WADDAICA: a webserver for aiding protein drug design by artificial intelligence and classical algorithm

Qifeng Bai[1]*, Jian Ma[3], Shuo Liu[3], Tingyang Xu[4], Antonio Jesús Banegas-Luna[2], Horacio Pérez-Sánchez[2]*, Yanan Tian[3], Junzhou Huang[4], Huanxiang Liu[3], Xiaojun Yao[1]*

[1] Key Lab of Preclinical Study for New Drugs of Gansu Province, Institute of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Lanzhou University, Lanzhou, Gansu 730000, P. R. China

[2] Structural Bioinformatics and High Performance Computing Research Group (BIO-HPC), Computer Engineering Department, UCAM Universidad Católica San Antonio de Murcia, Murcia, Spain

[3] School of Pharmacy, Lanzhou University, Lanzhou, Gansu 730000, P. R. China

[4] Tencent AI Lab, Shenzhen, P. R. China

* To whom correspondence should be addressed. Tel: +86-931-8912011; Fax: +86-931-8912011; Email: baiqf@lzu.edu.cn, hperez@ucam.edu or xjyao@lzu.edu.cn

Abstract

Artificial intelligence can train the related known drug data into deep learning models for drug design, while classical algorithms can design drugs through established and predefined procedures. Both deep learning and classical algorithms have their merits for drug design. Here, the webserver WADDAICA is built to employ the advantage of deep learning model and classical algorithms for drug design. The WADDAICA mainly contains two modules. In the first module, WADDAICA provides deep learning models for scaffold hopping of compounds to modify or design new novel drugs. The deep learning model which is used in WADDAICA shows a good scoring power based on the PDBbind database. In the second module, WADDAICA supplies functions for modifying or designing new novel drugs by classical algorithms. WADDAICA shows better Pearson and Spearman correlations of binding affinity than Autodock Vina that is considered to have the best scoring power. Besides, WADDAICA supplies a friendly and convenient web interface for users to submit drug design jobs. We believe that WADDAICA is a useful and effective tool to help researchers to modify or design novel drugs by deep learning models and classical algorithms. WADDAICA is free and accessible at https://bqflab.github.io.

Keywords: Drug design, Webserver, Artificial intelligence, Classical algorithm, Deep learning, Class D GPCR

1. Introduction

Deep learning has made rapid progress in image classification [1], speech recognition [2], natural language processing [3], drug discovery [4-6], etc. The traditional machine learning methods rely on manual features extraction, while deep learning allows models to learn the task-related features extraction automatically [7]. Deep learning is a subset of machine learning techniques that uses neural networks to solve complex and challenging problems. It includes a diversity of artificial neural network variants, such as deep convolutional neural networks (CNNs), deep recurrent neural networks (RNNs), graph neural networks (GNNs), and so on. The CNNs approach is one common way to train the deep learning model for predicting the binding affinity between proteins and small molecules [8, 9]. Deep learning has successfully been applied to the *de novo* drug design and ligand binding affinity prediction that can be further used for virtual drug screening. Ligdream is one excellent method for *de novo* drug design [10] which is used to train a deep learning model that could design novel functional groups and scaffolds based on the supplied seed molecule by long short-term memory (LSTM) [11] networks and CNNs. Two well-known examples of neural networks in the field of drug discovery are Pafnucy [12] and OnionNet [13]. Both of them perform well in predicting the binding affinity between proteins and ligands. Pafnucy and OnionNet are tailored for structure-based virtual drug screening by training CNNs models. Pafnucy extracts the chemical information around ligand atoms within 20 Å side length of cubic box to fit into a CNN model for predicting the binding affinity between proteins and ligands. OnionNet takes into account the element-pair-specific contacts between proteins and ligands, and divides the contacts into different distance ranges that cover the local and nonlocal interaction information for training the binding affinity model.

Although deep learning has been successfully applied to drug discovery, it cannot replace the classical algorithms and programs for drug design completely. Some studies point out that the classical scoring functions show higher and more stable performance than the machine learning-based methods at different similarity levels of training sets [14]. Currently, the classical algorithm of *de novo* drug design is easier to grow and search the 3D conformation of ligands in a 3D protein pocket than deep learning [15]. Deep learning and classical algorithm have different strengths and they complement each other well. Our developed program MolAICal [15, 16] is a drug design software tool based on both the deep learning model and classical algorithm. It uses the deep learning model to produce drug-like fragments or molecules, and then perform the *de novo* drug design or virtual drug screening based on the produced molecular set. Moreover, MolAICal can cluster and filter the designed drugs according to the K-means algorithm, Pan-assay interference compounds (PAINS) [17], Lipinski's rule of five [18], synthetic accessibility (SA), and other user-defined rules. Autodock Vina [19] is another typical example of a popular and classical molecular docking program that can find the suitable 3D pose of the ligand in the pocket of protein and carry out virtual drug screening.

Both deep learning and classical algorithms have their unique advantages for drug design. However, some deep learning models and classical programs need special libraries and operating environments and do not have a friendly interface for users to design or modify drugs. A webserver could supply a convenient way for the researchers to design drugs without any special software and hardware requirements via the browsers [20-22]. Here, the webserver named WADDAICA is built for designing or modifying drugs by deep learning and classical algorithm. WADDAICA uses the good scoring model that is trained on the PDBbind database [23, 24] by OnionNet [13]. For the scoring function of the classical algorithm, Autodock Vina is reported to have the best scoring power by evaluating the ten popular docking programs [25]. WADDAICA employs the Vinardo score of MolAICal [15, 16] that shows better Pearson and Spearman correlations than the score of Autodock Vina. Our server contains two drug design modules based on deep learning models and classical algorithms. WADDAICA can easily use independent and combinational functions to design or modify candidate compounds. We strongly believe that WADDAICA can be a very helpful tool for researchers to discover novel drugs.

2. Materials and methods

2.1. Principles and process of server

WADDAICA is built on trained deep learning models [10, 13] and our developed software tool MolAICal [15, 16] that is written for drug design with classical programs and deep learning models. This server aims to supply a friendly web interface for the job submission of drug design or modification conveniently. Figure 1 shows the overall workflow of the two modules of WADDAICA. In the first module, the deep learning model is trained based on 385593523 drug-like molecules of ZINC 15 database [26] by using ligdream source code [10]. The 26 tokens of molecular strings are preserved for training the deep learning model. The 3D conformations of ligands are generated and optimized by RDKit and MMFF94 force field [27]. The molecule is rotated randomly and translated 2 Å after voxelizing into 1 Å cubic grid of side size 24 Å. The value of every voxel is definitive by atom type and the distance r between its center and neighboring atoms (see equation 1).

$$n(r) = 1 - \exp[-(r_{vdw}/r)^{12}] \quad (1)$$

Where r_{vdw} corresponds to the van der Waals radius of an atom. The shape variational autoencoder (VAE) encodes the ligand representation via convolutional neural networks (CNNs). The SMILES strings are generated by long short-term memory (LSTM) [28] and CNNs. The deep learning model for binding affinity is trained based on the PDBbind database by OnionNet [13]. Eight element types are used to determine the atom contact types between proteins and ligands. A total of 60 shells are

picked up for evaluating short-range and long-range element-pair interaction. The distance of the first shell is 1.0 Å and a distance of 0.5 Å is kept between two neighbor shells. The cut-off of the maximum distance is 30.5 Å between the farthest boundary and the atoms of the ligand. A total number of 3840 features is considered into the local and nonlocal interactions between the ligand and protein. The CNNs are employed to train the prediction model of binding affinity. The loss function is shown in equation 2:

$$loss = \alpha(1 - R) + (1 - \alpha)RMSE \quad (2)$$

where R is the correlation coefficient, RMSE is the root-mean-squared error and α is a tunable parameter. In the first model, the WADDAICA server can generate the appointed number of new molecules based on the submitted seed molecule, and predict the binding affinity of generated ligands in the pocket of protein by deep learning model (see Figure 1).

In the second module, WADDAICA employs our developed classical program MolAICal for drug design or modification. The new drugs are grown on the submitted drug seed by using genetic algorithm (see Figure 1). The value of maximum populations is set to 2000. 10% of generated ligands are selected for the next evolved growth. The top 105 molecules of generated molecules are chosen as the parent molecules. Besides, other additional 45 molecules are randomly selected from the generated molecules to enhance the diversity of ligands. Both the operators of crossover and mutation are set to 0.5. According to the report about Lipinski's rule of five values [29], the values of XLOGP, hydrogen acceptors, hydrogen donors, rotatable bonds, and molecular weight are set to 6.0, 12, 7, 20, and 1000.0, respectively. The Pan-assay interference compounds (PAINS) is used to filter out the false-positive growth compounds. The synthetic accessibility (SA) scores of growth molecules are stored in the file of statistical results. The result data of the submitted job is saved in the WADDAICA storage system for one week.

2.2. Software

The web application is constructed based on Flask V1.1.2 by using Python programming language. Several software tools are implemented in the WADDAICA web application under the permitted licenses. JSmol (http://www.jmol.org) is used to visualize the new molecules generated by deep learning model or classical algorithm. Autodock Vina [19] can assist to obtain the complex of protein and ligand. JSME [30] is employed to edit molecules and SMILES strings. Open Babel [31] plays a role in molecular format conversion. Besides, our developed program MolAICal is used to calculate Lipinski's rule of five values [18], synthetic accessibility, and PAINS [17].

3. Results and discussion

3.1. Input

WADDAICA mainly picks up two ways to submit input files on the job request page. One is a molecular editor interface that can draw a molecule or input molecular SMILES string directly. The other is a molecular upload interface that can send the input files into the server for running. WADDAICA supplies the six selectable functions to design or modify drugs on the job submission page. Besides, every function has an independent way for drug design or property calculation. For example, the function of drug design by AI can produce the new ligands based on the submitted seed by invoking deep learning model. And the function of drug properties calculation can independently compute Lipinski's rule of five values, synthetic accessibility, and PAINS. The "Job title" and "Email" are the optional fields for the users on the job submission page. In the first module, WADDAICA provides the JSME interface for the users to draw a molecule or write molecular SMILES strings as the input data (see Figure 2A). The deep learning model will run the job of new molecular generation in the background when the input data is sent to the server. In the second module, the input data contains the simple configure file, protein, and seed files with PDB format (see Figure 3A). The configure file only contains four simple parameters that are the box length, the coordinates of the box center, the names of protein and seed. WADDAICA can upload these prepared materials to run the job of drug design in the background. WADDAICA also supplies a friendly upload interface for other four functions: "Binding affinity by AI", "Molecular docking", "Binding affinity by CA" and "Drug properties calculation" (see Figure 3B, 3C, 3D and 3E). Once the input files are uploaded, the molecular growth job will be carried out automatically. In addition, WADDAICA supplies the tutorial template at the bottom of the job submission page. The users can easily submit the new jobs of drug design or modification by replacing the tutorial template.

When the users click the button of submit and running, the job submission page will skip to the status page (see Figure 2B). The status page shows the job title, job ID, status, and created time of the job. The value of status is queued, started, finished, or failed according to the actual task state. The status page will refresh to show the results at intervals until the job is finished. The users can wait for the final results on the status page or save the offline URL that can be loaded to check the final results when the status page is closed.

3.2. Output

When the job of drug design is complete by running a deep learning model or classical algorithm in the background, the generated molecular files are compressed into the zip file that can be downloaded from the results page. Meanwhile, the result page also shows the 3D structures of generated molecules by using the JSmol plugin (see Figure 2C). The users can selectively load the protein structure into JSmol interface to check the binding pose of the newly generated ligand in the protein pocket. In the

first module, the drugs are designed or modified by deep learning model. The zip file on the results page contains the appointed number of produced 3D molecules in the mol2 format and the file that stores the SMILES strings of generated molecules. The users can judge the newly generated ligands according to the experience in pharmaceutical chemistry or further evaluate the binding affinity between generated ligand and protein by deep learning model. The results page will show the pK_x (pK_d or pK_i) that can be used to assess the binding affinities of ligands quantitatively. WADDAICA also shows the results of binding free energy with equation 3:

$$\text{binding free energy} = RT * \log_e(10^{-\text{p}K_x}) \quad (3)$$

where R and T are the gas constant and temperature, respectively. In addition, the users can check Lipinski's rule of five, synthetic accessibility, and PAINS of generated ligand by submitting a job into the function of drug properties calculation in WADDAICA. In the second module, the drugs could be designed or modified by our developed program MolAICal with genetic algorithm. The output results are compressed into a zip file that contains generated 3D molecules in the mol2 format and result record file. The result record file consists of the items of ID, name, cluster, affinity, formula, inChIKey, and synthetic accessibility of generated ligands (see Figure 2D). The cluster item employs the K-means algorithm to classify the generated ligands. Affinity is the binding score between generated ligand and protein. The formula and inChIKey can help the users to retrieve and distinguish generated ligands. The users can select the wanted ligands according to items of the cluster, affinity, and synthetic accessibility. In this process, WADDAICA can filter out the PAINS and ligands that are not in accordance with the setting cut-off of Lipinski's rule of five, automatically.

3.3. Validation and case study

The assessment of scoring function based the PDBbind database can show the state-of-the-art of the deep learning model and the classical function of binding affinities that are used in the WADDAICA server. In the first module of WADDAICA, the deep learning model, which is trained by OnionNet [13] based on the PDBbind database, is employed to evaluate the binding affinity between the protein and ligand. Table 1 shows the performance comparison between OnionNet model and other three popular machine learning models (*k*NN-Score [32], RF-Score-v3 [33], and Pafnucy [33]). The OnionNet model shows lower standard deviations (SD) and better Pearson correlation coefficients (R_p) between the experimental pK_x and predicted pK_x than other three models. It indicates that the OnionNet model is a relatively good binding affinity model that can be further used for virtual drug screening. In the second module of WADDAICA, the classic scoring function Vinardo [15], which is trained on basis of the score function of AutoDock Vina, is employed to calculate the binding affinity between protein and ligand. Table 2 shows the performance comparison between Vinardo and the scoring functions of AutoDock (LGA) [25], AutoDock (PSO) [25], AutoDock Vina [25], LeDock

[25], rDock [25], and UCSF DOCK [25] based on the PDBbind database. The Vinardo has the best Pearson's correlation coefficient (R_p of 0.582) and Spearman's rank correlation coefficient (R_s of 0.592). It indicates the Vinardo has the best scoring power for predicting the binding affinity between protein and ligand. The rDock has the negative values of R_p and R_s that indicate the worse correlation between the experimental scores and predicted scores. In addition, the UCSF DOCK has not very good R_p and R_s between the experimental scores and predicted scores. It indicates our trained Vinardo can be used to design or modify drugs well by the classical algorithm.

To illustrate the two modules of WADDAICA, the structure of *Saccharomyces cerevisiae* pheromone receptor Ste2, which is determined by cryogenic electron microscopy (cryo-EM) [34], is selected for drug design by deep learning and classical algorithm. The *Saccharomyces cerevisiae* pheromone receptor Ste2 that belongs to one member of the class D GPCRs family exists as an essential dimer for signaling and functional endocytosis [35] in yeast cells. The drugs targeted to Ste2 can be used to treat intractable fungal diseases. The cryo-EM structure of Ste2 contains the high-affinity agonist tridecapeptide pheromone α-factor (WHWLQLKPGQPMY) in the orthosteric binding site. The residue Y13 in the C terminus of α-factor has the most contacts to the pocket of Ste2. The mutations F204C and F204S of Ste2 can cause the decrease of the ligand binding and signal transduction, and amidation in the C terminus of α-factor results in a 160-fold decrease of binding affinity in the Ste2 [34]. In this case, the residues Y13 and M12 of α-factor are chosen as the seed structure for drug design or modification by deep learning and classical algorithm. The A chain of Ste2 dimer and α-factor are selected for this case study. The center coordinates of the binding box in the pocket of Ste2 are set to 130.560, 120.576, and 128.238 Å, respectively. The lengths of the binding box in the pocket of Ste2 are set to 30.0, 30.0, and 30.0 Å, respectively. Figure 4 shows the binding affinity and XLOGP of ligands that are generated by the deep learning model and classical algorithm. In the first module of WADDAICA, when the seed ligand is submitted to the server, WADDAICA will invoke the deep learning model to generate the new ligands based on the structure of the seed ligand. The binding affinity of the submitted seed is -9.46 kcal/mol. It is obvious that some newly generated ligands have better binding affinity than the submitted seed ligand. The XLOGP values of generated ligands look like a normal distribution. The ligands in the range 1.25~3.5 of XLOGP account for the majority of the total generated ligands (see Figure 4A) indicating that the deep learning model can generate the new potential ligands of class D GPCR Ste2. In the second module of WADDAICA, the seed ligand and protein coordinates are submitted to the server. In the current example, the binding affinity of the initial seed ligand is -2.5 kcal/mol. After a cycle of molecular growth by genetic algorithm, it produces some new ligands that have better binding affinities than the submitted seed ligand. The ligands in the range -4.74~-4.0 kcal/mol of binding affinity take up the majority of the total generated ligands (see Figure 4B). In addition, the classical algorithm also grows the good binding ligands with the binding affinity of -5.1 kcal/mol in the pocket of Ste2. The XLOGP values of generated ligands by the classical algorithm are more discrete than by

the deep learning model (see Figure 4A and 4B) what indicates that the classical algorithms have a different style of drug design with deep learning. The deep learning models and classical algorithms have their specific advantages for drug design or modification in the WADDAICA server. If the users want to skip the patent protection and generate new similar drugs, they can use the deep learning model of WADDAICA. On the contrary, if the users want to modify or design new ligands fragment by fragment based on the submitted seed ligand, the classical algorithm module of WADDAICA is a good choice.

4. Conclusion

In this paper, the webserver WADDAICA is introduced for drug design or modification by deep learning and classical algorithm. The WADDAICA provides a friendly and convenient interface for users to submit the jobs by drawing or uploading the molecular files. In its first module, the deep learning models are employed to modify or design new novel drugs by convolutional neural networks. The deep learning model in WADDAICA shows a better scoring power for predicting the binding affinity of ligands. In the second module, the classical algorithms are used to modify or design new ligands in the protein pocket. The comparisons of scoring power show WADDAICA has better Pearson and Spearman correlations between the experimental scores and predicted scores. This fact indicates that WADDAICA can design new ligands in the protein pocket very well. In general terms, we strongly believe that this webserver is helpful and useful to researchers who are interested in drug design and they can take great advantage of it.

Availability

The documentation, related data and materials of WADDAICA can be obtained on https://heisenberg.ucam.edu:5000 or https://bqflab.github.io

CRediT authorship contribution statement

Qifeng Bai: Conceptualization, Software, Validation, Data Curation, Methodology, Writing - original draft, Supervision, Writing - reviewing and Editing, Funding acquisition, Project administration. **Jian Ma:** Software, Formal analysis, Data Curation. **Shuo Liu:** Validation, Resources, Data Curation. **Tingyang Xu:** Software, Validation, Data Curation, Investigation. **Antonio Jesús Banegas-Luna:** Data Curation, Writing - review & Editing. **Horacio Pérez-Sánchez:** Supervision, Writing - review & Editing, Methodology, Project administration. **Yanan Tian:** Resources, Validation, Methodology. **Junzhou Huang:** Investigation, Validation, Writing - review & Editing. **Huanxiang Liu:** Resources,

Methodology. **Xiaojun Yao:** Conceptualization, Supervision, Writing - review & Editing, Project administration. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We thank the Supercomputing Center of Lanzhou University for supplying the computing resource. We are grateful to "Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202004)" who supports the grant for this project. This work was partially supported by grants from the Fundación Séneca del Centro de Coordinación de la Investigación de la Región de Murcia (Spain) under Project 20988/PI/18, Spanish Ministry of Science and Innovation under Project CTQ2017-87974-R, and by European Project Horizon 2020 SC1-BHC-02-2019 [REVERT, ID:848098].

References

[1] Simonyan K, Zisserman AJapa. Very deep convolutional networks for large-scale image recognition. arXiv preprint. 2014:arXiv:1409.556.
[2] Dahl GE, Yu D, Deng L, Acero AJIToa, speech,, processing l. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing. 2011;20(1):30-42.
[3] Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85-117.
[4] Gawehn E, Hiss JA, Schneider G. Deep Learning in Drug Discovery. Mol Inform. 2016;35(1):3-14.
[5] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23(6):1241-50.
[6] Sliwoski G, Kothiwale S, Meiler J, Lowe EW, Jr. Computational methods in drug discovery. Pharmacol Rev. 2014;66(1):334-95.
[7] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.
[8] Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, et al. Deep-Learning-Based Drug-Target Interaction Prediction. J Proteome Res. 2017;16(4):1401-9.
[9] Ozturk H, Ozgur A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics. 2018;34(17):i821-i9.
[10] Skalic M, Jimenez J, Sabbadin D, De Fabritiis G. Shape-Based Generative Modeling for *de Novo* Drug Design. J Chem Inf Model. 2019;59(3):1205-14.
[11] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-80.
[12] Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. Bioinformatics. 2018;34(21):3666-74.
[13] Zheng L, Fan J, Mu Y. OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein-Ligand Binding Affinity Prediction. ACS Omega. 2019;4(14):15956-65.
[14] Li Y, Yang J. Structural and Sequence Similarity Makes a Significant Impact on Machine-Learning-Based Scoring Functions for Protein-Ligand Interactions. J Chem Inf Model. 2017;57(4):1007-12.
[15] Bai Q, Tan S, Xu T, Liu H, Huang J, Yao X. MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. Brief Bioinform. 2020.
[16] Bai Q. Research and development of MolAICal for drug design via deep learning and classical programming. arXiv preprint. 2020:arXiv:2006.09747.

[17] Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem. 2010;53(7):2719-40.
[18] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001;46(1-3):3-26.
[19] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455-61.
[20] Feng Z, Chen M, Liang T, Shen M, Chen H, Xie XQ. Virus-CKB: an integrated bioinformatics platform and analysis resource for COVID-19 research. Brief Bioinform. 2020.
[21] Weng G, Wang E, Wang Z, Liu H, Zhu F, Li D, et al. HawkDock: a web server to predict and analyze the protein-protein complex based on computational docking and MM/GBSA. Nucleic Acids Res. 2019;47(W1):W322-W30.
[22] Kong R, Yang G, Xue R, Liu M, Wang F, Hu J, et al. COVID-19 Docking Server: a meta server for docking small molecules, peptides and antibodies against potential targets of COVID-19. Bioinformatics. 2020;36(20):5109-11.
[23] Li Y, Su M, Liu Z, Li J, Liu J, Han L, et al. Assessing protein-ligand interaction scoring functions with the CASF-2013 benchmark. Nat Protoc. 2018;13(4):666-80.
[24] Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem. 2004;47(12):2977-80.
[25] Wang Z, Sun H, Yao X, Li D, Xu L, Li Y, et al. Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. Phys Chem Chem Phys. 2016;18(18):12964-75.
[26] Sterling T, Irwin JJ. ZINC 15--Ligand Discovery for Everyone. J Chem Inf Model. 2015;55(11):2324-37.
[27] Halgren TA, Nachbar RB. Merck molecular force field. IV. conformational energies and geometries for MMFF94. J Comput Chem. 1996;17(5-6):587-615.
[28] Hochreiter S, Schmidhuber JJNc. Long short-term memory. 1997;9(8):1735-80.
[29] Shultz MD. Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs. J Med Chem. 2019;62(4):1701-14.
[30] Bienfait B, Ertl P. JSME: a free molecule editor in JavaScript. J Cheminform. 2013;5:24.
[31] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminform. 2011;3:33.
[32] Khamis MA, Gomaa W. Comparative assessment of machine-learning scoring functions on PDBbind 2013. Engineering Applications of Artificial Intelligence. 2015;45:136-51.
[33] Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki PJB. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. 2018;34(21):3666-74.
[34] Velazhahan V, Ma N, Pandy-Szekeres G, Kooistra AJ, Lee Y, Gloriam DE, et al. Structure of the class D GPCR Ste2 dimer coupled to two G proteins. Nature. 2021;589(7840):148-53.
[35] Yesilaltay A, Jenness DD. Homo-oligomeric complexes of the yeast alpha-factor pheromone receptor are functional units of endocytosis. Mol Biol Cell. 2000;11(9):2873-84.

Figures legends

Figure 1. The workflow of two modules of WADDAICA.

Figure 2. Example input and output results from the pages of WADDAICA. (A) The job submission interface of “Drug design by AI”. (B) The job information. (C) The visualization of desgined ligand. The designed ligand in the protein pocket is shown by JSmol molecule viewer (http://www.jmol.org). (D) The result information of ID, name, cluster, binding affinities, formula, InChIKey, and synthetic accessibility of designed ligands.

Figure 3. The interfaces of WADDAICA. (A) The interface of “Drug design by CA”. (B) The interface of “Binding affinity by AI”. (C) The interface of “Molecular docking”. (D) The interface of “Binding affinity by CA”. (E) The interface of “Drug properties calculation”.

Figure 4. The case of drug design based on class D GPCR Ste2 by deep learning model and classical algorithm. (A) Binding affinity versus XLOGP for ligands generated by deep learning model. (B) Binding affinity versus XLOGP for ligands generated by the classical algorithm.

Tables legends

Table 1. Comparison of scoring power of machine learning models

Table 2. Comparison of scoring power of classical score functions

Table 1. Comparison of scoring power of machine learning models

Scoring function	SD	R_p
OnionNet [13]	1.45	0.78
*k*NN-Score [32]	1.65	0.672
RF-Score-v3 [33]	1.51	0.74
Pafnucy [33]	1.61	0.70

Table 2. Comparison of scoring power of classical score functions

Scoring function	R_p	R_s
Vinardo	0.582	0.592
AutoDock (LGA) [25]	0.404	0.450
AutoDock (PSO) [25]	0.466	0.513
AutoDock Vina [25]	0.569	0.584
LeDock [25]	0.463	0.486
rDock [25]	-0.021	-0.005
UCSF DOCK [25]	0.276	0.323

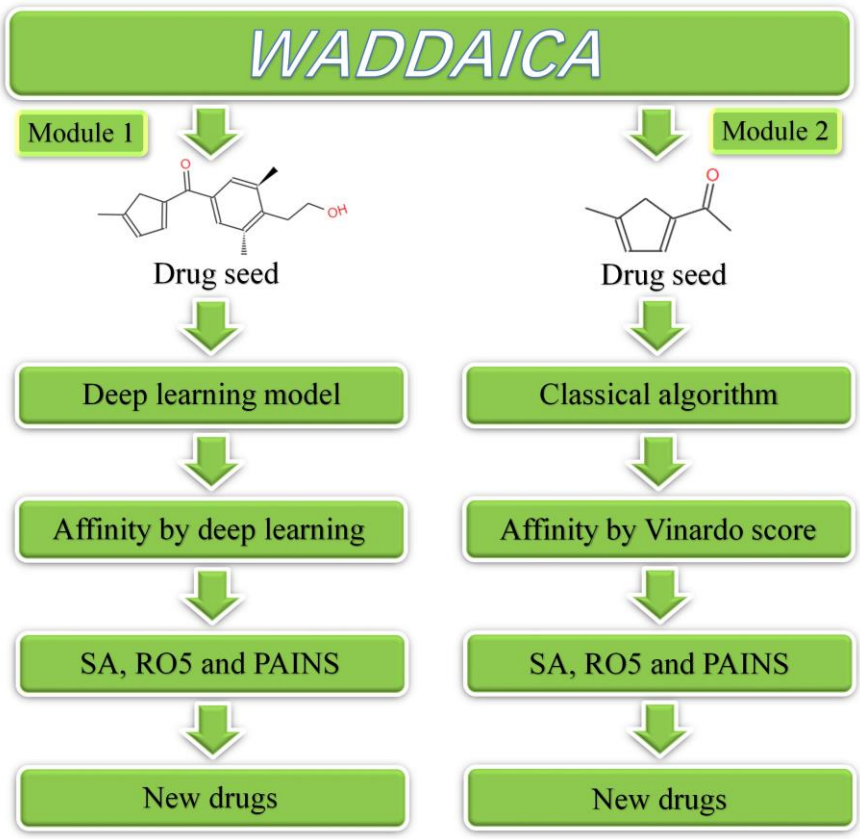
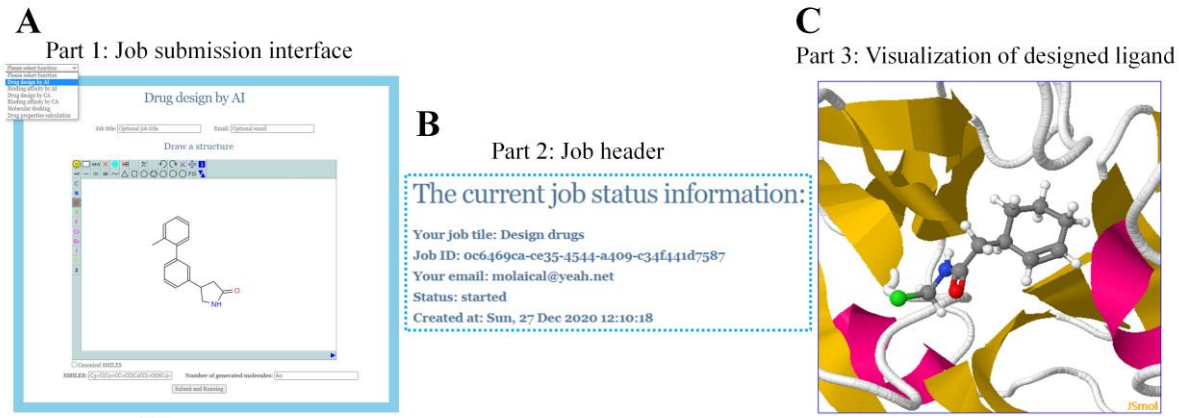

WADDAICA
Module 1
Module 2
Drug seed
Drug seed
Deep learning model
Classical algorithm
Affinity by deep learning
Affinity by Vinardo score
SA, RO5 and PAINS
SA, RO5 and PAINS
New drugs
New drugs

Figure 1



D Part 4: The result information of designed ligands

ID	Name	Cluster	Affinity(kcal/mol)	Formula	InChIKey	Synthetic_Accessibility
1	lig_1.mol2	[1]	-4.33	C12H14NOBr	AZLYWPHYKDFZEP-UHFFFAOYSA-N	89.21
2	lig_2.mol2	[2]	-4.25	C8H15NO3	HBEUBORJQWPWSV-UHFFFAOYSA-N	92.93
3	lig_3.mol2	[2]	-4.23	C10H18N3O3	QAQKNACALMUSEY-UHFFFAOYSA-O	88.06
4	lig_4.mol2	[2]	-4.22	C9H16N2O3	PTRHEBIEQFPOTP-UHFFFAOYSA-N	90.57
5	lig_5.mol2	[2]	-4.23	C11H18N2O2	MUCRREGYGMHREH-ZJUUUORDSA-N	87.75
6	lig_6.mol2	[2]	-4.20	C7H12N2O4	OUZZSTAXFUOJGY-UHFFFAOYSA-N	92.04
7	lig_7.mol2	[2]	-4.27	C11H14N4O3S	HCDPTRJIBOLPHF-UHFFFAOYSA-N	83.72
8	lig_8.mol2	[2]	-4.29	C13H15N3O2	ZFHNKSONFZKEFE-UHFFFAOYSA-N	77.70
9	lig_9.mol2	[2]	-4.32	C9H9N5O3S	FZNHFZAFSVYNDA-SCSAIBSYSA-N	77.06
10	lig_10.mol2	[2]	-4.26	C7H13NO3	IPSSKSAOBMREDR-UHFFFAOYSA-N	93.11

Figure 2

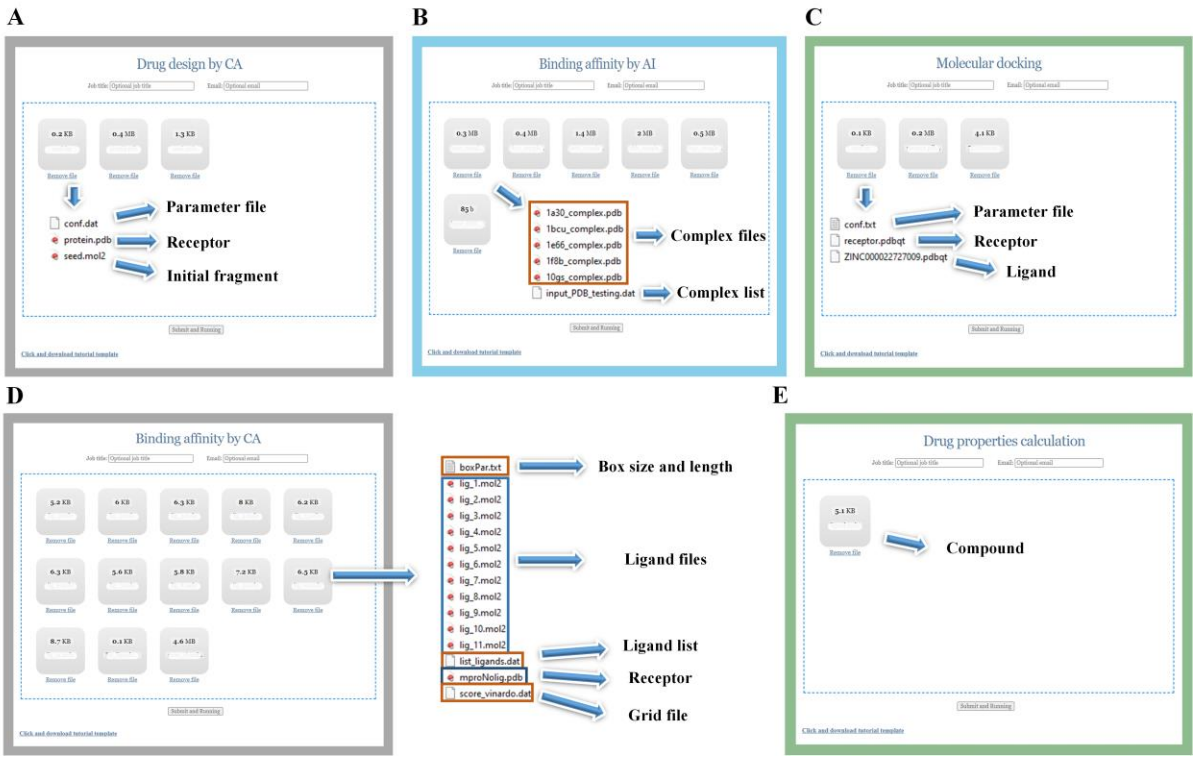

A
Drug design by CA
Parameter file
conf.dat
protein.pdb
Receptor
seed.mol2
Initial fragment
B
Binding affinity by AI
1a30_complex.pdb
1bcu_complex.pdb
1e66_complex.pdb
1f8b_complex.pdb
10gs_complex.pdb
Complex files
input_PDB_testing.dat
Complex list
C
Molecular docking
conf.txt
receptor.pdbqt
ZINC000022727009.pdbqt
Parameter file
Receptor
Ligand
D
Binding affinity by CA
boxPar.txt
Box size and length
lig_1.mol2
lig_2.mol2
lig_3.mol2
lig_4.mol2
lig_5.mol2
lig_6.mol2
lig_7.mol2
lig_8.mol2
lig_9.mol2
lig_10.mol2
lig_11.mol2
Ligand files
list_ligands.dat
Ligand list
mproNolig.pdb
Receptor
score_vinardo.dat
Grid file
E
Drug properties calculation
Compound

Figure 3

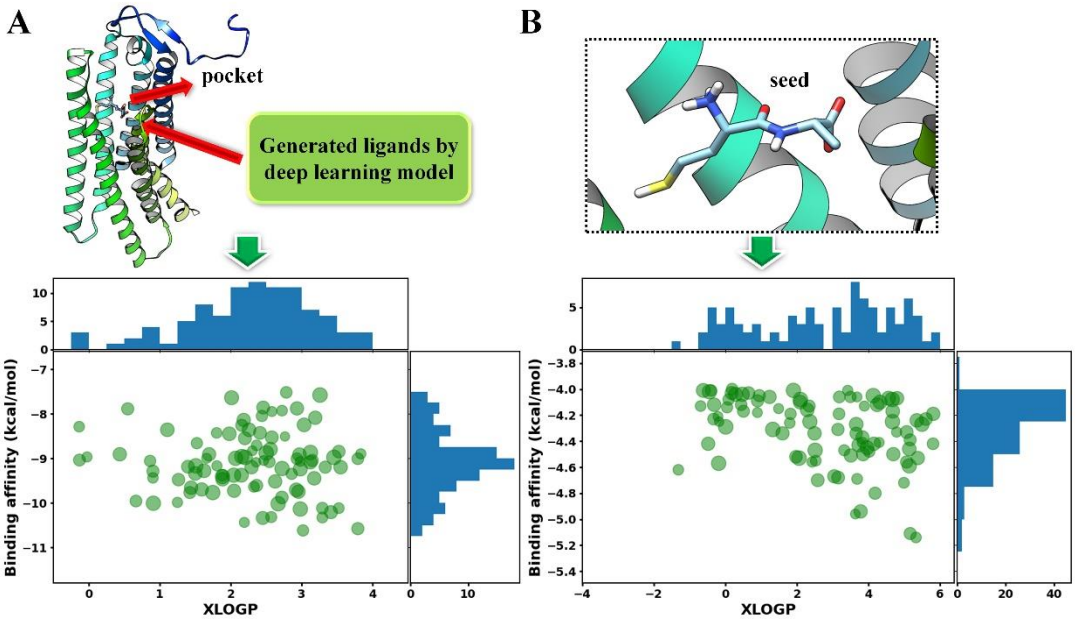

A
pocket
Generated ligands by deep learning model
B
seed
Binding affinity (kcal/mol)
XLOGP
Binding affinity (kcal/mol)
XLOGP

Figure 4