# Manual

## 1 Introduction

The case of ACM4 (Muscarinic acetylcholine receptor M4) agonists is presented here to show how to use MUBD-DecoyMaker 2.0 for the calculation. The users may have noted that GLL/GDD has included a ready-to-use benchmarking set for ACM4 agonists (http://cavasotto-lab.net/Databases/GDD/). For the case with available benchmarking data sets, the users are strongly suggested to check whether the data sets are biased or not, though it is not a must. The protocols below will guide the users to detect potential 2D bias in the benchmarking set for ACM4 agonists, build MUBD and validate its quality. Please note that the case here was run on a windows-based machine with Intel Core(TM) i7-7700 CPU@3.60GHz and RAM of 16 GB.

## 2 Instructions

### 2.1 Detect 2D Bias

(1) load the input: add the directory of two csv files that contain SMILES specifications of the ACM4 agonists and the ACM4 decoys (i.e. ACM4_Agonist_Ligands.csv and ACM4_Agonist_Decoys.csv) to the textbox. Here, the users are encouraged to make a working directory of Data_files in the D disk, i.e. D:\Data_files for all the input and output files.

(2) click the button ‘Detect’: The value of NLBScore that represents 2D bias appears in the text box. (Fig. 1)

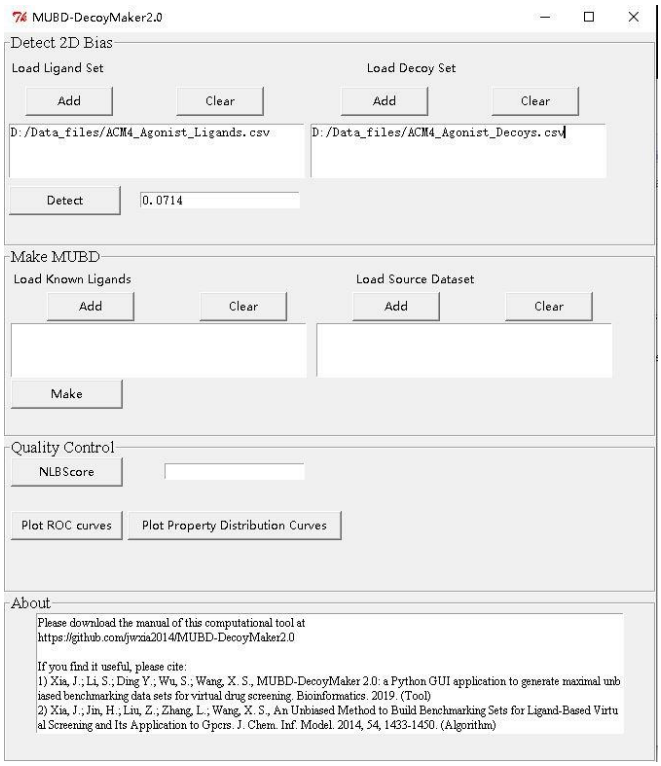


Fig. 1. The module of ‘Detect 2D Bias’

**2.2 Make MUBD**

(1) load the input: The users may directly use the ligands in the available benchmarking set (e.g. ACM4_Agonist_Ligands.csv) or provide a new ligand set with SMILES specifications. The source compound set bound to this computational tool is a prepared ZINC12 all-purchasable subset with the information of SMILES and six physicochemical properties (DataWithPs.csv, https://www.dropbox.com/sh/9c8ajswrbbppodo/AACgqVPEkNcEWgDxfE3E9eCva?dl=0). (Fig. 2A)

(2) click the button ‘Make’: The logs of the calculation are shown on a pop-up window. (Fig. 2B)

(3) Check output files: Once done, there will be two csv files that contain diverse ligands (Diverse_ligands_PS.csv) and maximal unbiased decoys

(Final_decoys.csv) and one folder named ‘Final decoys’ that includes the csv files that contain maximal unbiased decoys for every ligand (Final_decoys_*.csv). It should be noted that all compounds are represented by SMILES.

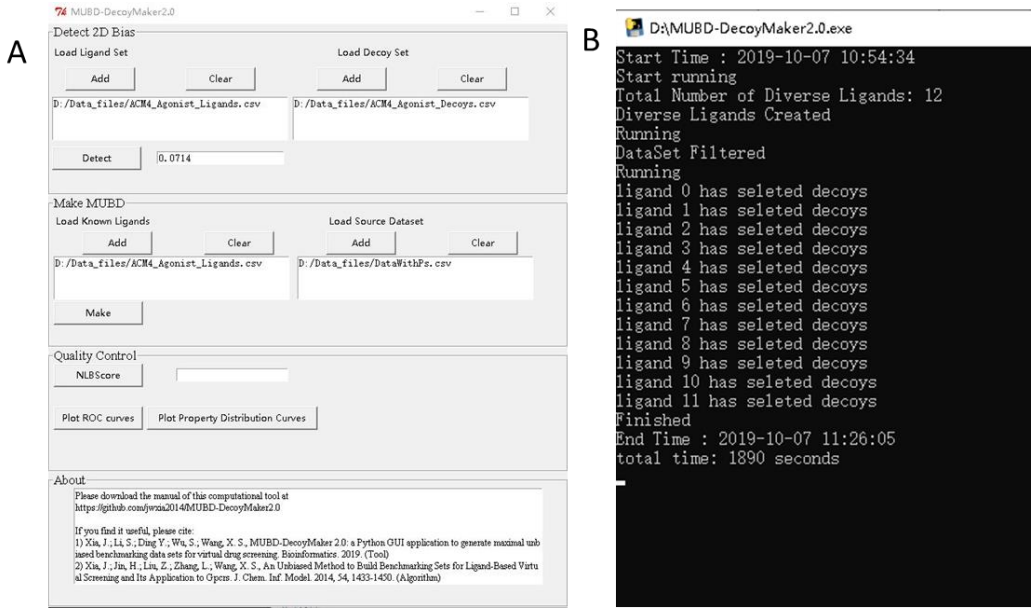


Fig. 2 (A)The module of ‘Make MUBD’ and (B) the pop-up window

**2.3 Quality Control**

(1) Click the button ‘NLBScore’: The value of NLBScore that represents 2D bias appears in the text box (Fig. 3A). You may note that NLBScore of the generated MUBD is less than the available GLL/GDD data set, i.e. 0.0455 vs. 0.0714, indicating the 2D bias has been reduced.

(2) Click the button ‘Plot ROC curves’: The figure that contains ROC curves based on leave-one-out cross-validation using simp-based and MACCS sims-based similarity search pops up and the mean(ROC AUCs) are also shown (Fig. 3B). The value of mean(ROC AUCs) close to 0.5 indicates of the unbiased feature of the MUBD.

(3) click the button ‘Plot Property Distribution Curves’: The figure that contains Property Distribution Curves for six physicochemical properties shows up. (Fig. 3C). The similar distribution of ligands and decoys is also an indicator of unbiased feature of the MUBD.

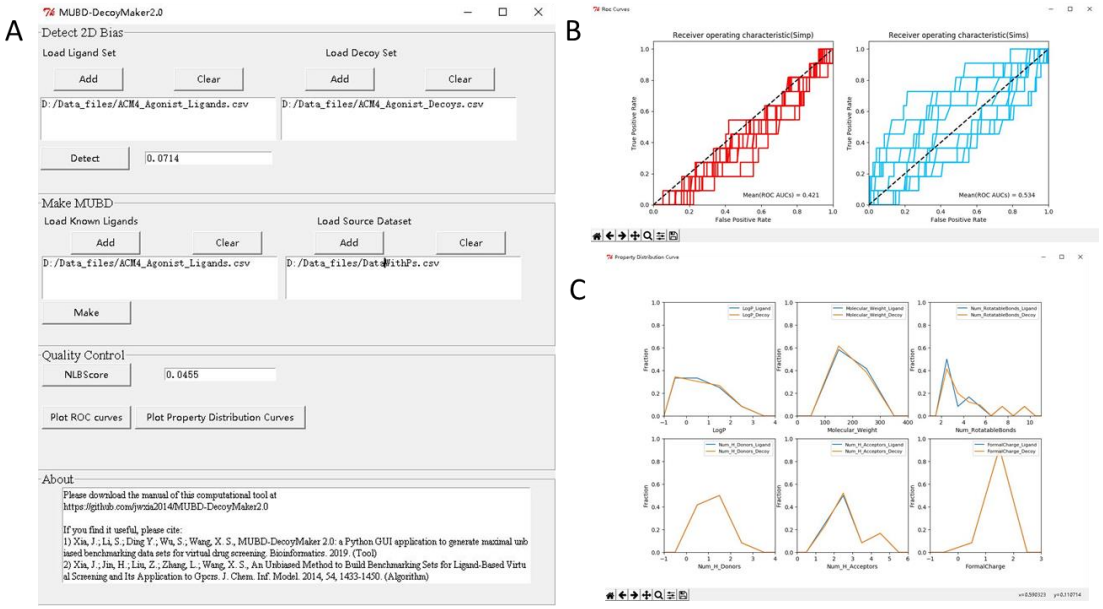


Fig. 3 The module of ‘Quality Control’ : (A) NLBScore, (B) ROC curves, (C) Property Distribution Curves

**<u>If you find it useful, please cite:</u>**


(1) Xia, J.; Li, S.; Ding Y.; Wu, S.; Wang, X. S., MUBD-DecoyMaker 2.0: a Python GUI application to generate maximal unbiased benchmarking data sets for virtual drug screening. Bioinformatics. 2019. (Tool)

(2) Xia, J.; Jin, H.; Liu, Z.; Zhang, L.; Wang, X. S., An Unbiased Method to Build Benchmarking Sets for Ligand-Based Virtual Screening and Its Application to Gpcrs. J. Chem. Inf. Model. 2014, 54, 1433-1450. (Algorithm)


Any question or feedback is welcome. Please send emails to jie.william.xia@hotmail.com or x.simon.wang@gmail.com.