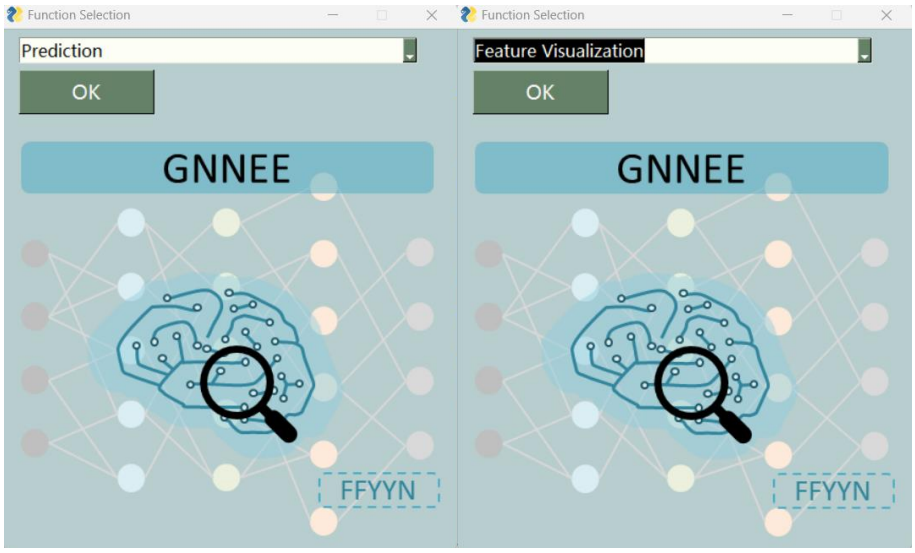
Documentation

1.Introduction

Environmental estrogens (EEs) have received extensive attention due to their frequent detection and biological activities. For chemical risk assessment purpose, applying deep learning (DL) model with defined applicability domain (AD) are important to accurate and reliable screening EEs. We developed a standalone software called GNNEEs (v1.0) via the python package auto-py-to-exe for prediction of potential EEs. The graph neural network (GNN) model and AD calculation were encapsulated in this software.

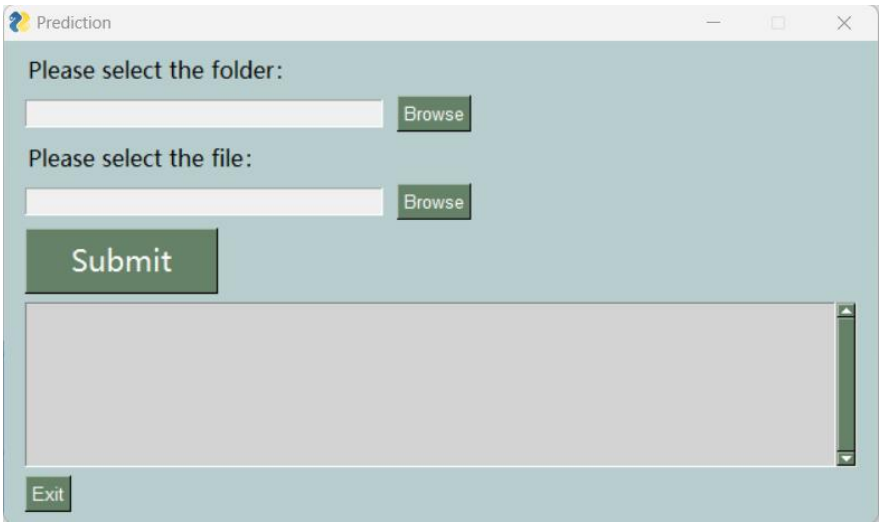
2.Guidance

Two functions are available in this software: reliable prediction and explainable prediction for a compound of interest. Users can select either function by clicking on it and then clicking the ‘OK’ button.

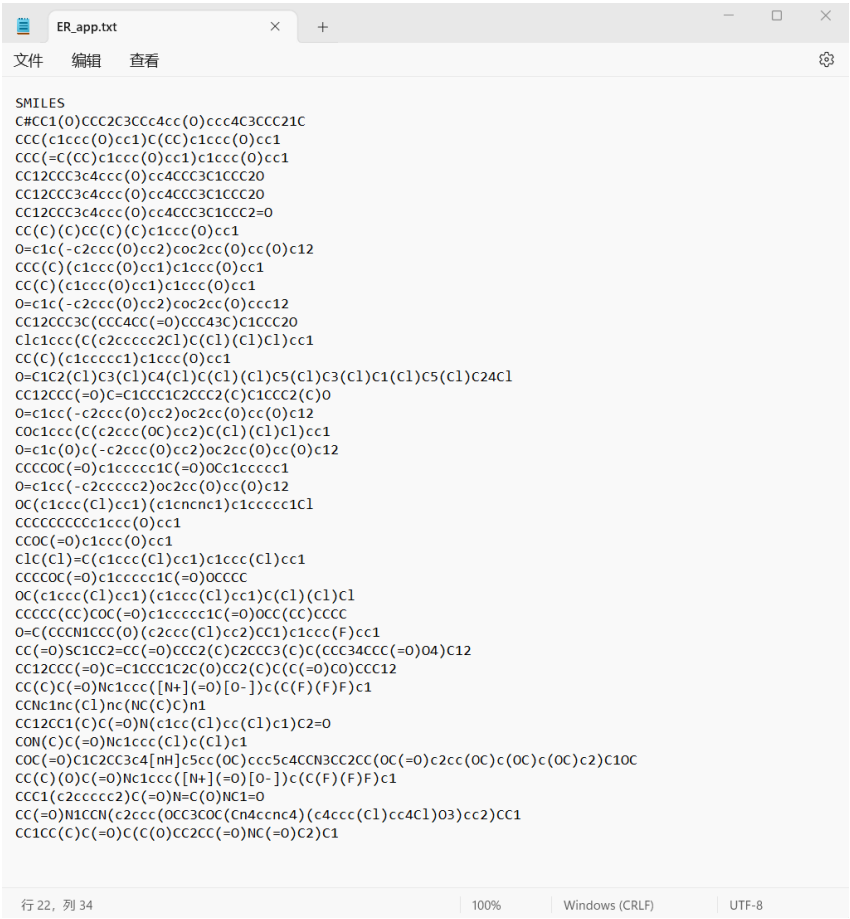


(1) Prediction

If the ‘prediction’ function is chosen, the users need to select the input path and file for the compounds of interest.



The file should be saved in TXT format, with a column titled ‘SMILES’ listing the SMILES of each compound.



After selecting the file, the user can click the ‘Submit’ button to obtain predicted results with AD indicators, which can help judge the reliability of the prediction. The predicted results will be saved to a CSV format file in the selected input path.

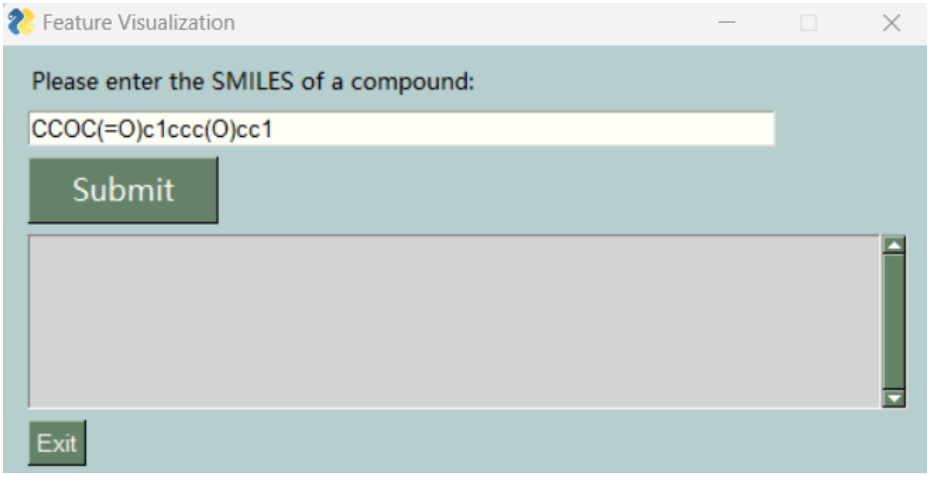
The output file named df_predict.csv

In the output file, the title of the table is the SMILES of each compound, predictive results of each compound, AD indicators. (DD represent AD_{FT} method, AD represent the AD_{FP} method)

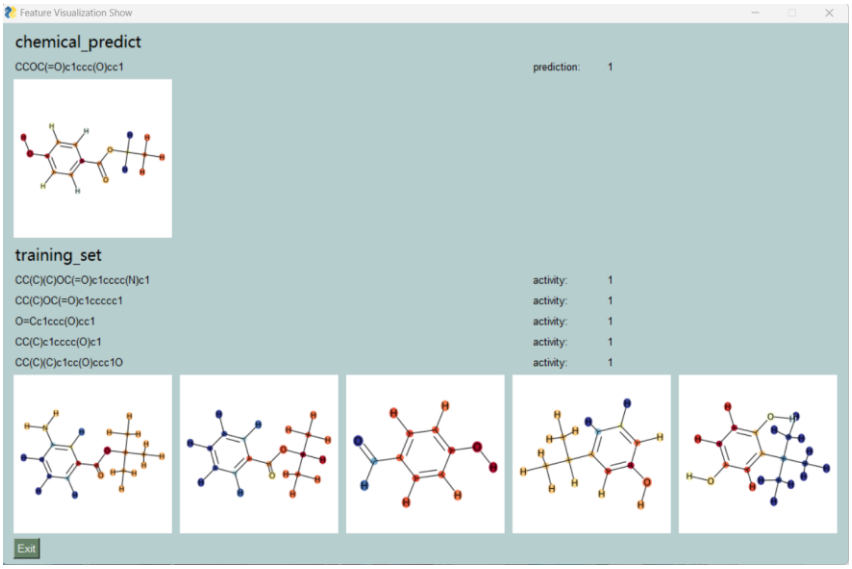
	SMILES	predictior	DD_N5	DD_N10	DD_N15	DD_N20	DD_N25	DD_N30	DD_N35	AD_0.65	AD_0.70	AD_0.75	AD_0.80	AD_0.85	AD_0.90	AD_0.95
0	C#CC1(O	1	3089.13	3457.36	3615.37	3386.42	3152.67	3041.68	2976.41	130	82	42	17	12	2	0
1	CCC(c1cc	1	3776.77	3294.94	3370.28	3570.07	3553.11	3614.01	3453.5	30	13	7	4	2	1	1
2	CCC(=C((	1	3265	3082.81	2888.65	3292.74	3257.35	3200.61	2971.63	18	9	6	4	3	1	1
3	CC12CCC	1	2000.63	2636.92	2760.45	2734.97	2743.69	2954.43	3178.31	135	96	56	19	13	7	1
4	CC12CCC	1	2000.63	2636.92	2760.45	2734.97	2743.69	2954.43	3178.31	135	96	56	19	13	7	1
5	CC12CCC	1	2126.71	2115.48	2760.29	2884.95	2798.82	2954.88	3112.1	114	54	30	22	8	2	0
6	CC(C)(C)(	1	2879.37	3715.2	3054.24	2760.63	3046.95	3104.95	3196.28	38	21	14	5	3	0	0
7	O=c1c(-c	1	1678.93	2635.65	2967.63	3215.49	3369.17	3362.75	3089.14	103	69	30	19	9	3	1
8	CCC(C)(c	1	3642.96	3582.07	2918.52	2804.07	2643.85	2649.37	2631.52	32	19	9	6	3	2	1
9	CC(C)(c1c	1	2529.32	2905.1	2993.3	3111.07	3160.14	3120.86	3145.02	47	21	12	6	4	1	0
10	O=c1c(-c	1	2456.05	2606.11	2992.32	3222.31	3317.5	3328.6	3151.25	105	66	36	17	9	2	0
11	CC12CCC	1	2651.7	2970.15	2978.71	2846.96	2869.41	3089.81	3057.29	160	127	93	42	16	6	3
12	Clc1ccc(C	1	3414.42	3912.81	3987.24	3692.47	3771.76	3417.02	3340.17	35	23	23	4	0	0	0
13	CC(C)(c1c	1	2814.55	2944.34	2862.38	2933.55	3047.28	2773.77	3006.27	35	16	6	2	0	0	0
14	O=C1C2(	0	3320.21	3437.58	3059.31	3244.55	3285.35	3290.13	3159.06	2	1	0	0	0	0	0
15	CC12CCC	1	2653.52	3473.06	3546.17	3133.94	3220.73	3324.27	3287.96	214	146	112	81	50	26	5
16	O=c1cc(-	1	2407.82	2994.08	2870.28	3151.51	3307.72	3458.39	3311.74	73	41	21	14	8	0	0
17	COc1ccc((	1	3097.72	3042.57	3262.94	3297.01	3277.68	3367.95	3256.33	8	5	0	0	0	0	0
18	O=c1c(O)	1	3615.15	2725.06	2564.25	2630.82	2386.05	2678.07	2894.61	101	58	29	13	8	5	3
19	CCCCOC(	0	3240.14	3499.17	3152.25	3169.29	3094.91	2970.73	3184.2	33	20	9	6	1	0	0
20	O=c1cc(-	1	2456.24	2321.49	2834.79	2967.99	2879.94	2898.32	3040.02	73	41	21	14	8	0	0
21	OC(c1ccc	1	3612.05	3782.23	4064.06	3949.05	3685.64	3509.46	3447.73	2	2	2	1	1	1	1
22	CCCCCC(	1	2214.18	3021.39	3236.06	3044.72	3170.27	3177.26	3266.67	24	16	7	4	4	3	2
23	CCOC(=C	1	3581.76	3423.37	3044.62	3116.07	3154.19	3244.92	3345.2	38	24	12	5	1	0	0
24	ClC(Cl)=C	0	3237.24	2728.09	2639.33	2361.56	2642.28	2601.74	2789.42	1	0	0	0	0	0	0
25	CCCCOC(	0	3709.66	2609.44	3006.58	3144.61	3259.69	3158.31	3135.08	48	32	20	16	9	0	0
26	OC(c1ccc	0	2971.6	2954.08	2854.12	2677.61	2571.8	2601.1	2683.3	4	1	1	1	1	0	0
27	CCCCC(C	0	3703.61	3305.3	3094.31	3301.36	3244.48	3117.95	3065.05	67	41	24	17	15	12	3
28	O=C(CCC	0	3843.12	3296.14	3006.22	3123.16	3178.47	3213.12	3281.62	21	9	6	3	2	2	2
29	CC(=O)S(	1	2956.83	2853.1	3119.93	3123.83	3400.55	2964.48	3082.67	56	28	8	3	2	1	1
30	CC12CCC	0	2800.41	2699.22	2818.67	2614.04	2852.31	2808.88	2696.36	194	146	63	31	17	6	0
31	CC(C)C(=	0	3180.59	3053.2	2662.45	2905.11	2931.74	2676.72	2586.89	30	9	2	0	0	0	0
32	CCNc1nc(	0	2994.61	2559.78	2299.78	2628.58	2968.48	2924.47	2769.34	23	12	6	3	3	2	0
33	CC12CC1	0	2699.6	2822.98	2339.26	2741.11	2896.28	3004.22	2730.41	3	0	0	0	0	0	0
34	CON(C)C	0	2509.53	2703.96	2762.51	2789.47	2976.96	2912.88	3105.94	4	3	3	3	3	2	1
35	COC(=O)	0	2827.81	2747.22	3127.9	2926.86	3092.33	3056.52	3191.98	96	36	12	5	4	4	3
36	CC(C)(O)(	0	3259.31	3228.4	2709.67	2801.72	3016.32	2783	2927.17	10	0	0	0	0	0	0
37	CCC1(c2c	0	3225.9	3098.38	2645.37	3062.25	2810.14	2965.88	3007.43	8	5	2	0	0	0	0
38	CC(=O)N	0	2491.17	3043.49	3471.64	3667.5	3360.23	3205.69	3224.4	99	30	3	2	0	0	0
39	CC1CC(C	0	2459.14	3161.24	3298.89	3234.97	3067.85	3444.73	3368.51	1	0	0	0	0	0	0

(2) Feature Visualization

If the ‘feature visualization’ function is selected, the users need to enter the SMILES of a compound and click the ‘Submit’ button.



This will generate a feature visualization of the input compound, highlighting critical substructures related to estrogenic agonist activity. Additionally, the feature visualization plots of similar compounds in the training set will be displayed simultaneously in the window, helping the user to better understand the model's prediction.



3.Developer Information

Maintainer: Dr. Fan, Professor Geng

Contact US : fanfan20201031@163.com

State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing 210023, Jiangsu, PR China