Supporting Information for

Absolute binding free energy calculation and design of a subnanomolar inhibitor of phosphodiesterase-10

Zhe Li[a, b, #], Yiyou Huang[a, #], Yinuo Wu[a], Jingyi Chen[a], Deyan Wu[a], Chang-Guo Zhan[b,*], and Hai-Bin Luo[a,*]

[a] School of Pharmaceutical Sciences, Sun Yat-Sen University, Guangzhou 510006, P.R. China

[b] Department of Pharmaceutical Sciences, College of Pharmacy, University of Kentucky, 789 South Limestone Street, Lexington, KY, 40536

Contents:

Section S1. PDE4, PDE5, and PDE9 inhibitors used in this study

PDE4 inhibitors: There are 20 PDE4 inhibitors examined in this study. The structures of the first 8 molecules share certain similarity, whereas the rest structures are quite different from each other. Structures of all the PDE4 inhibitors are given in Fig. S1.

PDE5 inhibitors: 11 PDE5 inhibitors were used in this study. Their structures are given in Fig. S2.

PDE9 inhibitors: Seven PDE9 inhibitors were used in this study. Their structures are given in Fig. S3.

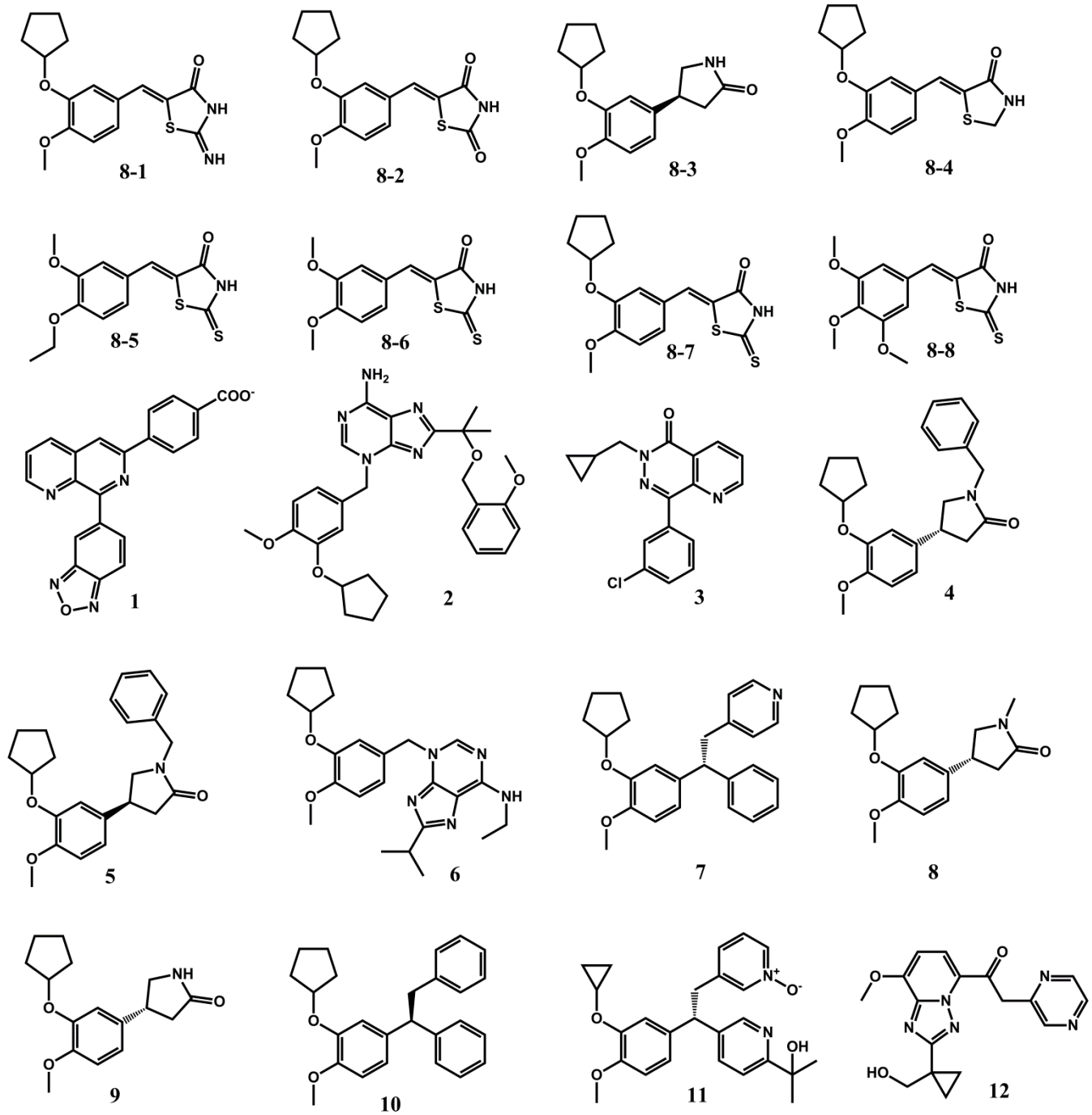


Fig. S1. Structures of PDE4 inhibitors used in this study.

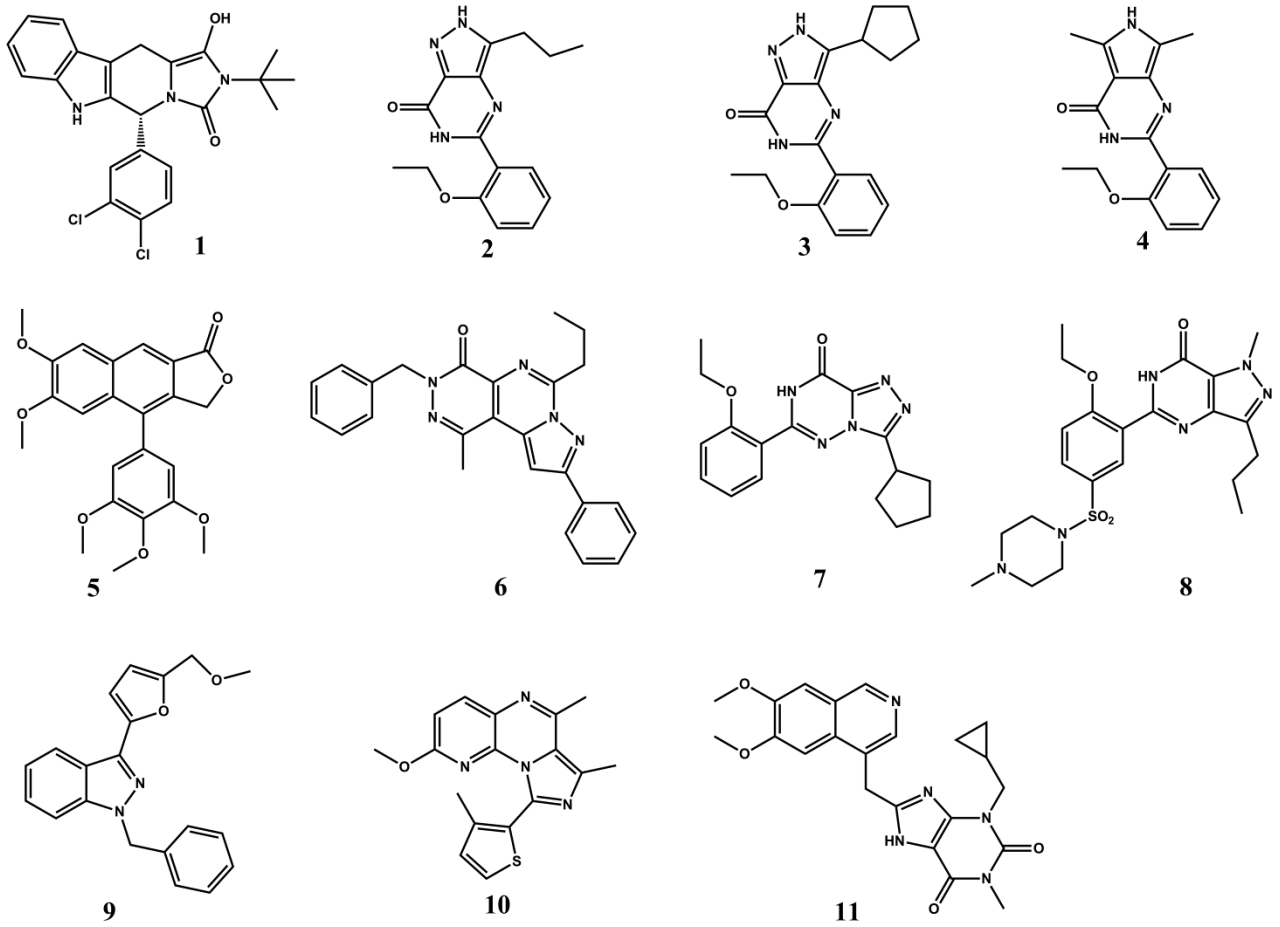


Fig. S2. Structures of PDE5 inhibitors used in this study.

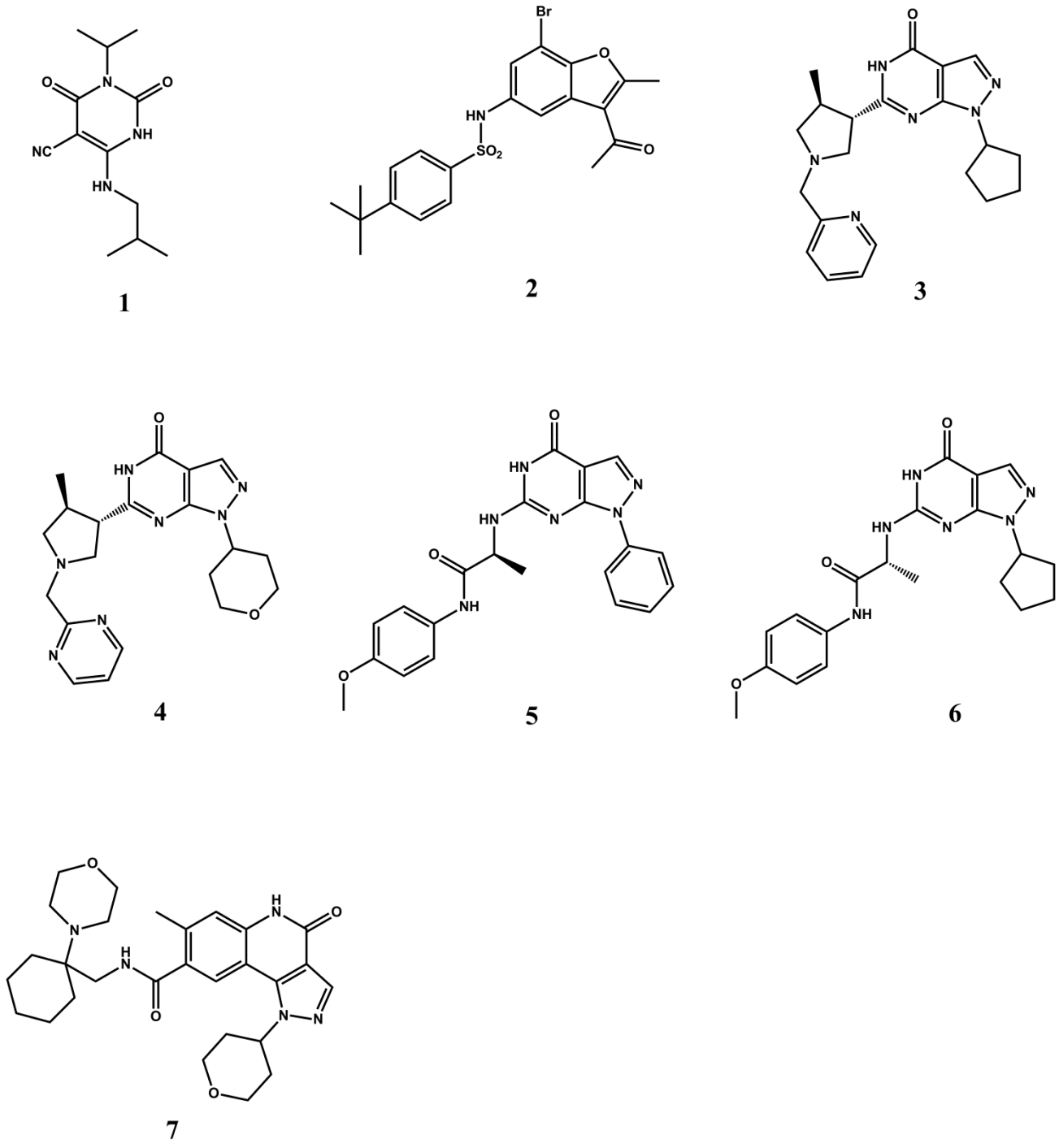


Fig. S3. Structures of PDE9 inhibitors used in this study.

Section S2. An example to show how the GA-FEP calculations work

By using lig_16 of CDK2 as an example, the fitted probability distribution $P(\Delta U)$ is shown as Fig. S4. Picture (a) represents the complex system (Rec-Lig), and picture (b) represents the Lig system. The probability distributions of both forward ΔU and backward ΔU are included, and shown in the same color for the same alchemical state. The 10 probability distributions depicted by black, red, green, blue and cyan in Fig. S4 correspond to the first 5 states that annihilated the electrostatic interactions. The rest probability distributions in the center of the figure correspond to the following 5 states that deal with annihilating vdW interactions. To improve calculation accuracy, for each state, the forward and backward perturbation energies (calculated by equation (2)), and BAR(2, 3) energies that combines the forward and backward results (calculated by equation (6) and (7)) were calculated based on the fitted probability distribution. For each state, if the energies of the forward and backward perturbations differ too much, the final result would be unreliable, and more simulations should be considered. Table S1 listed the calculation result of lig_16 of CDK2. In this table, c_fwd, c_bwd, l_fwd, l_bwd represent the forward and backward energy results calculated by basic FEP equation (2), c_dif and l_dif represents the energy difference between forward energy and backward energy, c_bar and l_bar represents the energy results calculated by BAR, and ene represents the final energy calculation result. As seen from the data in the table, for all the states, the forward and backward energies were similar to each other. The energy differece between the final state and the state prior to it was calculated by using the basic FEP equation instead of BAR due to the backward energy calculation from the final target state will face with 'end-point catastrophes'(4). The c_tot and l_tot are the summations of the energetic results using BAR method (except for the final state which was calculated by using the basic FEP equation), and the final binding free energy was the difference between l_tot and c_tot. The sandard deviation of the electrostatic energy was evaluated based on the forward, backward and BAR energies. The sandard deviation of the total interaction energies was evaluated based on the forward and BAR energies, because the backward energy may be calculated for the final state.

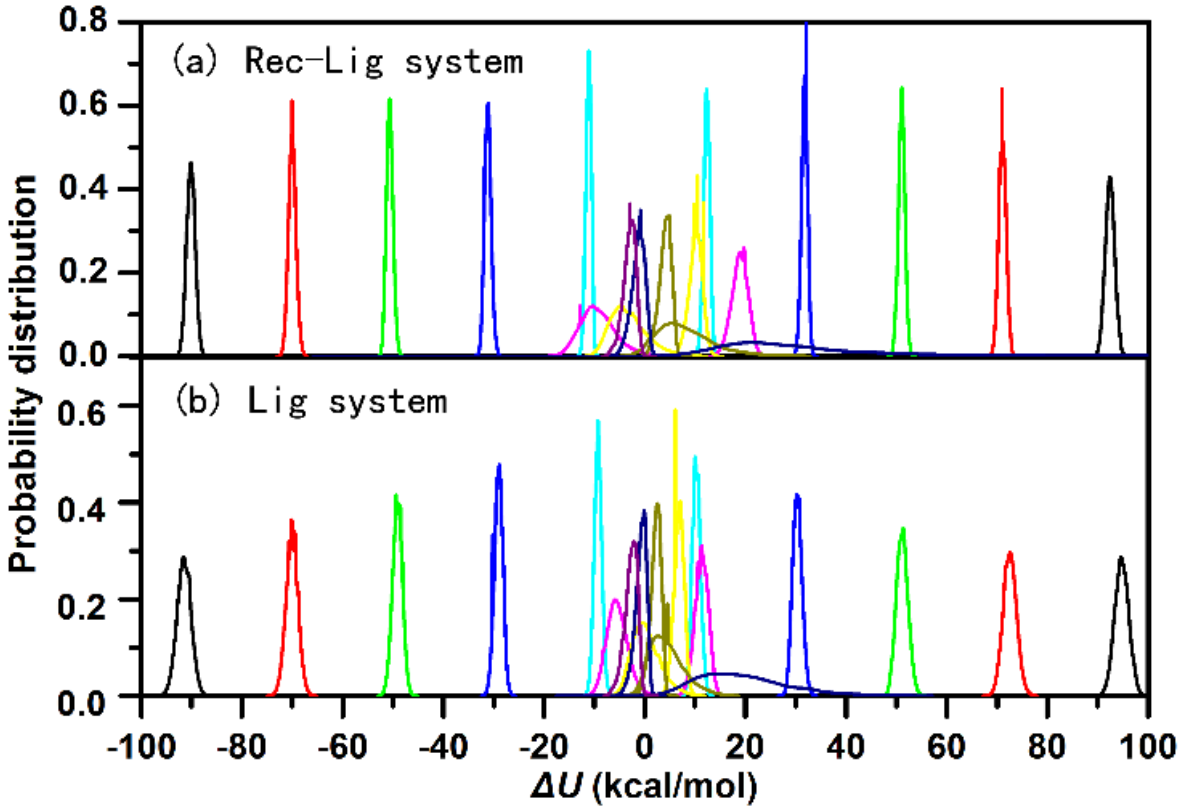


Fig. S4. The fitted P(Δ*U*) of all the states of lig_16 of CDK2. The P(Δ*U*) for both the forward and backward calculations are included in this figure, and they are almost symmetric to each other with respect to the origin.

Table S1. Calculation result of lig_16 of CDK2. In this table, c_fwd, c_bwd, l_fwd, l_bwd represent the forward and backward energy results calculated by basic FEP equation, c_dif and l_dif represent the energy difference between the forward and backward energies, c_bar and l_bar refer to the energy results calculated by BAR, and ene represents the final energy calculation result. c_tot and l_tot refer to the summations of the energy results using BAR method (except for the final state which was calculated by using basic FEP equation), and the final binding free energy was the difference between l_tot and c_tot.

state	c_fwd[1]	c_bwd	c_bar	c_dif	l_fwd[2]	l_bwd	l_bar	l_dif	c_tot	l_tot	ene
00_reference	91.608	-90.593	91.14	1.015	93.106	-93.139	93.122	-0.033	91.14	93.122	1.982
01_charge_0.2	70.56	-70.351	70.454	0.209	71.138	-71.308	71.262	-0.17	161.594	164.384	2.79
02_charge_0.4	50.662	-50.883	50.787	-0.221	50.023	-49.95	50.084	0.073	212.381	214.468	2.087
03_charge_0.6	31.443	-31.434	31.429	0.009	29.6	-29.632	29.642	-0.032	243.81	244.11	0.3
04_charge_0.8	11.992	-11.352	11.688	0.64	9.772	-9.77	9.768	0.002	255.498	253.878	-1.62
05_charge_1.0	16.762	-16.75	15.359	0.012	9.688	-8.182	8.789	1.506	270.857	262.667	-8.19
06_vdw_0.2	8.361	-8.719	7.769	-0.358	5.845	-3.807	4.386	2.038	278.626	267.053	-11.573
07_vdw_0.4	2.624	-1.597	0.936	1.027	1.024	-0.782	0.122	0.242	279.562	267.175	-12.387
08_vdw_0.6	-3.193	0.751	-5.566	-2.442	-2.373	2.019	-4.361	-0.354	273.996	262.814	-11.182
09_vdw_0.8	-4.721	---	---	-4.721	-4.493	---	---	---	269.275	258.321	-10.954

[1]: c in c_fwd, c_bwd, c_bar and c_tot represents the Rec-Lig complex

[2]: l in l_fwd, l_bwd, l_bar and l_tot represents the Lig system

Section S3. Using 10 lambda windows could reach to similar accuracy with that of 20 or 39 lambda windows.

The overall alchemical transformation of either Rec-Lig system or Lig system contains 10 alchemical states (lambda windows), in which the first 5 states were used to decouple the electrostatic interactions and the rest were used to decouple the vdW interactions. To show the effect of increasing the number of intermediate states to the energy calculation results, we calculated ABFE for 3 CDK2 inhibitors, including **30**, **28** and **1oiy**, with doubled and quadrupled number of lambda windows. Table S2 shows the calculation details of ligand **30**, and calculation details of all other ligands could be found in Supporting Dataset S1. As could be seen from this table, the convergence of the FEP-ABFE calculations were pretty good. When lambda of charge interaction reached to 1.0 and lambda of vdW interaction reached to 0.999936, the energies for (Rec-Lig system, Lig system) calculated with 39, 20 and 10 lambda values were (320.002, 305.880) kcal/mol, (319.376, 305.848) kcal/mol and (320.444, 306.577) kcal/mol, respectively, indicating the reliability of 10-lambda-windows FEP calculation. The energy for the last step of 10-lambda-windows calculation (lambda of vdW interaction from 0.999936 to 1.0) was calculated based on forward perturbation instead of BAR, and the calculated FEP-ABFE result (-14.119 kcal/mol) was similar to that of 20-lambda-windows calculation (-13.988 kcal/mol) and 39-lambda-windows calculation (-14.234 kcal/mol). Calculation results for **28** and **1oiy** (see Supporting Dataset S1 for details) also showed the good convergence and reliability of the 10-lambda-windows FEP calculation.

We further increased the simulation time from 4 ns to 20 ns for each lambda window for these three ligands, and the sampled probability distributions of the last 18 ns were used for energy calculation, and the ABFE results wasn't affected too much (Supporting Dataset S1).

Table S2. Calculation details of CDK2 inhibitor 30 using different number of λ values. com_ene represents the energy of Rec-Lig system; lig_ene represents the energy of Lig system; FEP_ene represents the FEP energy calculated by equation (1). In the first several steps, charge interactions were decoupled with λ_chg gradually increased to 1.00, while the vdW interactions were gradually decoupled in the following steps. Part of the table was omitted, and the full table could be found in Supporting Dataset S1.

λ values		39 λ			20 λ			10 λ		
λ_chg	λ_vdw	com_ene	lig_ene	FEP_ene	com_ene	lig_ene	FEP_ene	com_ene	lig_ene	FEP_ene
0.05	0.000000	28.791	29.525	0.734						
0.10	0.000000	56.044	57.329	1.285	56.001	57.294	1.293			
0.15	0.000000	81.770	83.523	1.753						
0.20	0.000000	105.936	108.121	2.185	105.936	108.093	2.157	105.992	109.034	3.042
…	…	…	…	…	…	…	…	…	…	…
1.00	0.000000	294.846	293.617	-1.229	294.988	293.590	-1.398	295.142	294.826	-0.316
1.00	0.264908	301.030	296.720	-4.310						
1.00	0.468559	306.490	299.561	-6.929	306.217	299.525	-6.692			
1.00	0.622850	311.184	302.071	-9.113						
1.00	0.737856	315.151	304.124	-11.027	314.820	303.962	-10.858	315.194	305.726	-9.468
1.00	0.822021	318.501	305.902	-12.599						
1.00	0.882351	321.411	307.568	-13.843	320.790	307.188	-13.602			
1.00	0.924581	323.666	308.980	-14.686						
1.00	0.953344	325.305	309.982	-15.323	324.558	309.309	-15.249	325.283	311.295	-13.988
1.00	0.972319	326.316	310.592	-15.724						
1.00	0.984375	326.908	310.863	-16.045	326.328	310.396	-15.932			
1.00	0.991696	327.036	310.759	-16.277						
1.00	0.995904	326.775	310.452	-16.323	325.972	310.460	-15.512	326.552	311.584	-14.968
1.00	0.998162	326.100	309.856	-16.244						
1.00	0.999271	324.825	308.937	-15.888	323.653	308.961	-14.692			
1.00	0.999756	322.830	307.680	-15.150						
1.00	0.999936	320.002	305.880	-14.122	319.376	305.848	-13.528	320.444	306.577	-13.867
1.00	0.999989	316.871	303.339	-13.532						
1.00	0.999999	313.305	299.746	-13.559	313.125	299.812	-13.313			
1.00	1.000000	309.043	294.809	-14.234	308.863	294.875	-13.988	315.126	301.007	-14.119

Section S4. Comparison between ABFE and RBFE for the same data set

To show the accuracy of the ABFE calculation method and compare it with the reported RBFE method, the calculation result of each target is given here separately, as shown in Fig. S5. On the left side are the results calculated by our ABFE method, and on the right side are the results calculated by RBFE method reported by Abel and Wang (1). The pairwise comparison between the two methods shows that they have comparable accuracy. For some targets, such as CDK2, the ABFE method performed even better than the RBFE method.

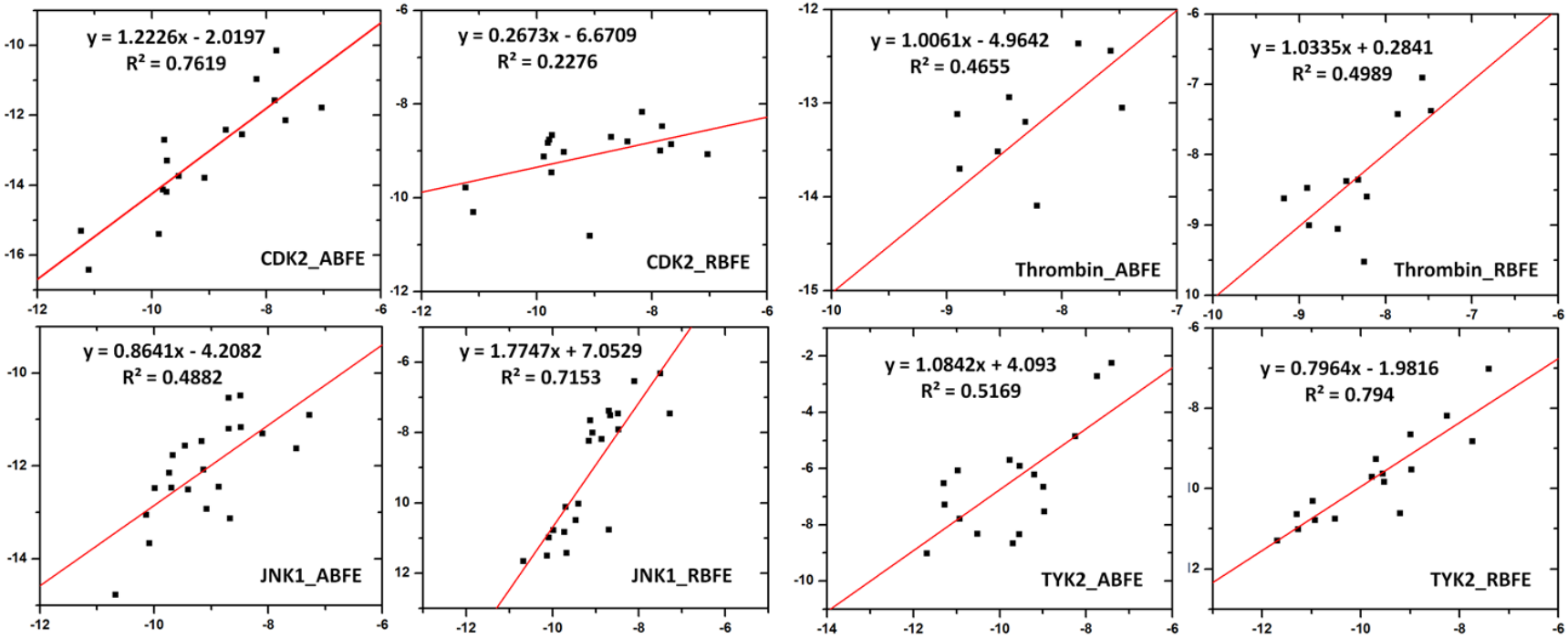


Fig. S5. Comparison between the ABFE and RBFE data for the same data set.

Section S5. Residual distribution of the GA-FEP/ABFE method.

to find out the distribution property of the calculated results of our GA-FEP method, we tried to put all the residual distributions of the 100+ results in Figure 5 together, and the distribution is just like the histogram in Fig. S6 below. We have also added a Gaussian distribution with the same deviation (SD = 1.03) and AUC on this figure, and we found that the distribution of the calculated result is just very much like a Gaussian distribution.

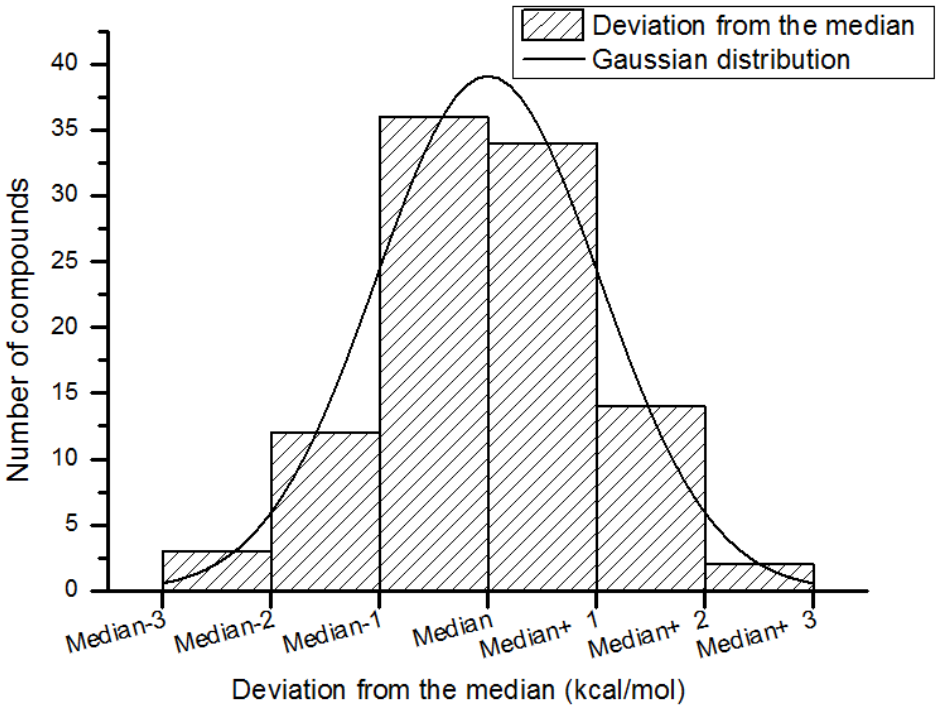


Fig. S6. Residual distributions of all the > 100 ABFE results.

Section S6. Designed PDE10 inhibitors. Molecular structures of designed PDE10 inhibitors and their inhibitory activity (IC_{50}) are given in Table S3. The FEP calculated ABFE value are given in Table S4.

Table S3. Molecular structures of designed PDE10 inhibitors and their inhibitory activity (IC_{50})

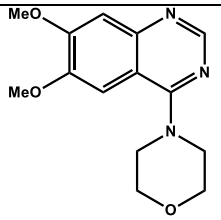
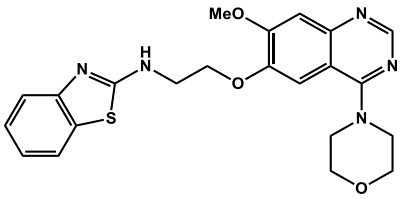
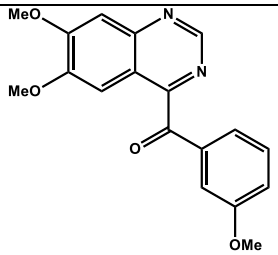
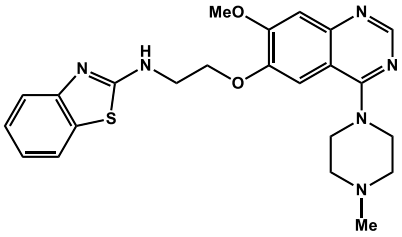
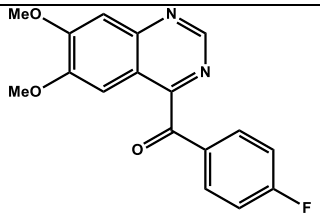
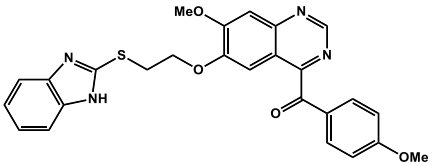
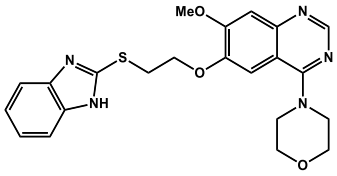
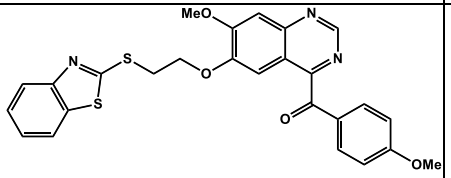
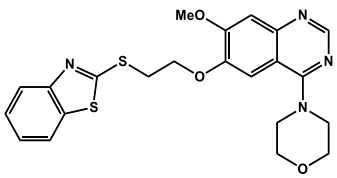
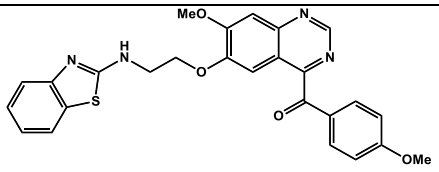
Ligand	Structure	IC_{50} (nM)	Ligand	Structure	IC_{50} (nM)
LHB-1		**1800**	**LHB-6**		**1.7**
LHB-2		**890**	**LHB-7**		**1.1**
LHB-3		**403**	**LHB-8**		**2.0**
LHB-4		**5.9**	**LHB-9**		**21**
LHB-5		**73**	**LHB-10**		**0.87**

Table S4. Calculation and experimental results of designed PDE10 inhibitors. exp stands for Experimental results, and FEP-cal stands for the energy calculated by FEP. Correlation constant R between calculation and experimental results is 0.86.

ligand_ID	IC50(nM)	exp (kcal/mol)	FEP-cal (kcal/mol)	ene_corrected_by_linear_regression (kcal/mol)[1]
LHB-1	1800	-7.855	-9.208	-6.37
LHB-2	890	-8.273	-12.121	-8.39
LHB-3	403	-8.744	-12.603	-8.73
LHB-4	5.9	-11.252	-17.07	-11.83
LHB-5	73	-9.758	-16.209	-11.23
LHB-6	1.7	-11.991	-16.779	-11.63
LHB-7	1.1	-12.250	-17.648	-12.24
LHB-8	2	-11.895	-18.039	-12.51
LHB-9	21	-10.498	-16.394	-11.36
LHB-10	0.87	-12.389	-15.328	-10.62

[1]: The linear regression function between exp(x) and FEP-cal(y) is y =1.4379x – 0.0545. ene_corrected_by_linear_regression is the corresponding predicted x value for each ligand based on the linear regression function.

Section S7. Crystal structure of PDE10 with LHB-6 was determined to verify the predicted binding mode. The coordinate and structure factors have been deposited in the Protein Data Bank with PDB ID **5ZNL**. The diffraction data is given in Table S5.

Table S5. Diffraction data and structure refinement statistic for PDE10A-LHB-6 structure

Data collection	PDE10A-**LHB-6**
Wavelength (Å)	1.5418
Temperature (K)	100
Resolution (Å)	24.11-2.80
Space group	$P2_12_12_1$
Unit Cell	
a, b, c (Å)	49.151,81.328,158.207
α, β, γ (°)	90.00, 90.00, 90.00
No. reflections	73182(15928)
Completeness (%)	97.86(97.69)
R_{merge}	0.107(0.274)
$\langle I/\sigma(I) \rangle$	11.7(3.7)
Redundancy	4.6(3.34)
Structure refinement	
R-factor/R-free	0.23/0.30
RMS deviations	
Bond lengths	0.0081 Å
Bond angles	1.1754
Average B-factor (Å^2)(atoms)	
Protein	34.8(5003)
Inhibitor	30.6(62)
Zn	33.6(2)
Mg	17.6(2)
Waters	16.8(80)
Ramachandran plot	
Preferred	95.70%
Allowed	4.14%

Section S8. Details about fitting probability distribution. In the post-simulation processing steps, for each FEP state, the last 2 ns were saved into production MD trajectory files with an interval of 100 fs, which resulted in 20,000 snapshots. As shown in Fig. S7, all the 20,000 data points of ΔU was transformed to its distribution $P(\Delta U)_{data}$ with 150 bins. As a result, we get 150 data points of $P(\Delta U)_{data}$. Since there are 15 parameters to fit, the number of data points is 10 times of that of the parameters, and thus it's unlikely to be overfitted.

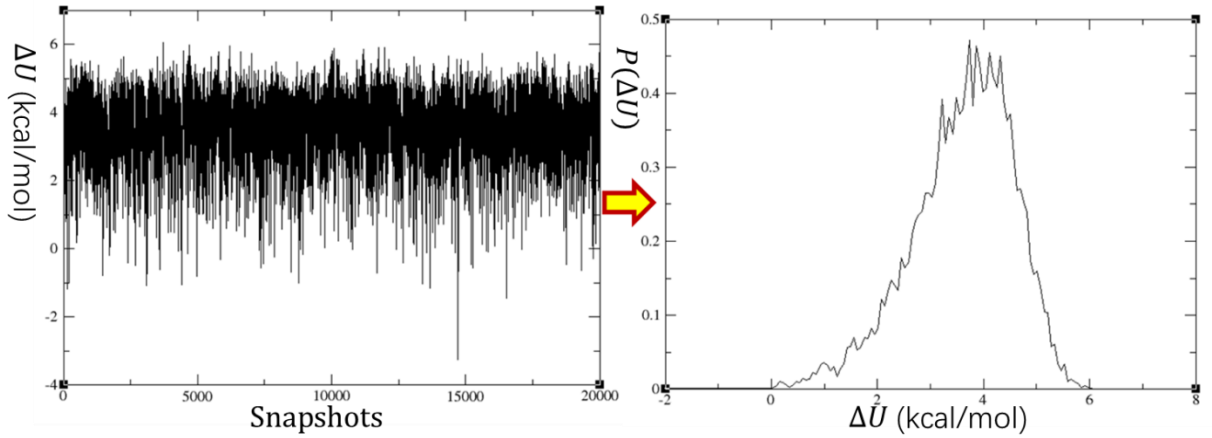


Fig. S7. ΔU data points transformed to the distribution $P(\Delta U)_{data}$

To fit these parameters, we used the following least square procedures:

1. Since the function to be fitted is in the following form:

$$P(\Delta U)_{func} = \sum_{i=1}^{5} c_i \exp\left(-\frac{(\Delta U - \mu_i)^2}{2{\sigma_i}^2}\right)$$

We designed a cost function in the form of

$$Cost\ (c_i, \mu_i, \sigma_i, \Delta U, P(\Delta U)_{data}) = \sum_{over\ all\ data\ points} \left(P(\Delta U)_{data} - P(\Delta U)_{func}\right)^2$$

$$= \sum_{over\ all\ data\ points} \left(P(\Delta U)_{data} - \sum_{i=1}^{5} c_i \exp\left(-\frac{(\Delta U - \mu_i)^2}{2{\sigma_i}^2}\right)\right)^2,$$

where $P(\Delta U)_{data}$ is the distribution data of the original FEP simulation.

2. We minimized the value of the cost function by using Stochastic Steepest Descent (SGD) method and get the optimized parameters for the fitted function. Of course, other

minimization method can also be applied, which include Mini-batch Steepest Descent, Stochastic Steepest Descent with momentum, Nesterov Accelerated Gradient, Adagrad, Adadelta, RMSprop, etc. However, SGD method just worked well in this situation.

SGD method optimize the 15 parameters c_i, μ_i, σ_i where i = 1, 2, 3, 4, 5 by the following procedure:

$$(c_i, \mu_i, \sigma_i)_t = (c_i, \mu_i, \sigma_i)_{t-1} - \lambda \nabla_{c_i,\mu_i,\sigma_i} Cost\,(c_i, \mu_i, \sigma_i, \Delta U, P(\Delta U)_{data})$$

Loop until the parameters in the tth cycle $(c_i, \mu_i, \sigma_i)_t$ are very close to $(c_i, \mu_i, \sigma_i)_{t-1}$, and the resulted $(c_i, \mu_i, \sigma_i)_t$ are the optimized parameters for the fitted function.

The fitted probability distribution will be like Fig. S8:

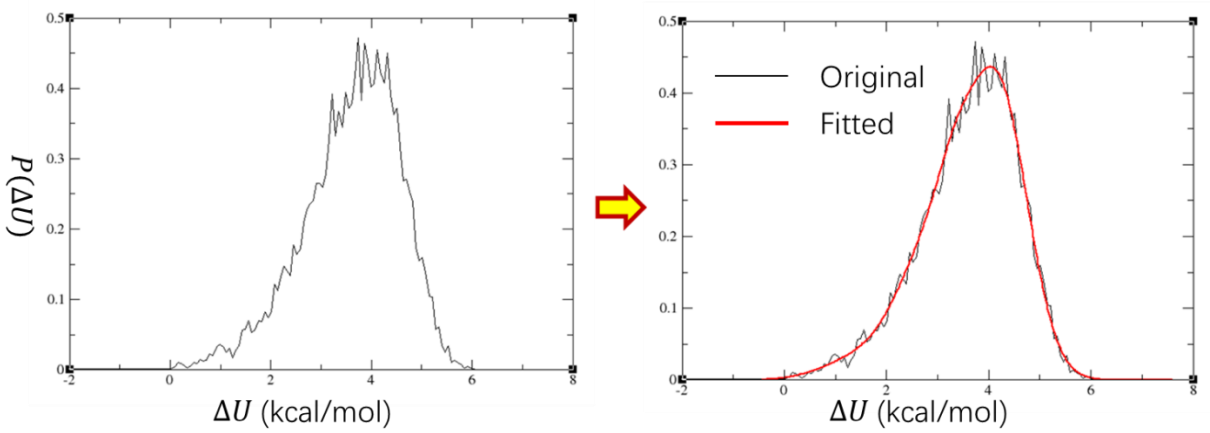


Fig. S8. Fitting the probability distribution.

Note: The probability distribution used as the example is a little bit "skewed" because this is a perturbation of ligand system from (λ_chg = 1.00, λ_vdw = 0.953344) to (λ_chg = 1.00, λ_vdw = 0.995904). The ligand is nearly totally annihilated, and the surrounding water molecules can get closer to the atoms of the ligand. The unphysical state will skew the resulted probability distribution. We are using this example here just want to show that 5 Gaussians can work well even for skewed distributions near the end of the perturbation.

Here are some other examples of fitted probability distributions in Fig. S9:

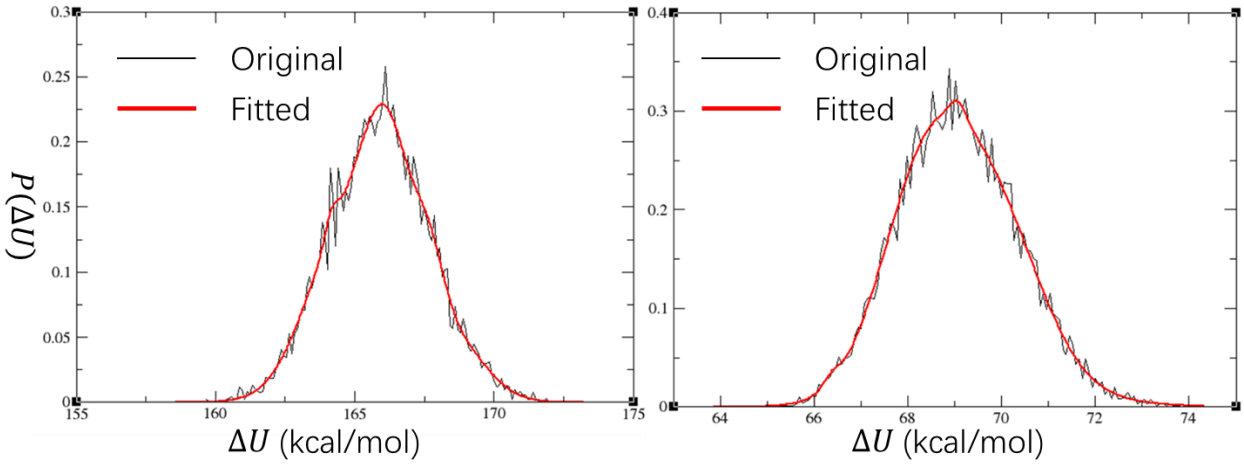


Fig. S9. Examples of some fitted probability distributions

3. Based on the fitted function, we generate 500,000 to 1,000,000 new data points. The newly generated data points are used for further energy calculations based on either BAR method (Eq. 6 and 7) or traditional FEP method (Eq. 2).

Section S9. ^{1}H NMR and ^{13}C NMR spectrums of the designed PDE10 inhibitors

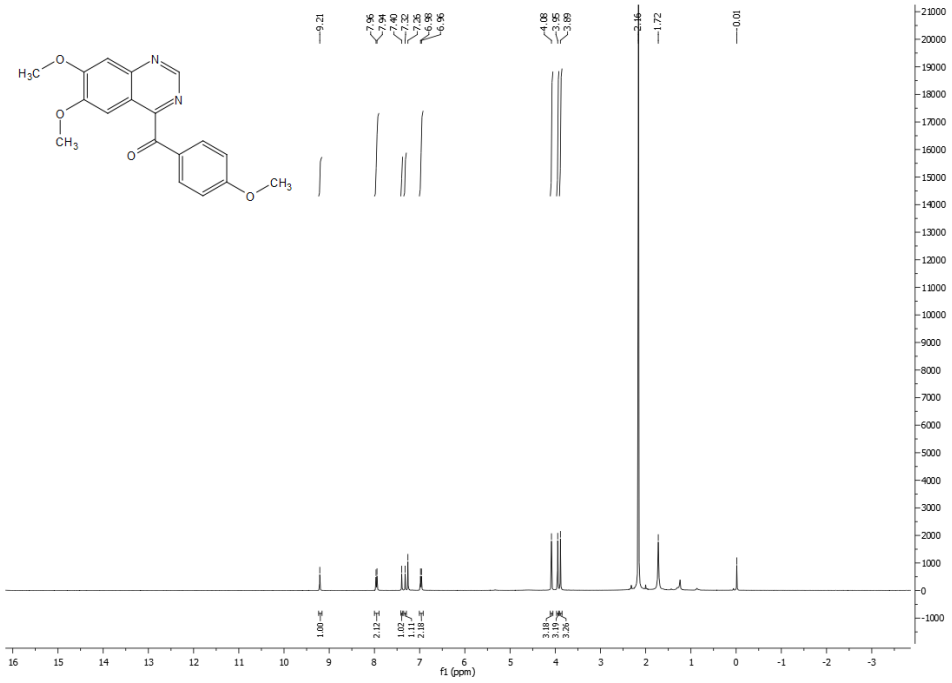


Fig. S10. ^{1}H-NMR spectrum of compound **LHB-2**

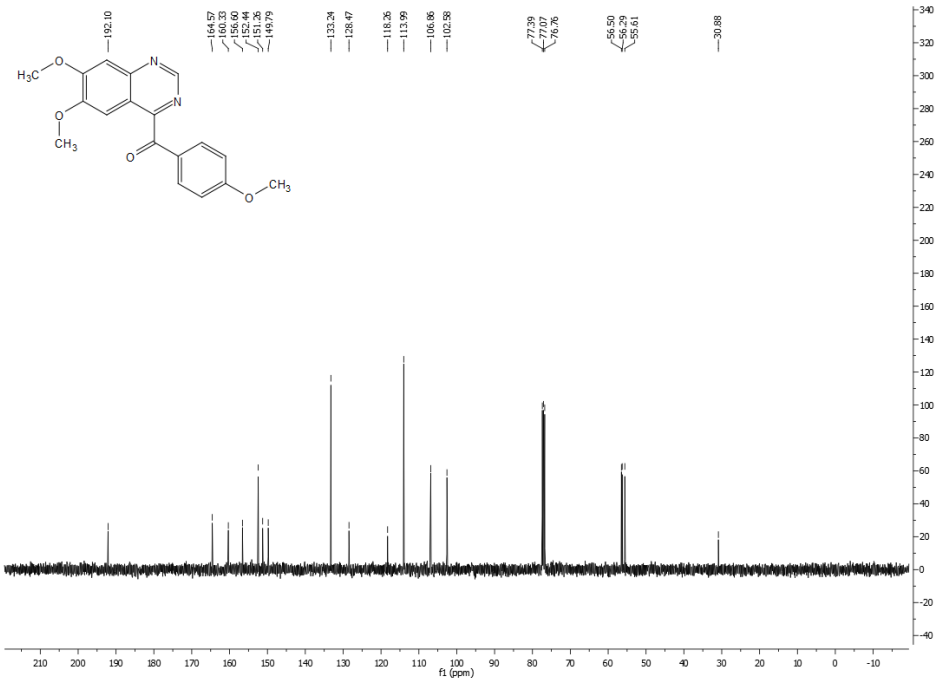


Fig. S11. ^{13}C-NMR spectrum of compound **LHB-2**

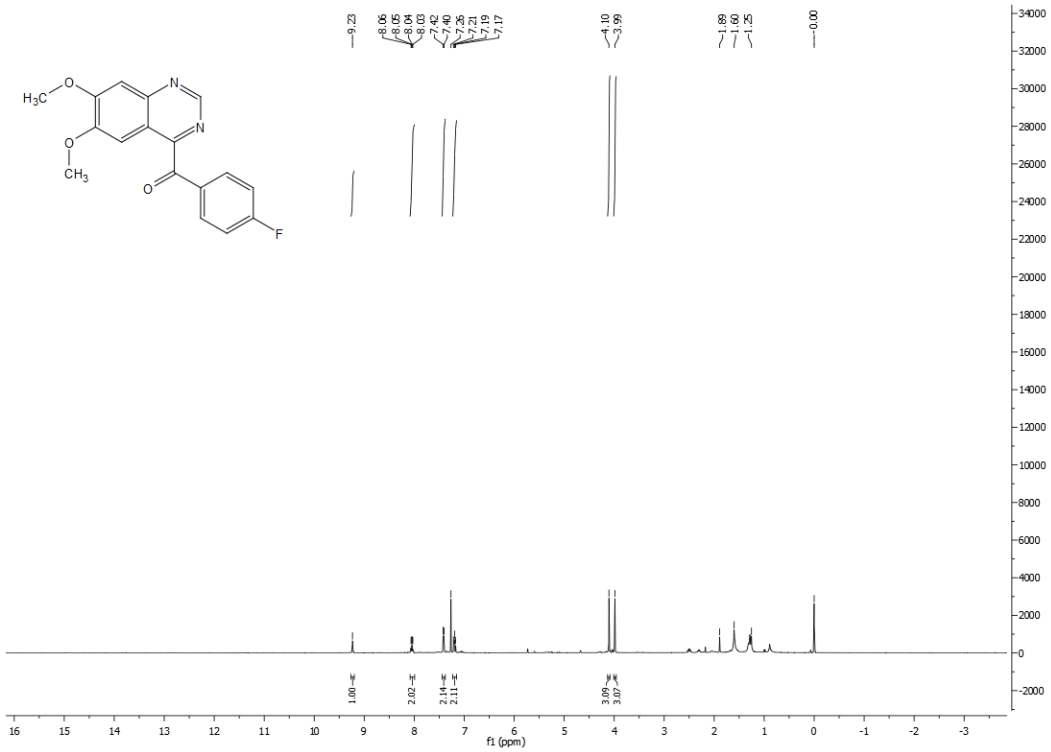


Fig. S12. ^{1}H-NMR spectrum of compound **LHB-3**

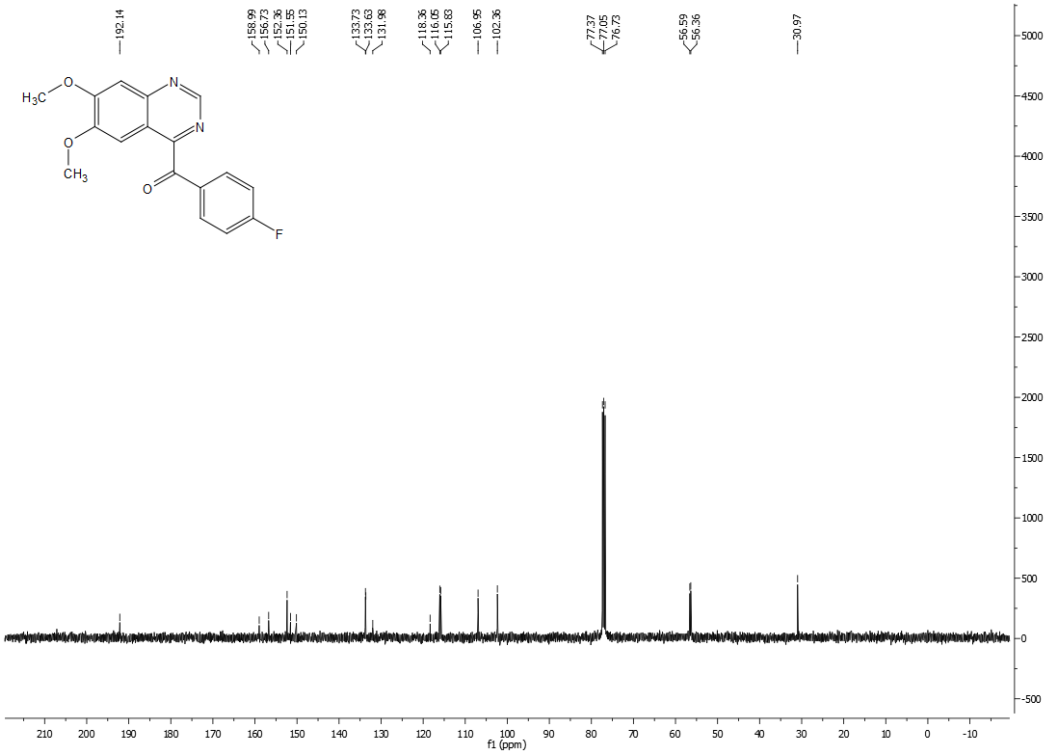


Fig. S13. ^{13}C-NMR spectrum of compound **LHB-3**

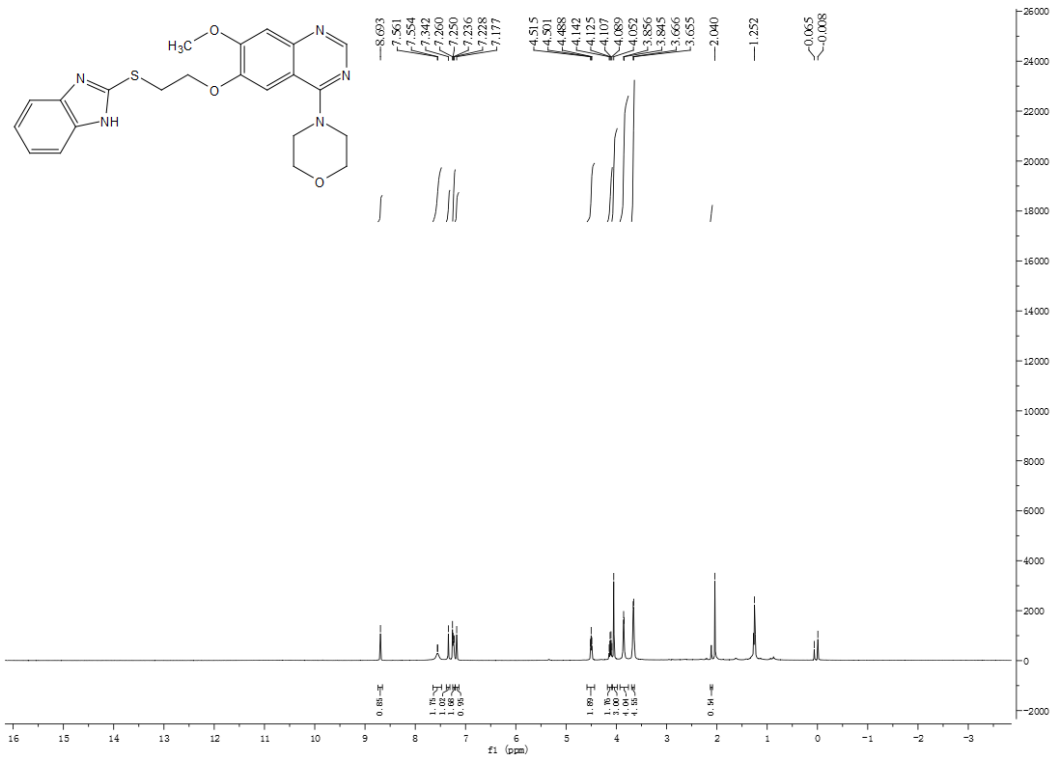


Fig. S14. ^{1}H-NMR spectrum of compound **LHB-4**

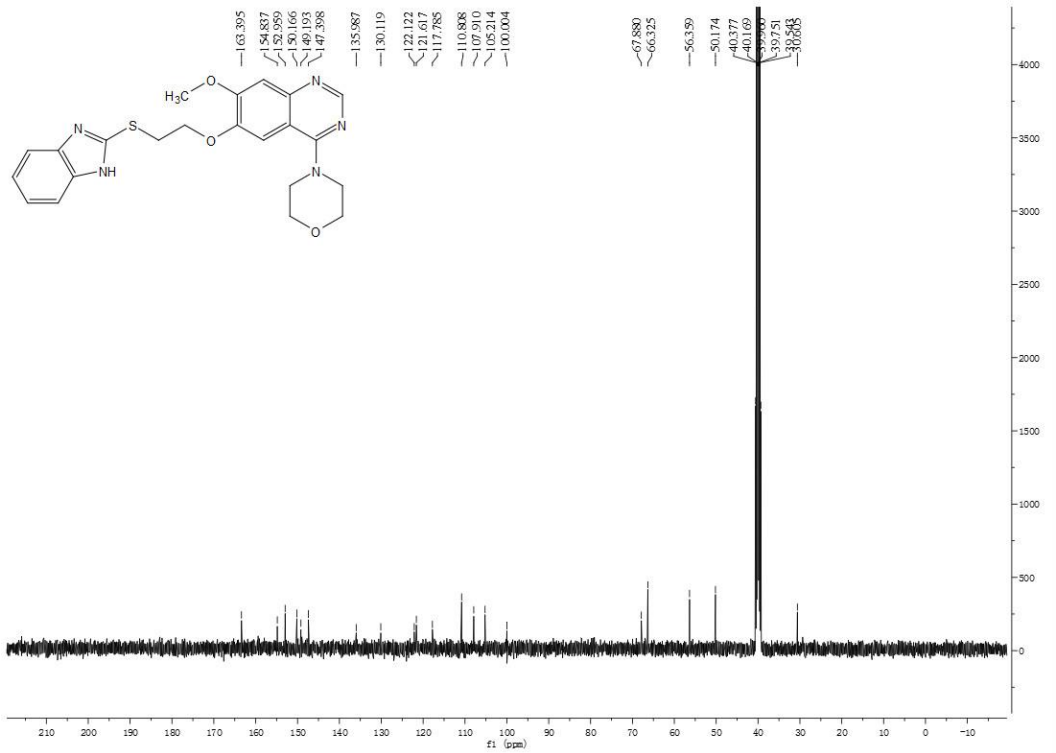


Fig. S15. ^{13}C-NMR spectrum of compound **LHB-4**

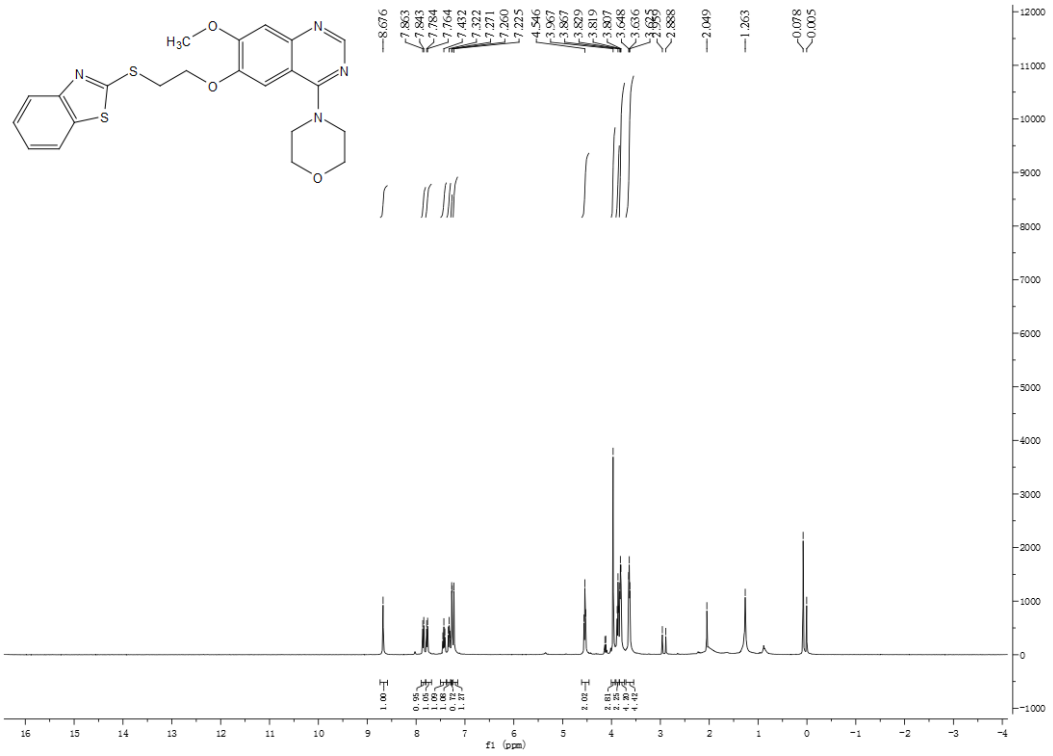


Fig. S16. ^{1}H-NMR spectrum of compound **LHB-5**

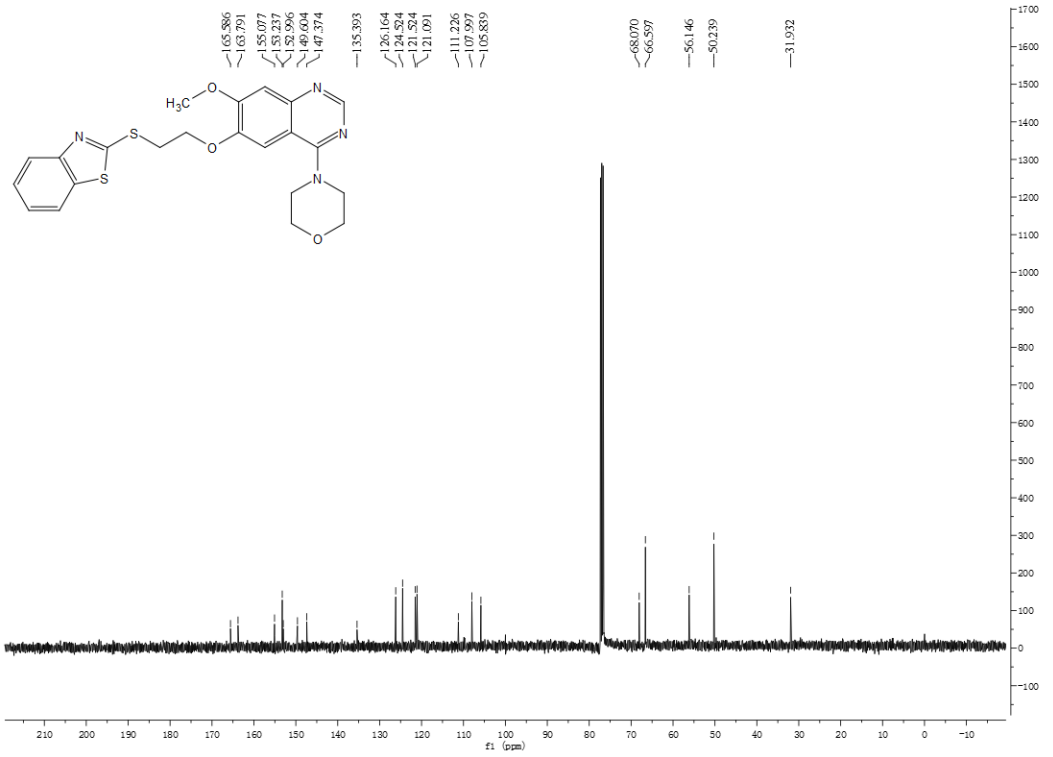


Fig. S17. ^{13}C-NMR spectrum of compound **LHB-5**

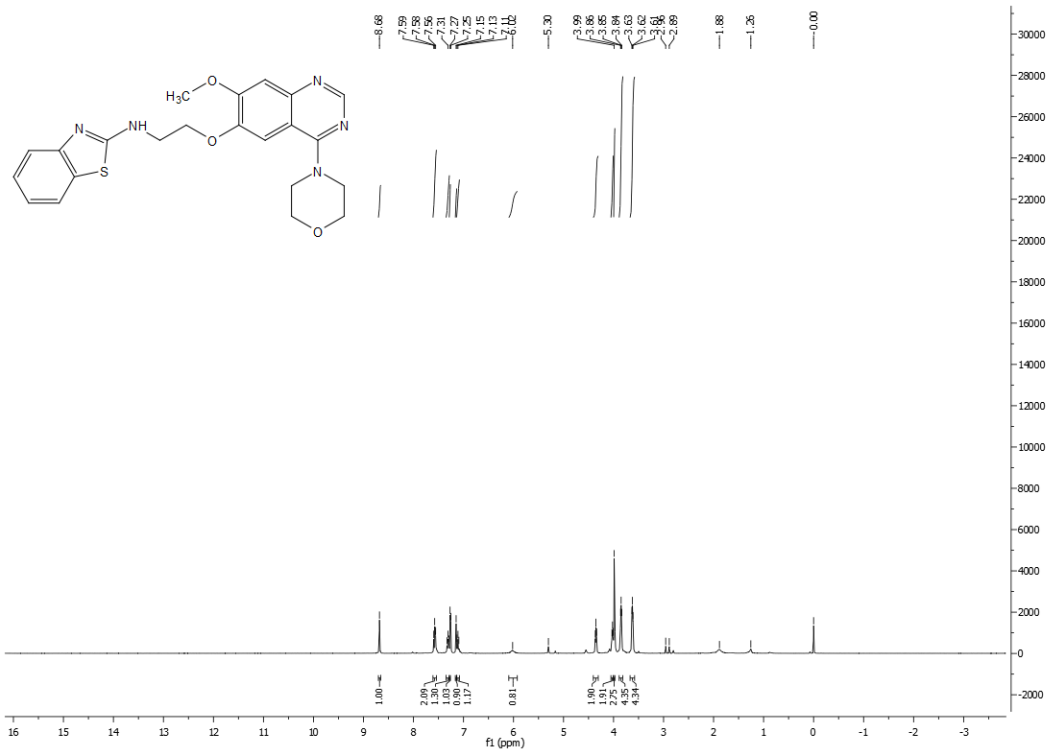


Fig. S18. ^{1}H-NMR spectrum of compound **LHB-6**

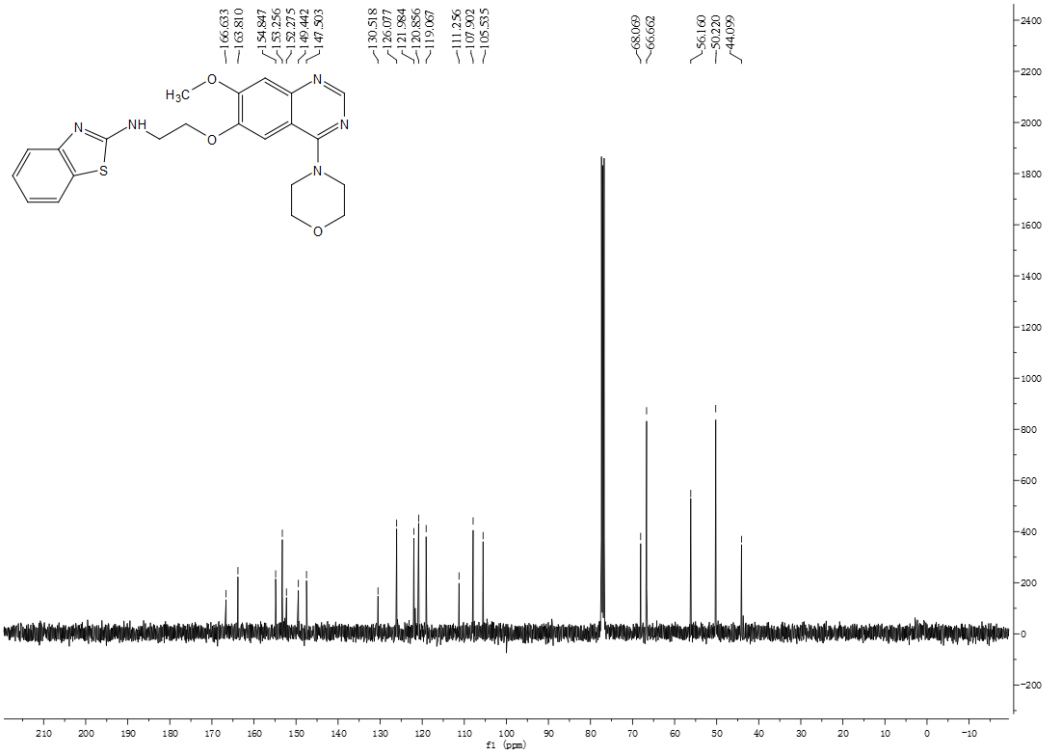


Fig. S19. ^{13}C-NMR spectrum of compound **LHB-6**

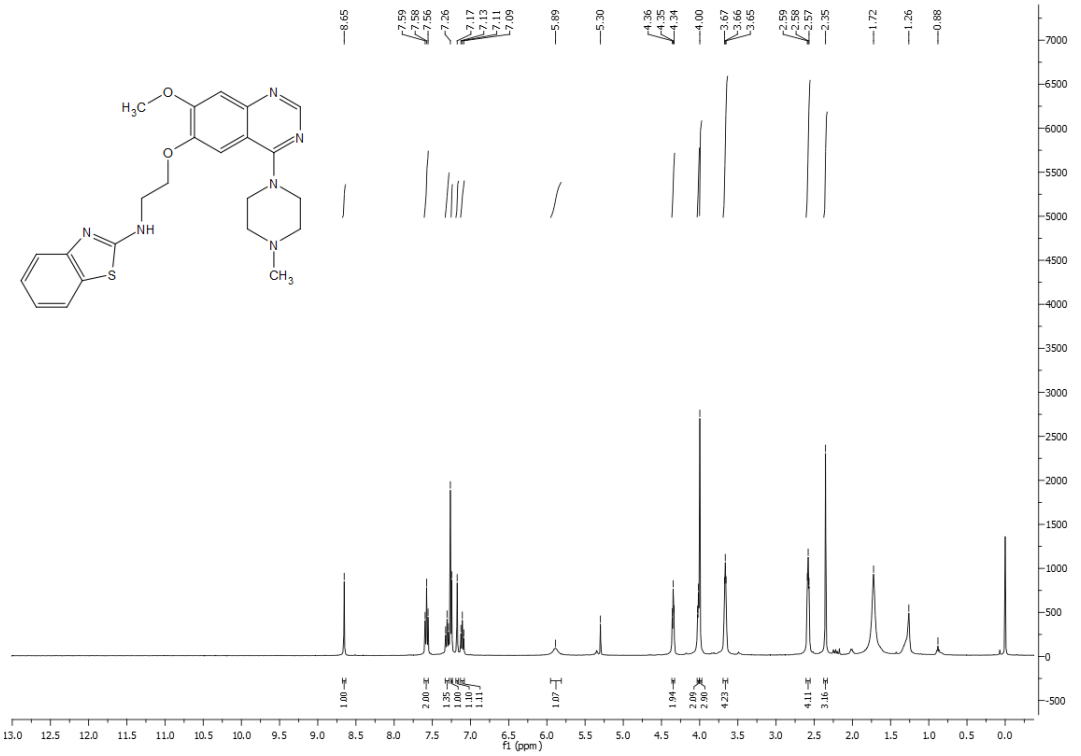


Fig. S20. ^{13}C-NMR spectrum of compound **LHB-7**

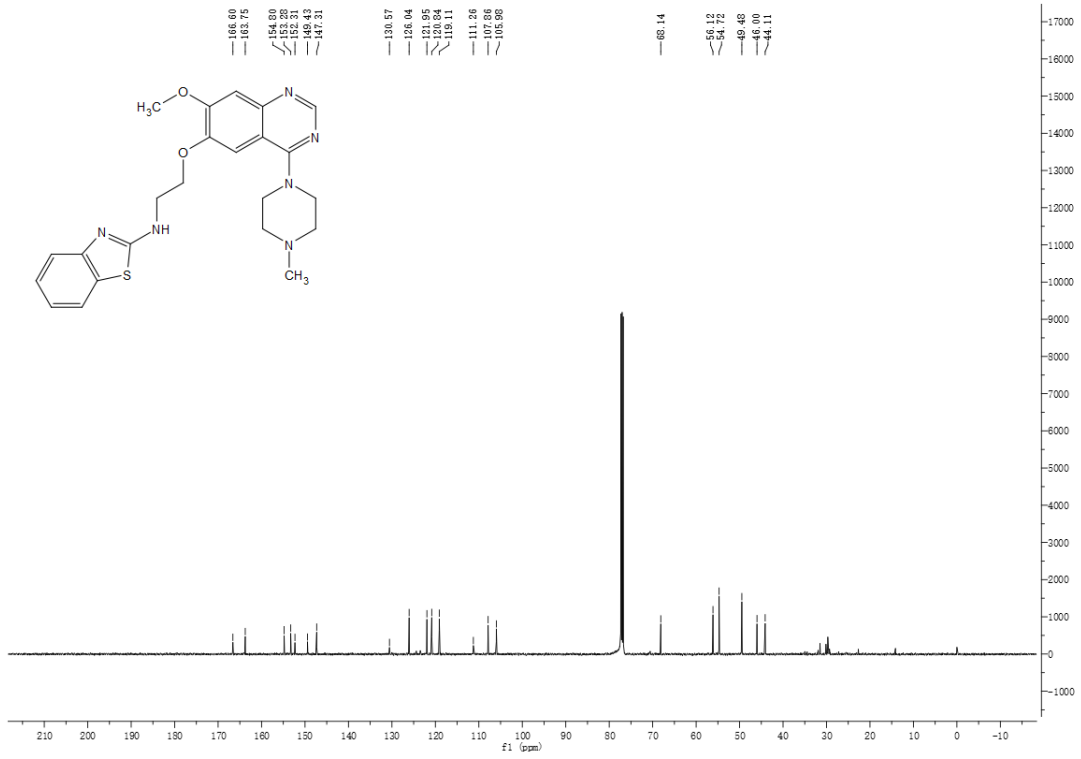


Fig. S21. ^{13}C-NMR spectrum of compound **LHB-7**

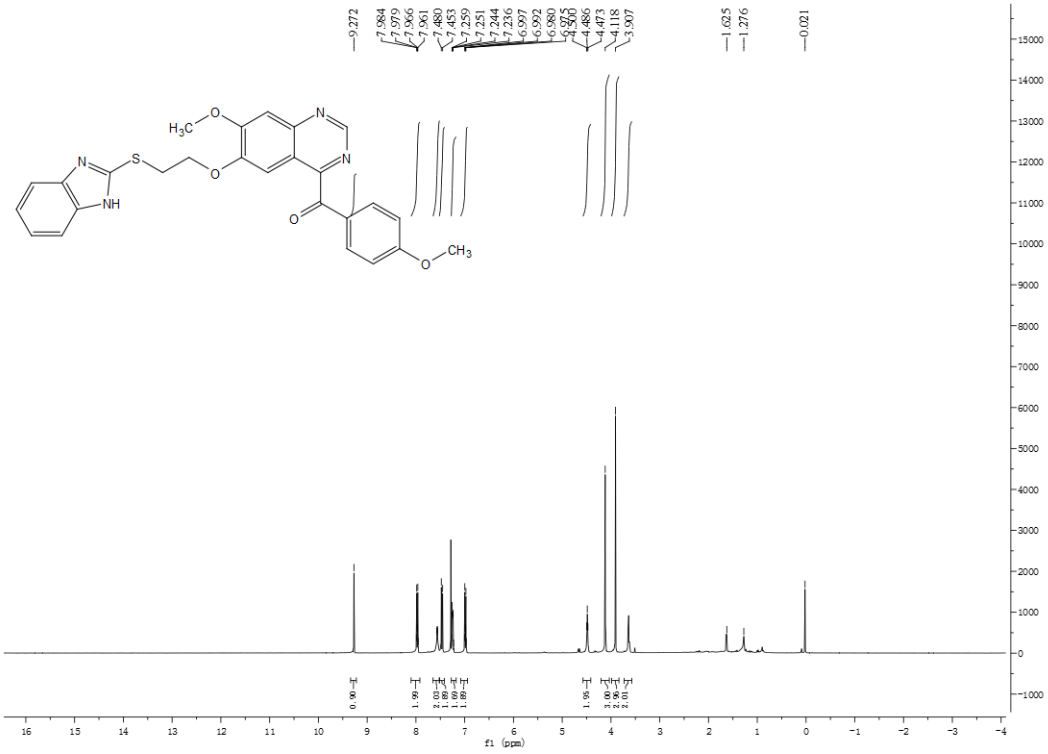


Fig. S22. ^{1}H-NMR spectrum of compound **LHB-8**

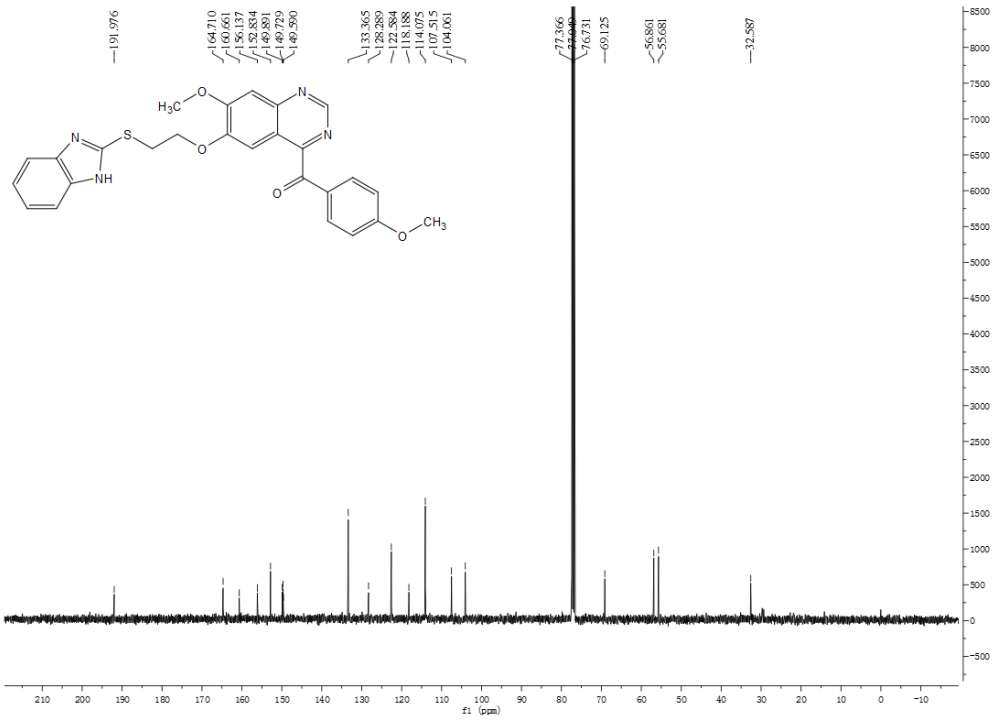


Fig. S23. ^{13}C-NMR spectrum of compound **LHB-8**

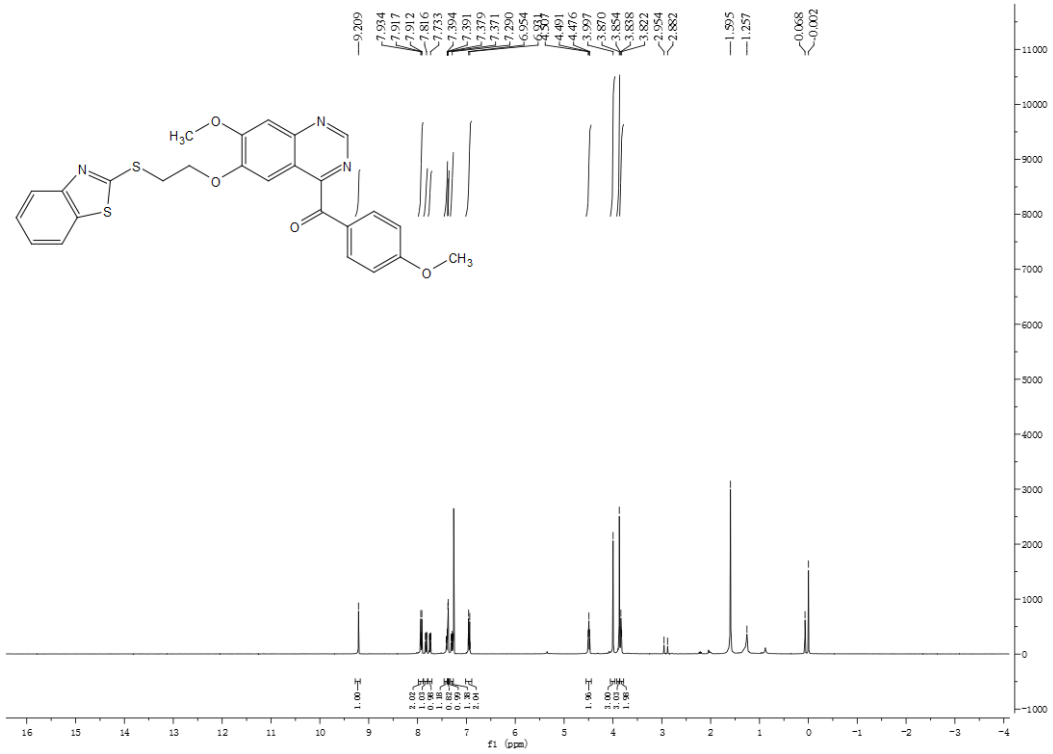


Fig. S24. ^{1}H-NMR spectrum of compound **LHB-9**

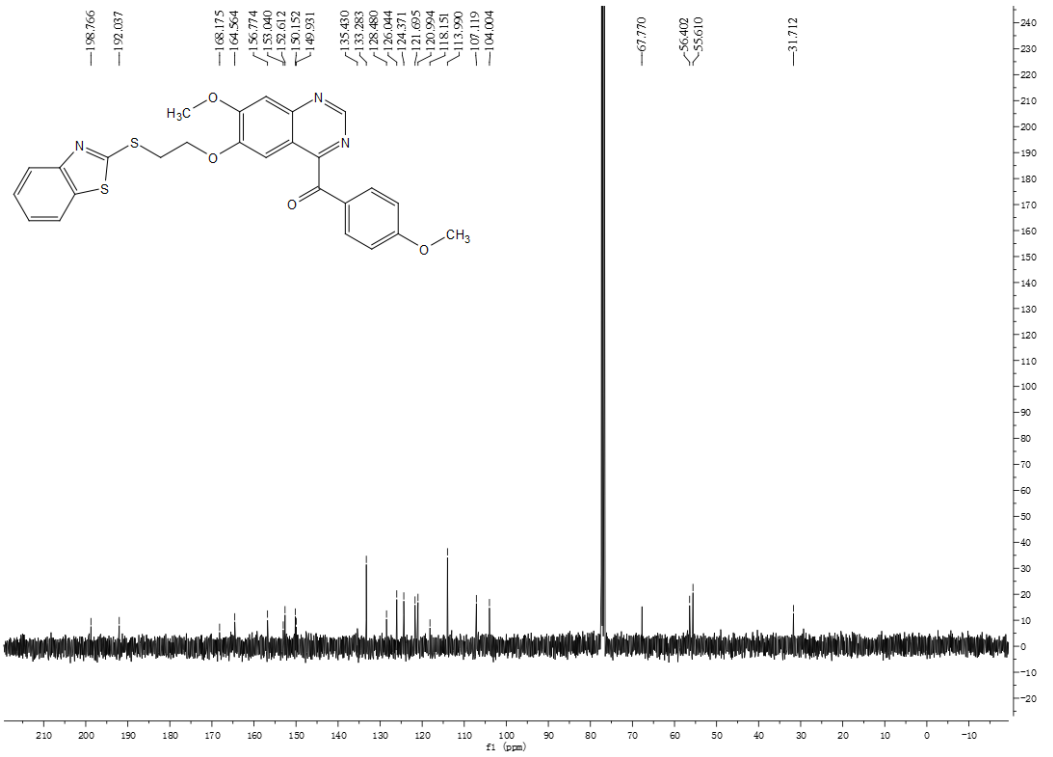


Fig. S25. ^{13}C-NMR spectrum of compound **LHB-9**

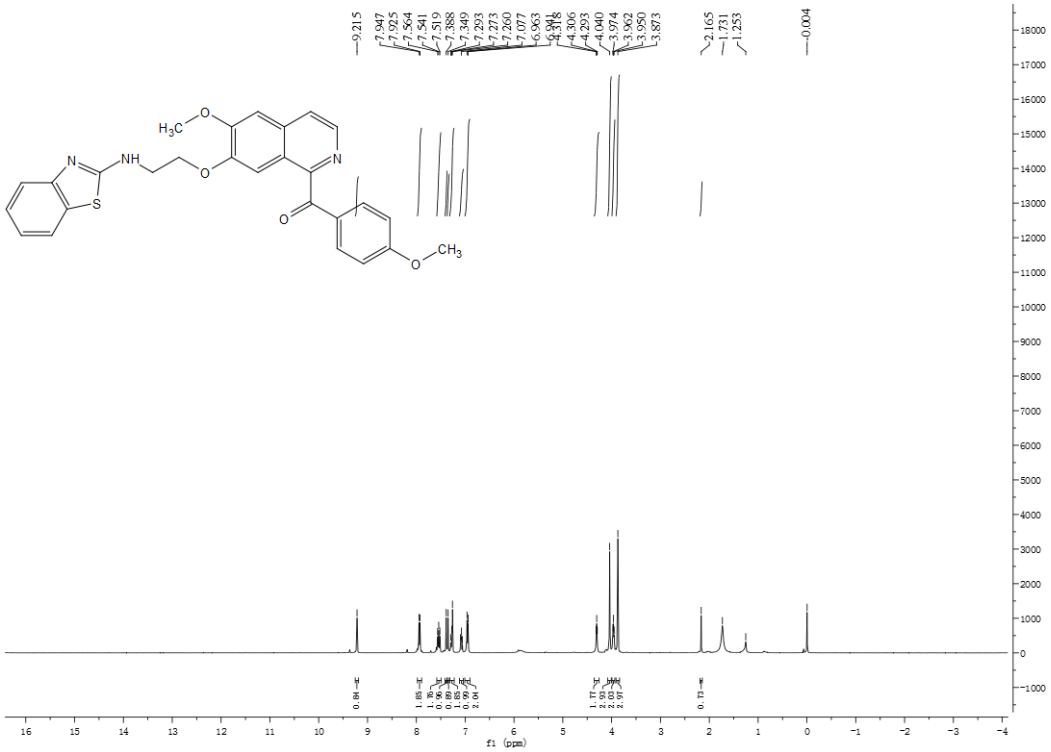


Fig. S26. ^{1}H-NMR spectrum of compound **LHB-10**

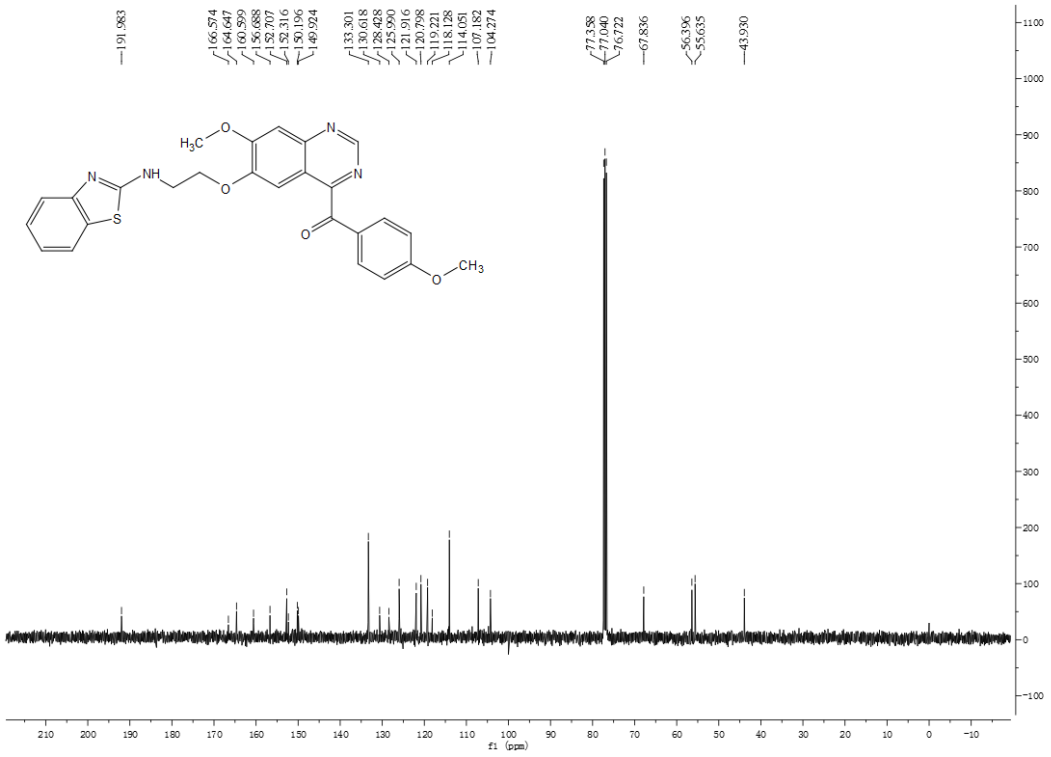
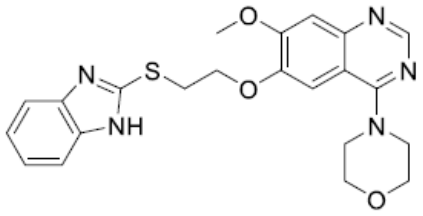


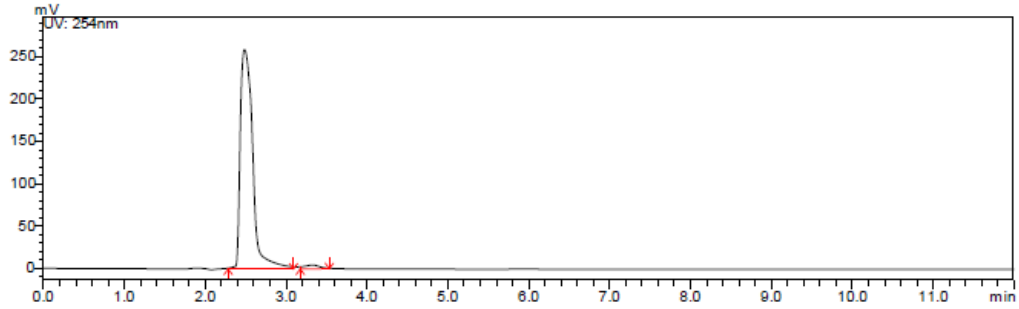
Fig. S27. ^{13}C-NMR spectrum of compound **LHB-10**

Section S10. HPLC spectrum for the purity of representative target compounds

SHIMADZU LC-20AT (column, Hypersil BDS C18, 5.0 μm, 4.6 mm×150 mm (Elite); detector, SPD-20A UV/vis detector, UV detection at 254 nm; elution, MeOH in water (50%,v/v); T=25°C; flow rate = 1.0 mL/min).



LHB-4

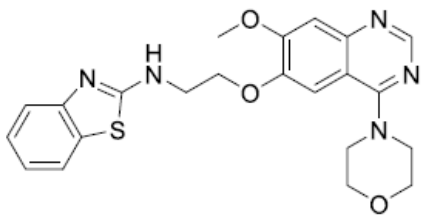


Peak#	Retention Time	Peak area	Peak height	Peak area%
1	2.481	2780685	259226	97.8756
2	3.320	60355	4508	2.1244

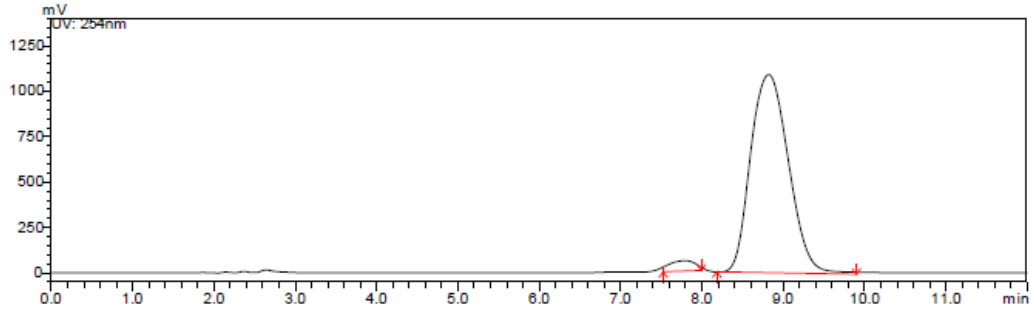
Purity: 97.9%

Fig. S28. HPLC spectrum for the purity of **LHB-4**

SHIMADZU LC-20AT (column, Hypersil BDS C18, 5.0 μm, 4.6 mm×150 mm (Elite); detector, SPD-20A UV/vis detector, UV detection at 254 nm; elution, MeOH in water (50%,v/v); T=25°C; flow rate = 1.0 mL/min).



LHB-6

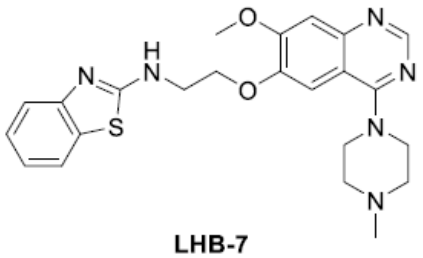


Peak#	Retention Time	Peak area	Peak height	Peak area%
1	7.788	1255231	56914	3.4254
2	8.817	35389720	1091410	96.5746

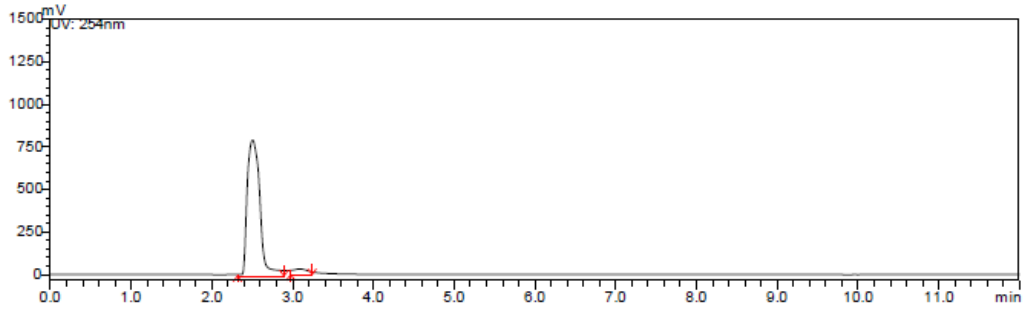
Purity: 96.6%

Fig. S29. HPLC spectrum for the purity of **LHB-6**

SHIMADZU LC-20AT (column, Hypersil BDS C18, 5.0 μm, 4.6 mm×150 mm (Elite); detector, SPD-20A UV/vis detector, UV detection at 254 nm; elution, MeOH in water (50%,v/v); T=25°C; flow rate = 1.0 mL/min).



LHB-7

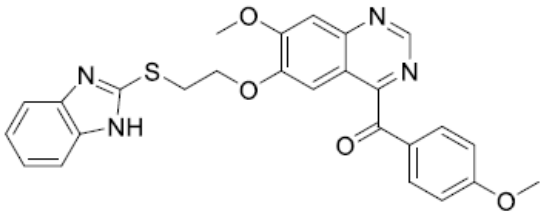


Peak#	Retention Time	Peak area	Peak height	Peak area%
1	2.500	8769921	797416	95.3530
2	3.085	427404	33065	4.6470

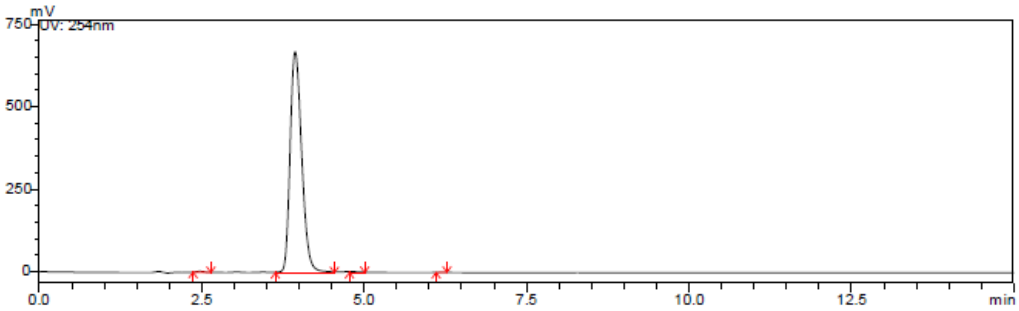
Purity: 95.4%

Fig. S30. HPLC spectrum for the purity of **LHB-7**

SHIMADZU LC-20AT (column, Hypersil BDS C18, 5.0 μm, 4.6 mm×150 mm (Elite); detector, SPD-20A UV/vis detector, UV detection at 254 nm; elution, MeOH in water (60%,v/v); T=25°C; flow rate = 1.0 mL/min).



LHB-8

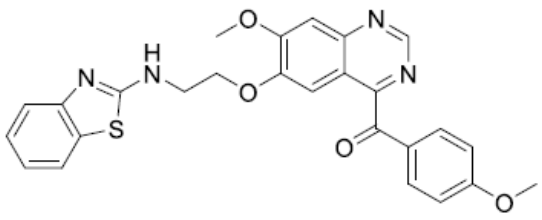


Peak#	Retention Time	Peak area	Peak height	Peak area%
1	2.468	24300	2893	0.2897
2	3.935	8334977	669950	99.3730
3	4.782	26419	2668	0.3150
4	6.181	1874	319	0.0223

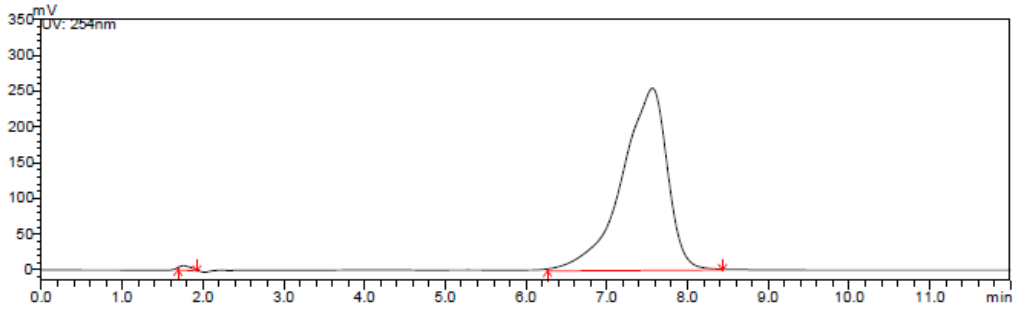
Purity: 99.4%

Fig. S31. HPLC spectrum for the purity of **LHB-8**

SHIMADZU LC-20AT (column, Hypersil BDS C18, 5.0 μm, 4.6 mm×150 mm (Elite); detector, SPD-20A UV/vis detector, UV detection at 254 nm; elution, MeOH in water (55%,v/v); T=25°C; flow rate = 1.0 mL/min).



LHB-10



Peak#	Retention Time	Peak area	Peak height	Peak area%
1	1.756	65065	6282	0.6450
2	7.562	10022295	254729	99.3550

Purity: 99.4%

Fig. S32. HPLC spectrum for the purity of **LHB-10**

Section S11. High-resolution mass spectra (HRMS) spectrums of representative compounds

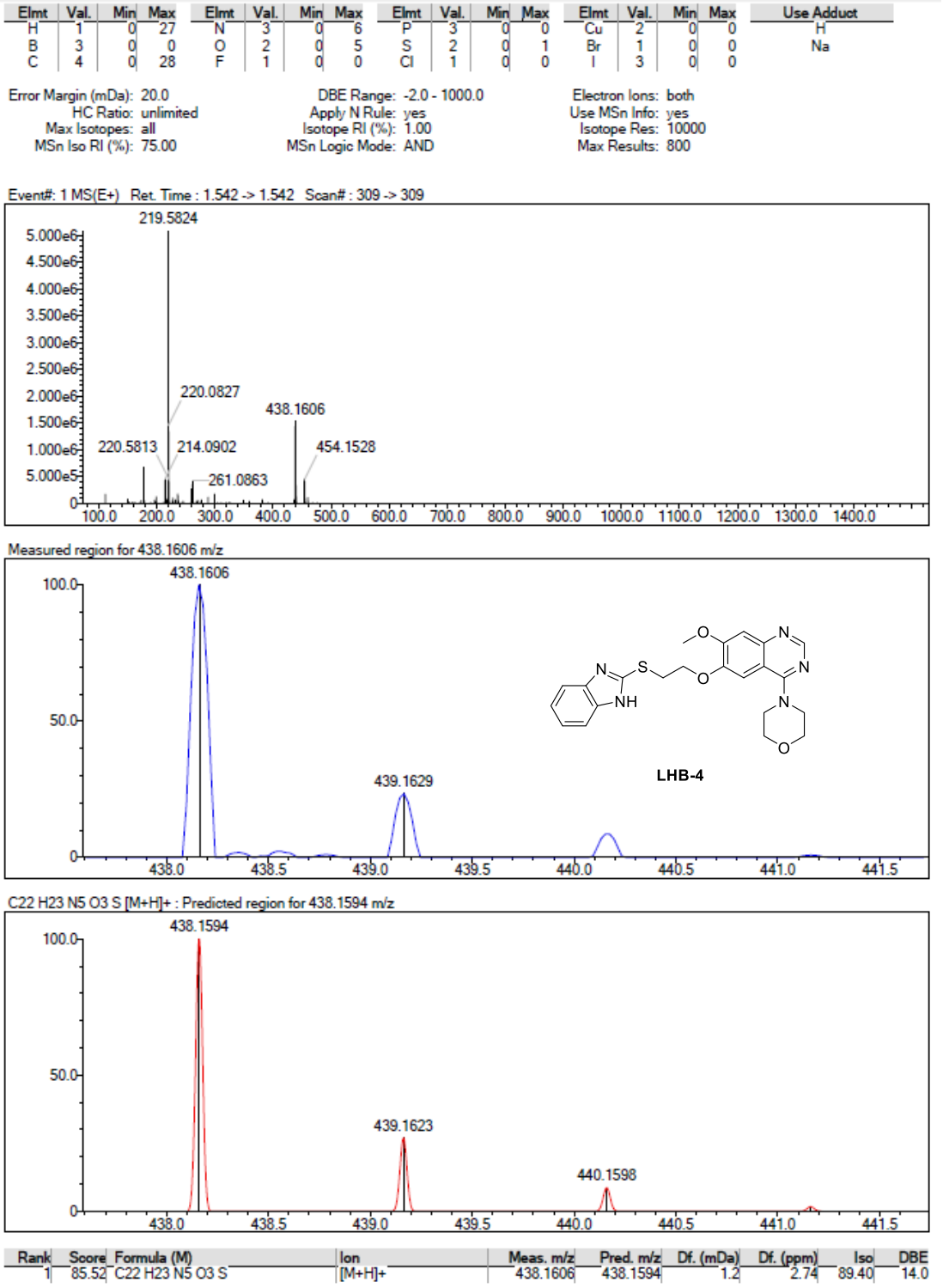


Fig. S33. HRMS spectrum of **LHB-4**

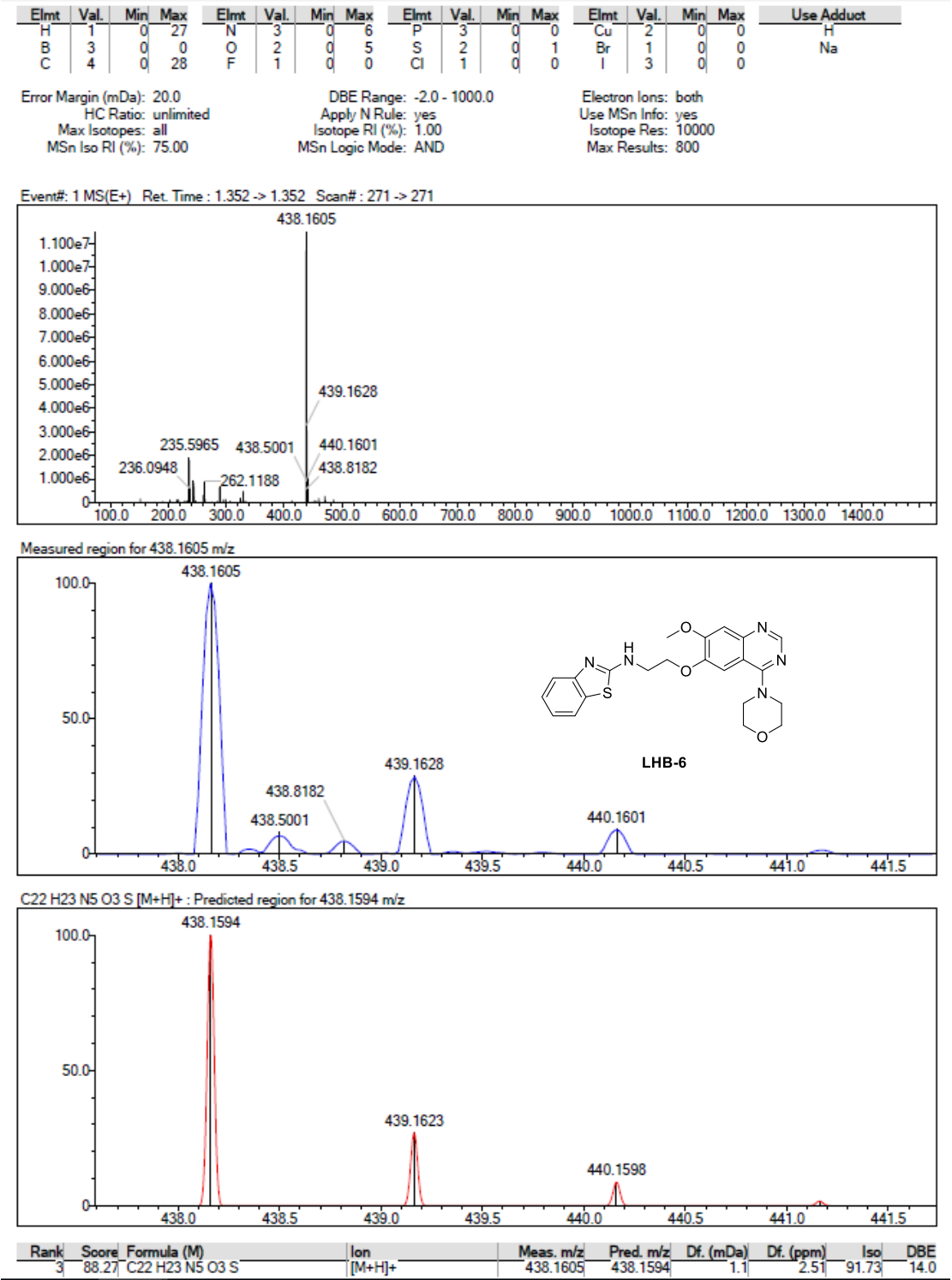


Fig. S34. HRMS spectrum of **LHB-6**

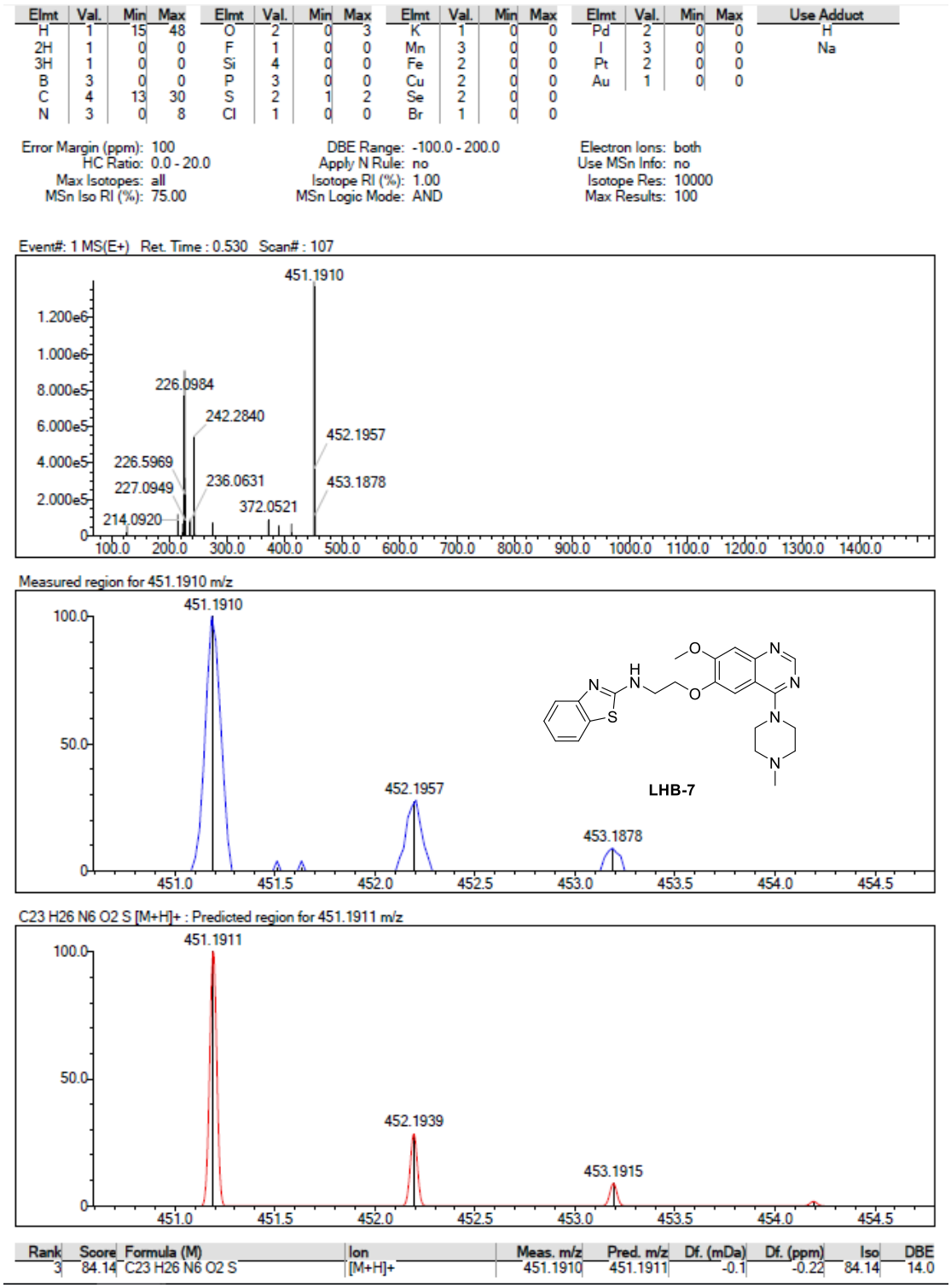


Fig. S35. HRMS spectrum of **LHB-7**

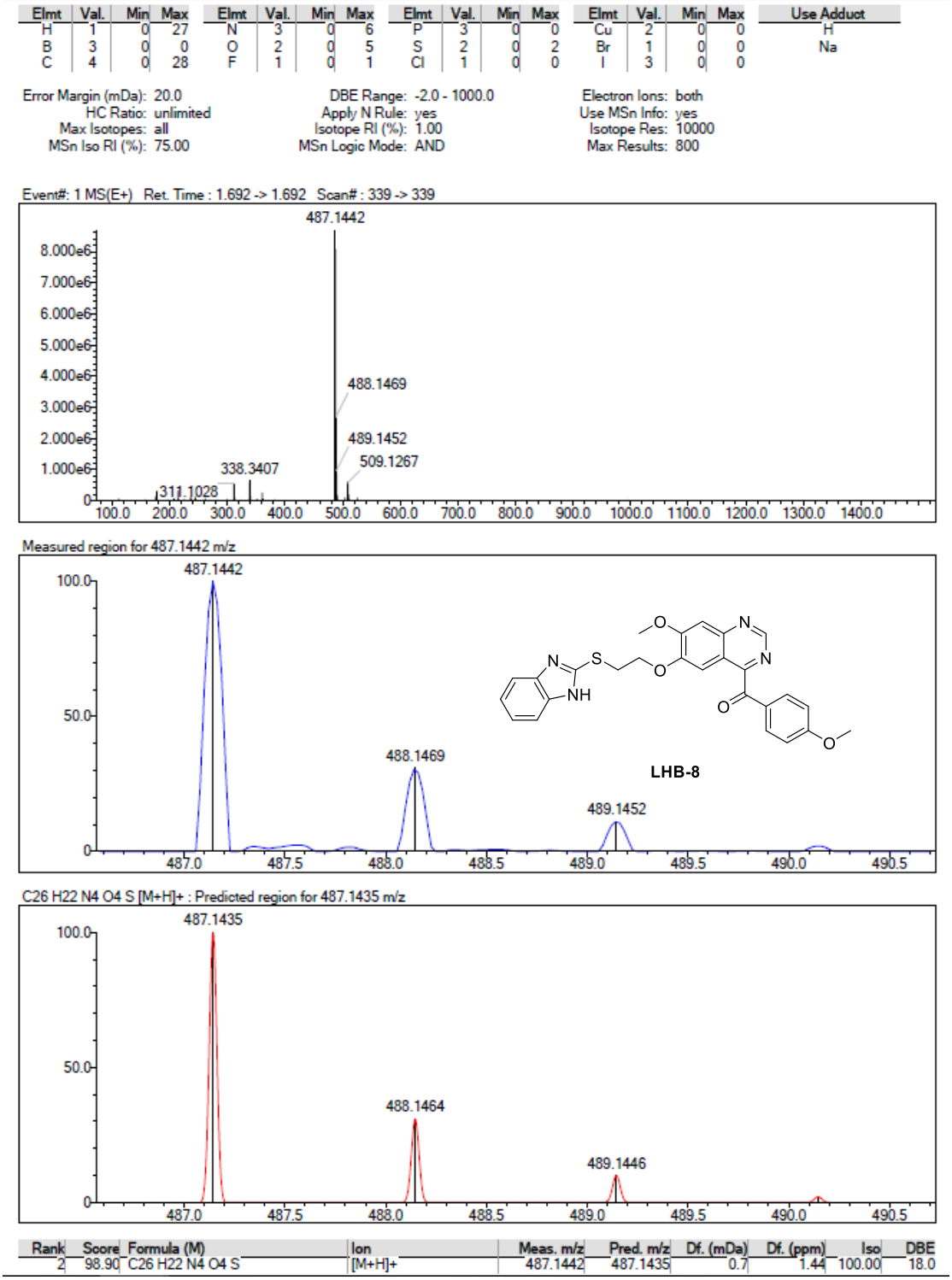


Fig. S36. HRMS spectrum of **LHB-8**

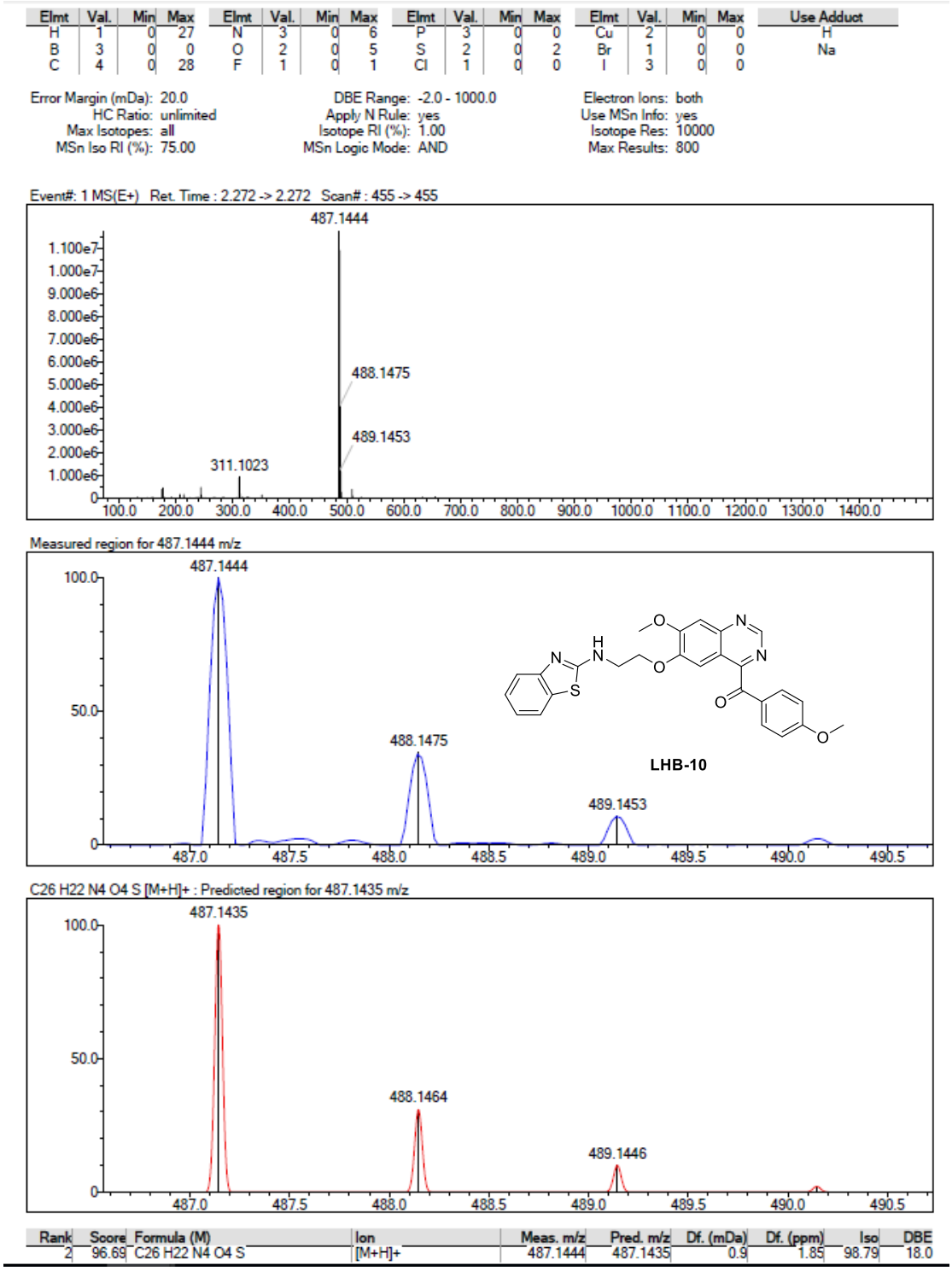


Fig. S37. HRMS spectrum of **LHB-10**

Supporting References

1. Neuhaus J-M, Sitcher L, Meins F, Jr, Boller T (1991) A short C-terminal sequence is necessary and sufficient for the targeting of chitinases to the plant vacuole. *Proc Natl Acad Sci USA* 88:10362–10366.
2. Bennett CH (1976) Efficient estimation of free energy differences from Monte Carlo data. *J Comput Phys* 22(2):245-268.
3. Shirts MR, Chodera JD (2008) Statistically optimal analysis of samples from multiple equilibrium states. *J Chem Phys* 129(12):124105.
4. Pitera JW, van Gunsteren WF (2002) A comparison of non-bonded scaling approaches for free energy calculations. *Mol Simul* 28(1-2):45-65.