Jiménez-Rosés, M.; Morgan, B.A.; Jimenez Sigstad, M.; Tran, T.D.Z.; Srivastava, R.; Bunsuz, A.; Borrega-Román, L.; Hompluem, P.; Cullum, S.A.; Harwood, C.R.; et al. Combined Docking and Machine Learning Identify Key Molecular Determinants of Ligand Pharmacological Activity on B2 Adrenoceptor. *Pharmacol. Res. Perspect.* **2022**, *10*, e00994. [CrossRef] [PubMed]
54. Zhou, F.; Yin, S.; Xiao, Y.; Lin, Z.; Fu, W.; Zhang, Y.J. Structure–Kinetic Relationship for Drug Design Revealed by a PLS Model with Retrosynthesis-Based Pre-Trained Molecular Representation and Molecular Dynamics Simulation. *ACS Omega* **2023**, *8*, 18312–18322. [CrossRef] [PubMed]
55. Amangeldiuly, N.; Karlov, D.; Fedorov, M.V. Baseline Model for Predicting Protein-Ligand Unbinding Kinetics through Machine Learning. *J. Chem. Inf. Model.* **2020**, *60*, 5946–5956. [CrossRef] [PubMed]
56. Liu, H.; Su, M.; Lin, H.-X.; Wang, R.; Li, Y. Public Data Set of Protein-Ligand Dissociation Kinetic Constants for Quantitative Structure-Kinetics Relationship Studies. *ACS Omega* **2022**, *7*, 18985–18996. [CrossRef]