Electronic Supplementary Material (ESI) for ChemComm.
This journal is © The Royal Society of Chemistry 2021

Supporting Information

A deep learning-driven scaffold hopping in the discovery of Akt kinases inhibitors

Zuqin Wang,[a,#] Ting Ran,[b,#] Fang Xu,[a,#] Chang Wen, [b] Shukai Song,[a] Yang Zhou,[a,*] Hongming Chen[b,*] and Xiaoyun Lu[a,*]

[a] College of Pharmacy, Jinan University, 601 Huangpu Avenue West, Guangzhou 510632, China

[b] Bioland Laboratory (Guangzhou Regenerative Medicine and Health - Guangdong Laboratory) Guangzhou 510530, China

[#]These authors contributed equally to this work.

Corresponding authors: For Y.Z., zhouyang@jnu.edu.cn; For H. C., chen_hongming@grmh-gdl.cn; For X. L., luxy2016@jnu.edu.cn

Computational Methods

Deep conditional transformer neural networks

SyntaLinker is a deep conditional generative model with transformer architecture[1]. The flow-chart of the architecture was shown in Fig. S1. The encoder layer consists of multi-head self-attention sub-layers and position-wise feed forward network (FFN) sub-layers. The decoder layer has masked self-attention, encoder-decoder self-attention and FFN sub-layers. These two layers respectively process a source SMILES sequence and a target SMILES sequence, which is different to graph generative model[2] that employs graph to describe a molecular structure. The source SMILES sequence is thus embedded to a latent representation L. Given L, the decoder layer maps the target sequence to a probability distribution for sampling a SMILES token, and then generates an output sequence. The loss function is calculated between the target sequence and the output sequence. Training the transformer model on a source and target sequence pair can learn the transformation rule from the source sequence to the target sequence. For SyntaLinker, the source

sequence is a terminal fragment pair and the target sequence is a complete structure that is assembled by the two terminal fragments and a linker fragment. When a SyntaLinker model is ready after a number of training epochs, it can be sampled to output a complete molecular structure given an input fragment pair. In order to build a SyntaLinker model, the ChEMBL compounds were chosen as the training set. About 5 million of terminal fragment pairs were generated by splitting the compounds and as the source. The details about data preparation and model construction refers to our previous report[3].

In this study, we applied this method to scaffold hopping of AKT inhibitors. With the ChEMBL model, the terminal fragments extracted from a known AKT inhibitor AZD5363 was used as a model input for sampling, which generated thousands of structures still containing the terminal fragments, but the linker fragment in AZD5363 was replaced by new fragments derived from sampling of the model. The generated structures are different from the seed one just regarding the linker fragment. So, scaffold hopping can be achieved based on the SyntaLinker algorithm under the condition that the linker fragment is the core functional scaffold in an active compound.

Molecular docking

The Glide docking workflow in Schrodinger software (Schrödinger, LLC, New York, NY, 2020) [4,5] was employed for molecular docking. The crystal structure coded 4GV1 of Akt1 was retrieved for docking from the RCSB PDB database. The protein was prepared with hydrogen added and crystal waters deleted using the Protein Preparation Wizard. Ligands for docking were prepared with all tautomers and isomers generated using the LigPrep module. The ionized states of function groups were determined by the Epik algorithm.[6] Then, a grid file with the information of docking site was generated with the center of the crystal ligand at the ATP-binding pocket as the reference. Other settings were maintained default except for the number of docking pose, which was set to 10 for each ligand. Finally, all docked poses were ranked by their docking score, but only the docking poses similar to that of AZD5363 were retained.

Experimental Methods

***In vitro* kinase inhibition activity assays**

Inhibition activities of compounds against Akt1, Akt2, Akt3 were determined using the FRET-based

Z'-Lyte assay systems according to the manufacturer's instructions (Life Technologies, Carlsbad, CA, USA). The reactions were carried out in 384-well plates in a 10 µL of reaction volume with appropriate amounts of kinases in 50 mM HEPES (pH 7.5), 5 mM $MgCl_2$, 2 mM $MnCl_2$, 1 mM EGTA, and 0.01% Brij-35.The reactions were incubated 1 h at room temperature in the presence of 2 µM of Ser/Thr 06 Peptide substrate with 75, 200 and 100 µM of ATP for kinases Akt1, 2 and 3, respectively, and in the presence of various concentrations of the compounds, then 5 µL development reagent was added for further 1 hours room temperature incubation followed by the addition of 5 µL of stop solution. Fluorescence signal ratio of 445 nm (coumarin)/520 nm (fluorescein) was examined with EnVision Multilabel Reader (Perkin Elmer, Inc.). The data were analyzed using Graphpad Prism 5 (Graphpad Software, Inc).

PKACα and PKCα enzymatic activities were performed by ICE Bioscience Inc. using homogeneous time-resolved fluorescence resonance energy transfer (HTRF) assay. The compounds were 4-fold serial diluted from 2 mM stock for 10 doses in DMSO. 2 µL compounds add into 78µL 1X assay buffer (20 mM Hepes, 5 mM $MgCl_2$, 0.01% Triton X-100). Shake plates for 20 min on plate shaker. Transfer 1µL compounds solution into each well of assay plate (784075, Greiner). Prepare 2.5X kinase in 1x kinase buffer (Cisbio, 62EZBFDD), 5mM $MgCl_2$, 1mM $MnCl_2$, 1mM DTT). Transfer 2 µL 2.5X Kinase solution into each well of assay plates (784075, Greiner). Centrifuge plates at 1000g for 30s, RT for 10 min. Prepare 2.5x STK-substrate-biotin and ATP mixture in 1X kinase buffer. Start the reaction by adding 2µl STK-substrate-biotin and ATP. Centrifuge plates at 1000g for 30s. Seal the assay plate, RT for 30min. Prepare 2X Sa-XL 665 and STK-antibody-Cryptatein HTRF detection buffer. Add 5µl Sa-XL 665 and STK-antibody-Cryptate into each well of the assay plate. Centrifuge plate at 1000g for 30s, RT for 1h. Read fluorescence signal at 615 nm (Cryptate) and 665 nm (XL665) on Biotek. A Ratio (665/615nm) is calculated for each well. Data was analyzed with Graphpad 7.0 software.

Cell culture and cell proliferation inhibition assay

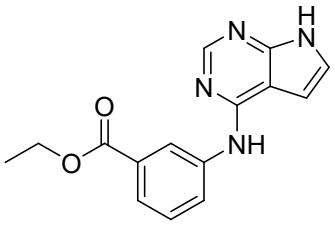
HEPG-2, HEK293, HCT116, and U937 were purchased from the American Type Culture Collection (ATCC, Rockville, MD) or Shanghai Cell Bank (Type Culture Collection, Chinese Academy of Sciences). The cell lines were maintained in RPMI-1640 supplemented with 10% FBS, 100 U/mL

penicillin, 50 mg/mL streptomycin, and 2 mmol/L glutamine in a humidified CO_2 incubator at 37 °C. All cells were passaged for less than 3 months before renewal from frozen, early-passage stocks obtained from the indicated sources. Cell proliferation was evaluated by CCK-8 assay (Cell Counting Kit 8, Dojindo Laboratories, Kumamoto, Japan). Cells were treated with indicated concentration of compounds for 72 h and the IC_{50} values were determined as previously reported.[7]

General Methods for Chemistry.

All reagents and solvents were purchased from commercial sources without further purification. Flash chromatography was performed using 300 mesh silica gel. All reactions were monitored by thin-layer chromatography (TLC) using silica gel plates with fluorescence F254 and UV light visualization. ^{1}H NMR spectra were recorded on a Bruker AV-400 spectrometer at 400 MHz. ^{13}C NMR spectra were recorded on a Bruker AV-600 spectrometer at 151 MHz. Coupling constants (*J*) are expressed in hertz (Hz). Chemical shifts (δ) of NMR are reported in parts per million (ppm) units relative to an internal standard (TMS). Low resolution ESI-MS were recorded on an Agilent 1200 HPLC-MSD mass spectrometer and high resolution ESI-MS on an Applied Biosystems Q-STAR Elite ESI-LC-MS/MS mass spectrometer. Purity of compounds was determined by reverse-phase high performance liquid chromatography [HPLC, Dionex Summit HPLC (Column: Diamonsil C18, 5.0 µM, 4.6 × 250 mm (Dikma Technologies); detector: PDA-100 photodiode array; inJector: ASI-100 autoinJector; pump: p-680A)] to be >95 %. A flow rate of 1.0 mL/min was used with mobile phase of 100% MeOH and 0.1% TEA aqueous solution.

Ethyl-3-((7H-pyrrolo[2,3-d]pyrimidin-4-yl)amino)benzoate (4)

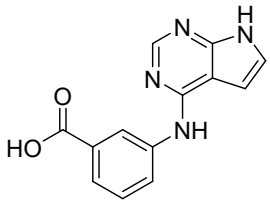


Under argon atmosphere, ethyl 3-aminobenzoate **2** (3.23 g, 19.54 mmol) and 4-chloro-7H-pyrrolo[2,3-d] pyrimidine **3** (3.0 g, 19.54 mmol) was dissolved in 25 mL ethanol, then HCl (1.81 mL, 58.61 mmol) was added, the mixture was heated at 90 °C and refluxed overnight, the solution

turned into blue-purple. Solvent was removed, then adjusted the pH to neutral with saturated $NaHCO_3$, the aqueous layer was extracted with dichloromethane (50 mL × 4). The combined organic extracts were washed with brine, dried over anhydrous Na_2SO_4, and removing the solvent, a crude material was obtained. The material was purified by column chromatography (DCM/MeOH, 50:1, v/v) to afford compound **4** (4.2 g, yield: 76.16%) as pink solid: MS (ESI) m/z 283.1 $[M+H]^+$.^{1}H NMR (400 MHz, DMSO-*d*6) δ 11.83 (s, 1H), 9.56 (s, 1H), 8.44 (d, *J* = 2.0 Hz, 1H), 8.38 (s, 1H), 8.34 (s, 1H), 7.63 – 7.56 (m, 1H), 7.49 (d, *J* = 8.0 Hz, 1H), 7.28 (dd, *J* = 3.4, 2.4 Hz, 1H), 6.85 (dd, *J* = 3.5, 1.8 Hz, 1H), 4.33 (t, *J* = 7.1 Hz, 2H), 1.34 (t, *J* = 7.1 Hz, 3H).

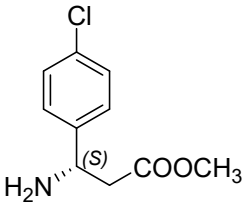
.

3-((7H-pyrrolo[2,3-d]pyrimidin-4-yl)amino)benzoic acid (5)



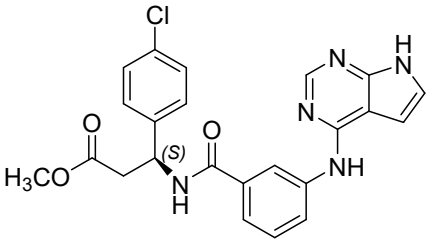
Sodium hydroxide (850 mg, 21.25 mmol) was added to a solution of **4** (2.0 g, 7.08 mmol) in a mixture of THF and H_2O (20 mL, 1:1, v/v). The resulting mixture was heated at 70 °C for 2 hours. Then the reaction mixture was acidified to pH 2~3, and the precipitate was filtered off and washed with water to afford compound **5** (1.7 g, yield: 94.38 %) as an off-white solid: MS (ESI) m/z 254.9 $[M+H]^+$.^{1}H NMR (400 MHz, DMSO-*d*6) δ 12.79 (s, 1H), 11.37 (s, 1H), 8.42 (s, 1H), 8.20 (t, *J* = 1.9 Hz, 1H), 7.96 (dd, *J* = 8.3, 1.9 Hz, 1H), 7.87 (d, *J* = 7.8 Hz, 1H), 7.61 (t, *J* = 7.9 Hz, 1H), 7.44 (dd, *J* = 3.5, 2.3 Hz, 1H), 7.00 (s, 1H).

Methyl (S)-3-amino-3-(4-chlorophenyl) propanoate (7)



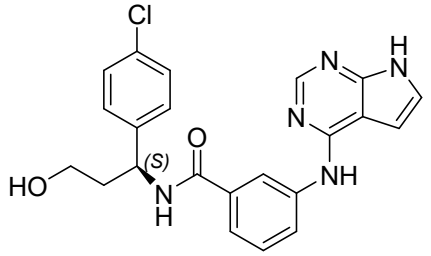
(S)-3-amino-3-(4-chlorophenyl) propanoic acid **6** (1 g, 5.01 mmol) was added to the solution of anhydrous methanol (10 mL). And then thionylchloride (0.73 mL, 10.02 mmol) was added dropwise to the reaction mixture at a rate to maintain the temperature at 0 °C. The reaction was warmed to 65 °C for 4 h, and concentrated to dryness in vacuo. Then the mixture was washed with saturated $NaHCO_3$ and the aqueous layer was extracted with EtOAc (20 mL×3), final the yellow oil **7** (850 mg, yield: 79.42 %) was obtained and used to next step directly. HRMS (ESI) for $C_{10}H_{12}ClNO_2$ $[M+H]^+$, calcd: 214.0629; found: 214.0613.

Methyl-(S)-3-(3-((7H-pyrrolo[2,3-d]pyrimidin-4-yl)amino)benzamido)-3-(4-chlorophenyl)propanoate (8)



To a mixture of compound **5** (500 mg, 1.96 mmol) and HATU (744.8 mg, 1.96 mmol) in DMF (10 mL) was added DIPEA (0.55 mL, 3.28 mmol) at room temperature. After 30 min, compound **7** (350 mg, 2.25 mmol) was added in the reaction mixture and stirred for 8 h, then extracted with EtOAc (3×30 mL), and washed brine (20 mL). The organic layers were dried over Na_2SO_4, filtered, and concentrated in vacuo to give the crude product. Then it was purified by column chromatography on silica gel (DCM/MeOH, 60:1, v/v) to afford **8** (615 mg, yield: 83.45 %) as a yellowish white solid: MS (ESI) m/z 450.2 $[M+H]^+$.^{1}H NMR (400 MHz, DMSO-*d*6) δ 12.15 (s, 1H), 10.05 (s, 1H)，8.95 (d, *J* = 8.1 Hz, 1H), 8.33 (s, 1H), 8.14 (d, *J* = 2.0 Hz, 1H), 8.04 (d, *J* = 8.1 Hz, 1H), 7.62 (d, *J* = 8.0 Hz, 1H), 7.57 – 7.47 (m, 1H), 7.49 – 7.37 (m, 4H), 7.35 (t, *J* = 2.8 Hz, 1H), 6.85 – 6.79 (m, 1H), 5.50-5.41 (m, *J* = 8.7, 6.2 Hz, 1H), 3.58 (s, 3H), 3.07 – 2.84 (m, 2H).

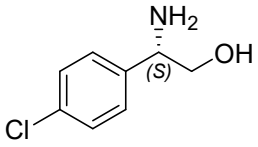
(S)-3-((7H-pyrrolo[2,3-d]pyrimidin-4-yl)amino)-N-(1-(4-chlorophenyl)-3-hydroxypropyl)benzamide (1a)



A solution of **8** (590 mg, 1.31 mmol) in THF (15 mL) was added to a mixture of $NaBH_4$ (497 mg, 13.14 mmol) and LiCl (550 mg, 13.14 mmol) in EtOH (10 mL). The reaction mixture was refluxed at 60 °C and stirred for 6 h, cooled, then diluted with EtOAc (3×20 mL), and washed brine (15 mL). The organic layers were dried over Na_2SO_4, filtered, and concentrated in vacuo to give the crude product. Then the material was purified by column chromatography on silica gel (DCM/MeOH, 30:1, v/v) to afford **1a** (280 mg, yield: 40.67 %) as a white solid: ^{1}H NMR (400 MHz, DMSO-*d*6) δ 11.77 (s, 1H), 9.47 (s, 1H), 8.80 (d, *J* = 8.1 Hz, 1H), 8.29 (s, 1H), 8.23 (t, *J* = 1.9 Hz, 1H), 8.19 (d, *J* = 8.0 Hz, 1H), 7.51 (d, *J* = 8.0 Hz, 1H), 7.46 – 7.38 (m, 5H), 7.25 (t, *J* = 2.9 Hz, 1H), 6.82 (dd, *J* = 3.5, 1.9 Hz, 1H), 5.17 (q, *J* = 8.3 Hz, 1H), 4.63 (t, *J* = 4.9 Hz, 1H), 3.52 – 3.40 (m, 2H), 2.08-2.00 (m, 1H), 1.93-1.84 (m, 1H). ^{13}C NMR (151 MHz, DMSO-*d*6) δ 166.64, 153.87, 151.36, 151.16, 143.60, 141.00,

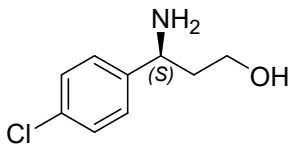
135.58, 131.59, 128.92, 128.79, 128.64, 123.27, 122.81, 120.96, 119.88, 104.22, 99.22, 58.24, 50.34, 39.15. $[\alpha]_D^{25}$ = 4.27 (c= 0.015, MeOH) HPLC analysis: MeOH-H_2O (80:20), 4.49 min, 98.83 % purity. HRMS (ESI) for $C_{22}H_{20}ClN_5O_2$ [M+H] $^+$, calcd: 422.1378; found: 422.1320.

(S)-2-amino-2-(4-chlorophenyl) ethan-1-ol (10a)



(S)-2-amino-2-(4-chlorophenyl) acetic acid **9** (3.0 g, 16.16 mmol) and sodium borohydride (1.5 g, 40.40 mmol) were added to a 100 mL dried flask, and then the flask was cooled in an ice bath and anhydrous THF (30 mL) was added. After 10 min, the iodine (4.9 g, 19.39 mmol) was dissolved in THF (10 mL) solution was added dropwise to the flask. Then the reaction was heated to 70 ºC reflux. After 14 h, the reaction was cooled to room temperature and methanol was added dropwise until the suspension turned to a clear solution. Then the reaction was stirred at room temperature for 30 min, followed by adding sodium hydroxide solution (20 mL, 3 M in water) and the reaction was stirred at room temperature. After an additional 4 h, the solution was concentrated under reduced pressure and the remaining suspension was extracted with DCM (3×40 mL). The combined organic solution was washed with brine, dried by anhydrous Na_2SO_4, and concentrated under reduced pressure. The resulting compound **10a** (2.3g, yield: 82.1 %) was obtained as a yellow solid, which was used without further purification: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.31 (d, *J* = 8.6 Hz, 2H), 7.26 (d, *J* = 8.6 Hz, 2H), 4.04 (dd, *J* = 8.2, 4.2 Hz, 1H), 3.70 (dd, *J* = 10.9, 4.2 Hz, 1H), 3.53 (dd, *J* = 11.0, 8.2 Hz, 1H).

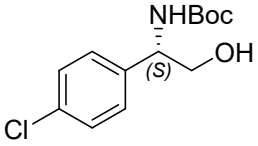
(S)-3-amino-3-(4-chlorophenyl) propan-1-ol (10b)



A solution of (S)-3-amino-3-(4-chlorophenyl) propanoic acid **6** (5.0 g, 25.05 mmol) in dry THF (40 mL) was cooled to 0 °C and $LiAlH_4$ (2.85 g, 75.14 mmol) was added in small portions. The reaction mixture was refluxed for 5 h under an argon atmosphere. The reaction mixture was then cooled to 0 °C, and quenched with H_2O (2.8 mL). The mixture was treated with 15 wt% NaOH in H_2O (2.8 mL) followed by H_2O (8.4 mL) and stirred for over 30 min at room temperature. The yellow mixture was filtered off and washed with organic solvents, the filtrate was concentrated. Then the mixture

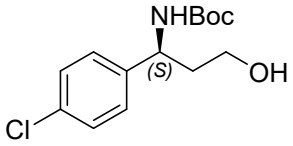
was extracted in EtOAc (3×40 mL) and the organic phases were combined, washed with brine (2×30 mL), dried (Na_2SO_4) and evaporated under reduced pressure. Purification by column chromatography through silica gel (DCM/MeOH, 40:1, v/v) to get compound **10b** (3.5 g, yield: 75.27%). ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.33-7.30 (m, 2H), 7.25-7.23 (m, 2H), 4.13 (t, J = 6.8 Hz, 1H), 3.81 – 3.72 (m, 2H), 1.89-1.83 (m, 2H). HRMS(ESI) for $C_9H_{12}ClNO$ $[M+H]^+$, calcd: 186.0680; found: 186.0669.

Tert-butyl (S)-(1-(4-chlorophenyl)-2-hydroxyethyl)carbamate (11a)



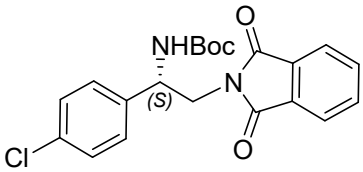
Di-tert-butyl dicarbonate (0.80 mL, 3.51 mmol) and triethylamine (0.75 mL, 5.26 mmol) was added in one portion to **10a** (600 mg, 3.50 mmol) in DCM (10 mL) under argon. The resulting solution was stirred at room temperature for 5 hours then evaporated to dryness and purification by column chromatography through silica gel to get compound **11a** as a white solid (689mg, yield: 72.53%):^{1}H NMR (400 MHz, DMSO-d_6) δ 7.36 (d, *J* = 8.5 Hz, 2H), 7.30 (d, *J* = 8.5 Hz, 2H), 7.25 (d, *J* = 8.3 Hz, 1H), 4.81 (t, *J* = 5.8 Hz, 1H), 4.51 (q, *J* = 7.3 Hz, 1H), 3.53 – 3.42 (m, 2H), 1.36 (s, 9H).

Tert-butyl (S)-(1-(4-chlorophenyl)-3-hydroxypropyl)carbamate (11b)



In a similar manner to that described for **11a**, by use of compound **10b** (1.47g, 7.96mmol), **11b**(1.6g, yield:70.71%) was obtained as a white solid: ^{1}H NMR (400 MHz, DMSO-d_6) δ 7.39 (s, 1H), 7.37 (d, *J* = 8.4 Hz, 2H), 7.29 (d, *J* = 8.2 Hz, 2H), 4.60 (d, *J* = 7.5 Hz, 1H), 4.51 (t, *J* = 4.9 Hz, 1H), 3.41 – 3.35 (m, 1H), 3.32-3.25 (m, 1H), 1.85-1.77 (m, 1H), 1.71-1.63 (m, 1H), 1.35 (s, 9H).

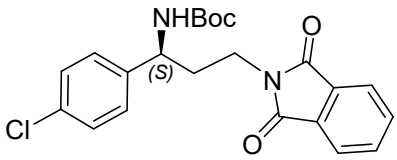
Tert-butyl(S)-(1-(4-chlorophenyl)-2-(1,3-dioxoisoindolin-2-yl)ethyl)carbamate(12a)



To a solution of **11a** (680 mg,2.51 mmol), PPh_3 (789 mg, 3.01 mmol) and phthalimide (406 mg, 2.76 mmol) in THF (10 mL) at room temperature was added DEAD (0.4 mL, 2.51 mmol). After 3h at room temperature, the reaction solution was concentrated and purified on silica to give the compound

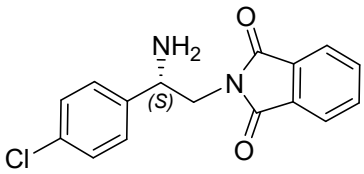
12a (460 mg, yield: 45.86%) as a white solid: ^{1}H NMR (400 MHz, DMSO-d_6) δ 7.89 – 7.79 (m, 4H), 7.60 (d, *J* = 9.6 Hz, 1H), 7.38 (s, 4H), 4.97 (q, *J* = 8.1 Hz, 1H), 3.81 (d, *J* = 7.4 Hz, 2H), 1.22 (s, 9H).

Tert-butyl (S)-(1-(4-chlorophenyl)-3-(1,3-dioxoisoindolin-2-yl) propyl)carbamate (12b)



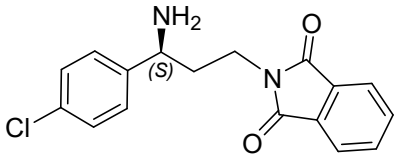
In a similar manner to that described for **12a**, by use of compound **11b** (1.0g，3.50 mmol), **12b** (1.05 g, yield: 72.32%) was obtained as a white solid: ^{1}H NMR (400 MHz, DMSO-d_6) δ 7.83 (s, 4H), 7.53 (d, *J* = 8.7 Hz, 1H), 7.32 (q, *J* = 8.5 Hz, 4H), 4.55 – 4.46 (m, 1H), 3.65-3.53 (m, 2H), 2.05-1.93 (m, 2H), 1.34 (s, 9H).

(S)-2-(2-amino-2-(4-chlorophenyl)ethyl)isoindoline-1,3-dione(13a)



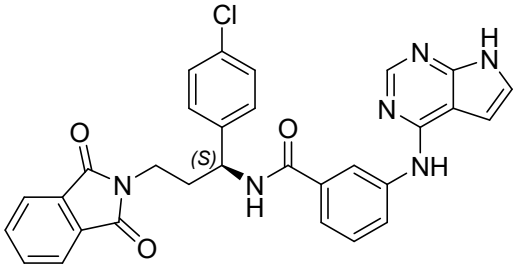
TFA (1 mL) was added to a suspension of **12a** (380 mg, 0.95 mmol) in DCM (10 mL) under argon. The resulting solution was stirred for 4 hours at room temperature, then the mixture was washed with saturated sodium bicarbonate and extracted with DCM(3×20 mL), the organic layers were dried over Na_2SO_4, filtered, and concentrated in vacuo to give the **13a** (241.4 mg, yield: 84.7 %) as a white solid: ^{1}H NMR (400 MHz, Chloroform-*d*) δ 7.86 (dd, *J* = 5.4, 3.1 Hz, 2H), 7.74 (dd, *J* = 5.5, 3.0 Hz, 2H), 7.41 – 7.30 (m, 4H), 4.37 (dd, *J* = 8.1, 6.1 Hz, 1H), 3.90 – 3.79 (m, 2H). HRMS (ESI) for $C_{16}H_{13}ClN_2O_2$ $[M+H]^+$, calcd: 301.0738; found: 301.0741.

(S)-2-(3-amino-3-(4-chlorophenyl)propyl)isoindoline-1,3-dione(13b)



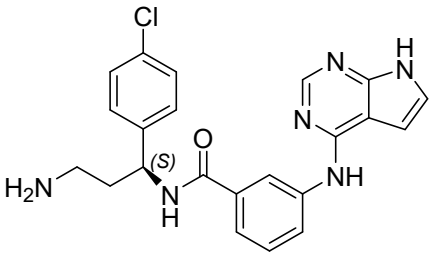
In a similar manner to that described for 13a, by use of compound **12b** (1.0 g，2.41 mmol), **13b** (510 mg, yield: 67.22 %) was obtained as a white solid. HRMS (ESI) for $C_{17}H_{15}ClN_2O_2$ $[M+H]^+$, calcd: 315.0895; found: 315.0873.

(S)-3-((7H-pyrrolo[2,3-d]pyrimidin-4-yl)amino)-N-(1-(4-chlorophenyl)-3-(1,3-dioxoisoindolin-2-yl)propyl)benzamide(14)



To a mixture of compound **5** (416.8 mg, 1.64 mmol) and HATU (625 mg,1.64 mmol) in DMF (15 mL) was added DIPEA (0.45 mL, 2.86 mmol), at room temperature. After 30 min, compound **13b** (430 mg, 1.37 mmol) was added in the reaction mixture and stirred overnight, then diluted with EtOAc (3×20 mL), and washed brine (20 mL). The organic layers were dried over Na_2SO_4, filtered, and concentrated in vacuo to give the crude product. Then it was purified by column chromatography on silica gel (DCM/MeOH, 50:1, v/v) to afford **14** (350 mg, yield:46.50 %) as a white solid: MS (ESI) m/z 550.7 $[M+H]^+$. ^{1}H NMR (400 MHz, DMSO-*d*6) δ12.07 (s, 1H), 8.91 (d, *J* = 8.3 Hz, 1H), 8.32 (s, 1H), 8.14 (s, 1H), 8.08 (s, 1H), 7.87 – 7.79 (m, 4H), 7.60 (d, *J* = 7.8 Hz, 1H), 7.52 (dd, *J* = 8.4, 4.6 Hz, 1H), 7.44 (d, *J* = 8.3 Hz, 2H), 7.37 (d, *J* = 8.3 Hz, 2H), 7.33 (s, 1H), 6.83 (d, *J* = 2.9 Hz, 1H), 5.07 (q, *J* = 7.7 Hz, 1H), 3.69 (q, *J* = 7.4 Hz, 2H), 2.22 (q, *J* = 7.1 Hz, 2H).

(S)-3-((7H-pyrrolo[2,3-d]pyrimidin-4-yl)amino)-N-(3-amino-1-(4-chlorophenyl)propyl)benzamide(1b)

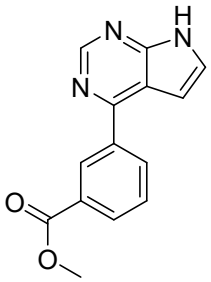


NH_2NH_2(0.18 mL, 3.62 mmol) was added to a THF/MeOH (4 mL/4 mL) solution of **14** (200 mg, 0.36 mmol) and stirred for 8-9 hours. The solution was concentrated and purified by column chromatography (DCM/MeOH, 10:1, V/V) to get the title compound **1b** (120 mg, yield: 78.55%) as a white solid: ^{1}H NMR (400 MHz, DMSO-d_6) δ 11.79 (s, 1H), 9.48 (s, 1H), 9.07 (d, *J* = 8.0 Hz, 1H), 8.29 (s, 1H), 8.25 (d, *J* = 1.9 Hz, 1H), 8.20 (dd, *J* = 8.1, 2.2 Hz, 1H), 7.51 (d, *J* = 7.6 Hz, 1H), 7.45 – 7.39 (m, 5H), 7.25 (d, *J* = 3.4 Hz, 1H), 6.82 (d, *J* = 3.5 Hz, 1H), 5.20 – 5.12 (m, 1H), 2.68-2.56 (m, 2H), 2.00 – 1.91 (m, 1H), 1.91 – 1.81 (m, 1H).^{13}C NMR (151 MHz, DMSO-*d*6) δ 166.65, 153.88, 151.37, 151.16, 143.56, 141.05, 135.53, 131.58, 128.90, 128.80, 128.63, 123.25, 122.80, 120.96, 119.87, 104.24, 99.27, 52.47, 51.31, 38.92. $[\alpha]_D^{25}$ = 4.48(c= 0.025, MeOH). HPLC analysis: MeOH-H_2O (80:20), 6.79

min, 98.76 % purity.

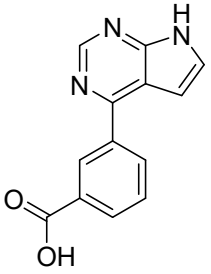
HRMS (ESI) for $C_{22}H_{21}ClN_6O$ $[M+H]^+$, calcd: 421. 1538; found: 421.1530.

Methyl 3-(7H-pyrrolo[2,3-d]pyrimidin-4-yl)benzoate(16)



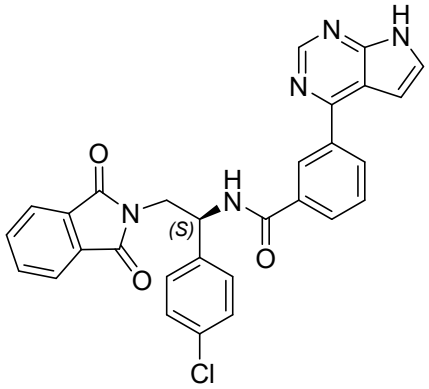
(3-(methoxycarbonyl)phenyl)boronic acid **15** (3 g, 16.67 mmol), compound **3** (2.56 g, 16.67mmol), $PdCl_2$(dppf) (609 mg, 0.8335 mmol), K_2CO_3 (4.61 g, 33.34 mmol) were mixed in a flask containing argon-saturated 1,4-dioxane (30 mL) and water (6 mL), then the solution was heated to 80 ºC and reacted for 5-6 hours. After cooling to room temperature, the reaction mixture was extracted with ethyl acetate (3×40 mL). The combined organic extracts were dried with anhydrous Na_2SO_4 and evaporated. The crude product was purified by column chromatography on silica gel (DCM/MeOH, 50/1, v/v) to obtain a beige solid **16** (3.22 g, yield: 76.37%): MS (ESI) m/z 251.9 $[M-H]^-$.^{1}H NMR (400 MHz, DMSO-d_6) δ 12.35 (s, 1H), 8.88 (s, 1H), 8.78 (t, *J* = 1.8 Hz, 1H), 8.46 (d, *J* = 8.0 Hz, 1H), 8.13 (d, *J* = 7.9 Hz, 1H), 7.76 (t, *J* = 7.8 Hz, 1H), 7.73 – 7.70 (m, 1H), 6.90 (dd, *J* = 3.7, 1.6 Hz, 1H), 3.92 (s, 3H).

3-(7H-pyrrolo[2,3-d]pyrimidin-4-yl)benzoic acid(17)



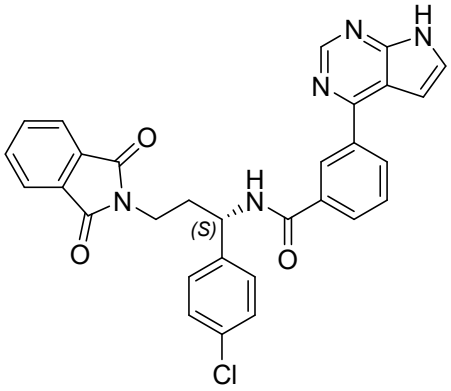
In a similar manner to that described for **5**, by use of compound **16** (3.22 g，3.50 mmol), **17** (2.81 g, yield: 92.38 %) was obtained as a light green solid: ^{1}H NMR (400 MHz, DMSO-d_6) δ 13.11 (s, 1H), 9.01 (s, 1H), 8.69 (d, *J* = 1.8 Hz, 1H), 8.40 (d, *J* = 7.8 Hz, 1H), 8.18 (d, *J* = 7.7 Hz, 1H), 7.89 (t, *J* = 2.7 Hz, 1H), 7.79 (t, *J* = 7.7 Hz, 1H), 7.01 – 6.96 (m, 1H). HRMS (ESI) for $C_{13}H_9N_3O_2$ $[M+H]^+$, calcd: 240.0768; found: 240.0749.

(S)-N-(1-(4-chlorophenyl)-2-(1,3-dioxoisoindolin-2-yl)ethyl)-3-(7H-pyrrolo[2,3-d]pyrimidin-4-yl)benzamide(18a)



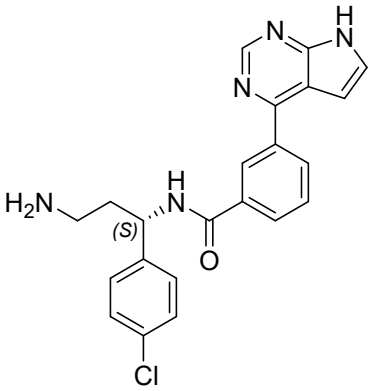
In a similar manner to that described for **14**, by use of compound **13a** (251 mg, 0.83 mmol) and compound **17** (219 mg, 0.91 mmol), **18a** (166 mg, yield: 38.11%) was obtained as a beige solid: MS (ESI) m/z 519.6 $[M-H]^-$. ^1H NMR (400 MHz, DMSO-d_6) δ 12.32 (s, 1H), 9.20 (d, J = 8.9 Hz, 1H), 8.87 (s, 1H), 8.53 (s, 1H), 8.30 (d, J = 7.7 Hz, 1H), 7.93 (d, J = 7.8 Hz, 1H), 7.88-7.81 (m, 4H), 7.71 (t, J = 3.0 Hz, 1H), 7.68 (t, J = 7.8 Hz, 1H), 7.52 (d, J = 8.4 Hz, 2H), 7.41 (d, J = 8.5 Hz, 2H), 6.88 (d, J = 3.5 Hz, 1H), 5.55 (q, J = 7.7 Hz, 1H), 4.08 (d, J = 7.3 Hz, 2H).

(S)-N-(1-(4-chlorophenyl)-3-(1,3-dioxoisoindolin-2-yl)propyl)-3-(7H-pyrrolo[2,3-d]pyrimidin-4-yl)benzamide (18b)



In a similar manner to that described for **14**, by use of compound **13b** (310mg, 0.987 mmol) and compound **17** (247.2 mg, 1.03 mmol), **18b** (150 mg, yield: 30.37 %) was obtained as a white solid: MS (ESI) m/z 533.8 $[M-H]^-$.^1H NMR (400 MHz, DMSO-d_6) δ 12.32 (s, 1H), 9.10 (d, *J* = 8.3 Hz, 1H), 8.87 (s, 1H), 8.60 (d, *J* = 1.8 Hz, 1H), 8.30 (d, *J* = 7.7 Hz, 1H), 7.97 (d, *J* = 7.7 Hz, 1H), 7.82-7.76 (m, 4H), 7.71 (t, *J* = 2.9 Hz, 1H), 7.66 (t, *J* = 7.8 Hz, 1H), 7.45 (d, *J* = 8.6 Hz, 2H), 7.37 (d, *J* = 8.5 Hz, 2H), 6.93 (dd, *J* = 3.7, 1.6 Hz, 1H), 5.09 (q, *J* = 7.7 Hz, 1H), 3.79-3.64 (m, 2H), 2.25 (q, *J* = 7.1 Hz, 2H).

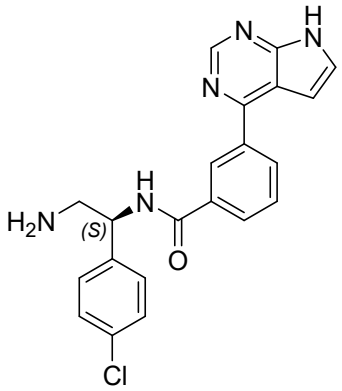
(S)-N-(3-amino-1-(4-chlorophenyl)propyl)-3-(7H-pyrrolo[2,3-d]pyrimidin-4-yl)benzamide(1c)



In a similar manner to that described for **1b**, by use of compound **18b** (110 mg; 0.20 mmol), **1c** (51 mg, yield: 61.22%) was obtained as a beige solid: ^{1}H NMR (400 MHz, DMSO-d_6) δ 9.26 (d, *J* = 7.9 Hz, 1H), 8.88 (s, 1H), 8.63 (s, 1H), 8.32 (d, *J* = 7.8 Hz, 1H), 8.03 (d, *J* = 7.8 Hz, 1H), 7.73 – 7.67 (m, 2H), 7.42 (q, *J* = 8.5 Hz, 4H), 6.89 (d, *J* = 3.6 Hz, 1H), 5.19 (q, *J* = 7.4 Hz, 1H), 2.64-2.57 (m, 2H), 1.99-1.91 (m, 1H), 1.88-1.80 (m, 1H). ^{13}C NMR (151MHz, DMSO-*d*6) δ 166.10, 155.34, 153.08, 151.39, 143.52, 138.49, 135.59, 131.68, 131.58, 129.40, 129.25, 128.90, 128.64, 128.51, 127.83, 115.05, 100.25, 79.63, 51.40, 39.10. $[\alpha]_D^{25}$= -15.41(c=0.011, MeOH). HPLC analysis: MeOH-H_2O (80:20), 9.45 min, 100 % purity.

HRMS (ESI) for $C_{22}H_{20}ClN_5O$ $[M+H]^+$, calcd: 406.1429; found: 406.1418.

(S)-N-(2-amino-1-(4-chlorophenyl)ethyl)-3-(7H-pyrrolo[2,3-d]pyrimidin-4-yl)benzamide(1d)



In a similar manner to that described for **1b**, by use of compound **18a** (125 mg，0.24 mmol) , **1d** (45 mg, yield: 47.95 %) was obtained as a yellow solid:^{1}H NMR (400 MHz, DMSO-*d*6) δ 8.95 (d, *J* = 7.8 Hz, 1H), 8.88 (s, 1H), 8.66 (d, *J* = 2.0 Hz, 1H), 8.32 (d, *J* = 7.8 Hz, 1H), 8.08 (d, *J* = 7.7 Hz, 1H), 7.74 – 7.66 (m, 2H), 7.46 – 7.37 (m, 4H), 6.89 (d, *J* = 3.6 Hz, 1H), 5.02 – 4.95 (m, 1H), 2.98 – 2.83 (m, 2H).^{13}C NMR (151 MHz, DMSO-*d*6) δ 166.67, 155.40, 153.08, 151.39, 141.97, 138.45, 135.68, 131.72, 131.67, 129.38, 129.33, 129.06, 128.62, 128.49, 127.95, 115.07, 100.27, 57.02, 47.47. $[\alpha]_D^{25}$= -14.88(c= 0.006, MeOH). HPLC analysis: MeOH-H_2O (80:20), 11.45 min, 100 % purity.

HRMS (ESI) for $C_{21}H_{18}ClN_5O$ $[M+H]^+$, calcd: 392.1273; found: 392.1263.

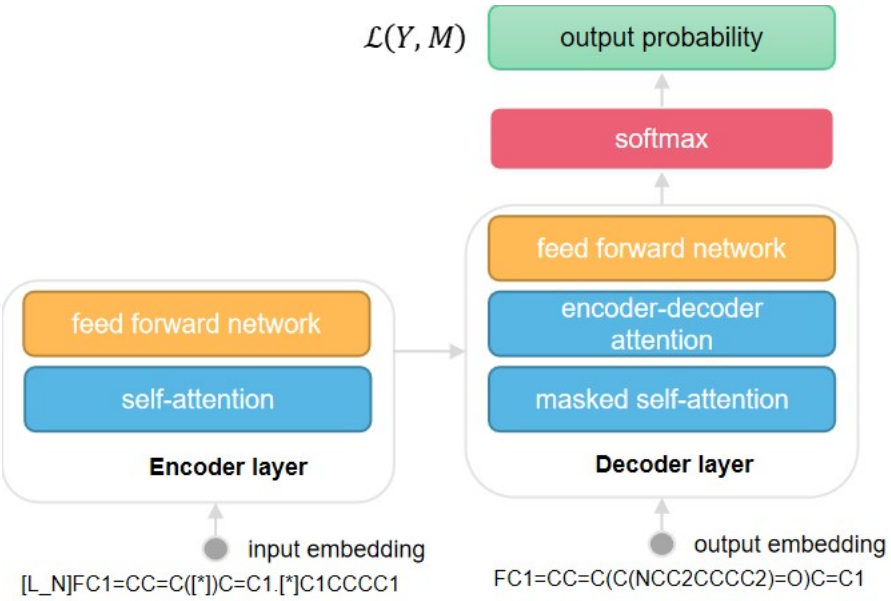


Fig. S1 A simple flowchart of SyntaLinker's transformer architecture mainly comprising encoder and decoder layers. The SMILES[8] string at the lower left represents a terminal fragment pair, while that at the lower right represents the original compound. In the input and output embedding layers, the SMILES strings are both coded as a one-hot matrix for processing in encode and decode layer. The decoder output is fitted to a distribution using the softmax function. Additionally, we introduce a character L_N as a distance constraint at the beginning of the source sequence to regulate the length of linker fragment in the output structure. N is the shortest linker bond number between the two terminal fragments in the original compound. $L(Y,M)$ represents the loss function, where Y is the target sequence and M is the output sequence.

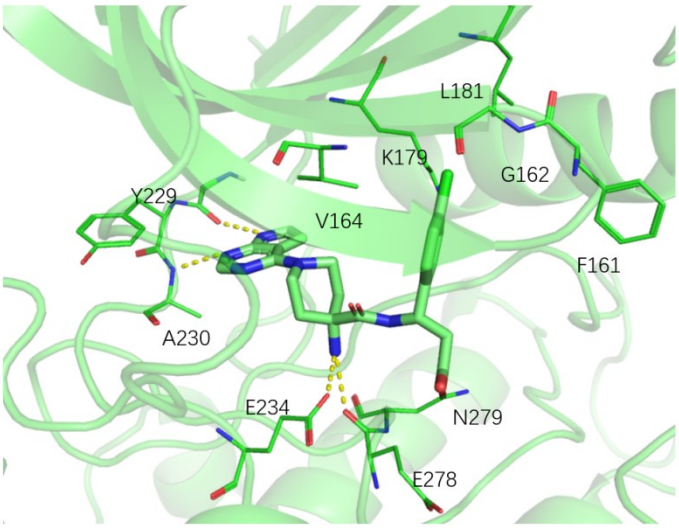


Fig. S2 Crystal structure of AZD5363 in complexed with Akt1 (PDB: 4GV1). The ligand was shown as green sticks. Cartoon represents the secondary structure of protein, in which residues were shown as yellow lines. Yellow dashes represented hydrogen bonds.

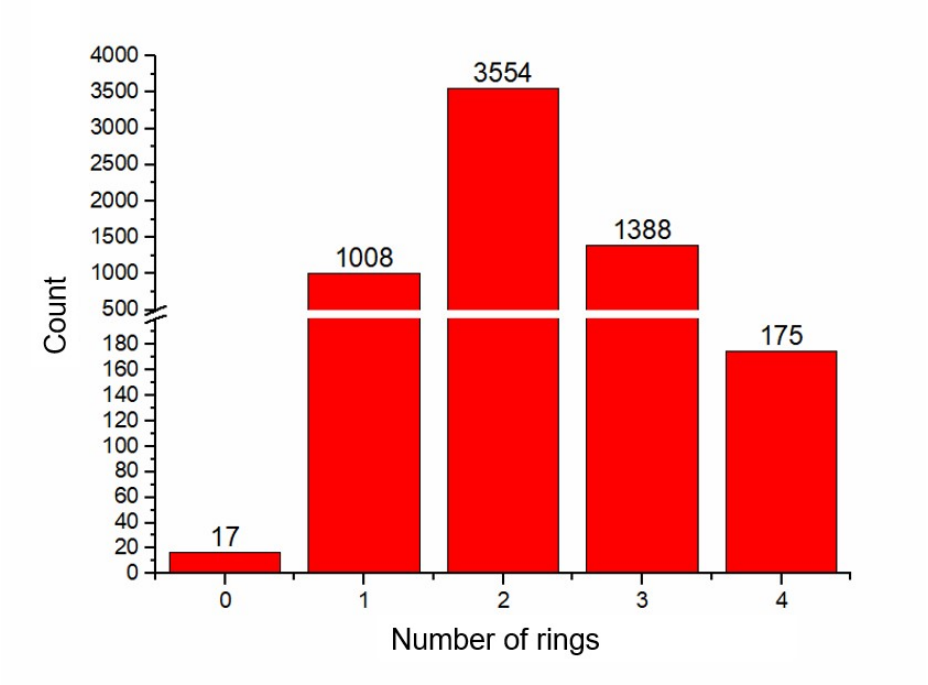


Fig. S3 Number of rings in the generated linkers.

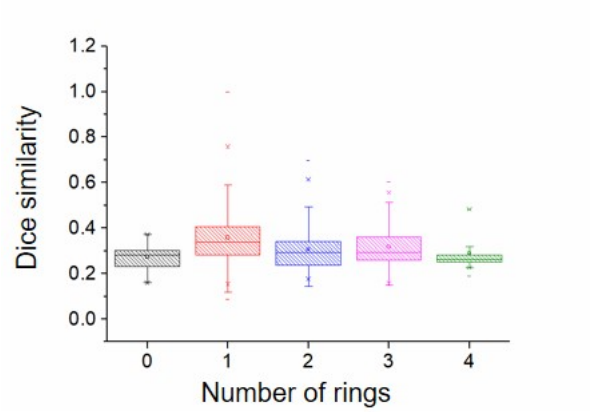


Figure S4. Dice similarity between generated linker fragments and the 4-aminopiperidine-4-carboxamide linker of AZD5363. The X axis represents the number of rings contained in the generated linker fragments. The Y axis represents the Dice similarity coefficients, which were calculated based on molecular fingerprint using RDKit[9] package in Python.

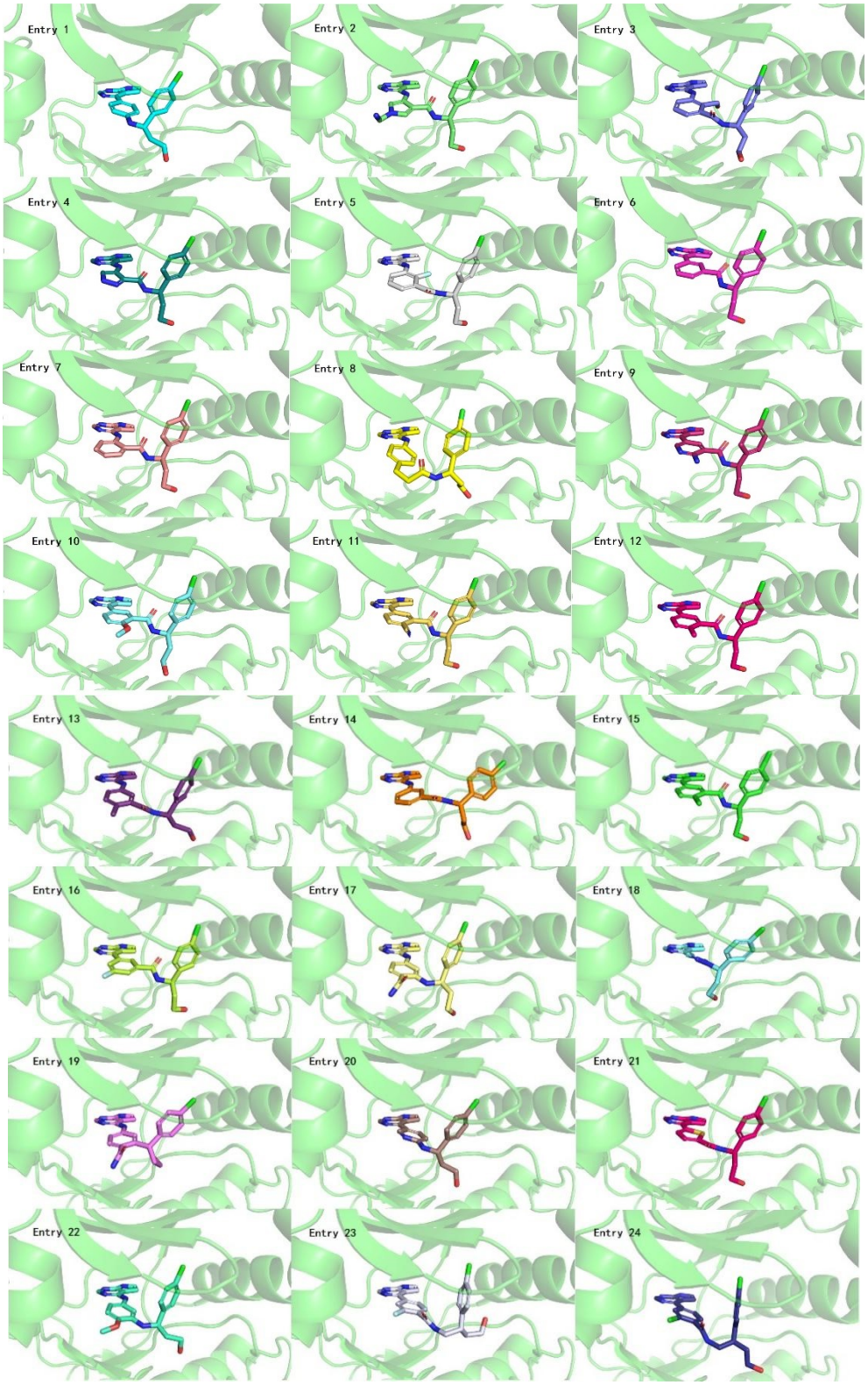
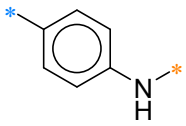
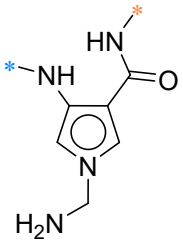
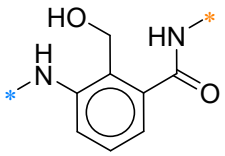
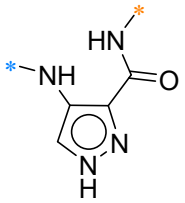
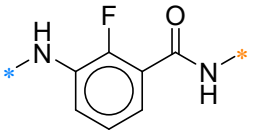
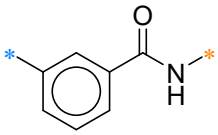
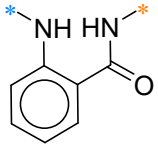
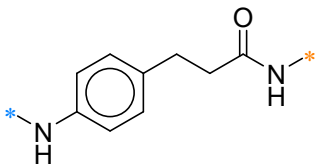
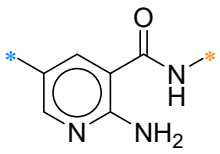
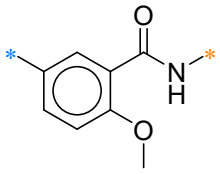
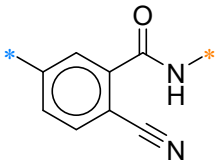
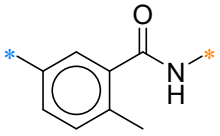
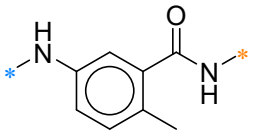
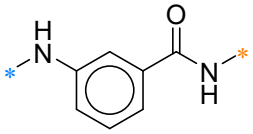
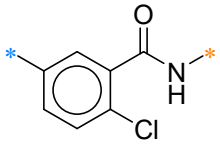
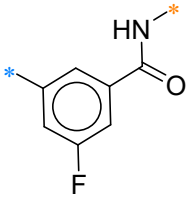
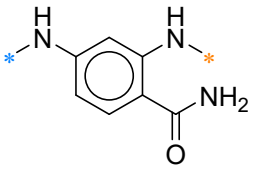
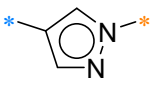
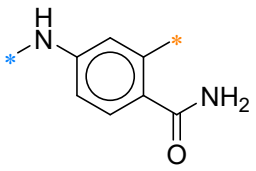
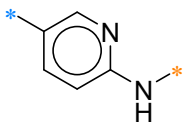
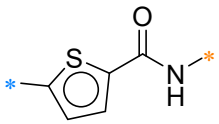
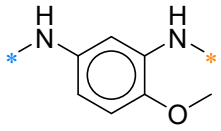
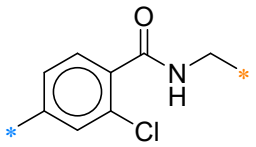
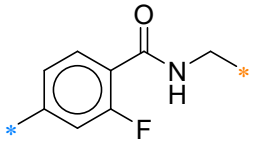


Figure S5. Predicted binding modes of the 24 generated structures in complexed with Akt1 (PDB: 4GV1). The ligand was shown as green sticks.

Table S1. SMILES strings and structures of 24 privileged aromatic linker fragments in the one ring group.

Entry	Smiles	Structure
1	[*]Nc1ccc(cc1)[*]	
2	[*]NC(=O)c1cn(cc1N[*])CN	
3	[*]NC(=O)c1cccc(c1CO)N[*]	
4	[*]NC(=O)c1n[nH]cc1N[*]	
5	[*]NC(=O)c1cccc(c1F)N[*]	
6	[*]NC(=O)c1cccc(c1)[*]	
7	[*]NC(=O)c1ccccc1N[*]	
8	[*]NC(=O)CCc1ccc(cc1)N[*]	

9	[*]NC(=O)c1cc(cnc1N)[*]	
10	[*]NC(=O)c1cc(ccc1OC)[*]	
11	[*]NC(=O)c1cc(ccc1C#N)[*]	
12	[*]NC(=O)c1cc(ccc1C)[*]	
13	[*]NC(=O)c1cc(ccc1C)N[*]	
14	[*]NC(=O)c1cccc(c1)N[*]	
15	[*]NC(=O)c1cc(ccc1Cl)[*]	
16	[*]NC(=O)c1cc(cc(c1)F)[*]	
17	[*]Nc1ccc(c(c1)N[*])C(=O)N	
18	[*]c1cnn(c1)[*]	

19	[*]Nc1ccc(c(c1)[*])C(=O)N	
20	[*]Nc1ccc(cn1)[*]	
21	[*]NC(=O)c1ccc(s1)[*]	
22	[*]Nc1ccc(c(c1)N[*])OC	
23	[*]CNC(=O)c1ccc(cc1Cl)[*]	
24	[*]CNC(=O)c1ccc(cc1F)[*]	

* represents the anchoring sites on the linker fragments in charge of connection with terminal fragments.

References

1 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Adv. Neural Inf. Process. Syst.*, 2017, **2017**-**Decem**, 5999–6009.

2 J. Lim, S. Y. Hwang, S. Moon, S. Kim and W. Y. Kim, *Chem. Sci.*, 2020, **11**, 1153–1164.

3 Y. Yang, S. Zheng, S. Su, C. Zhao, J. Xu and H. Chen, *Chem. Sci.*, 2020, **11**, 8312–8322.

4 *,† Richard A. Friesner, ‡ Jay L. Banks, ‡ Robert B. Murphy, ‡ Thomas A. Halgren, ‡,‖ Jasna J. Klicic, ‡ Daniel T. Mainz, ‡ Matthew P. Repasky, † Eric H. Knoll, § Mee Shelley, § Jason K. Perry, # David E. Shaw, § and Perry Francis and P. S. Shenkin‡, *J. Med. Chem.*, 2004, **47**, 1739–1749.

5 *,† Thomas A. Halgren, † Robert B. Murphy, ‡ Richard A. Friesner, † Hege S. Beard, § Leah L. Frye, † and W. Thomas Pollard and J. L. Banks†, *J. Med. Chem.*, 2004, **47**, 1750–1759.

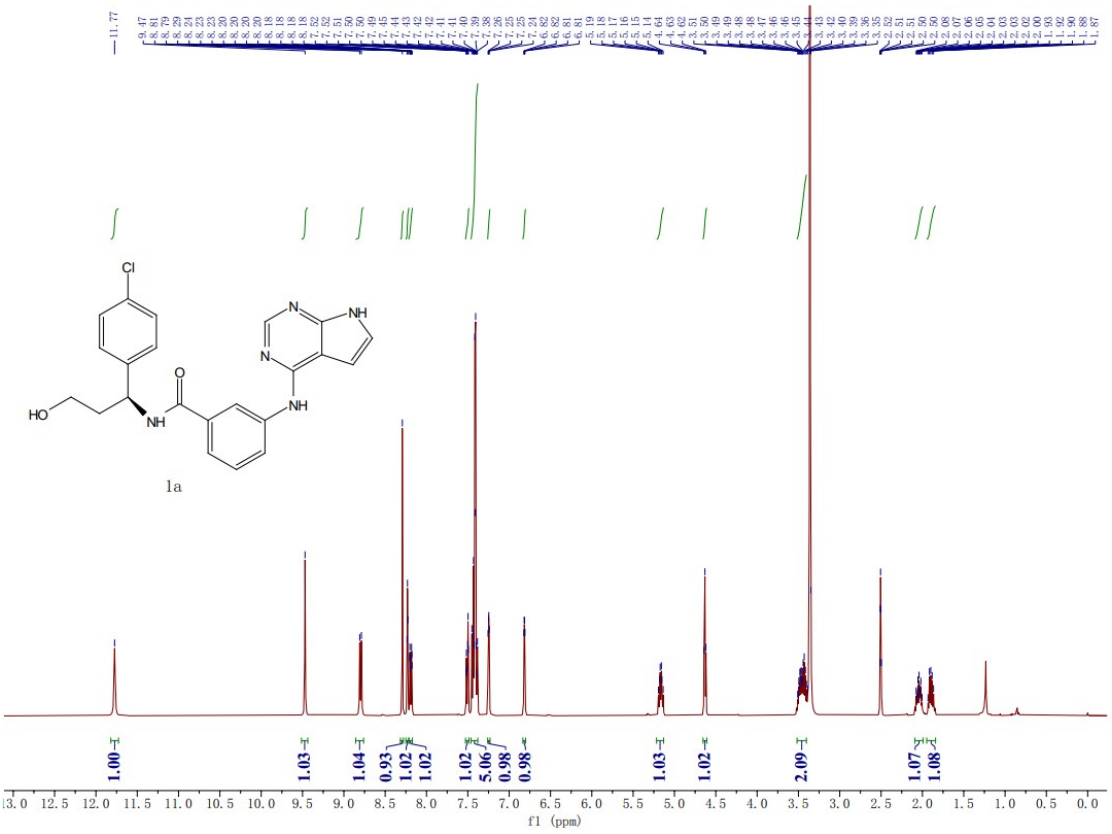
6 J. R. Greenwood, D. Calkins, A. P. Sullivan and J. C. Shelley, *J. Comput. Mol. Des. 2010 246*, 2010, **24**, 591–604.

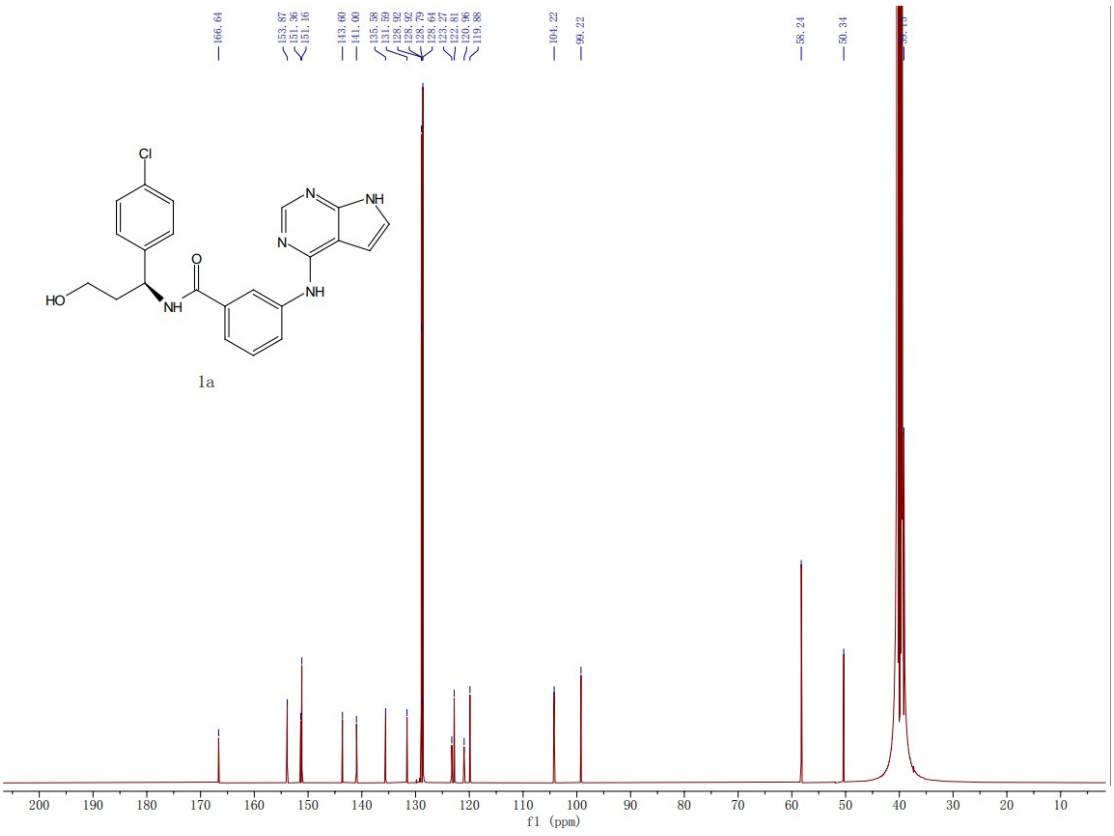
7 J. Zou, Z. Zhang, F. Xu, S. Cui, C. Qi, J. Luo, Z. Wang, X. Lu, Z. Tu, X. Ren, L. Song and K. Ding, *J. Drug Target.*, 2019, **27**, 442–450.

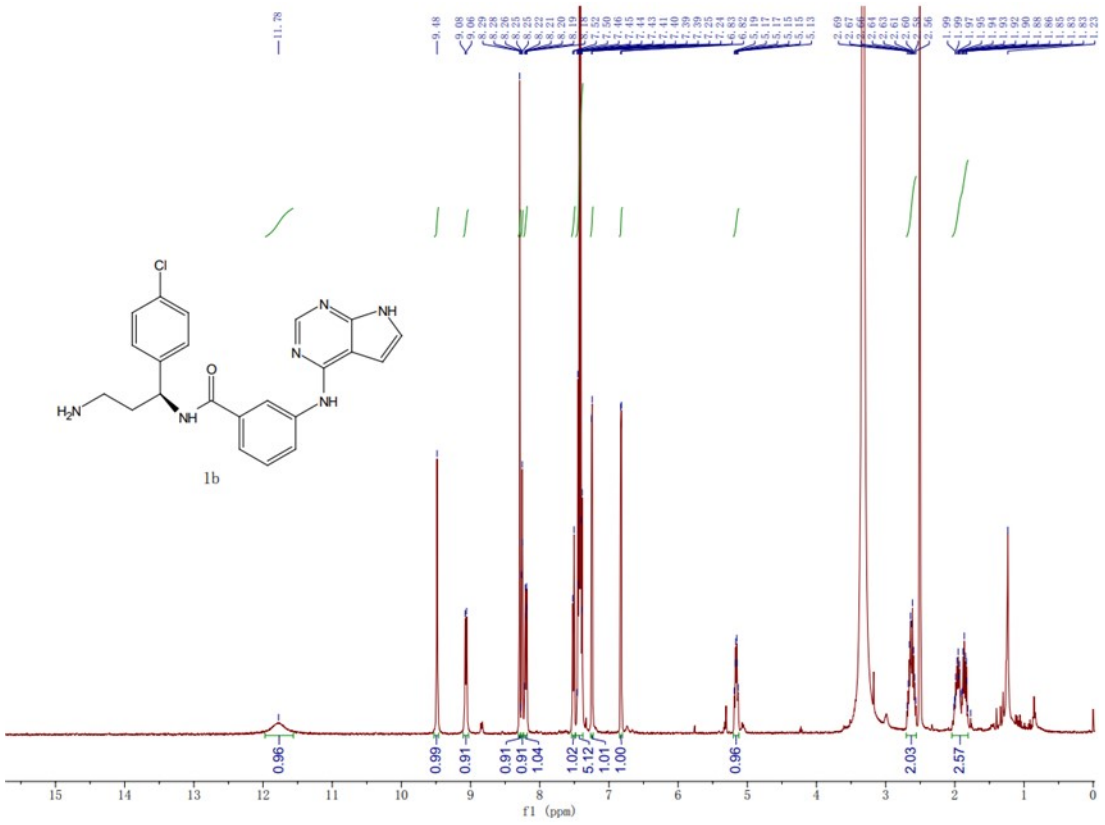
8 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.

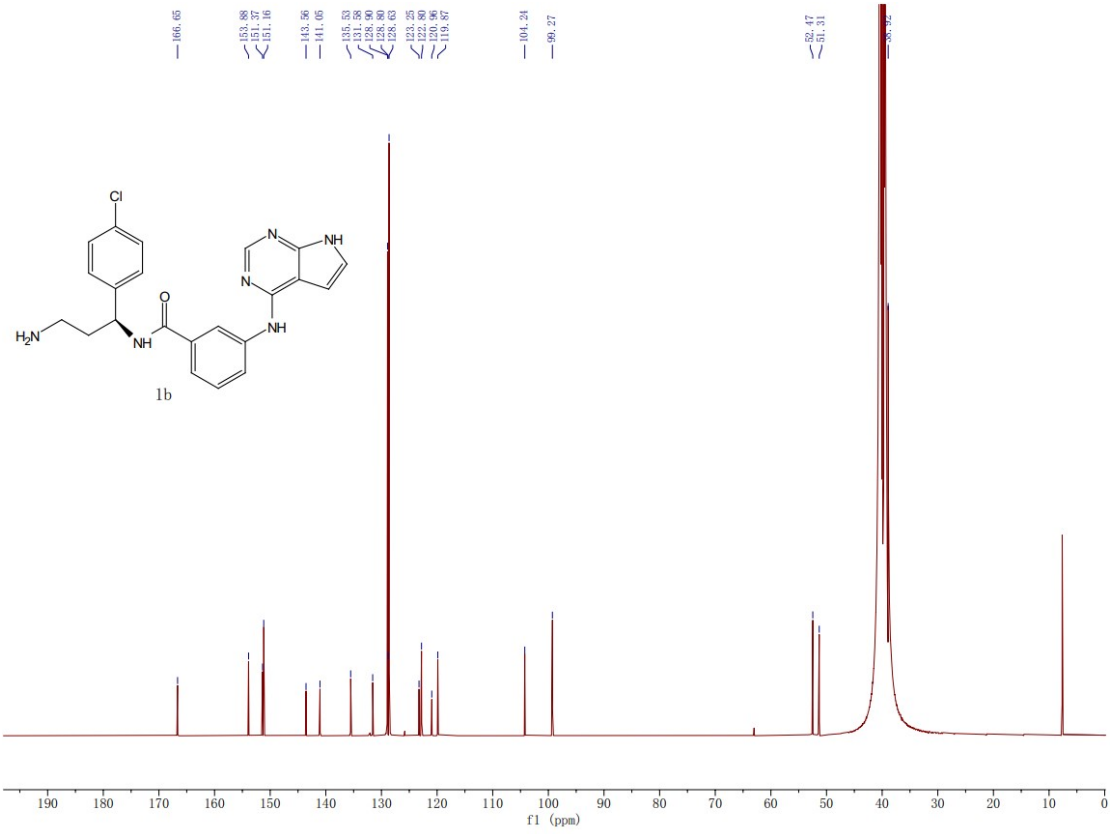
9 RDKit: Open-source cheminformatics; http://www.rdkit.org

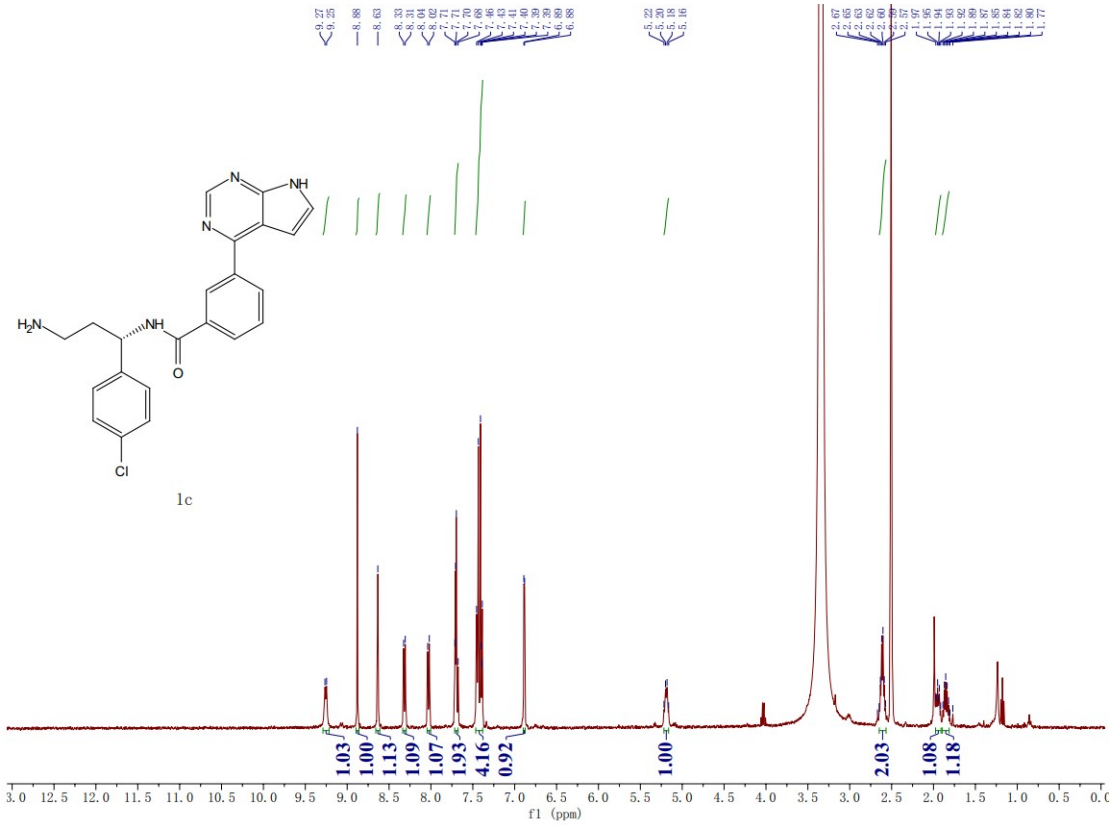
The ^{1}H and ^{13}C NMR spectra of compounds 1a-1d

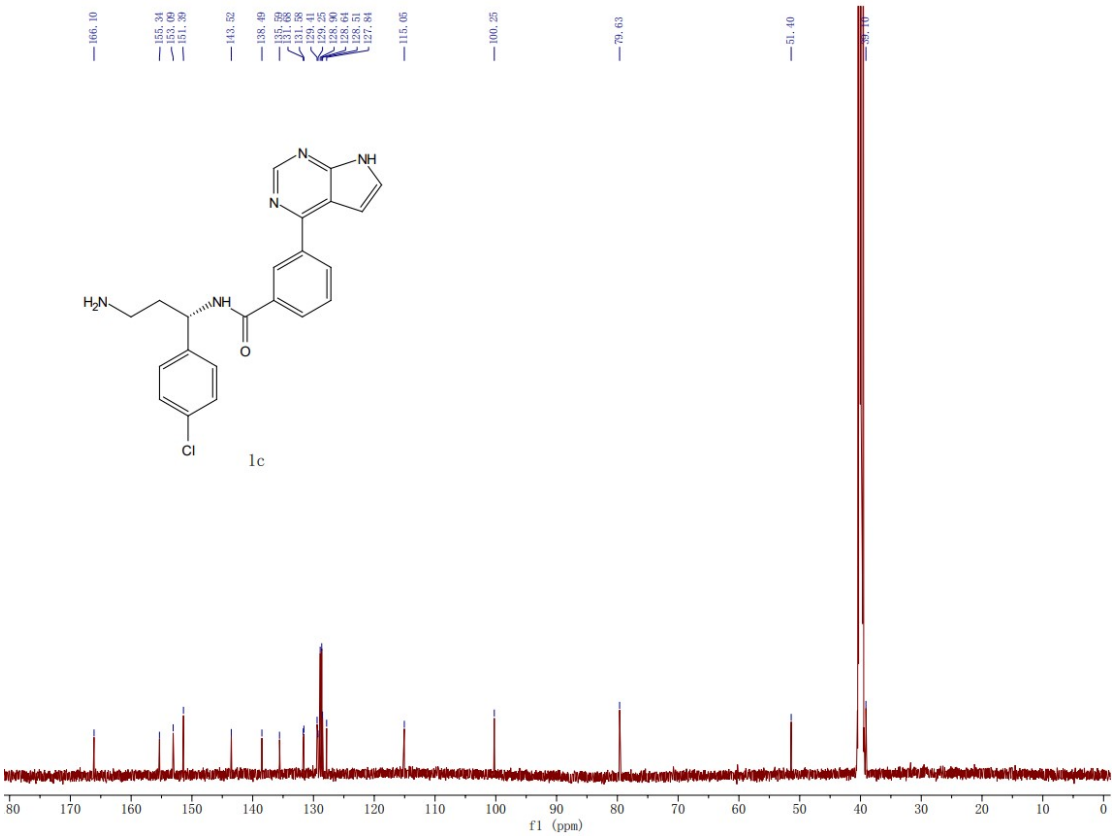


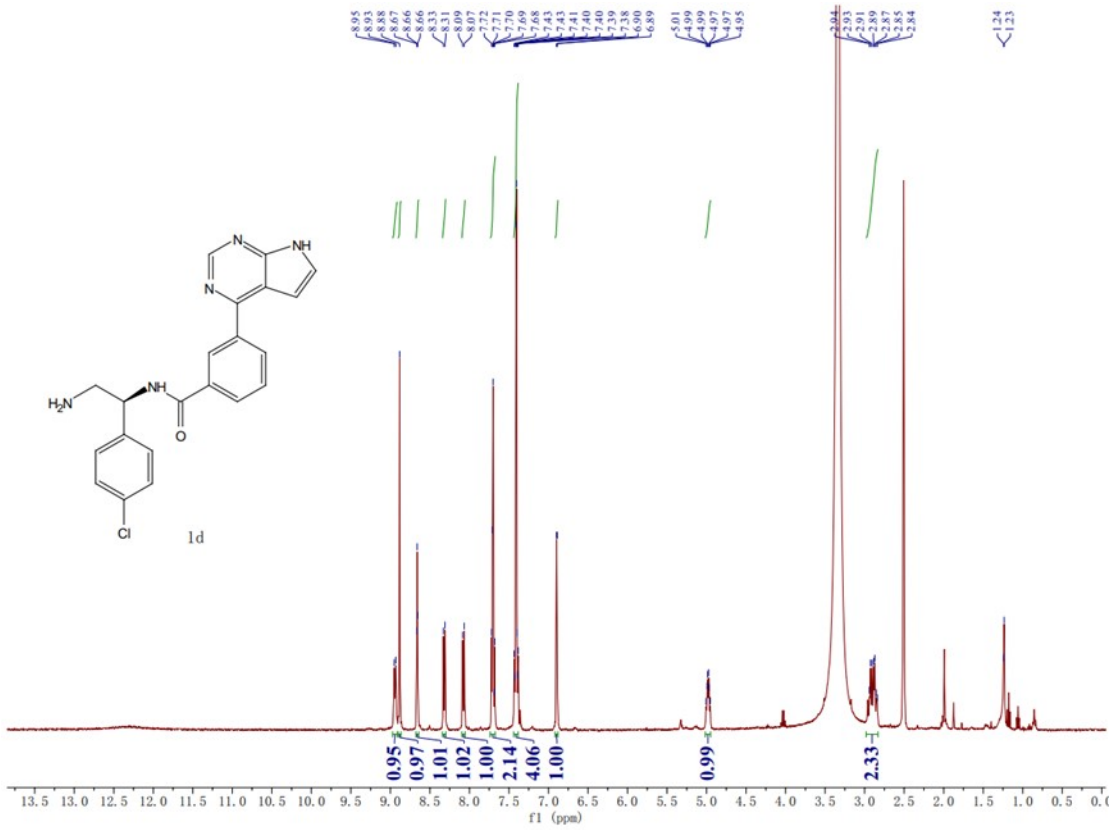


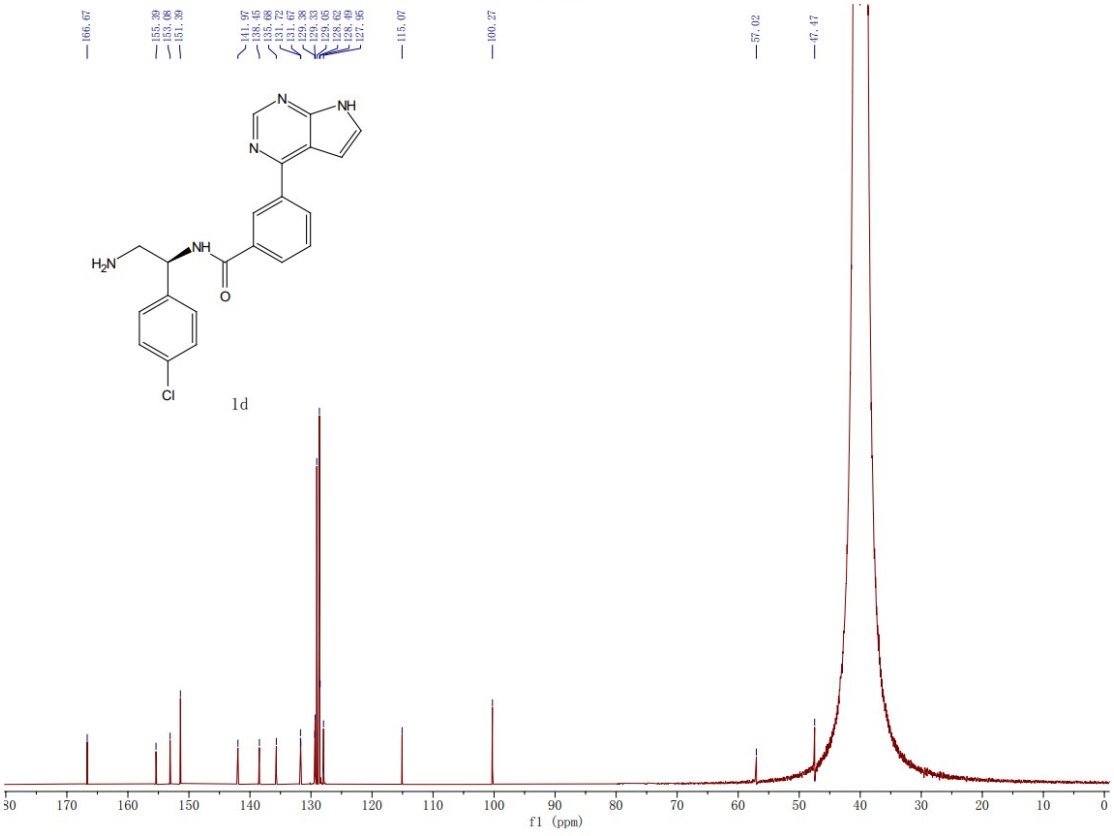

11.78
9.48
9.08
8.29
8.26
8.25
8.23
8.21
8.20
8.19
7.53
7.50
7.46
7.45
7.44
7.43
7.41
7.40
7.39
7.35
7.24
6.83
6.82
5.19
5.17
5.15
5.15
5.13
2.69
2.67
2.66
2.64
2.63
2.61
2.60
2.59
2.56
1.99
1.97
1.95
1.94
1.93
1.90
1.88
1.86
1.85
1.83
1.23
Cl
N
NH
N
O
NH
NH
H2N
1b
0.96
0.99
0.91
0.91
0.91
1.04
1.02
5.12
1.01
1.00
0.96
2.03
2.57
f1 (ppm)


166.65
153.88
151.57
151.16
143.36
141.05
135.53
131.38
128.90
128.80
128.63
123.25
122.80
120.96
119.87
104.24
99.27
52.47
51.31
38.32
Cl
N
NH
N
O
NH
NH
H2N
1b
f1 (ppm)

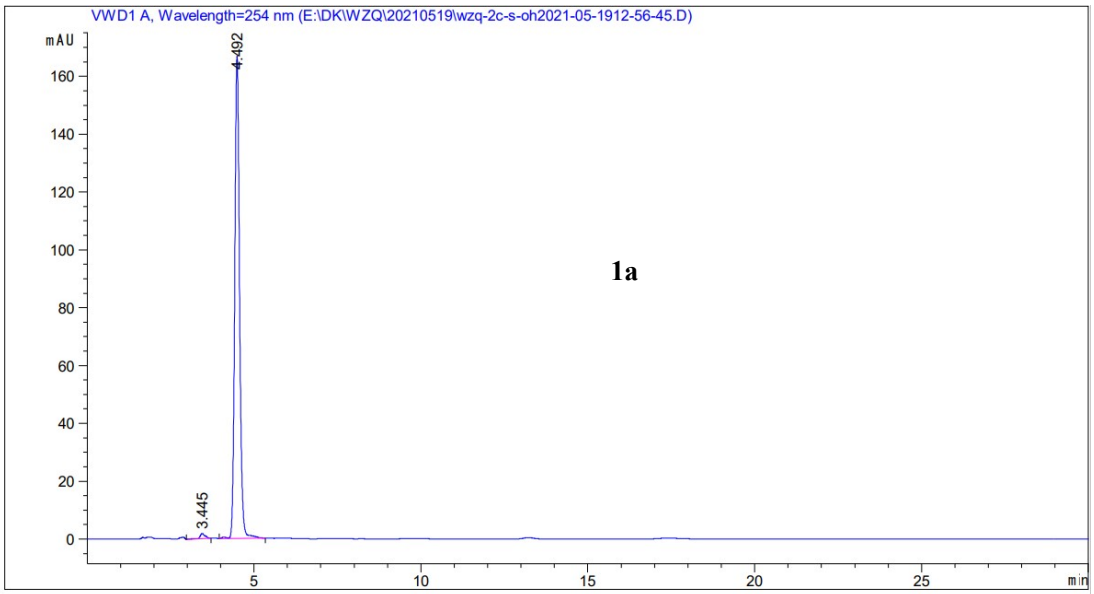

9.27
9.25
8.88
8.63
8.33
8.31
8.04
8.02
7.71
7.70
7.68
7.66
7.65
7.41
7.40
7.39
7.39
6.89
6.88
5.22
5.20
5.18
5.16
2.67
2.65
2.63
2.62
2.60
2.59
2.57
1.97
1.96
1.94
1.93
1.92
1.89
1.87
1.85
1.84
1.82
1.80
1.77
N
NH
N
H2N
NH
O
Cl
1c
1.03
1.00
1.13
1.09
1.07
1.93
4.16
0.92
1.00
2.03
1.08
1.18
f1 (ppm)


166.10
155.34
153.89
151.39
143.32
138.49
135.59
131.68
131.38
129.41
129.15
128.90
128.64
128.51
127.84
115.05
100.25
79.63
51.40
39.10
N
NH
N
H2N
NH
O
Cl
1c
f1 (ppm)


8.95
8.93
8.88
8.87
8.66
8.66
8.33
8.31
8.09
8.07
7.72
7.71
7.70
7.69
7.68
7.43
7.43
7.41
7.41
7.40
7.40
7.39
7.38
6.90
6.89
5.01
4.99
4.99
4.97
4.97
4.95
4.95
2.94
2.93
2.91
2.89
2.87
2.85
2.84
1.24
1.23
0.95
0.97
1.01
1.02
1.00
2.14
4.06
1.00
0.99
2.33
f1 (ppm)
1d


166.67
155.39
153.68
151.39
141.97
138.45
135.68
131.72
131.67
129.38
129.33
129.05
128.82
128.49
127.95
115.07
100.27
57.02
47.47
f1 (ppm)
1d

The HPLC and HRMC Spectra of Compounds 1a-1d



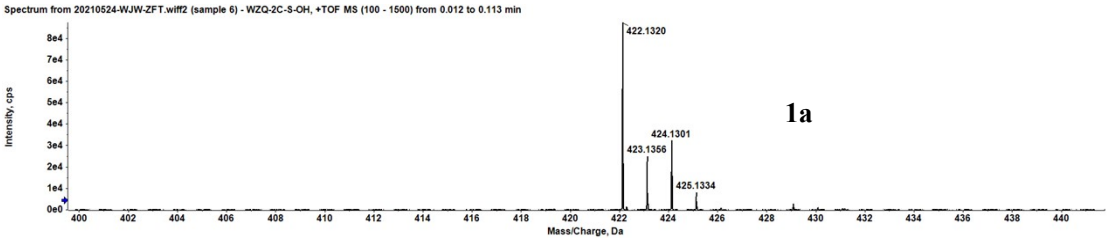
Area Percent Report

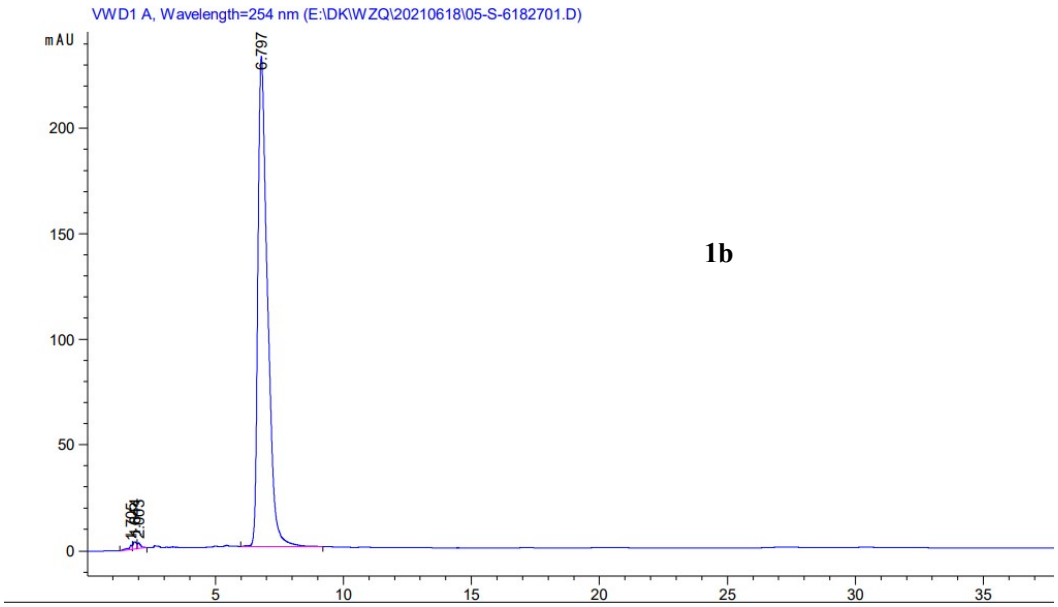
Sorted By : Signal
Multiplier : 1.0000
Dilution : 1.0000
Sample Amount: : 5.00000 [ng/ul] (not used in calc.)
Use Multiplier & Dilution Factor with ISTDs

Signal 1: VWD1 A, Wavelength=254 nm

Peak #	RetTime [min]	Type	Width [min]	Area [mAU*s]	Height [mAU]	Area %
1	3.445	BB	0.1537	19.61727	1.79211	1.1603
2	4.492	VB R	0.1529	1671.03589	166.75253	98.8397

Totals : 1690.65315 168.54464





1b

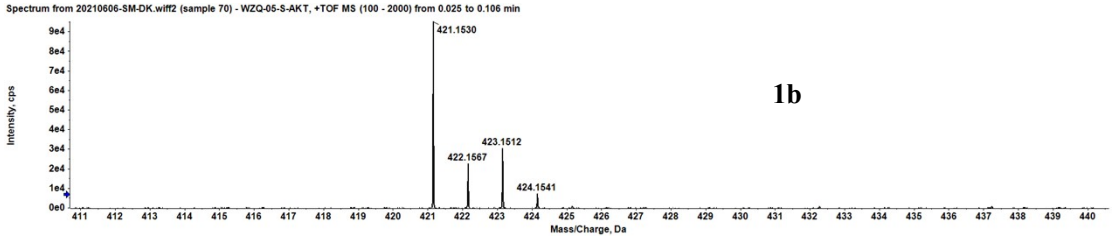
Area Percent Report

==========

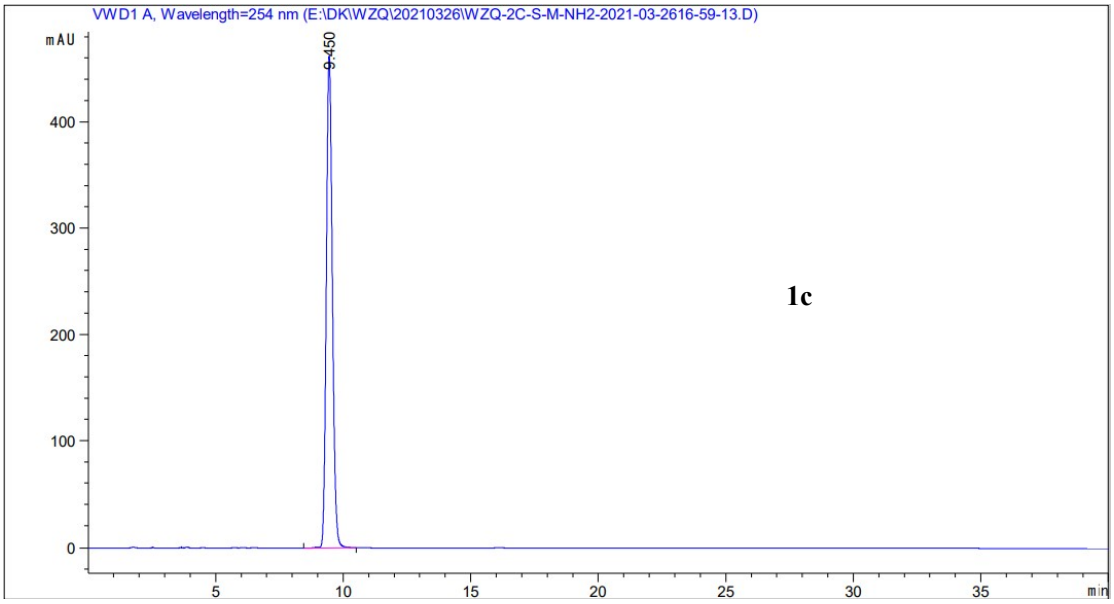
Sorted By : Signal
Multiplier : 1.0000
Dilution : 1.0000
Sample Amount: : 5.00000 [ng/ul] (not used in calc.)
Use Multiplier & Dilution Factor with ISTDs

Signal 1: VWD1 A, Wavelength=254 nm

Peak #	RetTime [min]	Type	Width [min]	Area [mAU*s]	Height [mAU]	Area %
1	1.705	BV	0.1320	19.47442	2.02897	0.3190
2	1.844	VV	0.1307	33.13063	3.45901	0.5427
3	2.003	VB	0.1167	23.00822	2.63992	0.3769
4	6.797	VB R	0.3823	6029.39893	231.93478	98.7615

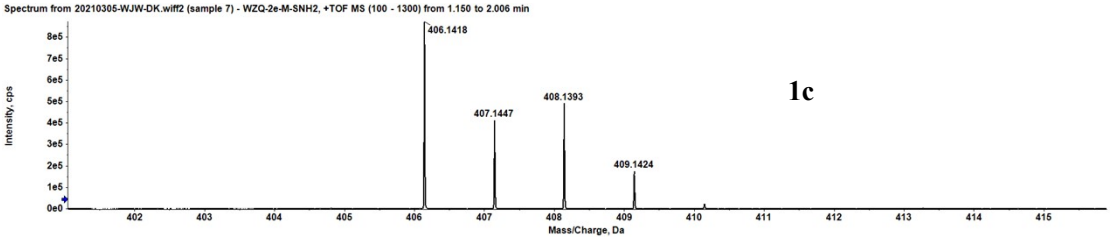


1b

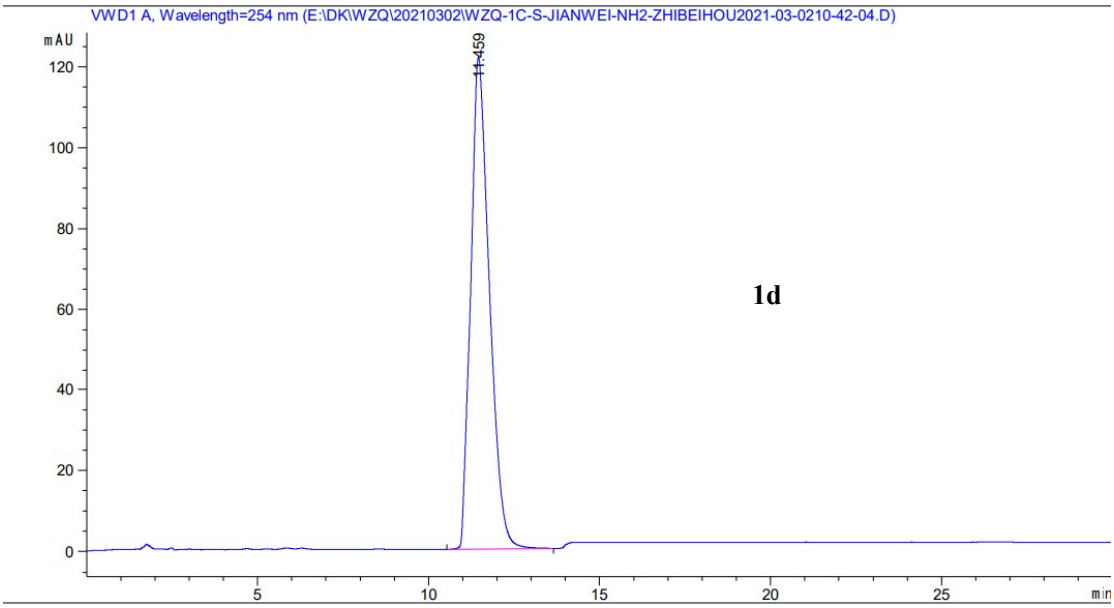


1c

```
================================================================================
                          Area Percent Report
================================================================================

Sorted By             :      Signal
Multiplier            :      1.0000
Dilution              :      1.0000
Sample Amount:                :      5.00000  [ng/ul]    (not used in calc.)
Use Multiplier & Dilution Factor with ISTDs


Signal 1: VWD1 A, Wavelength=254 nm

Peak RetTime Type  Width     Area      Height     Area
  #   [min]        [min]    [mAU*s]     [mAU]       %
----|-------|----|-------|----------|----------|--------|
   1   9.450 BB    0.2546 7583.10693  461.81387 100.0000

Totals :                   7583.10693  461.81387
```



1c




```
================================================================================
                          Area Percent Report
================================================================================

Sorted By             :      Signal
Multiplier            :      1.0000
Dilution              :      1.0000
Sample Amount:                :     10.00000  [ng/ul]    (not used in calc.)
Use Multiplier & Dilution Factor with ISTDs


Signal 1: VWD1 A, Wavelength=254 nm

Peak RetTime Type  Width     Area      Height     Area
  #   [min]        [min]   [mAU*s]     [mAU]       %
----|-------|----|-------|----------|----------|--------|
   1  11.459 BB    0.5513 4618.96631  121.84123 100.0000

Totals :                  4618.96631  121.84123
```