Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Medical Engineering—Review

The Application of Artificial Intelligence Accelerates G Protein-Coupled Receptor Ligand Discovery

Wei Chen [a,b,*], Chi Song [a,b], Liang Leng [a,b], Sanyin Zhang [a], Shilin Chen [a,b,*]

[a] *State Key Laboratory of Southwestern Chinese Medicine Resources, Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China*
[b] *Institute of Herbgenomics, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China*

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
G protein-coupled receptor
Ligand
Artificial intelligence
Multi-omics
Drug discovery

ABSTRACT

G protein-coupled receptors (GPCRs) are crucial players in various physiological processes, making them attractive candidates for drug discovery. However, traditional approaches to GPCR ligand discovery are time-consuming and resource-intensive. The emergence of artificial intelligence (AI) methods has revolutionized the field of GPCR ligand discovery and has provided valuable tools for accelerating the identification and optimization of GPCR ligands. In this study, we provide guidelines for effectively utilizing AI methods for GPCR ligand discovery, including data collation and representation, model selection, and specific applications. First, the online resources that are instrumental in GPCR ligand discovery were summarized, including databases and repositories that contain valuable GPCR-related information and ligand data. Next, GPCR and ligand representation schemes that can convert data into computer-readable formats were introduced. Subsequently, the key applications of AI methods in the different stages of GPCR drug discovery were discussed, ranging from GPCR function prediction to ligand design and agonist identification. Furthermore, an AI-driven multi-omics integration strategy for GPCR ligand discovery that combines information from various omics disciplines was proposed. Finally, the challenges and future directions of the application of AI in GPCR research were deliberated. In conclusion, continued advancements in AI techniques coupled with interdisciplinary collaborations will offer great potential for improving the efficiency of GPCR ligand discovery.

1. Introduction

G protein-coupled receptors (GPCRs) are a diverse and essential class of cell surface receptors involved in various physiological processes [1]. They belong to the largest family of membrane proteins, with more than 800 members identified in humans [2], including approximately 400 olfactory receptors [3]. GPCRs play fundamental roles in cellular signaling by transducing extracellular signals into intracellular responses and regulating a wide range of biological functions, such as sensory perception [4], neurotransmission [5], hormone regulation [6], immune response [7], and cell proliferation [8]. Accordingly, GPCRs are of immense therapeutic importance and attractive targets for drug discovery. It is estimated that more than 30% of all currently marketed drugs target GPCRs [9]. Because of their involvement in numerous physiological processes and diseases, GPCRs offer great potential for developing novel therapeutics with high specificity and efficacy [10–13].

Despite significant progresses have been made in GPCR research, numerous challenges remain associated with understanding the complex functions and regulatory mechanisms of GPCRs. For example, numerous GPCRs have unknown endogenous ligands [14] or ligands with limited potency and selectivity [15]. Understanding the precise ligand–receptor interactions is another ongoing area of study. GPCRs exhibit diverse binding modes and undergo conformational changes once bound by ligands [16]. Determining the precise structural details of the interactions and elucidating the key residues involved are crucial for rational drug design. Furthermore, the classification and characterization of GPCR subfamilies pose challenges owing to their high sequence diversity and varying functions [17]. Although experimental techniques can address these challenges, they are time-consuming and expensive.

[*] Corresponding authors.
E-mail addresses: greatchen@ncst.edu.cn (W. Chen), slchen@icmm.ac.cn (S. Chen).

https://doi.org/10.1016/j.eng.2023.09.011
2095-8099/

Please cite this article as: W. Chen, C. Song, L. Leng et al., The Application of Artificial Intelligence Accelerates G Protein-Coupled Receptor Ligand Discovery, Engineering, https://doi.org/10.1016/j.eng.2023.09.011

Advances in computational approaches, particularly artificial intelligence (AI)-based methods, have played a pivotal role in addressing these challenges and expediting GPCR research [18]. AI has emerged as a powerful tool that enables researchers to navigate the complex landscape of GPCR biology and ligand discovery with unprecedented efficiency. AI-based methods, such as machine learning and deep learning algorithms, have revolutionized GPCR research by analyzing large-scale data and extracting meaningful patterns. Machine learning algorithms have been used to analyze vast datasets of chemical structures, molecular properties, and biological activity to build predictive models [19–21]. Deep learning algorithm-based methods have assisted ligand optimization by generating novel molecular structures with desired properties [22,23]. Generative models can learn the underlying distribution of GPCR ligands and generate new molecules with similar properties [24,25]. By leveraging the power of AI, researchers have unraveled novel insights into the intricate world of GPCRs, paving the way for the development of innovative therapies to treat a wide range of diseases [26].

Given the extensive number of GPCRs in humans and the staggering diversity of millions of natural small molecule ligands, potential combinations of these receptors and ligands are astronomical. It is impossible to test or screen every possible interaction experimentally. To accelerate the process of identifying and confirming GPCR-ligand interactions, the development of AI-based automatic systems is indispensable. Hence, in this study, we propose an AI-based automatic system for GPCR ligand discovery, as shown in Fig. 1.

By addressing the essential aspects of the field, the present study provides a comprehensive survey of the application of AI in GPCR ligand discovery. We begin by discussing the online resources available for GPCR and ligand information, which provide researchers with valuable resources to access comprehensive information about GPCRs and their associated ligands. Subsequently, various strategies for GPCR ligand representation using AI-based approaches have been explored. Furthermore, this review highlights the significant contributions of AI to the different aspects of GPCR ligand discovery. In addition, the integration of AI with multi-omics data for GPCR ligand discovery and measurement is thoroughly discussed. The primary objective of this review is to provide readers with a clear landscape of recent developments in AI-driven GPCR ligand discovery. By bridging the gap between AI and GPCR research, this review will facilitate the development of novel therapeutics targeting GPCRs and contribute to advancements in drug discovery.

2. Strategies of applying AI in GPCR ligand discovery

Strategies for applying AI in GPCR ligand discovery involve several key steps, namely, data collation, data representation, model selection, and application, which are summarized in Fig. 2. These steps form a systematic process that leverages AI techniques to accelerate the discovery and optimization of GPCR ligands.

The first step is to gather relevant data, such as GPCRs, ligands, and information on their interactions. These data were obtained from various sources and formed the foundation for training AI models and guiding drug discovery. Once data are collected from these sources, they must be curated, cleaned, and standardized before proceeding to the next step.

The second step is data representation, which converts the curated data into numerical formats that can be effectively processed using AI algorithms. The choice of the representation scheme depends on the specific task and type of data available. Different types of data require different representation approaches to convey information effectively to AI models.

Model selection is another crucial step in which researchers select the most suitable AI model based on task complexity, available data, feature space, interpretability, computational resources, and performance metrics. The chosen model should strike a balance between predictive accuracy, interpretability, and resource requirements, leading to more effective drug discovery and the identification of potential GPCR ligands with therapeutic potential. K-nearest neighbors, support vector machine (SVM), random forest (RF), recurrent neural network (RNN), and convolutional neural network (CNN) [27] are commonly used algorithms in AI-based GPCR ligand discovery.

The ultimate stage is the application step, during which the selected AI model is deployed to execute specific tasks in GPCR ligand discovery, such as binding affinity prediction, de novo design, and other critical aspects. The following sections present the key applications of AI in GPCR ligand discovery, highlighting its profound impact on the identification of potential GPCR ligands.

3. Online resources for GPCR and ligand

Online data resources play a critical role in AI-based GPCR ligand discovery and offer numerous advantages that significantly enhance the efficiency and success of drug development. These resources provide access to vast databases, allowing AI algorithms to perform virtual screening and ligand discovery on an unprecedented scale, leading to the identification of novel GPCR ligands.

Several databases have been established to centralize and curate GPCR-related information. These resources compile data on GPCR sequences, structures, functional annotations, ligand–receptor interactions, and signaling pathways. Ligand databases contain a vast collection of small molecules that target GPCRs. These resources store information about ligand structures, properties, activities, and binding affinities. For a quick reference, Table 1 lists the datasets comprising these invaluable repositories.

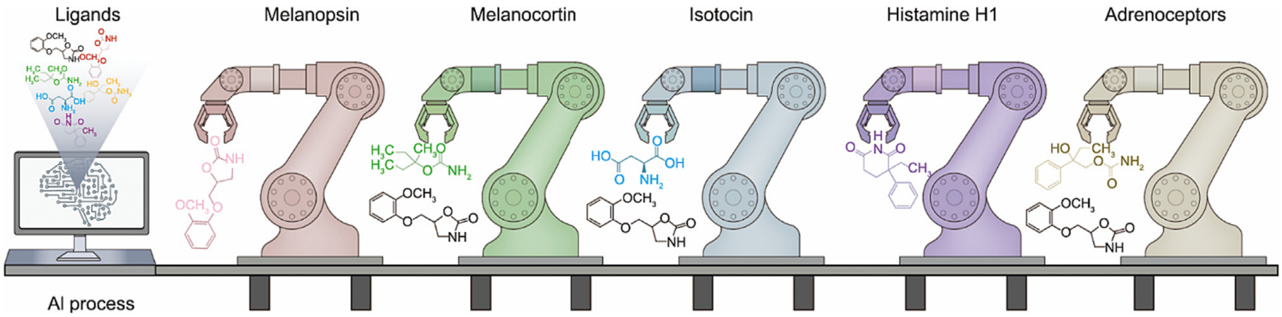


Fig. 1. AI-based automatic systems for GPCR ligand discovery. The ligands were processed by using AI techniques and represented by boxes with different colors. The robot arms, each adorned with a unique color, symbolize diverse AI-based models designed to specific GPCRs. As the ligands move along the conveyer belt, the robot arms will select out specific ligands.

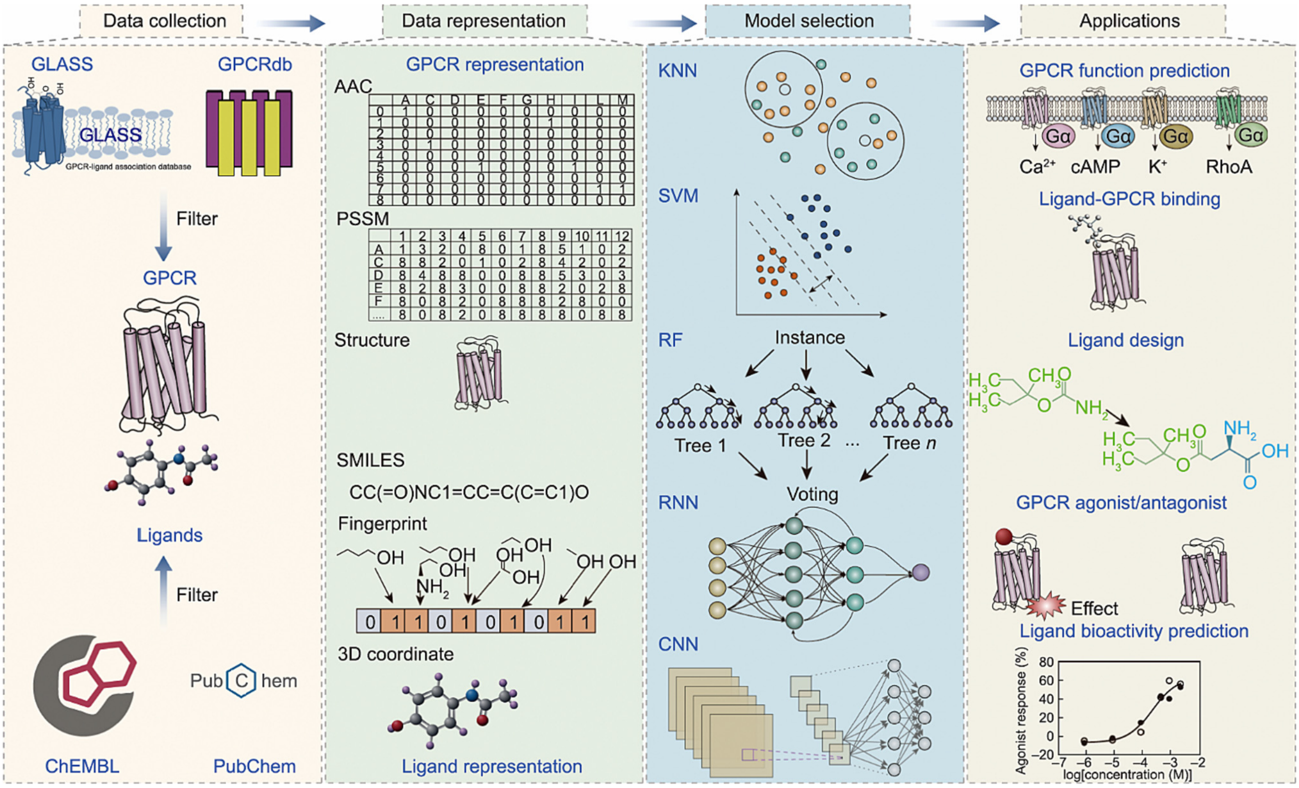


Fig. 2. Framework of applying AI in GPCR ligand discovery. The sequential steps include data collation, data representation, model selection, and application. AAC: amino acid composition; PSSM: position-specific scoring matrix; KNN: k-nearest neighbor; SVM: support vector machine; RF: random forest; RNN: recurrent neural network; CNN: convolutional neural network; SMILES: simplified molecular input line entry system; 3D: three-dimensional; cAMP: cyclic adenosine monophosphate; RhoA: ras homolog family member A; Gα: alpha subunit of G protein.

Table 1
Representative databases for GPCR ligand discovery.

Database	Website URL	Routinely updated	Reference
IUPHAR/BPS	https://www.guidetopharmacology.org	Yes	[28]
GLASS	https://zhanggroup.org/GLASS/	No	[29]
GPCR-I-TASSER	https://zhanggroup.org/GPCR-I-TASSER/	No	[30]
GPCR-RD	https://zhanggroup.org/GPCR-RD/	No	[31]
GPCR-EXP	https://zhanggroup.org/GPCR-EXP/	No	[32]
GPCRdb	https://gpcrdb.org/	Yes	[17]
GpDB	https://bioinformatics.biol.uoa.gr/gpDB	No	[33]
GPCR-ModSim	https://gpcr-modsim.org/	No	[34]
GOMoDo	https://molsim.sci.univr.it/gomodo	No	[35]
RCSB PDB	https://www.rcsb.org/	Yes	[36]
Uniprot	https://www.uniprot.org/	Yes	[37]
MPStruc	https://blanco.biomol.uci.edu/mpstruc/	Yes	[38]
MemProtMD	https://memprotmd.bioch.ox.ac.uk/	Yes	[39]
PubChem	https://pubchem.ncbi.nlm.nih.gov/	Yes	[40]
ChEMBL	https://www.ebi.ac.uk/chembl/	Yes	[41]
ZINC	https://zinc.docking.org/	Yes	[42]
Drugbank	https://go.drugbank.com/	Yes	[43]
BindingDB	https://www.bindingdb.org/rwd/bind/index.jsp	Yes	[44]
DUD-E	https://dude.docking.org/	No	[45]

3.1. Resources of GPCR

The IUPHAR/BPS Guide to PHARMACOLOGY is a valuable resource for GPCR research, as it provides detailed information on the pharmacology and function of GPCRs [28], including receptor classification, expression profiles, physiological functions, and functional assays.

Zhang Lab has developed a series of GPCR-related resources, including the GLASS [29], GPCR-I-TASSER [30], GPCR-RD [31], and GPCR-EXP [32]. GLASS is a manually curated database of experimentally validated GPCR-ligand associations collected from the literature and crosschecked with five primary pharmacological datasets. GPCR-I-TASSER is a computational method designed to predict the three-dimensional (3D) structures of GPCRs. GPCR-RD is a database that contains GPCR restraints gathered from scientific literature through an automated text-mining algorithm combined with manual validation. GPCR-EXP is a database of GPCR structures that provides structure-related data, such as resolution, publication information, and biological ligands.

GPCRdb is a comprehensive and widely used online resource that provides extensive information and tools for studying GPCRs [17]. The GPCRdb offers a wealth of information on various aspects of GPCRs, including their sequences, structures, functions, pharmacology, and ligand interactions. A significant feature of GPCRdb is its comprehensive collection of GPCR ligands and their interactions. It contains a vast repository of ligand-receptor binding data, including binding affinities, potencies, and functional activities. Additionally, the GPCRdb offers various analysis and visualization tools to assist researchers in studying GPCRs. These tools enable sequence analysis, alignment, and comparison of GPCR sequences as well as the prediction of functional sites.

GpDB is a database of G proteins and their associations with GPCRs and effector molecules [33], where G proteins and GPCRs are hierarchically classified into different classes, families, subfamilies, and types.

GPCR-ModSim is a web server for the computational modeling and simulation of GPCRs based on their amino acid sequences using homology modeling techniques [34]. GPCR-ModSim can also be used to equilibrate the GPCR structure with an all-atom molecular dynamic simulation protocol.

GOMoDo is an intuitive web server that enables the seamless modeling of GPCR structures and the docking of corresponding ligands to these models, all within a single, unified pipeline [35].

The RCSB PDB [36] provides a repository of experimentally determined three-dimensional structures of GPCRs, allowing researchers to search for specific GPCR structures, visualize them in various formats, and analyze their interactions with ligands and other molecules.

In addition to the above-mentioned databases, Uniprot [37], MPStruc [38], and MemProtMD [39] are also noteworthy databases that provide valuable information and resources related to GPCRs and will aid researchers in understanding the characteristics and roles of GPCRs.

3.2. Resources of ligands

PubChem is a publicly available database provided by the National Center for Biotechnology Information [40]. It contains information on the physicochemical properties and commercial availability of the chemical ligands. Users can search for specific GPCRs and explore the chemical ligands associated with their biological activities and properties.

ChEMBL is a database of curated bioactive molecules with drug-like properties [41]. It provides information about compounds and their interactions with biological targets, including GPCRs. Researchers can search for specific GPCRs and access information on compounds that have been tested against them, including antibodies and other small molecules.

ZINC is a widely utilized database that houses millions of commercially available chemical compounds specifically curated for virtual screening purposes [42]. In addition to providing comprehensive information regarding compound structures, physical properties, and availability, ZINC offers external links to facilitate further exploration of these compounds.

DrugBank is a widely used database that contains information on drugs and drug targets, including their chemical structures, pharmacological properties, mechanisms of action, indications, and clinical data [43]

BindingDB is a database dedicated to protein–ligand binding data [44]. It provides information on the measured binding affinities and other relevant details for protein–ligand complexes, enabling researchers to explore and analyze ligand–target interactions.

DUD-E provides a curated collection of ligands (active compounds) and decoys (inactive compounds) that can be used to assess the performance of virtual screening methods [45].

Detailed information on the ligands (agonists, antagonists, and modulators) that interact with GPCRs, along with data on their binding affinities, activities, and structures, can also be found in the IUPHAR/BPS Guide to PHARMACOLOGY.

4. Representation strategies

Representation strategies are essential for AI-based GPCR ligand discovery, because they enable efficient data processing, enhance chemical space exploration, and facilitate feature extraction. They also contribute to interpretability, support data fusion and integration, and accelerate the screening processes. The core objective is to convert GPCR sequences and molecular structures into meaningful numerical representations that can be processed using AI algorithms.

4.1. GPCR representation strategies

Converting GPCR into computer-readable features is a key step in AI-aided GPCR discovery and involves converting GPCR into numerical features that can be processed by AI models.

Amino acid composition (AAC) analysis is one of the most straightforward methods for encoding GPCRs. AAC represents the relative frequency of amino acids in a sequence and captures the overall distribution of amino acids [46]. However, AAC cannot capture sequential or positional information of amino acids in GPCRs. In such cases, a position-specific scoring matrix method has been proposed to consider neighboring or context-dependent amino acid information [47].

GPCRs can also be represented by encoding schemes, such as one-hot encoding, and embedding techniques, such as word2vec or transformers. One-hot encoding represents each amino acid as a binary vector that is set to 1 at the corresponding position and 0 at all other positions [48]. For example, considering 20 unique amino acids, each amino acid in the GPCR sequence is represented by a binary vector of length 20, with only one position being 1 and the remainder being 0.

Word2vec is a popular algorithm for learning word embeddings that aims to capture semantic and syntactic relationships between words from a large corpus of text data [49]. The main idea behind word2vec is to train a neural network model to predict the context words surrounding a target word within a given window size [50]. Thus, the model learns to represent words as dense vectors such that similar words have similar vector representations, allowing the similarity between words to be measured based on the distance or cosine similarity between their vectors. Continuous Bag-of-Words and Skip-Gram are two primary architectures used in word2vec [50]. By representing the amino acids of GPCR as word embeddings, the algorithm can capture meaningful relationships between amino acids.

Unlike word2vec, transformers employ attention mechanisms to weigh the importance of different words in a sentence when computing their embeddings [51]. This allows transformers to effectively capture the relationships between words that are further apart and enables the model to consider the entire sentence or document context. Popular transformer-based embedding models, such as ProtBERT [51] and TAPE [52], leverage transformer-based embeddings to capture important features and relationships in protein sequences.

In addition to sequence-based features, other structural properties of GPCRs, such as secondary structure, solvent accessibility, and physicochemical properties of amino acids, have been used to encode GPCRs [53]. These features provide additional information about receptor structure and can be helpful in predicting various aspects of GPCR function.

To correlate specific locations within GPCR structures across different sequences, indexing systems have been proposed to compare information regarding structures, ligand-binding sites, and functional motifs across different GPCRs. The Ballesteros-Weinstein system is a commonly used indexing system that assigns generic residue numbers to specific positions in GPCR. It assigns residue numbers by considering the most conserved residues within transmembrane helices and other pivotal structural components, ensuring uniform numbering throughout GPCRs [54].

4.2. Ligand representation strategies

Effective representation of ligands, such as in virtual screening and ligand-based drug discovery, is crucial for various tasks in drug discovery. In Fig. 3, using acetaminophen as an example, we list the commonly used ligand representation strategies, namely simplified molecular input line entry system (SMILES), fingerprinting, and molecular graphs.

SMILES is a line notation system that represents the structure of a molecule using American Standard Code for Information Interchange (ASCII) characters [55]. SMILES provides a textual representation of the molecular structure, including atom type, bond type, and connectivity information, which can be used for structure searching and similarity calculations. SMILES is usually converted into the aforementioned one-hot-encoded representation. Suppose we have a SMILES string "CC(=O)NC1=CC=C(C=C1)O", the one-hot encoding representation was shown in Fig. 3. The resulting one-hot encoded vector represents the presence or absence of each character in the SMILES string and can be used as an input for the AI. Notably, SMILES differs from isomeric SMILES, primarily by incorporating stereochemical information. Although both notations represent chemical structures using simplified strings of characters, isomeric SMILES considers the stereochemistry, ensuring a more precise representation of molecular configurations and potentially distinct stereoisomers.

In contrast to SMILES notation, which may lead to non-unique representations, representing ligands using the International Chemical Identifier (InChI) ensures a unique and standardized code for each chemical structure. The InChI of a ligand can be either obtained from public databases or generated using the InChI Trust software. InChI provides a comprehensive representation of ligands by incorporating multiple layers of information, such as atoms and their bond connections, tautomeric forms, isotope variations, and stereochemical arrangements. For example, the InChI of acetaminophen is "1S/C8H9NO2/c1-6(10)9-7-2-4-8(11)5-3-7/h2-5,11H,1H3,(H,9,10)".

Although SMILES and InChI provide simplified representations of ligands, they cannot capture the 3D topological arrangement of atoms in the molecule. Molecular fingerprints are binary representations of ligands that encode the presence or absence of specific substructural features or chemical fragments and generate fixed-length binary vectors that represent the presence or absence of specific chemical features in the ligand. Extended connectivity fingerprints (ECFP) [56], Molecular Access System (MACCS) keys [57], and Morgan fingerprints [58] are commonly used fingerprint-based methods. ECFP generates circular fingerprints by considering atom neighborhoods and bond types within a defined radius from each atom and captures structural information up to a certain distance from each atom. MACCS keys are based on a predefined set of structural fragments and capture the presence or absence of predefined structural fragments or patterns in a molecule using a binary value. Different keys capture various aspects of the molecular structure, such as ring systems, functional groups, and specific bond arrangements. Morgan fingerprints encode the local chemical environment around each atom in a molecule using circular substructures with increasing radii. The fingerprints were generated by iteratively extending each atom and hashing the encountered substructures. The resulting fingerprints are binary or integer vectors that represent the occurrence or count of substructures.

A molecular graph is a representation of the structural formula of a ligand based on graph theory, where atoms correspond to nodes and bonds correspond to edges [59]. Graph-based representations capture the connectivity and topological relationships between atoms in a ligand, thereby providing a comprehensive depiction of its structure. The obtained molecular graph can be used for downstream tasks.

5. AI-based GPCR ligand discoveries

The integration of AI into GPCR research has the potential to accelerate the discovery and development of new drugs targeting these important receptors. According to a recent review, AI approaches were mentioned in only 3.6% of published GPCR research by 2022 [60], indicating that the integration of AI into GPCR-related drug discovery has not seen a similar level of advancement as observed in other target areas. Notably, AI methods have the potential to contribute to multiple stages of the GPCR drug discovery process, augmenting our comprehension and expediting breakthroughs in related fields, such as GPCR function prediction, ligand–GPCR binding prediction, ligand design and bioactivity prediction, and agonist identification (Fig. 4).

5.1. GPCR gene ontology (GO) function prediction

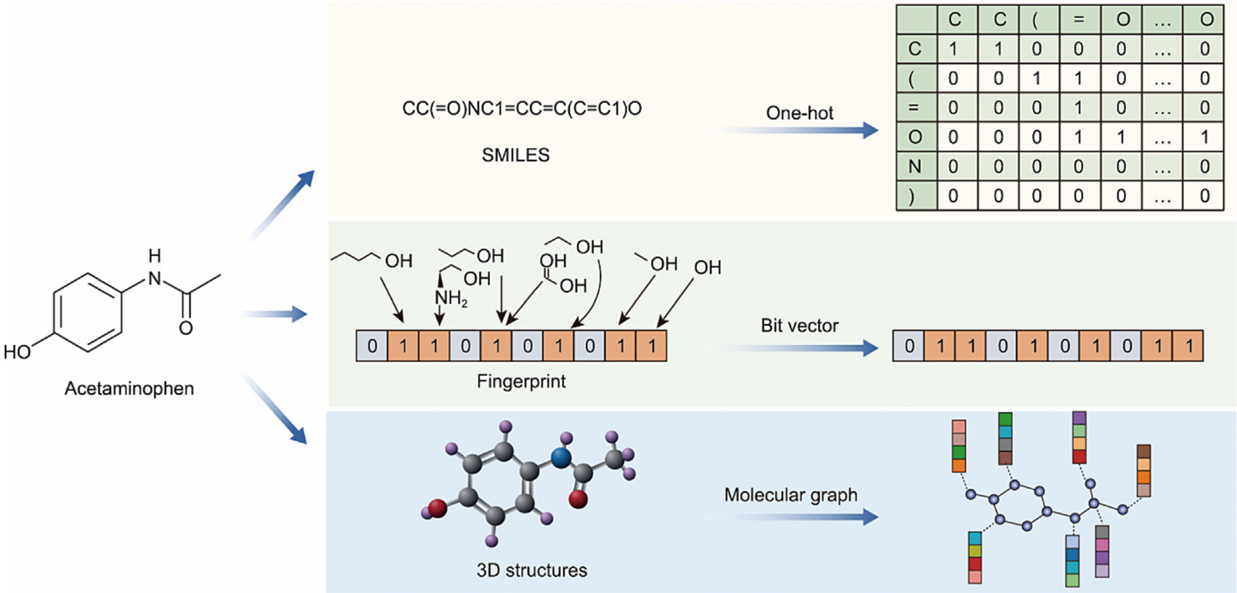
GPCR GO function prediction is significantly important in the field of drug discovery and development. By understanding the GO functions of GPCRs, researchers can gain insights into their roles in signaling pathways and cellular processes. This knowledge will help identify novel drug targets, design therapeutics, develop personalized treatment strategies, and even aid in drug repurposing and targeting orphan GPCRs.

GO [61] is a widely used vocabulary for annotating and describing the functions of genes and gene products and is an essential resource for functional annotation. Wu et al. [53] proposed a three-stage approach, namely text mining (TM)-inductive matrix completion (IMC), which combines TM and IMC, to predict GO terms associated with GPCR proteins. TM-IMC begins by encoding the textual information of GPCR and GO terms using the word2vec algorithm, representing each GPCR or GO as a bag of instances that describe specific function terms. Next, the bag of instances of GPCR or GO was converted into a single vector using the multi-instance learning algorithm based on the fisher vector representation (miFV) [62]. Finally, the IMC method was used to predict the GPCR functions (GO terms) by treating them as a problem of completing the protein function association matrix. The source code of TM-IMC is accessible free on GitHub[†], through which users can predict the GO terms of GPCRs.

5.2. Ligand-GPCR binding prediction

Ligand-GPCR binding refers to the interaction between a ligand molecule and a GPCR and is highly selective and specific, with different ligands having varying affinities for specific GPCRs. The study of ligand–GPCR binding is essential for understanding the mechanisms of GPCR signaling and for drug discovery. Although

[†] https://zhanglab.ccmb.med.umich.edu/TM-IMC.



	C	C	(	=	O	…	O
C	1	1	0	0	0	…	0
(	0	0	1	1	0	…	0
=	0	0	0	1	0	…	0
O	0	0	0	1	1	…	1
N	0	0	0	0	0	…	0
)	0	0	0	0	0	…	0

Fig. 3. Ligand representation strategies. Acetaminophen was taken as an example to show the commonly used ligand representation strategies.

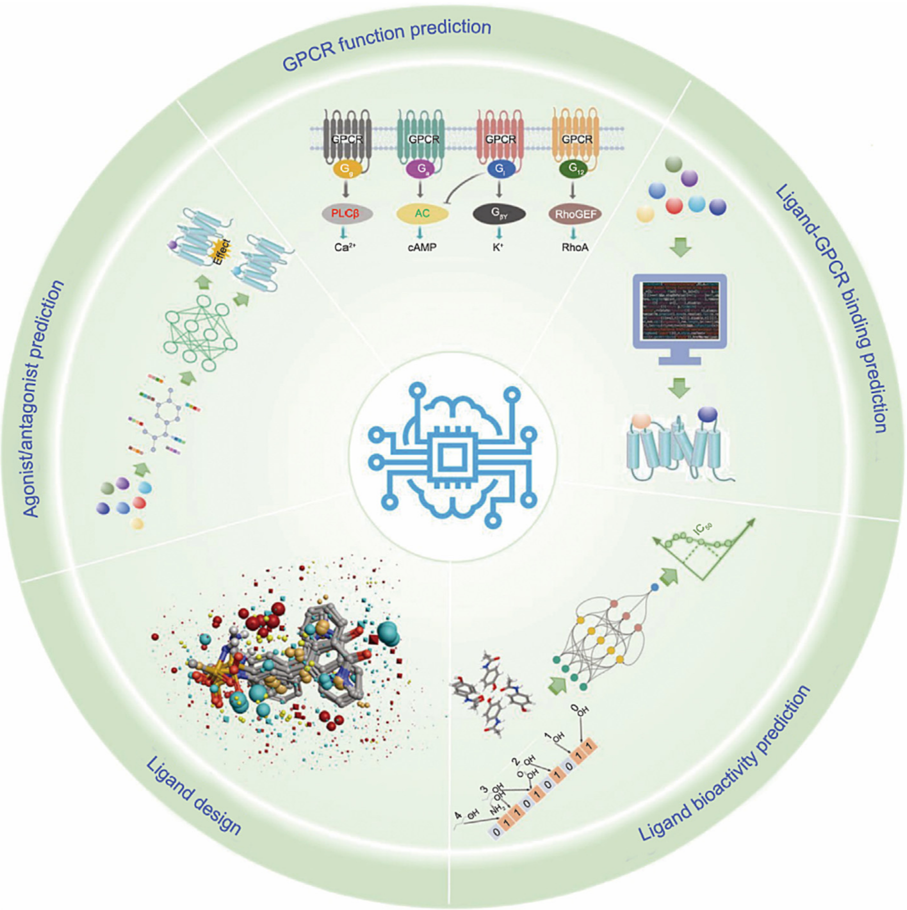


Fig. 4. Illustrative examples of applying AI in GPCR ligand discovery. G_q: G protein subfamily q; G_s: G protein subfamily s; G_i: G protein subfamily i; G_{12}: G protein subfamily 12; PLCβ: phospholipase C-beta; AC: adenylyl cyclase; $G_{\beta\gamma}$: beta and gamma subunits of G protein; RhoGEF: Rho specific guanine nucleotide exchange factor.

experimental techniques, such as X-ray crystallography, cryo-electron microscopy, and nuclear magnetic resonance spectroscopy can determine the ligand–GPCR interactions, they are costly and time-consuming. To overcome these limitations, computational approaches, particularly AI algorithms, have emerged as valuable tools for predicting ligand–GPCR interactions in a more efficient and cost-effective manner. AI algorithms are superior to experimental methods in handling high-dimensional data, capturing complex relationships, and making large-scale predictions. Multiple AI models have been proposed to predict ligand–GPCR interactions.

Based on a dataset containing 303,587 ligand–GPCR interactions, Seo et al. [63] proposed an RF model to predict ligand–GPCR interactions. In their model, GPCRs were encoded using AAC, while motif sequences and ligands were encoded using hub and cycle structures obtained from molecular graphs. The proposed model not only performed better than the empirical affinity predictions of Cyscore [64], but also accurately identified novel ligand–GPCR interactions. Although the source code of the proposed methods was not provided, this study provides valuable insights into computationally predicting ligand–GPCR interactions, which will be helpful for further studies on structure-unknown GPCRs and orphan GPCRs.

In addition to computationally predicting ligand–GPCR interactions, the determination of ligand-specific binding site regions on GPCRs is crucial for understanding the mechanisms of ligand–GPCR interactions and facilitating drug design. Di Rienzo et al. [65] introduced a 3D Zernike polynomial-based approach to identify ligand-binding sites on GPCRs, achieving an area under the receiver operating characteristic curve of 0.77. They further applied this approach to predict ligand-binding sites in olfactory neuron GPCRs in *Caenorhabditis elegans*. This study not only provides a computational tool with broad applicability for predicting binding sites on GPCRs but also enhances our understanding of olfactory GPCRs.

5.3. Ligand bioactivity prediction

Ligand bioactivity was quantitatively measured using half-maximal inhibitory concentration (IC_{50}), half-maximal effective concentration (EC_{50}), inhibition constant (K_i), and dissociation constant (K_d) values, which provided information on the potency, affinity, and efficacy of a ligand in its interaction with GPCRs. Considering these values, researchers can focus on ligands that exhibit the desired levels of potency and affinity, thereby increasing the chances of identifying promising lead compounds. Hence, determining ligand bioactivity is indispensable for virtual screening and drug discovery.

Although biological high-throughput assays offer robust compound screening capabilities, they are not without limitations, notably their time-consuming and labor-intensive nature. In this context, computational methods have proven invaluable by enabling the prioritization of candidates and effectively directing experimental efforts and resources towards the most promising subset of compounds.

In a pioneering study, Wu et al. [66] proposed a weighted deep learning (WDL)-RF method for predicting the bioactivity of GPCR-associated ligands. The WDL-RF is a two-stage computational model that can handle ligands of arbitrary sizes. The first stage of WDL-RF generates molecular fingerprints using a WDL method, while the second stage employs an RF algorithm to perform bioactivity calculations. Large-scale benchmark tests showed that WDL-RF obtained an average root mean square error (RMSE) of 1.33 and a correlation coefficient of 0.80 for predicting the bioactivities of ligands of 26 different GPCRs. Notably, owing to the log transformation of raw bioactivities to p-activities, the resulting RMSE was also presented on a logarithmic scale. To facilitate accessibility and further development, the datasets and source codes of WDL-RF are provided[†], through which users can develop new models based on their own in-house data and predict the bioactivities of new GPCR-associated ligands.

The identification of key substructures that govern ligand bioactivity is of equal importance for virtual screening and drug discovery. By identifying the specific substructures or functional groups responsible for bioactivity, researchers can gain insights into the ligand's mode of action and potentially use this knowledge to design new ligands with improved activity.

Accordingly, Wu et al. [67] developed a deep neural network model called SED to predict ligand bioactivity and detect the substructures that determine ligand bioactivity. The SED method surpassed WDL-RF in predicting the bioactivity of GPCR-associated ligands by encoding ligands using the optimal bits of ECFP. It also has the potential to identify key substructures relevant to ligand bioactivity. The dual capabilities of the SED method provide a novel and promising approach for drug discovery.

Later, Velloso et al. [68] proposed a pdCSM-GPCR model to quantitatively predict the bioactivity of ligands associated with 36 different GPCRs. In pdCSM-GPCR, the ligands are represented using graph-based signatures. This representation captured the structural information of the ligands and allowed the extraction of relevant features for bioactivity prediction. Results from the 10-fold cross-validation and blind tests showed that pdCSM-GPCR outperformed WDL-RF in predicting the bioactivity of the ligands of almost all GPCRs. They also identified important common features of potent GPCR ligands that tended to have bicyclic rings, leading to high levels of aromaticity. The pdCSM-GPCR model is expected to serve as a valuable tool for screening efforts in drug discovery.

5.4. Ligand design

The design of novel ligands that can selectively and effectively modulate GPCR activity has great potential for the treatment of diseases. Because GPCRs are one of the most successful drug target classes, the exploration and design of novel ligands for these receptors can pave the way for the discovery of innovative drugs with improved pharmacological properties [69–71]. Advancements in deep learning have significantly revolutionized the field of chemical structure generation, offering powerful tools for generating novel molecules, and driving innovations in drug design.

Using reinforcement learning techniques, Liu et al. [72] proposed an RNN-based model called DrugEx to identify novel bioactive ligands against GPCRs. DrugEx incorporates an exploration strategy to enhance exploration and promote the generation of diverse molecules. This combination of the exploration strategy with the RNN-based generation model empowered DrugEx to explore a broader range of chemical spaces, facilitating the identification of novel and promising drug candidates. Comparative results demonstrated that DrugEx outperformed REINVENT [73] in generating bioactive ligands against human adenosine A_{2A} receptors ($A_{2A}AR$).

Subsequently, by adding crossover and mutation operations of evolutionary algorithms to reinforcement learning, DrugEx was updated to a new version, DrugEx v2 [74]. Compared to the original version, DrugEx v2 can generate ligands for multiple targets and one specific target. The test results from both the multitarget and target-specific tasks showed that the SMILES generated by DrugEx v2 were chemically reasonable and represented diverse and unique desired molecules.

Both DrugEx and DrugEx v2 were trained with fixed objectives and lacked the capability of users to provide prior information, such as the desired scaffold. More recently, DrugEx v3 was developed, which introduced a novel positional encoding scheme specifically tailored for atoms and bonds, allowing the transformer model to effectively process molecular graph representations [75]. This enhancement enables simultaneous growth of multiple fragments within a given scaffold and facilitates their connection

[†] https://zhanglab.ccmb.med.umich.edu/WDL-RF/.

to generate entirely new molecules. To demonstrate its efficacy, DrugEx v3 was employed to design ligands targeting A_{2A}AR, and its performance was compared with that of the SMILES-based methods. The results revealed that all the generated molecules were chemically reasonable, showing a remarkable 100% validity rate. In addition, a significant proportion of these ligands exhibited high predicted affinities for A_{2A}AR with the given scaffolds. These findings highlighted the efficacy of DrugEx v3 in generating chemically reasonable ligands.

5.5. GPCR ligand identification and classification

Two primary mechanisms of ligand binding exist in the discovery of bioactive GPCR ligands: orthosteric and allosteric interactions [76]. Orthosteric binding occurs when a ligand binds directly to the active site of the receptor where endogenous ligands typically bind, thereby influencing the signaling pathway of the receptor. In contrast, allosteric binding involves ligand binding to a distinct site on the receptor, which induces conformational changes that modulate the binding and signaling of orthosteric ligands. While orthosteric ligands have traditionally been the focus, allosteric ligands offer selective modulation and the ability to fine-tune receptor activity. Hence, identification of allosteric ligands offers new avenues for drug discovery in the field of GPCRs.

In a recent study, Hou et al. [18] introduced an 11-class classification task that aimed to simultaneously distinguish allosteric ligands across the GPCR A, B, and C subfamilies and inactive ligands. To obtain an optimal model, various combinations of diverse molecular features and machine learning algorithms were employed during the model training process. Results from independent test showed that "SVM-ECFP6" was the best model and obtained satisfying performances across all GPCR classes. This study provides insights into the in-silico discovery of GPCR allosteric ligands.

In addition to orthosteric and allosteric ligands, agonists and antagonists are essential components of GPCR bioactive ligand discovery [77]. Agonists are ligands that bind to GPCRs and activate their signaling pathways, mimicking the effects of endogenous ligands. They promote cellular responses and physiological functions that are mediated by specific GPCR. In contrast, an antagonist is a ligand that counteracts the effects of an agonist by binding to a receptor without activating it [78]. This binding inhibits the response of the receptor to an agonist, suppressing signaling and biological effects [78]. Identifying GPCR-binding agonists and antagonists also holds significant importance and has significant implications for drug discovery, understanding cellular signaling, and developing therapeutic interventions for various diseases and disorders.

Using the RF algorithm and encoding molecules with ECFP fingerprints, another research group proposed a two-layer prediction model to classify polymerase chain reaction (PCR)-associated ligands [79]. The first layer identifies whether a query molecule is a GPCR ligand, while the second layer classifies the ligands as either agonists or antagonists. The proposed method achieved an accuracy of 70% for classifying food and drug administration (FDA)-approved GPCR drugs, indicating its potential as a useful tool for identifying GPCR-binding agonists and antagonists.

6. Multi-omics integration screening strategy

The multi-omics integration screening strategy combines and analyzes data from multiple omics platforms to gain a comprehensive understanding of biological systems [80]. By integrating multiple omics datasets, researchers can uncover complex relationships between different molecular layers and gain a more holistic view.

Although the multi-omics approach is valuable for understanding complex biological systems, its application in GPCR ligand discovery and measurement has not been widely reported or established. Most of the reported approaches for GPCR ligand discovery and measurement focus on identifying and characterizing ligands based on their chemical properties, molecular interactions, and functional activities.

In this section, we propose a multi-omics integration strategy for screening GPCR-associated ligands (Fig. 5). Integrating multi-omics data allows for the identification of novel ligands, understanding their mode of action, and predicting their efficacy or potential side effects. A general overview of this strategy is as follows:

First, endogenously expresses the GPCR of interest or creates a stable cell line expressing the GPCR of interest using genetic engineering techniques and then treats the selected cell line with a library of compounds or individual ligands known to interact with the GPCR. To ensure the reliability of experimental results, it is crucial to include appropriate controls. In the case of endogenously expressing cell lines, controls might involve comparing responses with cells lacking GPCR or using pharmacological agents known to modulate receptor activity. For genetically engineered cell lines, a cell line with the same genetic modifications but lacking the GPCR insert, as well as unmodified cells, was used as a control.

Second, the treated cell lines are processed at the desired time points, and sequencing or measurements are performed using high-throughput technologies to obtain raw data, such as genomic, transcriptomic, proteomic, and metabolomic approaches. These data provide a multidimensional view of GPCR–ligand interactions. Genomic data identifies genetic variations or mutations in GPCR genes that may affect ligand binding or receptor activity. Although heterologous systems strive for genomic equivalence, the choice of expression vectors and promoters introduced during the expression process may lead to variations in genomic data. Transcriptomic data revealed the expression patterns of GPCRs and their associated signaling molecules in different cell lines. Proteomic data can help identify proteins involved in GPCR signaling pathways and their post-translational modifications. Metabolomic data provided information on the metabolites associated with GPCR activity and downstream signaling.

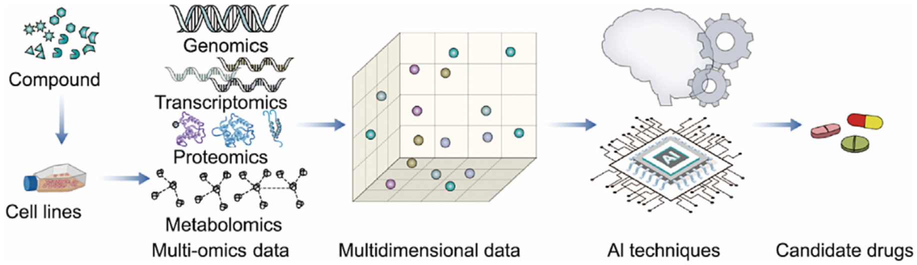


Fig. 5. Multi-omics integration screening strategy for GPCR ligand discovery.

Third, to ensure that multi-omics data are in a suitable format for subsequent analysis, it is necessary to perform quality control, normalization, filtering, and other preprocessing steps specific to the chosen omics field. Quality control evaluates the data integrity and removes low-quality reads and technical biases. Normalization adjusts the data to a common scale, accounting for various factors, such as sequencing depth or biases. Filtering eliminated noise, outliers, and irrelevant features. These steps minimize technical artifacts and variations in the data, enabling meaningful and reliable downstream analysis.

Fourth, analyze the processed multi-omics data to identify differentially expressed genes, proteins, and metabolites associated with the treatment and GPCR activation. In this process, AI methods can be employed to integrate multiple omics datasets by selecting relevant features, applying dimensionality reduction, choosing an appropriate integration method, training the AI model on preprocessed data, and analyzing and interpreting the integrated data for patterns and correlations.

Once the multi-omics data have been preprocessed, the subsequent step is to identify differentially expressed genes, proteins, or metabolites associated with compound treatment and GPCR activation. This analysis could be further enhanced by leveraging AI methods, particularly deep learning, to integrate multiple omics datasets. AI plays a pivotal role in identifying relevant features from high-dimensional multiomics datasets, reducing complexity, and extracting meaningful information. AI integration methods allow for the combination of different omics layers, enabling researchers to discover intricate associations and interactions between genes, proteins, and metabolites that may have been overlooked when considering each omics layer independently. Taken together, by utilizing AI to integrate diverse omics data, previously hidden relationships can be revealed, offering valuable insights into GPCR drug discovery.

7. Future perspectives

The application of AI in GPCR ligand discovery has shown great promise, bringing about a revolution in the process of identifying and optimizing GPCR ligands through large-scale data analysis and computational modeling. However, several challenges and perspectives must be considered for further advancement in this field.

One of the primary challenges is the availability and quality of the data. AI techniques rely heavily on high-quality data to train accurate models. Although there has been an increase in the amount of GPCR-related data, they are still relatively limited compared to other areas of research. GPCRs belong to a diverse family of receptors that exhibit significant structural and functional diversity [81]. Experimental data on GPCRs are not as abundant as those on other drug target classes. The availability of diverse and well-curated datasets encompassing a wide range of GPCR subtypes and ligand classes is essential for training robust AI models. To address this challenge, collaborative data-sharing efforts, diverse data collection from multiple sources, standardized experimental protocols, data annotation, and quality control measures should be implemented. Additionally, AI-driven data augmentation techniques can be utilized to generate synthetic data. Collectively, the implementation of these strategies will enable the development of more robust and accurate AI models for GPCR ligand discovery.

The interpretability of AI models and their predictions is another critical challenge. GPCRs possess complex structural and functional characteristics, and understanding the molecular basis of ligand receptor interactions is essential for rational drug design. AI models, particularly deep learning models, are often considered black boxes, making it difficult to interpret the underlying features and decision-making processes. To address this challenge, efforts should be focused on developing AI models that provide interpretable results. One approach is to incorporate attention mechanisms [48] into AI models, which will enable researchers to gain insights into the specific regions or features of GPCRs or ligands that are most relevant for model predictions. In addition to attention mechanisms, feature importance analysis can contribute to interpretability [82]. By analyzing the contribution of each input feature to the model's predictions, researchers can identify the most influential features or descriptors.

Structural information is crucial for GPCR ligand discovery because of the complex 3D nature of GPCRs [83]. Integrating structural data, including experimentally determined crystal structures and homology models, into AI models can significantly enhance their performance and enable more accurate structure-based ligand designs. The incorporation of structural information provides valuable insights into the topological arrangement of ligand-binding sites, key residues involved in the interactions, and conformational changes upon ligand binding. Therefore, the development of innovative methods that effectively integrate structural information into AI models represents an important future direction for GPCR ligand discovery, ultimately leading to more efficient and rational drug design strategies.

Scoring functions that encompass low-end empirical methods and high-end physics-based techniques also play an important role in drug discovery. By fusing scoring functions and AI, researchers can optimize ligand–receptor interactions and accelerate drug discovery effort [84]. There are several strategies for integrating scoring functions using AI protocols. These include data-driven scoring functions, where AI models learn from experimental data and molecular structures; hybrid approaches combining empirical and physics-based scores through machine learning; active learning strategies for iterative model refinement; and potential advances in deep learning and generative models for more accurate predictions.

The integration of AI with experimental methods is essential for the successful translation of AI-generated predictions into real-world applications. Although AI can accelerate the screening and optimization of GPCR ligands, experimental validation is necessary to confirm the predicted results and ensure the safety and efficacy of the identified ligands. Collaboration between computational and experimental researchers is crucial for bridging the gap between AI-based predictions and experimental validations.

Finally, the integration of multi-omics data sources, including genomic, transcriptomic, proteomic, and metabolomic data, holds tremendous potential for enhancing our understanding of GPCR biology and accelerating the discovery of novel ligands. These different types of data can provide complementary information regarding GPCR function, expression patterns, and signaling pathways, offering a more comprehensive view of GPCR biology. Hence, challenges related to data integration, standardization, and analytical methods must be addressed. Establishing standardized formats and ontologies, developing robust analysis methods, and promoting data sharing will facilitate the effective integration of multi-omics data, ultimately leading to a deeper understanding of GPCR biology and the discovery of novel ligands.

By addressing these challenges and leveraging the power of AI, we can continue to revolutionize the process of identifying and optimizing GPCR ligands, ultimately leading to the development of effective therapies and treatments for various diseases.

Acknowledgments

We extend our gratitude to the anonymous reviewers for their constructive comments, which have greatly enhanced the quality and clarity of the present work. This work was supported by Natural Science Foundation of Sichuan (2023NSFSC0683) and Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (ZYYCXTD-D-202209).

Authors' contributions

Wei Chen and Shilin Chen conceived, supervised, and reviewed the manuscript. Wei Chen, Chi Song, Liang Leng, Sanyin Zhang, and Shilin Chen wrote the manuscript. All of the authors have read and approved the final manuscript.

Compliance with ethics guidelines

Wei Chen, Chi Song, Liang Leng, Sanyin Zhang, and Shilin Chen declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Yang D, Zhou Q, Labroska V, Qin S, Darbalaei S, Wu Y, et al. G protein-coupled receptors: structure- and function-based drug discovery. Signal Transduction Targeted Ther 2021;6(1):7.

[2] Nieto GA, McDonald PH. GPCRs: emerging anti-cancer drug targets. Cell Signaling 2017;41:65–74.

[3] Hauser AS, Attwood MM, Rask-Andersen M, Schiöth HB, Gloriam DE. Trends in GPCR drug discovery: new agents, targets and indications. Nat Rev Drug Discov 2017;16(12):829–42.

[4] Julius D, Nathans J. Signaling by sensory receptors. Cold Spring Harbor Perspect Biol 2012;4(1):a005991.

[5] Hamm HE, Alford ST. Physiological roles for neuromodulation via $G_{i/o}$ GPCRs working through Gβγ-SNARE interaction. Neuropsychopharmacology 2020;45(1):221.

[6] Feng Z, Sun R, Cong Y, Liu Z. Critical roles of G protein-coupled receptors in regulating intestinal homeostasis and inflammatory bowel disease. Mucosal Immunol 2022;15(5):819–28.

[7] Ge YJ, Liao QW, Xu YC, Zhao Q, Wu BL, Ye RD. Anti-inflammatory signaling through G protein-coupled receptors. Acta Pharmacol Sin 2020;41(12):1531–8.

[8] Dorsam RT, Gutkind JS. G-protein-coupled receptors and cancer. Nat Rev Cancer 2007;7(2):79–94.

[9] Yasi EA, Kruyer NS, Peralta-Yahya P. Advances in G protein-coupled receptor high-throughput screening. Curr Opin Biotechnol 2020;64:210–7.

[10] Sriram K, Insel PA. G protein-coupled receptors as targets for approved drugs: how many targets and how many drugs? Mol Pharmacol 2018;93(4):251–8.

[11] Eiger DS, Pham U, Gardner J, Hicks C, Rajagopal S. GPCR systems pharmacology: a different perspective on the development of biased therapeutics. Am J Physiol Cell Physiol 2022;322(5):C887–95.

[12] Zhao P, Furness SGB. The nature of efficacy at G protein-coupled receptors. Biochem Pharmacol 2019;170:113647.

[13] Campbell AP, Smrcka AV. Targeting G protein-coupled receptor signalling by blocking G proteins. Nat Rev Drug Discov 2018;17(11):789–803.

[14] Raschka S. Automated discovery of GPCR bioactive ligands. Curr Opin Struct Biol 2019;55:17–24.

[15] Powers AS, Pham V, Burger WAC, Thompson G, Laloudakis Y, Barnes NW, et al. Structural basis of efficacy-driven ligand selectivity at GPCRs. Nat Chem Biol 2023;19(7):805–14.

[16] Frei JN, Broadhurst RW, Bostock MJ, Solt A, Jones AJY, Gabriel F, et al. Conformational plasticity of ligand-bound and ternary GPCR complexes studied by ^{19}F NMR of the β_1-adrenergic receptor. Nat Commun 2020;11(1):669.

[17] Pándy-Szekeres G, Caroli J, Mamyrbekov A, Kermani AA, Keserű GM, Kooistra AJ, et al. GPCRdb in 2023: state-specific structure models using AlphaFold2 and new ligand resources. Nucleic Acids Res 2023;51(D1):D395–402.

[18] Hou T, Bian Y, McGuire T, Xie XQ. Integrated multi-class classification and prediction of GPCR allosteric modulators by machine learning intelligence. Biomolecules 2021;11(6):870.

[19] Raschka S, Kaufman B. Machine learning and AI-based approaches for bioactive ligand discovery and GPCR-ligand recognition. Methods 2020;180:89–110.

[20] Rataj K, Kelemen ÁA, Brea J, Loza MI, Bojarski AJ, Keserű GM. Fingerprint-based machine learning approach to identify potent and selective 5-HT$_{2B}$R ligands. Molecules 2018;23(5):1137.

[21] Yadav P, Mollaei P, Cao Z, Wang Y, Farimani AB. Prediction of GPCR activity using machine learning. Comput Struct Biotechnol J 2022;20:2564–73.

[22] Yin Y, Hu H, Yang Z, Jiang F, Huang Y, Wu J. AFSE: towards improving model generalization of deep graph learning of ligand bioactivities targeting GPCR proteins. Brief Bioinform 2022;23(3):bbac077.

[23] Lee S, Kim S, Lee GR, Kwon S, Woo H, Seok C, et al. Evaluating GPCR modeling and docking strategies in the era of deep learning-based protein structure prediction. Comput Struct Biotechnol J 2022;21:158–67.

[24] Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: generative models for matter engineering. Science 2018;361(6400):360–5.

[25] Thomas M, Smith RT, O'Boyle NM, de Graaf C, Bender A. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. J Cheminform 2021;13(1):39.

[26] Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol 2019;37(9):1038–40.

[27] Chen W, Liu X, Zhang S, Chen S. Artificial intelligence for drug discovery: resources, methods, and applications. Mol Ther Nucleic Acids 2023;31:691–702.

[28] Alexander SPH, Christopoulos A, Davenport AP, Kelly E, Mathie A, Peters JA, et al. The concise guide to pharmacology 2021/22: G protein-coupled receptors. Br J Pharmacol 2021;178(Suppl 1):S27–S156.

[29] Chan WKB, Zhang H, Yang J, Brender JR, Hur J, Özgür A, et al. GLASS: a comprehensive database for experimentally validated GPCR-ligand associations. Bioinformatics 2015;31(18):3035–42.

[30] Zhang J, Yang J, Jang R, Zhang Y. GPCR-I-TASSER: a hybrid approach to G protein-coupled receptor structure modeling and the application to the human genome. Structure 2015;23(8):1538–49.

[31] Zhang J, Zhang Y. GPCRRD: G protein-coupled receptor spatial restraint database for 3D structure modeling and function annotation. Bioinformatics 2010;26(23):3004–5.

[32] Chan WKB, Zhang Y. Virtual screening of human class-A GPCRs using ligand profiles built on multiple ligand-receptor interactions. J Mol Biol 2020;432(17):4872–90.

[33] Theodoropoulou MC, Bagos PG, Spyropoulos IC, Hamodrakas SJ. gpDB: a database of GPCRs, G-proteins, effectors and their interactions. Bioinformatics 2008;24(12):1471–2.

[34] Esguerra M, Siretskiy A, Bello X, Sallander J, Gutiérrez-de-Terán H. GPCR-ModSim: a comprehensive web based solution for modeling G-protein coupled receptors. Nucleic Acids Res 2016;44(W1):W455–62.

[35] Sandal M, Duy TP, Cona M, Zung H, Carloni P, Musiani F, et al. GOMoDo: a GPCRs online modeling and docking webserver. PLoS One 2013;8(9):e74092.

[36] Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, et al. RCSB protein data bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. Nucleic Acids Res 2023;51(D1):D488–508.

[37] Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, Alpi E, et al. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res 2023;51(D1):D523–31.

[38] White SH. Biophysical dissection of membrane proteins. Nature 2009;459(7245):344–6.

[39] Newport TD, Sansom MSP, Stansfeld PJ. The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. Nucleic Acids Res 2019;47(D1):D390–7.

[40] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Res 2023;51(D1):D1373–80.

[41] Mendez D, Gaulton A, Bento AP, Chambers J, de Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 2019;47(D1):D930–40.

[42] Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, et al. ZINC20-a free ultralarge-scale chemical database for ligand discovery. J Chem Inf Model 2020;60(12):6065–73.

[43] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46(D1):D1074–82.

[44] Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 2007;35(Database issue):D198–201.

[45] 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):6582–94.

[46] Feng P, Liu W, Huang C, Tang Z. Classifying the superfamily of small heat shock proteins by using g-gap dipeptide compositions. Int J Biol Macromol 2021;167:1575–8.

[47] Khanh Le NQ, Nguyen QH, Chen X, Rahardja S, Nguyen BP. Classification of adaptor proteins using recurrent neural networks and PSSM profiles. BMC Genomics 2019;20(Suppl 9):966.

[48] Zhang G, Tang Q, Feng P, Chen W. IPs-GRUAtt: an attention-based bidirectional gated recurrent unit network for predicting phosphorylation sites of SARS-CoV-2 infection. Mol Ther Nucleic Acids 2023;32:28–35.

[49] Buchan DWA, Jones DT. Learning a functional grammar of protein domains using natural language word embedding techniques. Proteins 2020;88(4):616–24.

[50] Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. Comput Struct Biotechnol J 2021;19:1750–8.

[51] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2022;44(10):7112–27.

[52] Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, et al. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst 2019;32:9689–701.

[53] Wu J, Yin Q, Zhang C, Geng J, Wu H, Hu H, et al. Function prediction for g protein-coupled receptors through text mining and induction matrix completion. ACS Omega 2019;4(2):3045–54.

[54] Ballesteros JA, Weinstein H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. Methods Neurosci 1995;25: 366–428.

[55] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 1988;28(1):31–6.

[56] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50(5):742–54.

[57] Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 2002;42(6):1273–80.

[58] Zagidullin B, Wang Z, Guan Y, Pitkänen E, Tang J. Comparative analysis of molecular fingerprints in prediction of drug combination effects. Brief Bioinform 2021;22(6):bbab291.

[59] Wu Z, Wang J, Du H, Jiang D, Kang Y, Li D, et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. Nat Commun 2023;14(1):2585.

[60] Nguyen ATN, Nguyen DTN, Koh HY, Toskov J, MacLean W, Xu A, et al. The application of artificial intelligence to accelerate G protein-coupled receptor drug discovery. Br J Pharmacol 2023 May:bph.16140.

[61] Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, et al. The Gene Ontology knowledgebase in 2023. Genetics 2023;224(1):iyad031.

[62] Wei XS, Wu J, Zhou ZH. Scalable algorithms for multi-instance learning. IEEE Trans Neural Netw Learn Syst 2017;28(4):975–87.

[63] Seo S, Choi J, Ahn SK, Kim KW, Kim J, Choi J, et al. Prediction of GPCR-ligand binding using machine learning algorithms. Comput Math Methods Med 2018;2018:6565241.

[64] Cao Y, Li L. Improved protein-ligand binding affinity prediction by using a curvature-dependent surface-area model. Bioinformatics 2014;30(12):1674–80.

[65] Di Rienzo L, de Flaviis L, Ruocco G, Folli V, Milanetti E. Binding site identification of G protein-coupled receptors through a 3D Zernike polynomials-based method: application to C. elegans olfactory receptors. J Comput Aided Mol Des 2022;36(1):11–24.

[66] Wu J, Zhang Q, Wu W, Pang T, Hu H, Chan WKB, et al. WDL-RF: predicting bioactivities of ligand molecules acting with G protein-coupled receptors by combining weighted deep learning and random forest. Bioinformatics 2018;34 (13):2271–82.

[67] Wu J, Liu B, Chan WKB, Wu W, Pang T, Hu H, et al. Precise modelling and interpretation of bioactivities of ligands targeting G protein-coupled receptors. Bioinformatics 2019;35(14):i324–32.

[68] Velloso JPL, Ascher DB, Pires DEV. pdCSM-GPCR: predicting potent GPCR ligands with graph-based signatures. Bioinform Adv 2021;1(1):vbab031.

[69] Manglik A, Lin H, Aryal DK, McCorvy JD, Dengler D, Corder G, et al. Structure-based discovery of opioid analgesics with reduced side effects. Nature 2016;537(7619):185–90.

[70] Kampen S, Rodriguez D, Jørgensen M, Kruszyk-Kujawa M, Huang X, Collins Jr M, et al. Structure-based discovery of negative allosteric modulators of the metabotropic glutamate receptor 5. ACS Chem Biol 2022;17(10):2744–52.

[71] Roth BL, Irwin JJ, Shoichet BK. Discovery of new GPCR ligands to illuminate new biology. Nat Chem Biol 2017;13(11):1143–51.

[72] Liu X, Ye K, van Vlijmen HWT, IJzerman AP, van Westen GJP. An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A_{2A} receptor. J Cheminform 2019;11(1):35.

[73] Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. J Cheminform 2017;9(1):48.

[74] Liu X, Ye K, van Vlijmen HWT, Emmerich MTM, IJzerman AP, van Westen GJP. DrugEx v2: de novo design of drug molecules by Pareto-based multi-objective reinforcement learning in polypharmacology. J Cheminform 2021;13(1):85.

[75] Liu X, Ye K, van Vlijmen HWT, IJzerman AP, van Westen GJP. DrugEx v3: scaffold-constrained drug design with graph transformer-based reinforcement learning. J Cheminform 2023;15(1):24.

[76] Flor PJ, Acher FC. Orthosteric versus allosteric GPCR activation: the great challenge of group-Ⅲ mGluRs. Biochem Pharmacol 2012;84(4):414–24.

[77] Tyndall JDA, Sandilya R. GPCR agonists and antagonists in the clinic. Med Chem 2005;1(4):405–21.

[78] Sum CS, Murphy BJ, Li Z, Wang T, Zhang L, Cvijic ME. Pharmacological characterization of GPCR agonists, antagonists, allosteric modulators and biased ligands from HTS hits to lead optimization. In: Markossian S, Grossman A, Brimacombe K, Arkin M, Auld D, Austin C, et al, editors. Assay Guidance Manual [Internet]. Bethesda: Eli Lilly & Company and the National Center for Advancing Translational Sciences; 2004.

[79] Oh J, Ceong HT, Na D, Park C. A machine learning model for classifying G-protein-coupled receptors as agonists or antagonists. BMC Bioinf 2022;23 (Suppl 9):346.

[80] Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform 2022;23(1):bbab454.

[81] Lagerström MC, Schiöth HB. Structural diversity of G protein-coupled receptors and significance for drug discovery. Nat Rev Drug Discov 2008;7(4):339–57.

[82] Tang Q, Nie F, Zhao Q, Chen W. A merged molecular representation deep learning method for blood-brain barrier permeability prediction. Brief Bioinform 2022;23(5):bbac357.

[83] Odoemelam CS, Percival B, Wallis H, Chang MW, Ahmad Z, Scholey D, et al. G-protein coupled receptors: structure and function in drug discovery. RSC Adv 2020;10(60):36337–48.

[84] Guedes IA, Barreto AMS, Marinho D, Krempser E, Kuenemann MA, Sperandio O, et al. New machine learning and physics-based scoring functions for drug discovery. Sci Rep 2021;11(1):3198.