Drug design and repurposing with a sequence-to-drug paradigm

Lifan Chen[1,2,#], Zisheng Fan[1,3,#], Jie Chang[1,3,#], Ruirui Yang[1,2,4,#], Hao Guo[1], Yinghui Zhang[1,2], Tianbiao Yang[1,2], Chenmao Zhou[1,3], Zhengyang Chen[1,2], Chen Zheng[1,5], Xinyue Hao[1,3], Keke Zhang[1,3], Rongrong Cui[1], Yiluan Ding[6], Naixia Zhang[6], Xiaomin Luo[1,2], Hualiang Jiang[1,2,3,4,5, *], Sulin Zhang[1,2,*], Mingyue Zheng[1,2,3,4*]

[1]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China

[2]University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China

[3]School of Chinese Materia Medica, Nanjing University of Chinese Medicine, 138 Xianlin Road, Jiangsu, Nanjing 210023, China

[4]Shanghai Institute for Advanced Immunochemical Studies and School of Life Science and Technology, ShanghaiTech University, No. 393 Huaxia Middle Road, Shanghai 200031, China

[5]The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, 443 Huangshan Road, Hefei, Anhui 230027, China

[6]Department of Analytical Chemistry, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi

Road, Shanghai 201203, China

[#]Lifan Chen, Zisheng Fan, Jie Chang and Ruirui Yang contributed equally to this work.

*Correspondence should be addressed to:

Mingyue Zheng: myzheng@simm.ac.cn

Sulin Zhang: slzhang@simm.ac.cn

Hualiang Jiang: hljiang@simm.ac.cn

Abstract

Drug development based on target proteins has been a successful approach in recent decades. A conventional structure-based drug design pipeline is a complex, human-engineered pipeline with multiple independently optimized steps. Advances in end-to-end differentiable learning suggest the potential benefits of similarly reformulating drug design. Here, we proposed a new sequence-to-drug paradigm that discovers drug-like small-molecule modulators directly from protein sequences and validated this concept for the first time in three stages. First, we designed TransformerCPI2.0 as a core tool for the sequence-to-drug paradigm, which exhibited competitive performance with conventional structure-based drug design approaches. Second, we validated the binding knowledge that TransformerCPI2.0 has learned. Third, we applied a sequence-to-drug paradigm to discover new hits for E3 ubiquitin-protein ligases: speckle-type POZ protein (SPOP) and ring finger protein 130 (RNF130) which does not have a 3D structure. This first proof of concept shows that the sequence-to-drug paradigm is a promising direction for drug development.

Introduction

Drug development based on proteins has been a successful approach in the past decades, for diseases with well-defined protein targets[1-3]. A typical protein structure-based drug design (SBDD) project starts from the protein sequence, builds a three-dimensional (3D) structure by structure biology or structure prediction, identifies binding pockets (orthosteric sites or allosteric sites), and finally discovers active modulators through

virtual screening or de novo design[4,5] (Fig. 1a). It involves a complex, human-engineered pipeline with multiple independently optimized steps, and each step has its own limitations[5]. First, a large number of proteins do not have high-resolution structures. Recently, the great success of AlphaFold[6] and RoseTTAFold[7] in protein structure prediction has given an overoptimistic view that the issue has been fully solved. However, not all the predicted structures are suitable for SBDD[8,9], given that only 36% of all residues have very high confidence[10]. In particular, the precise structure prediction of active sites remains an unsolved challenge, since these local structures tend to break the 'protein-folding rules'[9]. Second, prior knowledge about binding pockets is essential for SBDD, and it is a nontrivial issue to define such pockets for novel targets with multiple domains[11]. Moreover, predicting allosteric sites is still challenging[12] due to the varied mechanisms of allosteric effects and high computational costs[13]. In addition, structural flexibility allows proteins to adapt to their individual molecular binders and undergo different internal motions[9,14,15], making pockets more difficult to define. Third, virtual screening can produce false positives[16] and may rapidly accumulate errors from the first two steps.

As an alternative to the conventional pipeline, we proposed a new sequence-to-drug paradigm that discovers modulators directly from protein sequences, jumping over the intermediate steps, by end-to-end differentiable learning (Fig. 1a). End-to-end differentiable deep learning has revolutionized computer vision and speech recognition[17] by replacing all components of complex pipelines with differentiable primitives to enable joint optimization from input to output[18]. The success of

89 AlphaFold[6] also relies heavily on the idea of end-to-end differentiability. Therefore, this
90 new paradigm is appealing because it performs the entire learning process in a self-
91 consistent and data-efficient manner, avoiding the error accumulation of complex
92 pipelines.

93 Many deep learning models have been proposed to use protein sequences as
94 input[19-28], but none have thoroughly verified the concept of the sequence-to-drug
95 paradigm. In this work, we address the issue for the first time through three stages (Fig.
96 1b). First, we designed TransformerCPI2.0, which exhibited competitive performance
97 against 3D structure-based approaches such as molecular docking, as a fundamental
98 tool of the sequence-to-drug paradigm. Second, we tested whether TransformerCPI2.0
99 learns knowledge as expected, rather than exhibiting only data bias[20]. Limited effort
100 has been devoted to investigating whether such a model learns in the manner we
101 expected. For example, in the interaction between a receptor and a ligand, a tiny
102 structural change in either the protein or the compound may lead to drastic changes in
103 the interaction, but we do not know whether existing compound-protein interaction
104 (CPI) models can capture such subtle changes and reflect them in the prediction results.
105 Our previous analyses revealed that although some models achieved good performance
106 on some datasets, they may not actually utilize protein-related information to make
107 predictions[20] and thus cannot generalize to mutants or new proteins. To judge whether
108 a model truly has the ability to generalize, we need to examine whether it merely
109 remembers the data distribution or learns the intermolecular recognition and regulation
110 we are interested in. Therefore, we designed two computational analyses, i.e., drug

resistance mutation analysis and molecular substitution effect analysis of the trifluoromethyl group, to validate the sensitivity of our model to single point mutations of proteins and activity cliffs of molecules. Third, we applied a sequence-to-drug paradigm to discover new hits for challenging targets, speckle-type POZ protein (SPOP) and ring finger protein 130 (RNF130) which does not have existing 3D structures. After the first proof of concept, the sequence-to-drug paradigm appears to be a promising direction for rational drug design.

Results

The sequence-to-drug paradigm achieves competitive performance against molecular docking.

To build a model that can implement the sequence-to-drug paradigm, we developed TransformerCPI2.0 based on our previous work[20], and its framework is shown in Fig. 1c. To train a general model for practical application, we constructed a chEMBL dataset containing 217,732 samples in the training set, 24,193 samples in the validation set and 10,199 in the test set. TransformerCPI[20], CPI-GNN[19], GraphDTA(GAT-GCN)[21], and GCN[21] were selected as baseline models, and all were retrained on the chEMBL dataset. We trained TransformerCPI2.0 and baseline models under the same criteria and compared their performance in terms of the area under the receiver operating characteristic curve (AUC) and the area under the precision recall curve (PRC) (Extended Data Fig. 1a~c). TransformerCPI2.0 achieves the best performance among all the models. In addition, we tested TransformerCPI2.0 and the baseline models on

133 the other three external datasets: GPCR set, Kinase set, and a large external set.
134 TransformerCPI2.0 also showed the greatest generalization ability among all the
135 models (Extended Data Fig. 1d~i and Supplementary Table 1~3). Furthermore, we
136 designed a time-split test set named the chEMBL27 dataset to predict new data that
137 were deposited online after the training set, and TransformerCPI2.0 still outperformed
138 the baseline models (Extended Data Fig. 1j, k). Since our training set was generated
139 from chEMBL23, this test suggested that our model can learn from past knowledge and
140 generalize to future data.

141 We also compared TransformerCPI2.0 with conventional SBDD approaches for a
142 better understanding of its capacity to enrich active molecules from large-compound
143 collections. On two benchmark docking datasets, DUD-E[29] and DEKOIS 2.0[30],
144 TransformerCPI2.0 achieved competitive performance against all the tested structure-
145 based docking programs except Glide SP (Fig. 1d, Supplementary Table 4~5).
146 Furthermore, TransformerCPI2.0 yielded satisfactory performance compared with
147 some well-known machine learning-based scoring functions (Fig. 1e, Supplementary
148 Tables 6~7). Overall, our method does not depend on the protein 3D structure yet
149 achieves performance comparable to that of 3D structure-based molecular docking, and
150 in subsequent studies, it will be used to solve virtual screening and target identification
151 tasks.

152

153 **Sensitivity to drug resistance-associated single point mutations of protein.**

154 To further test whether TransformerCPI2.0 is sensitive to tiny local modifications of

protein sequences, we proposed an analysis method named drug resistance mutation analysis, which mimics alanine scanning[31]. Briefly, we mutated each amino acid of the given protein sequence one by one and investigated whether the prediction score changed significantly. This analysis mimics the process of drug resistance in nature, where a single amino acid mutation may cause altered susceptibility to a drug.

We selected HIV-1 reverse transcriptase and its inhibitor doravirine as an example. Doravirine (formerly MK-1439) was approved by the FDA for the treatment of HIV-infected, treatment-naive individuals in combination with other antiretroviral drugs[32]. The cocrystal structure of HIV-1 reverse transcriptase and doravirine has been reported (PDB: 4NCG, Fig. 2a), and we conducted drug resistance mutation analysis and revealed the mutation effect of each amino acid position (Fig. 2c). It is encouraging that positions with a high relative activity change score (ΔR) are highly overlapped with the binding sites of doravirine (Fig. 2a, b), highlighting that TransformerCPI2.0 has implicitly learned information about binding site locations, since neither structure nor binding pocket information is included in the training phase. Among the mutations, P225 is predicted to be the most important site in HIV-1 reverse transcriptase (Fig. 2c) and actually plays an important role in the binding of doravirine (Fig. 2b). Additionally, P225, F227, L234 and P236 have been reported as drug resistance mutation sites[33-36] and are correctly retrieved as important sites by TransformerCPI2.0. Furthermore, we analysed the specific mutation patterns predicted by TransformerCPI2.0 and found that some predictions matched the ground truth (Fig. 2d): P225H, F227C/L/R and P236L were correctly predicted by TransformerCPI2.0.

Another possible concern is that the model might learn drug sensitivity-related residue information from the protein sequence alone rather than from the protein–ligand interactions that we expect it to learn[20]. We selected aspirin as a negative control (Extended Data Fig. 2a, b) and found that the pattern of ΔR was significantly different from that of doravirine. In conclusion, drug resistance analysis validated that TransformerCPI2.0 is sensitive to single point mutations of proteins and thus, similar to 3D structure-based docking methods, captures key residues for protein ligand binding.

Sensitivity to activity cliffs of small molecules.

To investigate whether TransformerCPI2.0 is sensitive to tiny modifications of ligands, we designed substitution effect analysis of the trifluoromethyl group as an example. Activity cliffs are generally understood as pairs or groups of similar compounds with large differences in potency[37,38]. Recently, Abula *et al*[39] proposed a dataset including 18,217 pairs of compounds and corresponding bioactivity data, with the only difference being that −CH3 is replaced by −CF3. Only 15.73% of substitutions of −CF3 for −CH3 could increase or decrease the biological activity by at least an order of magnitude, and an example is shown (Fig. 2e).

We computed the activity change score Δs_c and conducted substitution effect analysis on this part of the data. TransformerCPI2.0 reveals higher predictive ability than TransformerCPI, GraphDTA, GCN and random guessing (Fig. 2f and Supplementary Table 8). In addition, we investigated the performance of

199 TransformerCPI2.0 and baseline models on a subset where substitution of −CH3 by
200 −CF3 could increase or decrease the biological activity by at least three orders of
201 magnitude. TransformerCPI2.0 still outperformed the other models, and the accuracy
202 on this subset was higher than that on the whole dataset (Fig. 2g and Supplementary
203 Table 9). The predictive capacity of this test is much more challenging than that of the
204 whole dataset because the drastic biological activity change at this range involves
205 turning an active compound into an inactive one or vice versa.

206 Furthermore, we chose illustrative examples to demonstrate the capacity of
207 TransformerCPI2.0 (Extended Data Fig. 2c) to distinguish subtle structural differences
208 that produce drastic activity changes, where none of the protein targets and compounds
209 were in the training set. As shown, the activity change scores Δs_c predicted by
210 TransformerCPI2.0 are higher than 0.5. These four examples provide evidence that a
211 single substitution in compounds will significantly change the prediction score of
212 TransformerCPI2.0, which means that our model is sensitive enough to account for
213 activity cliffs of small molecules.

214

215 **Drug design targeting E3 ubiquitin-protein ligases.**

216 SPOP, functioning as an adaptor of cullin3-RING ubiquitin ligase, mediates substrate
217 protein recognition and ubiquitination[40,41]. A previous study validated SPOP as an
218 attractive target for the treatment of clear-cell renal cell carcinoma (ccRCC), but it is a
219 challenging target in terms of protein–protein interactions[42]. SPOP is overexpressed and
220 misallocated in the cytoplasm of ccRCC cells, which induces proliferation and

221 promotes kidney tumorigenesis[43]. Phosphatase and tensin homologue (PTEN) and dual
222 specificity phosphatase 7 (DUSP7) are two substrates of SPOP[43]. PTEN acts as a
223 negative regulator of the phosphoinositide 3-kinase/AKT pathway, and DUSP7
224 dephosphorylates extracellular signal-regulated kinase (ERK)[44]. Accumulation of
225 cytoplasmic SPOP in ccRCC cells decreases cellular PTEN and DUSP7 by mediating
226 the degradation of these two cytoplasmic proteins, leading to an increase in
227 phosphorylated AKT and ERK and thus promoting the proliferation of ccRCC cells[43].

228 Here, virtual screening with TransformerCPI2.0 was performed to discover new
229 scaffold compounds that directly target SPOP (Fig. 3a) to test the feasibility of the
230 sequence-to-drug paradigm on challenging targets. Four compounds were identified as
231 initial hits by a fluorescence polarization (FP) assay (hit rate ~5%), and 221C7 was the
232 most active compound with an IC_{50} of 4.30 μM (Fig. 3c). To further confirm that 221C7
233 disrupts SPOP-substrate interactions, an *in vitro* pull-down assay was performed. The
234 results revealed that compound 221C7 dose-dependently reduced PTEN protein
235 binding to the SPOP MATH domain ($SPOP^{MATH}$) (Fig. 3d). Then, a nuclear magnetic
236 resonance (NMR) experiment was conducted (Fig. 3e), and the result indicated direct
237 binding between $SPOP^{MATH}$ and 221C7. These results verified that 221C7 disrupts
238 SPOP-substrate interactions by directly binding to $SPOP^{MATH}$. Drug resistance mutation
239 analysis was applied to interpret the prediction of TransformerCPI2.0, and important
240 residues were mapped to the structure of full-length SPOP (PDB: 3HQI, Fig. 3f).
241 Important residues mostly located on the MATH domain, which consisted with
242 experiment result.

The initial hit 221C7 was inactive in cellular experiments, which might be attributable to poor cell permeability caused by its large topological polar surface area (TPSA)[45] of 214Å^2. Therefore, we conducted hit expansion and obtained 26 structural analogues of 221C7, 19 of which were active in the FP assay. Among them, 230D7 has a smaller TPSA (161Å^2) and the smallest IC_{50} of the FP assay (Fig. 4a~b), thus, it was selected for further validation. A protein thermal shift assay (PTS) revealed dose-dependent T_m shifts (Extended Data Fig. 3a), indicating that 230D7 could directly bind to $SPOP^{MATH}$ and destabilize the protein. Additionally, NMR experiments confirmed the direct binding between $SPOP^{MATH}$ and 230D7 (Extended Data Fig. 3b). An *in vitro* pull-down assay was performed to verify that 230D7 dose-dependently reduces PTEN binding to $SPOP^{MATH}$ (Extended Data Fig. 3c). After validating the molecular activity, we utilized 230D7 for the functional study at the cellular level.

We conducted a coimmunoprecipitation and *in vivo* ubiquitination experiment at the cellular level, and the results showed that 230D7 significantly disrupted the binding of PTEN and DUSP7 to SPOP in a dose-dependent manner (Fig. 4c, d), leading to decreases in PTEN and DUSP7 ubiquitination (Fig. 4e, f). Due to the inhibition of PTEN and DUSP7 ubiquitination under 230D7 treatment, accumulation of cellular PTEN and DUSP7 proteins was observed in 786-O cells treated with 230D7, causing decreases in phosphorylated AKT and ERK (Fig. 4g, h). Next, we tested the cell proliferation of three ccRCC cell lines (786-O, Caki-2, OS-CR-2) and two non-ccRCC cell lines (4T-1, MDA-MB-231) in the presence of 230D7 (Fig. 4i). 230D7 specifically inhibited the growth of ccRCC cell lines with an IC_{50} of approximately 20 μM

265 compared with non-ccRCC cell lines. To determine if 230D7 is suitable for *in vivo*
266 studies, we investigated the pharmacokinetics and acute toxicity profile of 230D7.
267 230D7 can be efficiently absorbed into the blood circulation following intraperitoneal
268 injection and has low acute toxicity (Extended Data Fig. 3d~h). A dose-dependent
269 reduction in 786-O tumour growth rate could be observed in NSG mice treated with
270 230D7 (Fig. 4j), revealing significant anti-ccRCC therapeutic effect of 230D7 *in vivo*.
271 No body weight loss in NSG mice was statistically noted during the entire
272 pharmacodynamics study of 230D7 (Extended Data Fig. 3i). Finally, we checked the
273 effect of 230D7 on oncogenic SPOP signalling in ccRCC xenograft tumours. As
274 expected, PTEN and DUSP7 are elevated in 230D7 treated groups, and p-AKT and p-
275 ERK levels are decreased (Fig. 4k). Moreover, we confirmed the high selectivity of
276 230D7 which does not target kinases (Extended Data Fig. 4). In conclusion, we
277 successfully discovered novel small-molecule inhibitors that target SPOP through a
278 sequence-to-drug paradigm, among which 230D7 showed therapeutic potential for
279 blocking SPOP activity to treat ccRCC.
280
281 **Drug design targeting E3 ubiquitin-protein ligases without 3D structure.**
282 After discovering novel inhibitors for SPOP, we applied this paradigm to discover hits
283 for a more challenging target RNF130 whose crystal structure is unknown. RNF130 is
284 an E3 ubiquitin-protein ligase without structural information, and no chemical binders
285 have been reported. Our recent study revealed that RNF130 plays an important role in
286 autoimmune inflammation, suggesting its inhibition could be of potential therapeutic

287 value (unpublished data, under review). We utilized TransformerCPI2.0 to screen
288 compounds that directly bind to RNF130 (Fig. 5a) and discovered iRNF130-63 to be a
289 binder of RNF130 (Fig. 5b~d). The direct binding between iRNF130-63 and RNF130
290 protein was confirmed through surface plasmon resonance (SPR), and this binding
291 exhibited a fast-on, fast-off kinetic pattern with a K_D of 9.36 μM (Fig. 5c). We also
292 performed a cellular thermal shift assay (CETSA), and the results supported that
293 iRNF130-63 directly binds to and thermally stabilizes the RNF130 proteins (Fig. 5d).
294 The success of discovering novel hits for SPOP and RNF130 demonstrated that the
295 sequence-to-drug paradigm is practicable for virtual screening with encouraging
296 prospects.
297

298 **Discussion**

299 Conventional structure-based drug design pipeline is a complex, human-engineered
300 pipeline with multiple independently optimized steps. Structural functional insights and
301 profound experience are needed to build 3D structures through structural biology or
302 structure prediction, identify the potential binding pockets, either orthosteric sites or
303 allosteric sites, and screen active molecules by different molecular docking protocols.
304 However, such multistep operation is error prone; for example, inaccurate protein
305 structures, the multiplicity and dynamics of binding pockets, incorrect pocket definition,
306 inappropriate selection of the scoring function, etc., and the errors or intrinsic accuracy
307 limitations of each step accumulate rapidly and significantly lower the success rate.
308 When little information about target proteins is known, this issue becomes more serious

309 and has formed a long-lasting obstacle to rational drug design.

310 To address this problem, we proposed a new sequence-to-drug paradigm that
311 discovers modulators directly from protein sequences, jumping over intermediate steps.
312 This proof-of-concept study analyses the sequence-to-drug paradigm in three aspects:
313 its competitive performance against 3D structure-based molecular docking, its learning
314 of binding knowledge, and drug design for challenging targets and drug repurposing.
315 Apart from methodology, the sequence-to-drug paradigm successfully discovered novel
316 inhibitors for SPOP, among which 230D7 showed potential to treat ccRCC, and
317 discovered the first binder of RNF130. Given that SPOP is an adaptor of E3 ligase and
318 RNF130 is an E3 ligase, the hits we found have the potential to serve as novel warheads
319 for proteolysis-targeting chimaeras (PROTACs). PROTACs have been successfully
320 developed for harnessing the ubiquitin–proteasome system to degrade a protein of
321 interest, receiving tremendous attention as a new and exciting class of therapeutic
322 agents that promise to significantly impact drug discovery[46]. Due to the limited
323 structural knowledge of E3 ligases and their substrate recognition and regulation of
324 ubiquitination, the hits we discovered may promote the development of PROTACs. All
325 these results demonstrate that the sequence-to-drug paradigm is a promising direction
326 for drug design.

327 Recently, due to the rapid growth of virtual libraries, such as DrugSpaceX[47] and
328 synthesis (REAL) combinatorial libraries[48] which cover spaces of 100 million to
329 multiple billions of chemicals, there is a high demand for developing computationally
330 efficient virtual screening approaches. The sequence-to-drug paradigm can be

331 combined with these large virtual libraries to rapidly discover new active scaffolds from
332 the unexplored chemical space. On the one hand, the explosion in available genomic
333 sequencing techniques and annotation techniques has revolutionized bioinformatics[49],
334 and the sequence-to-drug paradigm will likewise evolve rapidly by incorporating more
335 abundant multiple sequence alignment or functional annotation information. On the
336 other hand, we may perform more comprehensive proteome-wide virtual screening,
337 accelerating the discovery of new drug hits for novel but challenging biological targets.
338 Overall, our findings provide a proof of concept for the sequence-to-drug paradigm,
339 which we believe will become an essential component of future rational drug design
340 pipelines.

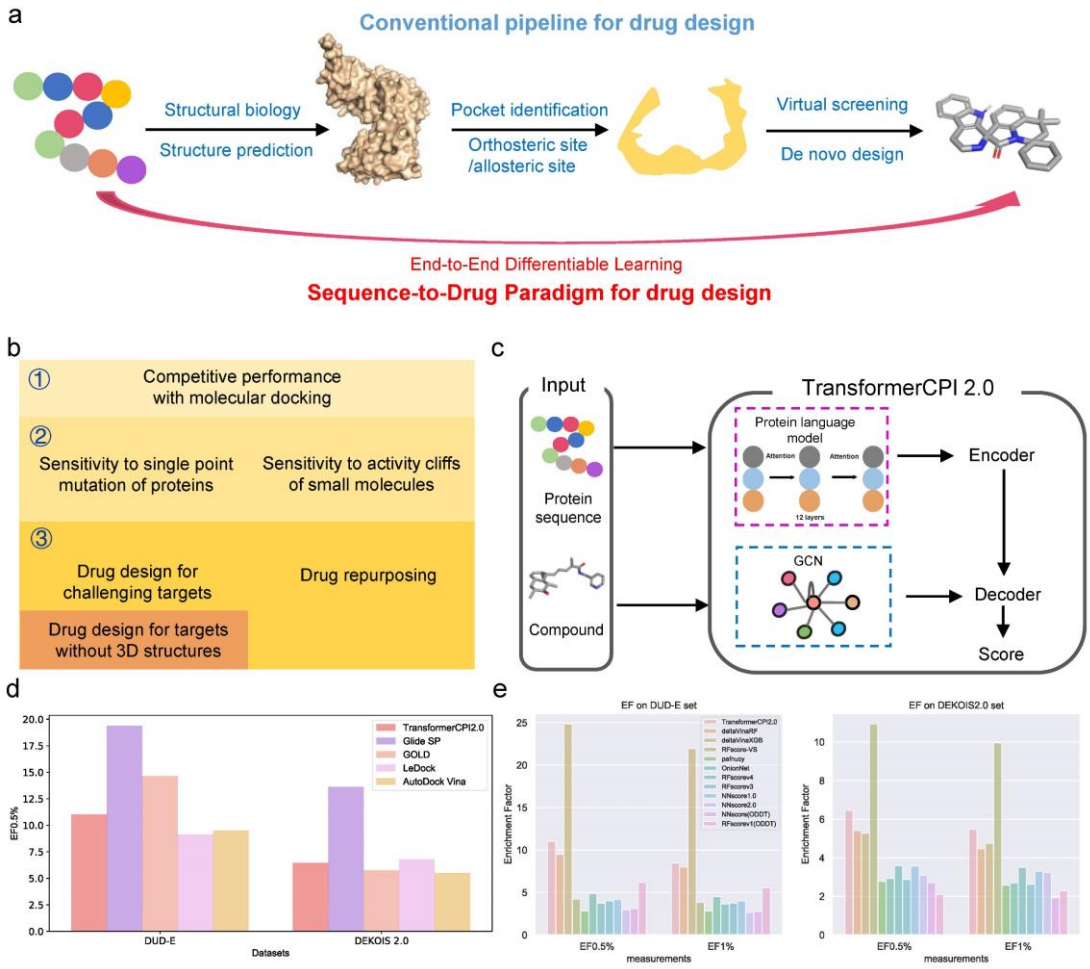


Fig. 1 | Competitive performance compared with 3D structure-based approaches. a, The conventional pipeline for target-based drug design and the sequence-to-drug paradigm. **b**, Three stages of the proof of concept of the sequence-to-drug paradigm, with each stage is labelled by different colours. **c**, The computational pipeline of TransformerCPI2.0. **d**, EF0.5% results of TransformerCPI2.0 and docking programs on DUD-E set and DEKOIS 2.0 set. **e**, EF0.5% and EF1% results of TransformerCPI2.0 and machine learning scoring functions on the DUD-E set and DEKOIS 2.0 set.

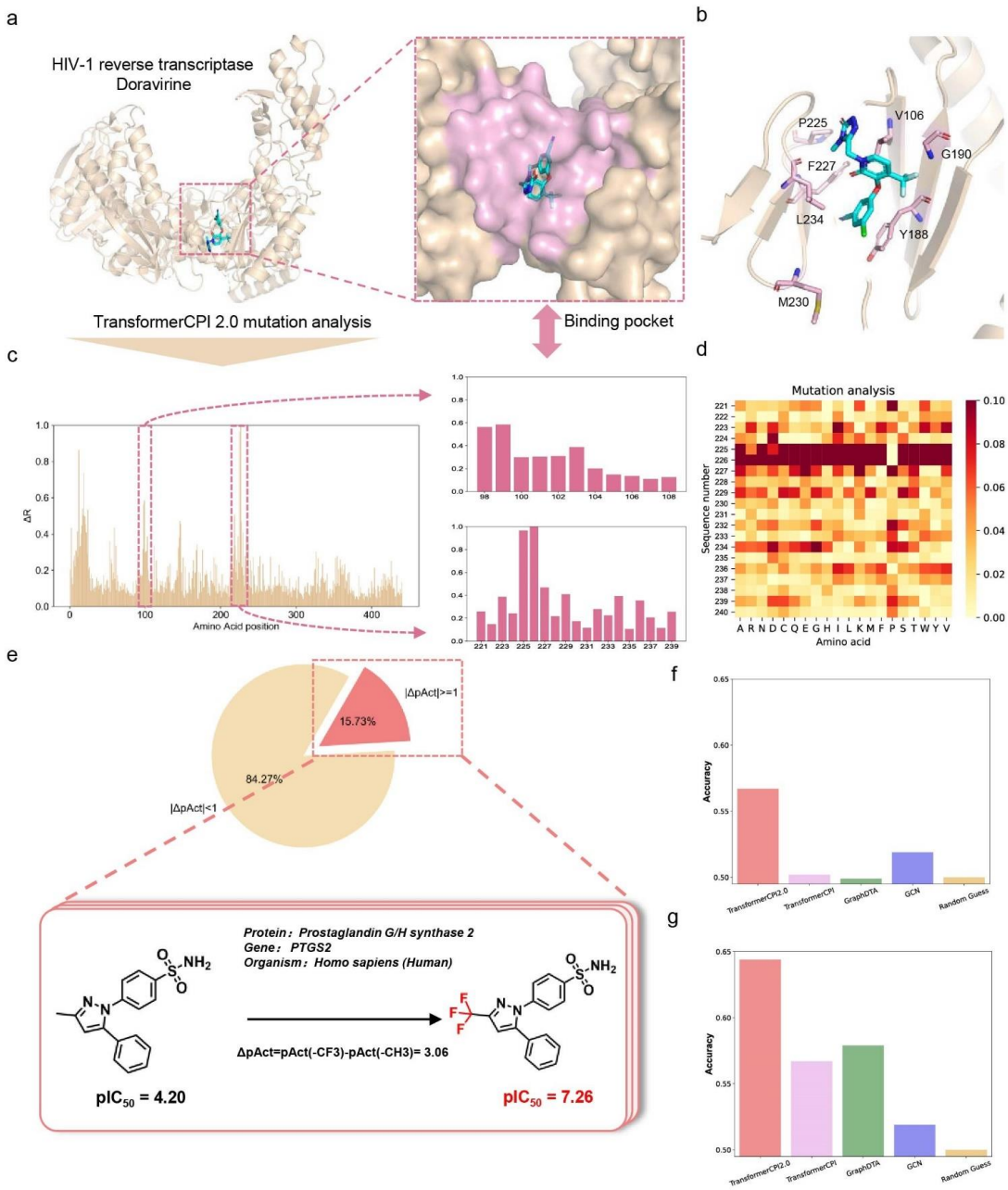


Fig. 2 | Drug resistance mutation analysis and substitution effect analysis of the trifluoromethyl group. a, The cocrystal structure of HIV-1 reverse transcriptase and doravirine (PDB: 4NCG). The binding pocket of doravirine is highlighted in pink. **b**, The binding mode of doravirine. The residues with drug-resistant mutations are coloured pink. **c**, Relative activity change score (ΔR) calculated by TransformerCPI2.0 at each amino acid position. The pink boxes mark the high ΔR regions, which are plotted in detail on the right. **d**, The heatmap plots the activity change score of positions 221 to 240, where each position is mutated to each of the 20 amino acids (including itself). The darker colour represents the higher activity change score caused by the mutation. **e**, Left, data distribution of the trifluoromethyl substitution dataset. Only 15.73% of substitution of −CH3 by −CF3 could increase or decrease the biological activity by at least an order of magnitude. We conducted substitution effect analysis on this part of data. Right, an example of −CH3 changed by −CF3 leads to a significant increase in biological activity. **f**, The accuracy of TransformerCPI2.0 and baseline models on the whole dataset. **g**, The accuracy of TransformerCPI2.0 and baseline models on the subset where the substitution of −CH3 by −CF3 could increase or decrease the biological activity by at least three orders of magnitude. **h**, An activity decrease example and an activity increase example are shown. The predictions of TransformerCPI2.0 are consistent with the ground truth for proteins and compounds not present in the training set.

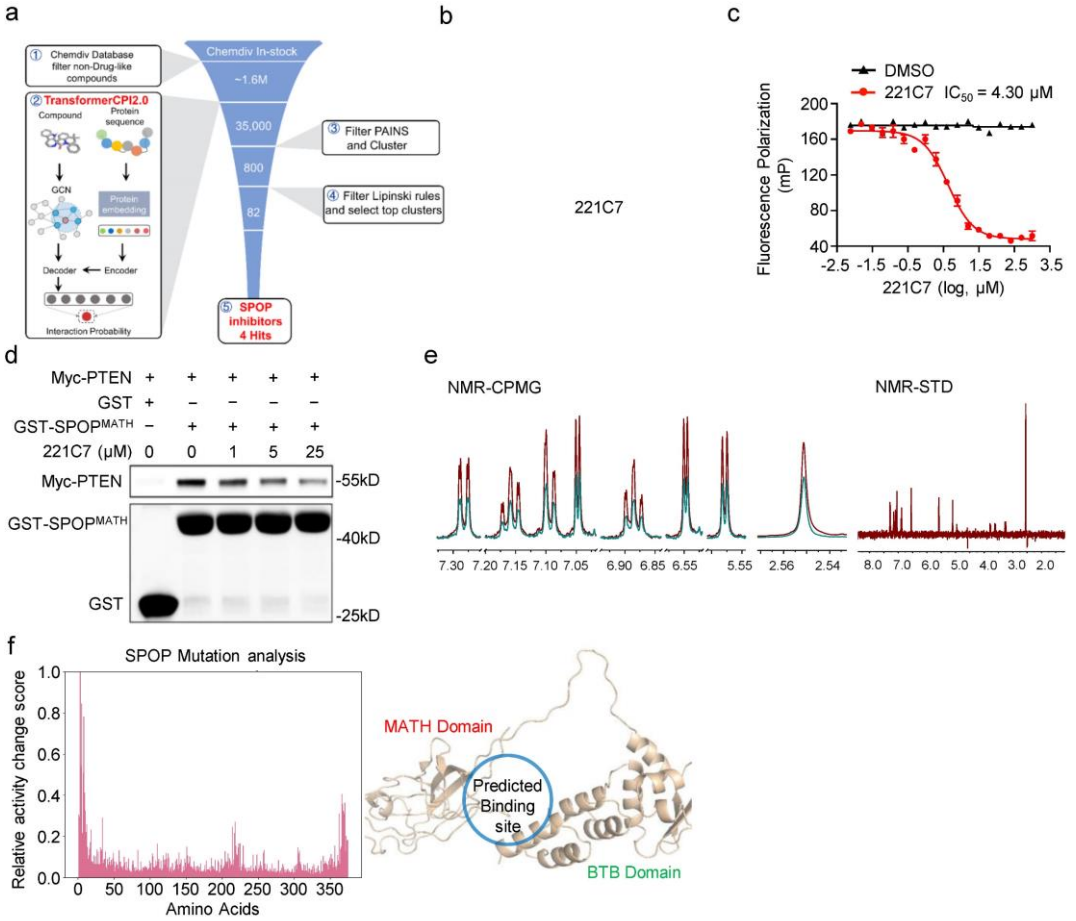


Fig. 3 | Discovering a novel scaffold hit of SPOP. a, Scheme of virtual screening protocol for small-molecule inhibitors of SPOP. **b**, Chemical structure of 221C7. **c**, 221C7 competitively inhibits puc_SBC1 peptide binding to $SPOP^{MATH}$, as measured by the FP assay. **d**, 221C7 disrupts protein binding between $SPOP^{MATH}$ and PTEN, as measured by *in vitro* pull-down assay. **e**, NMR measurement of direct binding between 221C7 and $SPOP^{MATH}$. CPMG NMR spectra for 221C7 (red), 221C7 in the presence of 5 μM $SPOP^{MATH}$ (green). The STD spectrum for 221C7 is recorded in the presence of 5 μM $SPOP^{MATH}$. **f**, Drug resistance mutation analysis interprets the prediction of TransformerCPI2.0, and important residues are mapped back to the structure of full-length SPOP (PDB:3HQI). The predicted binding site is located on the MATH domain and near the BTB domain.

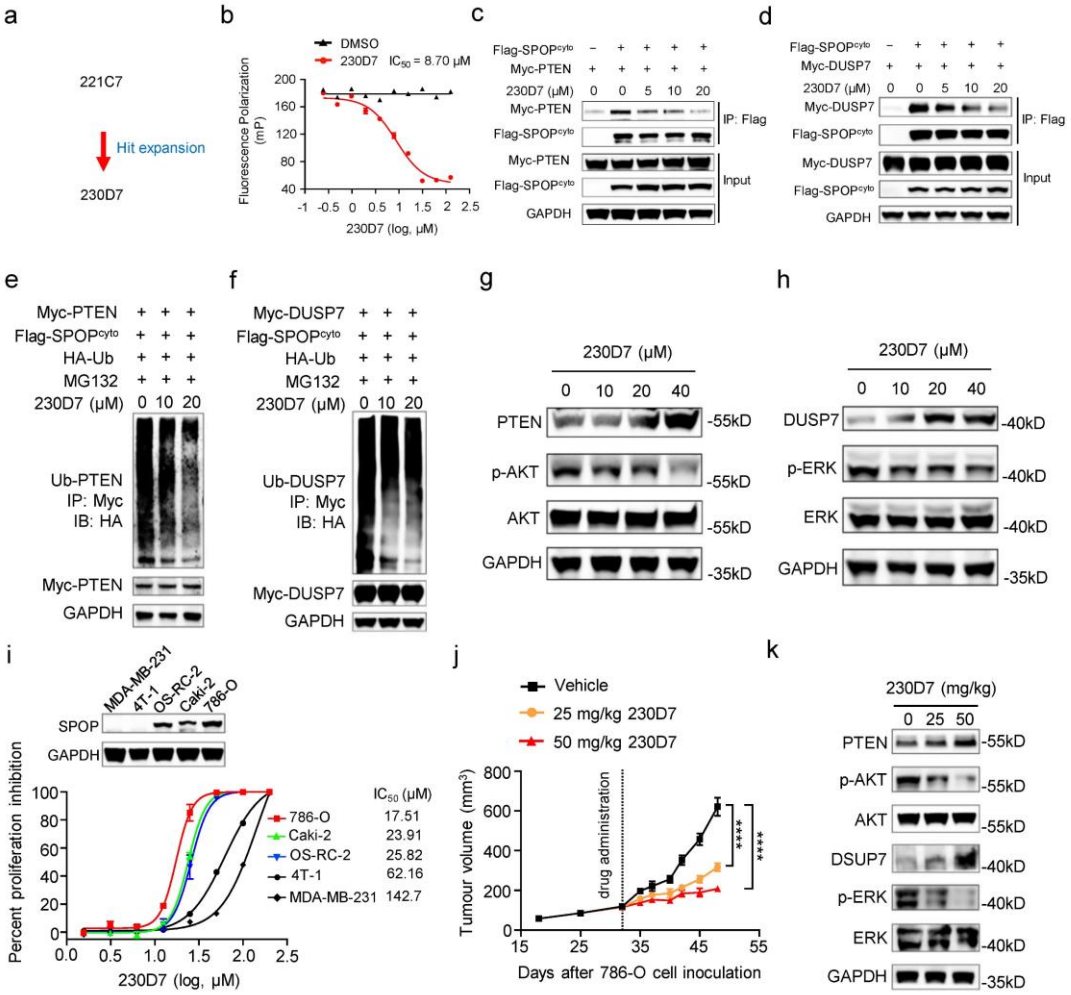


Fig. 4 | 230D7 showed therapeutic potential for blocking oncogenic SPOP activity to treat ccRCC. a, Hit expansion of 221C7 to 230D7. **b**, 230D7 competitively inhibits puc_SBC1 peptide binding to $SPOP^{MATH}$, as measured by the FP assay. **c~d**, SPOP-PTEN and SPOP-DUSP7protein interactions are inhibited in the presence of 230D7 in 293T cells. **e~f**, 230D7 inhibits the ubiquitination of PTEN and DUSP7 in 293T cells. **g~h**, 230D7 upregulates PTEN and DUSP7 protein levels in 786-O cells. The downstream p-AKT and p-ERK abundances are observed to decrease. **i**, Cell proliferations of three ccRCC cell lines and two non-ccRCC cell lines in the presence of 230D7. The abundance of cytoplasm SPOP protein was measured. **j**, *In vivo* anti-ccRCC efficacy of 230D7 in 786-O xenograft models in NSG mice. Mice were administrated 230D7 at 25 or 50 mg/kg daily for 16 days by intraperitoneal dosing. Error bars represent mean ± SEM, with n = 7 for each group. P values were evaluated using 2-tailed unpaired t-test, ****$P < 0.0001$. **k**, Quantification of PTEN and DUSP7 accumulation and repression of p-AKT and p-ERK at day 16 of 786-O xenograft tumours treated with vehicle or 230D7.

344
345
346

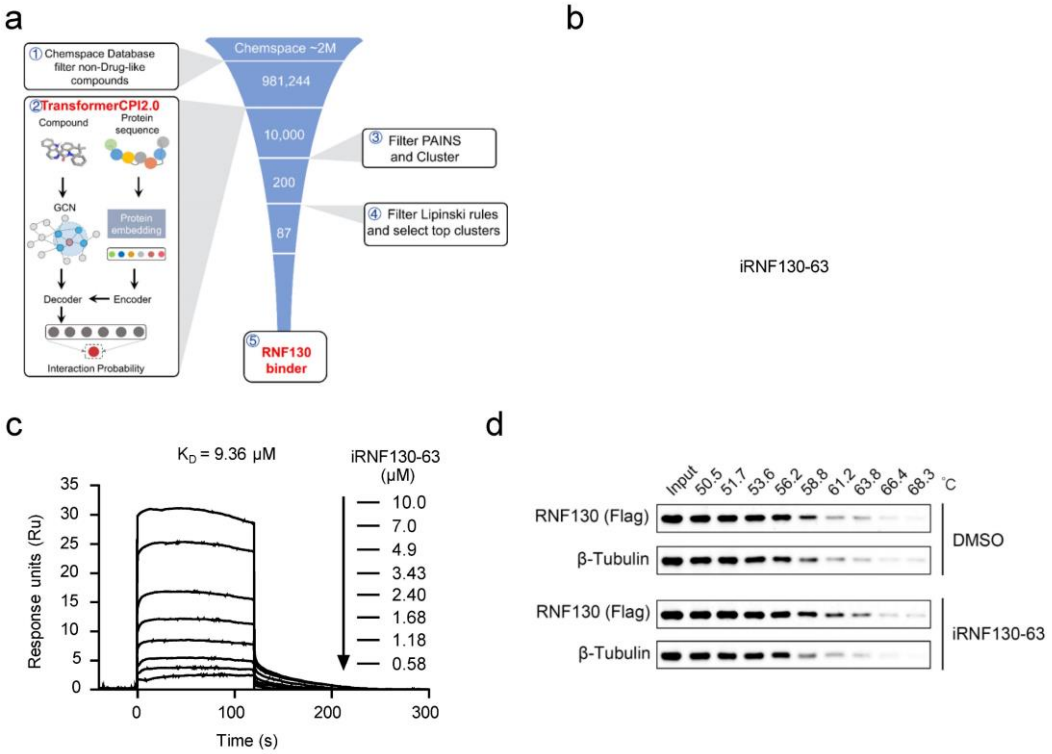


Fig. 5 | Discovering the chemical binder of RNF130. a, The virtual screening procedure of RNF130. **b**, Chemical structure of iRNF130-63. **c**, Surface plasmon resonance analysis examining the direct binding affinities of iRNF130-63 to RNF130. Graphs of equilibrium unit responses versus iRNF130-63 concentrations are plotted. The estimated KD is 9.36 μM. **d**, Representative western blots for the effect of 50 μM iRNF130-63 on the thermal stabilization of RNF130 protein. Cellular thermal shift assay (CETSA) was assayed in 293T cell lysate.

347

Method

TransformerCPI2.0 model and training details. Compared with TransformerCPI, TransformerCPI2.0 has been updated in the following four aspects: (1) removing 3-gram protein word embedding calculated by the Word2Vec algorithm, (2) computing protein sequence representation by a pretrained protein language model named TAPE-BERT, (3) replacing 1D convolutional neural networks and gated linear units with a self-attention-based transformer encoder, and (4) introducing a new atom vector into the atom sequence that carries the interaction information at the molecular level.

Pretraining the protein language model. Word2vec is an unsupervised technique to learn high-quality distributed vector representations that describe sophisticated syntactic and semantic word relationships and maps discrete words to low-dimensional real-valued vectors. However, the final embedding table of Word2vec is stationary, regardless of the upstream and downstream context information of the given word, which may lead to errors regarding the true meaning of the word in its local context. Since BERT[50] achieved great success in natural language processing (NLP), many efforts have been devoted to protein sequence representation learning. Many pretraining models based on long short-term memory (LSTM) or transformer architectures have been proposed, such as UniRep[51] and TAPE[52]. To maintain the model consistency and gain parallel computing efficiency, we chose the transformer model in TAPE (TAPE-BERT) to calculate protein sequence embedding. The TAPE-BERT model contains 12 self-attention encoder layers, 12 attention heads for each layer, 768 dimensions for the hidden state, and 3072 dimensions for feedforward layers. We first utilized the TAPE

Tokenizer from TAPE-BERT to encode the protein amino acid sequence into real values, where numbers 1~23 represent 23 common amino acids, 0 represents the token '<pad>', 24 represents the token '<cls>', and 25 represents the token '<sep>'. Then, we input this encoded real value sequence into the pretrained TAPE-BERT model and finally obtained protein embeddings with 768 dimensions. In TransformerCPI2.0, protein embedding from the TAPE-BERT model serves as the input to the encoder of TransformerCPI2.0, replacing the embedding calculated from the Word2Vec model.

Encoder of TransformerCPI2.0. Since protein embedding was calculated by the TAPE-BERT model, we replaced 1D convolutional neural networks and gated linear units with the original self-attention-based transformer encoder. Given that the position information of the amino acid sequence has been taken into consideration when computing protein embeddings, position embedding was removed from the transformer encoder. The encoder consists of 3 encoder layers, 8 attention heads for each layer, 768 dimensions for the hidden state, and 3072 dimensions for feedforward layers. To maintain the maximal performance of TAPE-BERT, the hidden state dimension of the encoder was exactly the same as that of TAPE-BERT, ensuring no information loss in this process. After the hyperparameter search, 12 attention heads showed little performance improvement compared to 8 attention heads but much higher training and inference time; therefore, only 8 attention heads were used in TransformerCPI2.0.

Atom embedding calculation. Each of the atom features was initially represented as a vector of size 34 using the RDKit python package, and the list of atom features can be found in our previous work. In TransformerCPI2.0, we additionally introduced a new

virtual atom that carries the information at the molecular level and does not exist in the given compound. This virtual atom was initialized as the average of atom features across the whole compound and was linked to all atoms. All the atom vectors together with the virtual atom vector were put into graph convolutional networks (GCNs)[53] to learn the representation by integrating their own neighbourhood information. Notably, only one GCN layer was used and recommended in this process, and more than two GCN layers harmed the performance of TransformerCPI2.0 to a great deal. Too many GCN layers may over smooth the atom features, causing different atom features to tend to be similar to each other. TransformerCPI2.0 then fails to learn compound-protein interaction features when the atom embedding carries excessively similar features.

Decoder of TransformerCPI2.0. Protein embedding and atom embedding serve as the target sequence and memory sequence of the transformer decoder, respectively. Consistent with the encoder, the decoder consists of 3 decoder layers, 8 attention heads for each layer, 768 dimensions for the hidden state, and 3072 dimensions for feedforward layers. In addition, the original transformer was designed to solve seq2seq tasks and utilize a causal mask operation to cover the downstream context of each word in the decoder. We removed the mask operation of the decoder to ensure that our model accesses the whole target sequence. Since we introduced a new virtual atom, as described above, we used the last layer representation of this virtual atom rather than the weighted sum of the last layer atom representation to predict the compound protein interaction probability. The last layer presentation of virtual atoms was fed into fully connected layers and finally returned the compound protein interaction probability.

414 **Training details.** The TransformerCPI2.0 model was trained by the RAdam[54] optimizer
415 with a learning rate of 1e-5 and a weight decay of 1e-3. The batch size of 1 was selected
416 to ensure that the longest protein sequence fit into the GPU memory. We employed the
417 gradient accumulation technique to expand the actual batch size to 64. Training was
418 performed on one NVIDIA Tesla V100 (16G) GPU. The TransformerCPI2.0 model was
419 trained for ~50 epochs or ~1.5 weeks of wall clock time.

420 **ChEMBL dataset construction.** Inheriting our previous work, we followed two rules
421 to construct a dataset: (i) CPI data was collected from an experimentally validated
422 database and (ii) each ligand should exist in both positive and negative classes. To build
423 a universal deep learning model for all types of proteins, we selected the chEMBL_23[55]
424 database to construct a universal dataset to train TransformerCPI2.0.

425 **Data cleaning.** The ChEMBL_23 database was released on 1 May 2017 and contains
426 2,101,843 compound records, 1,735,442 compounds, 14,675,320 activities, 1,302,147
427 assays, 11,538 targets and 67,722 source documents. We downloaded the whole
428 database and cleaned the data using the following procedure:

429 (1) The target type was set to 'SINGLE PROTEIN', and the molecule type was set
430 to 'Small molecule';

431 (2) Data with a confidence score of 9 and assay type of 'B' were reserved;

432 (3) Activity data with activity metrics of IC_{50}, EC_{50}, K_i in units of nM were selected.

433 **Dataset process.** After cleaning the data, we transformed IC_{50}, EC_{50}, and K_i to pIC_{50},
434 pEC_{50}, and pK_i and then split the dataset into a positive set and a negative set at the
435 threshold of 6.5. Samples with different labels were removed from the dataset. In the

early stage of drug discovery, only hit compounds whose IC_{50}, K_i or EC_{50} are at the μM or even nM level will be further optimized. Additionally, public data are prone to a certain experimental error, i.e., on average 0.5 log units for IC50 data[56,57]. The threshold of 6 considers those CPI pairs whose IC50, Ki or EC50 are smaller than 1 μM as positive samples, which causes the models to select CPI pairs with high activity. To decrease the risk of experimental error in public data, a stricter threshold of 6.5 was used in a previous work[58]. Therefore, we selected the threshold of 6.5 to define positive data and negative data. Data whose atom number was more than 60 or whose protein sequence length exceeded 4000 were filtered out, guaranteeing that all the data fit into GPU memory. Finally, we constructed a ChEMBL dataset including 3,348 proteins, 69,616 compounds, 117,513 positive CPIs, 134,611 negative CPIs and 252,124 samples in total.

Dataset split and label reversal experiment. We selected ligands that exist in both the positive and negative classes to make the compound distribution in positive samples and negative samples exactly the same, trying to eliminate the potential ligand bias as much as possible. Consequently, we used a label reversal experiment to split the ChEMBL dataset. For the ChEMBL dataset, we randomly selected 2,941 ligands and pooled all the negative CPI samples containing these ligands into the test set. Additionally, we selected another 2,900 ligands and pooled all their associated positive samples into the test set. The remaining datasets were split randomly into a training set and a validation set at a ratio of 10:1. The validation set was used to determine the hyperparameters, and the best model was evaluated on the test set. Under this

experimental design, we finally established a ChEMBL set containing 217,732 samples in the training set, 24,193 samples in the validation set and 10,199 in the test set.

External evaluation on external datasets. All the baseline models and TransformerCPI 2.0 were tested on the external test set, GPCR set and kinase set from our previous work. In addition to the GPCR and kinase set, we designed a large external set. This external set contains compounds that were not previously observed and 1,192 new protein targets that were not included in the training set. The total number of CPI pairs is 342,447, and the ratio of positive samples to negative samples is 1:3. This external set can evaluate the generalization ability of TransformerCPI2.0 and baseline models to the new compounds and new proteins. Another time-split dataset named the ChEMBL27 dataset contains compounds that were not previously observed and 637 new protein targets that are not included in the training set, and all the data were collected from the ChEMBL_27 database. The total number of CPI pairs is 92,919, and the ratio of positive samples to negative samples is 1:1.

Baseline models. All the baseline models, including TransformerCPI, CPI-GNN, GraphDTA(GAT-GCN), and GCN, were trained on the ChEMBL dataset with their own hyperparameters. Only the learning rate, weight decay rate and dropout rate were subjected to a hyperparameter search.

Drug resistance mutation analysis. Activity change score calculation**.** First, we input the wild-type protein and drug into TransformerCPI2.0 to calculate the original prediction score, denoted as s. Then, we mutated each amino acid of the protein sequence to all 20 amino acids (including itself) and calculated the prediction score s'.

Finally, we defined the activity change score $\Delta S \in \mathbb{R}^{l\times 20}$ as

$$\Delta S_{i,j} = |\, s - s'_{i,j} \,|, \qquad i = 1,2,\dots,l, \qquad j = 1,2,\dots,20.$$

Here, i corresponds to the position of the protein sequence, l corresponds to the length of the protein sequence, and j corresponds to 20 types of amino acids. Since an amino acid mutation will increase or decrease the prediction score s' and TransformerCPI2.0 may not be able to predict the trend of activity changes correctly, we calculated the absolute value of ΔS here. We analysed ΔS and found that TransformerCPI2.0 actually learns the key features of compound protein interactions because the pattern of ΔS revealed by heatmap analysis is consistent with that of drug resistance mutation.

Relative activity change score. After calculating the activity change score, we can evaluate whether a mutation at a specific position plays an important role in compound protein interactions. However, ΔS cannot quantify the contribution of each position on a protein sequence and rank the most important sites from the whole sequence. To quantify the contribution of each amino acid site to drug-protein interactions, we first computed the average score of each position $\Delta\bar{S} \in \mathbb{R}^{l}$ as

$$\Delta\bar{S}_i = \frac{\sum_{j=1}^{20} s'_{i,j}}{20}.$$

Furthermore, the value of $\Delta\bar{S}$ was normalized to between 0~1, and the relative activity change score $\Delta R \in \mathbb{R}^{l}$ was defined as

$$\Delta R_i = \frac{\Delta\bar{S}_i}{max(\Delta\bar{S}_i)}, i = 1,2,\dots,l.$$

The relative activity change score ΔR can then characterize the contribution of each amino acid position to TransformerCPI2.0 prediction and help researchers discover

502 novel and potential drug resistance mutation sites. On the other hand, ΔR can reflect
503 the compound–protein interactions in TransformerCPI2.0. An amino acid site whose
504 contribution to compound–protein binding is large can be revealed quantitatively by the
505 relative activity change score ΔR. Finally, we used the activity change score ΔS to
506 plot a heatmap to study the concrete pattern of drug resistance mutations and the relative
507 activity change score ΔR to rank the most important sites for compound protein
508 binding.

509 **Analysis of the substitution effect of the trifluoromethyl group.** The replacement of
510 methyl (Me or −CH3) with trifluoromethyl (TFM or −CF3) is frequently employed in
511 compound optimization. However, the exact effect of −CH3/–CF3 substitution on
512 bioactivity is still controversial. To further investigate whether TransformerCPI2.0
513 captures the key features of the compound and comprehensively understands
514 compound–protein interaction, we employed TransformerCPI2.0 to predict the
515 substitution effect of the trifluoromethyl group. We utilized a previously reported
516 dataset[39] containing 18,217 pairs of compounds and corresponding bioactivity data with
517 the only difference being that −CH3 is substituted by −CF3 to study this problem.

518 **Dataset analysis.** The statistical results showed that the replacement of −CH3 with
519 −CF3 does not improve bioactivity on average[39]. However, in 15.73% of cases,
520 substituting −CF3 for −CH3 increased or decreased the biological activity by at least
521 an order of magnitude. Therefore, we focused on these samples to study the prediction
522 capacity of TransformerCPI2.0. The actual bioactivity change $\Delta pAct$ was defined as

523 $$\Delta pAct = pAct(-CF3) - pAct(-CH3)\ .$$

Activity change score. First, we used TransformerCPI2.0 to calculate the prediction scores of compounds with trifluoromethyl substituents and denoted this score as $score_{-CF3}$. Then, we calculated the prediction scores of compounds with methyl substituents and denoted this score as $score_{-CH3}$. The activity change score Δs_c was defined as

$$\Delta s_c = score_{-CF3} - score_{-CH3} .$$

Considering that the distributions of $\Delta pAct$ and Δs_c were different from each other, we defined a correct prediction by a model as $\Delta pAct$ and Δs_c sharing the same sign. In other words, when the activity change trend predicted by the model matches the actual bioactivity change, the prediction is considered correct. After defining the evaluation metrics, we analysed the performance of TransformerCPI2.0 on the whole dataset. Furthermore, we investigated the prediction performance on the cases where the corresponding biological activities increased or decreased by at least three orders of magnitude because these cases are relevant to the activity cliff phenomenon in medicinal chemistry. Finally, we selected two cases that were not observed in the training set to show the power of TransformerCPI2.0.

Comparison with structure-based docking tools. TransformerCPI2.0 is designed to predict compound–protein interactions considering only protein sequence information; however, evaluating the predictive power of TransformerCPI2.0 on those targets whose crystal structures remain unknown is not practical. Although we can test the power of TransformerCPI2.0 in specific lead compound screening projects, we compared TransformerCPI2.0 with structure-based docking tools to verify its ability to enrich

positive molecules among compound libraries.

Docking dataset. We used two benchmark datasets, DUD-E and DEKOIS 2.0, to compare TransformerCPI2.0 and other docking tools. DUD-E is a widely used benchmark dataset for docking tool and scoring function comparisons. DUD-E includes a total of 102 proteins with 22886 clustered ligands drawn from ChEMBL, each with 50 property-matched decoys drawn from ZINC. DEKOIS2.0 is often used as a complementary benchmark for further validation. DEKOIS2.0 contains 81 structurally diverse targets. Each target has 40 active compounds extracted from BindingDB and 1200 decoys generated from ZINC. DUD-E and DEKOIS 2.0 both serve as external test sets for TransformerCPI2.0, which means that TransformerCPI2.0 was not trained or fine-tuned on these datasets.

Docking tools. Four docking programs, Glide SP (version 8.2)[59], GOLD (version 2019)[60] with Piecewise Linear Potential (CHEMPLP) SF, LeDock (version 1.0)[61] and AutoDock Vina[62], showed high docking power and computational efficiency and thus were chosen as the baseline models. We focused on the enrichment factor (EF0.5%, EF1%, EF5%) to evaluate different models, and the results of Glide SP, GOLD and LeDock on the two benchmark sets were extracted from Chao *et al.*[63]. In addition to the above four docking tools, many machine learning scoring functions have been tested on DUD-E and DEKOIS 2.0[63]; thus, we compared these machine learning scoring functions. We used the whole protein sequence as the default input of TransformerCPI2.0 and calculated the average EF0.5%, EF1% and EF5% of DUD-E and DEKOIS 2.0.

568 **Virtual screening of SPOP.** First, after filtering non-drug-like compounds from the
569 ChemDiv Library (San Diego, CA, USA), which contains approximately 1.6 million
570 in-stock compounds, TransformerCPI2.0 was applied to score the compounds, and the
571 top 35,000 molecules (~top 2%, ensuring compound diversity) were selected by
572 screening. Second, we filtered panassay interference compounds (PAINS) and clustered
573 these molecules automatically based on their extended-connectivity fingerprints
574 (ECFP), obtaining approximately 800 clusters. Third, we filtered these compounds by
575 the Lipinski rules and manually selected representative compounds from each cluster.
576 Finally, a total of 82 candidates were purchased for further experimental evaluation.
577 **Virtual screening of RNF130.** First, after filtering non-drug-like compounds from the
578 Chemspace Library (Monmouth Junction, NJ 08852, USA), which contains
579 approximately 2 million in-stock compounds, TransformerCPI2.0 was applied to score
580 the compounds, and the top 10,000 molecules (~top 0.5%, ensuring compound diversity)
581 were selected. Second, we filtered panassay interference compounds (PAINS) and
582 clustered these molecules automatically based on their extended-connectivity
583 fingerprints (ECFP), obtaining approximately 200 clusters. Third, we filtered these
584 compounds by the Lipinski rules and manually selected representative compounds from
585 each cluster. Finally, a total of 87 candidates were purchased for further experimental
586 evaluation.
587 **Plasmid construction.** Wild-type, truncated or mutant versions of the human proteins
588 were used in this study: SPOP (UniProt accession number: O43791-1), PTEN (UniProt
589 accession number: P60484-1), DUSP7 (UniProt accession number: Q16829-1),

590 RNF130 (UniProt accession number: Q86XS8-1). For plasmid construction, $SPOP^{WT}$
591 and $SPOP^{cyto}$ (residues 1-366) were subcloned into the pcDNA 3.1 vector with a Flag-
592 tag, $SPOP^{MATH}$ (residues 28-166) and $ARNO^{Sec7}$ (residues 50-250) was subcloned into
593 the pGEX 6p-1 vector with a GST-tag, PTEN and DUSP7 were subcloned into the
594 pcDNA 3.1 vector with a Myc-tag, RNF130 (residues 1-304) was subcloned into the
595 pcDNA3.1 vector with a C-terminal His_8-Flag-tag. All of the above plasmids were
596 synthesized by Sangon Biotech (Shanghai) Co., Ltd. pCMV-HA-Ub plasmid (CAT#.
597 kl-zl-0513) was purchased from Shanghai Kelei Biological Technology Co., Ltd.
598 **Recombinant protein expression and purification.** For expression of $SPOP^{MATH}$,
599 GST-tagged $SPOP^{MATH}$ plasmid was transformed into BL21-CodonPlus (DE3)-RIPL
600 Cells (CAT#. EC1007, Shanghai Weidi Biotechnology Co., Ltd.), and then the cells
601 were grown in lysogeny broth (LB) medium and induced by isopropyl β-D-1-
602 thiogalactopyranoside (IPTG) at a final concentration of 0.5 mM at 16 °C overnight.
603 Cells were harvested and lysed in solution (20 mM HEPES pH 7.4, 200 mM NaCl, 1
604 mM dithiothreitol) by sonication and then centrifuged at 16,000 rpm for 1 hour at 4 °C.
605 The supernatants were filtered by 0.22 μM syringe filters and purified on GST Trap
606 columns (GE Healthcare) by elution with 10 mM glutathione. The eluted components
607 were loaded onto desalting columns (GE Healthcare) to remove glutathione and
608 incubated with PreScission Protease for 6-8 hours at 4 °C. The components were
609 reloaded onto GST Trap columns to remove the GST tags and further purified by a
610 Superdex 75 10/300 GL column. The purified GST-$SPOP^{MATH}$ protein was concentrated
611 and stored in buffer (20 mM HEPES pH 7.4, 200 mM NaCl) at -80 °C.

612 RNF130 protein was expressed in Expi-293F (Invitrogen) using Expifectamine
613 transfection reagent according to the manufacturer's instructions. Cells were collected
614 3 days after transfection. Proteins were first captured by Ni^{2+}-Sepharose 6 Fast Flow
615 resin (GE healthcare) and then further purified by gel filtration chromatography with a
616 Superdex S200 column (GE Healthcare). The purified protein was concentrated and
617 stored in buffer (20 mM HEPES pH 7.5, 150 mM NaCl and 1 mM TCEP) at -80 °C.
618 **Fluorescence polarization (FP).** Fluorescence polarization experiments were
619 conducted in a 384-well black plate (Corning, 3575) using a 42 μL reaction system.
620 FITC-labelled SPOP substrate puc_SBC1 (FITC-LACDEVTSTTSSSTA)[64,65]
621 (synthesized by GL Biochem (Shanghai) Ltd) was used for the probe. Then, 20 μL of
622 reaction buffer (20 mM HEPES, pH 7.4) containing 200 nM $SPOP^{MATH}$ protein was
623 incubated with 2 μL of compound for 30 minutes at room temperature, and 20 μL of
624 reaction buffer containing 200 nM probe was added. Fluorescence polarization (mP)
625 signals were measured by a fluorescence mode (excitation filter 480 nm, emission filter
626 535 nm) in Spark microplate reader (Tecan).
627 **Protein thermal shift (PTS).** The Bio–Rad CFX96 RealTime PCR Detection System
628 was utilized to monitor the thermal stability of $SPOP^{MATH}$ protein. PTS experiments
629 were performed in a 96-well PCR plate (DN Biotech (Hong Kong) Co., Ltd.) with a 20
630 μL reaction system**.** A total of 20 μL of reaction buffer containing protein (5 μM for
631 $SPOP^{MATH}$), 5 × SYPRO Orange Protein Gel Stain (Sigma, S5692) and indicated
632 concentration of compound. The signals of all reaction systems were continuously
633 monitored and recorded from 25 °C to 90 °C for approximately 45 minutes. The T_m

634 values of $SPOP^{MATH}$ were measured using CFX manager software version 3.1.

635 **Nuclear magnetic resonance (NMR).** NMR spectroscopy experiments were
636 performed using a 600 MHz spectrometer (AVANCE III, Bruker) to validate protein–
637 ligand interactions. In Carr-Purcell-Meiboom-Gill (CPMG) and saturation transfer
638 difference (STD) NMR experiments, compound was dissolved to a final concentration
639 of 200 μM in a solution of PBS formed with D_2O containing 5 μM $SPOP^{MATH}$ protein
640 and 5% DMSO-d^6.

641 **Surface plasmon resonance (SPR) for RNF130.** The SPR binding assay was
642 performed using a Biacore T200 instrument (GE Healthcare). The purified RNF130
643 protein was covalently immobilized onto a CM5 sensor chip (Cytiva) by a standard
644 amine-coupling procedure in 10 mM sodium acetate (pH 4.5) with running buffer HBS
645 (50 mM HEPES pH 7.4, 150 mM NaCl). iRNF130-63 was serially diluted and injected
646 onto the sensor chip at a flow rate of 30 μL/min for 120 s (contact phase), followed by
647 120 s of buffer flow (dissociation phase). The equilibrium dissociation constant (K_D)
648 value was derived using Biacore T200 Evaluation software (version 1.0, GE
649 Healthcare).

650 **Cell lines.** 293T, MDA-MB-231 and 4T-1 cells were obtained from the American Type
651 Culture Collection (ATCC). OS-RC-2 and Caki-2 cells were purchased from the
652 National Biomedical Laboratory Cell Resource Bank. 786-O cells were kindly provided
653 by the Cell Bank/Stem Cell Bank, Chinese Academy of Sciences. 293T and MDA-MB-
654 231 cells were cultured in DMEM medium (BasalMedia, L110KJ) supplemented with
655 10% foetal bovine serum (FBS, Gibco, 10099141C) and 1% Penicillin-Streptomycin

656 (PS, Gibco, 2321118). 786-O, OS-RC-2 and 4T-1 cells were cultured in RPMI-1640
657 medium (BasalMedia, L210KJ) supplemented with 10% FBS and 1% PS. Caki-2 cells
658 were cultured in McCoy's 5A medium (Gibco, 2193071) supplemented with 10% FBS
659 and 1% PS. All cells were incubated at 37 °C under a 5% (v/v) CO_2 atmosphere.
660 **Western blot.** Total proteins from cells were lysed in RIPA lysis buffer (Beyotime,
661 P0013C) containing phosphatase inhibitor (Bimake, B15001) and protease inhibitor
662 (Bimake, B14001) on ice. Cell lysates were centrifuged at 13,000 × g for 15 minutes at
663 4 °C. The BCA protein assay kit (Thermo Scientific, 23225) was used to quantify the
664 protein concentration. Equal amounts of total proteins were separated by 10% SDS–
665 PAGE and then transferred onto nitrocellulose membranes. The membranes were
666 blocked with 5% skim milk in TBST for 1 hour at room temperature and then incubated
667 with primary antibodies overnight at 4 °C. Then membranes were incubated with HRP-
668 conjugated anti-rabbit antibody (secondary antibody, Promega, W4011) for 1 hour at
669 room temperature. Finally, the immune complexes were detected with an ECL kit
670 (Meilun, MA0186) and visualized as well as quantified using GenGnome XRQ NPC.
671 The following primary antibodies were used: anti-Flag (Cell Signaling Technology,
672 14793), anti-myc (Cell Signaling Technology, 2278), anti-GAPDH (Cell Signaling
673 Technology, 5174), anti-PTEN (Cell Signaling Technology, 9559), anti-HA (Cell
674 Signaling Technology, 3724), anti-p-ERK$^{T202/Y204}$ (Cell Signaling Technology, 4376),
675 anti-p-AKT^{Thr308} (Cell Signaling Technology, 4056), anti-ERK (Cell Signaling
676 Technology, 9102), anti-AKT (Cell Signaling Technology, 9272), anti-DUSP7
677 (ABGENT, AP8450a), anti-SPOP (Abcam, ab192233), anti-GST (Absin, abs830010)

678 and anti-β-Tubulin (Cell Signaling Technology, 15115).

679 ***In vitro* GST pull-down.** The plasmids (Myc-PTEN or Myc-DUSP7) were transiently
680 transfected into 293T cells. After transfection for 24 hours, the cells were harvested and
681 lysed in cell lysis buffer for Western and IP (Beyotime, P0013) containing protease
682 inhibitor on ice. GST or GST-SPOP^{MATH} proteins bound to GST magnetic beads
683 (GenScript, L00327) were incubated with the cell lysates (Myc-PTEN or Myc-DUSP7)
684 in the presence of different doses of compound for 2 hours at room temperature. The
685 beads were washed 3 times with PBST, and the precipitated proteins were eluted with
686 1 × SDS loading buffer (Beyotime, L00327) at 100 °C for 5 minutes and analysed by
687 Western Blot.

688 **Cellular thermal shift assay (CETSA).** 293T cells transfected with Flag-RNF130 for
689 48 hours were collected and lysed in 20 mM Tris pH 7.5, 150 mM NaCl and 1% Triton
690 X-100. Then, 50 μM iRNF130-63 or DMSO was added to the supernatant and
691 incubated at 25 °C for 30 minutes. After denaturing at various temperatures for 3
692 minutes on a temperature gradient PCR instrument (Eppendorf), the samples were
693 centrifuged at 20000 × g for 30 minutes at 4 °C, and the supernatants were analysed by
694 Western blot.

695 **Coimmunoprecipitation (Co-IP).** The plasmids (Flag-SPOP^{cyto}, Myc-PTEN or Myc-
696 DUSP7) were transiently cotransfected into 293T cells. After transfection for 24 hours,
697 293T cells were treated with different doses of compound for another 24 hours. The
698 cells were harvested and lysed in cell lysis buffer for Western and IP containing protease
699 inhibitor on ice. Approximately 80% of the total lysates were immunoprecipitated with

700 anti-Flag-conjugated magnetic beads (Bimake, B26102) for 2 hours at room
701 temperature, and other lysates were used as input. The magnetic beads were then
702 washed 3 times with PBST, and the immunoprecipitated proteins were eluted with 1 ×
703 SDS loading buffer at 100 °C for 5 minutes. The IP and lysate samples were analysed
704 by Western Blot.

705 ***In vivo* ubiquitination.** The plasmids (Myc-PTEN or Myc-DUSP7, Flag-SPOP^{cyto},
706 HA-Ub) were transiently cotransfected into 293T cells. After transfection for 24 hours,
707 293T cells were treated with different doses of compound for 24 hours. The cells were
708 then treated with 10 μM protease inhibitor MG132 (MedChemExpress (Monmouth
709 Junction, NJ, USA), HY-13259) for another 4 hours before harvesting. Next, the cells
710 were lysed in cell lysis buffer for Western and IP containing protease inhibitor on ice.
711 Approximately 80% of the total lysates were immunoprecipitated with anti-Myc-
712 conjugated magnetic beads (Bimake, B26302) for 2 hours at room temperature, and the
713 other lysates were used as input. The magnetic beads were then washed 3 times with
714 PBST, and the immunoprecipitated proteins were eluted with 1 × SDS loading buffer
715 at 100 °C for 5 minutes. The ubiquitination levels were detected using a Western Blot
716 assay.

717 **Cell proliferation.** Cells were seeded in 96-well plates and incubated with serially
718 diluted compounds for 72 hours. Cell viability was determined using the CellTiter-
719 Glo® Luminescent Cell Viability Assay kit (Promega, G7573) following the
720 manufacturer's instructions. IC_{50} values were determined by nonlinear regression
721 (curve fit) using a variable slope (four parameters) in GraphPad Prism (9.0).

722 **Animals.** All procedures performed on animals were in accordance with regulations
723 and established guidelines and were reviewed and approved by the Institutional Animal
724 Care and Use Committee at the Shanghai Institute of Materia Medica, Chinese
725 Academy of Sciences (IACUC Issue NO. 2022-01-JHL-27 for NSG mice; IACUC
726 Issue NO. 2021-03-JHL-22 for BALB/c mice). NSG mice were obtained from
727 Shanghai Model Organisms Center, Inc; BALB/c mice were obtained from Beijing
728 Huafukang Biotechnology Co. Ltd (Beijing, China). Six- to eight-week-old mice were
729 used for the studies and were maintained with free access to pellet food and water in
730 plastic cages ate 21 ± 2°C and kept on a 12 hours light/dark cycle.
731 **Pharmacokinetics.** The pharmacokinetic profiles of compound 230D7 were
732 determined in male BALB/c mice. The test compound 230D7 was dissolved in solution
733 containing DMSO, PEG400, PBS (5/5/90, v/v) and administered via intraperitoneal
734 administration (i.p.) at 10 mg/kg. Serial blood samples were collected at 0.25, 0.5, 1, 2,
735 4, 8, 24 hours after dosing and centrifuged to obtain the plasma fraction. A 10 μL aliquot
736 of plasma was deproteinized with 100 μL acetonitrile/methanol (1/1, v/v) containing
737 internal standard. After centrifugation, the supernatant was diluted with a certain
738 proportion of acetonitrile/water (1/1, v/v), mixed and centrifuged at 4,000 rpm for 10
739 minutes. Finally, the aliquots of the diluted supernatant were injected into LC-MS/MS
740 system.
741 **Acute toxicity.** BALB/c mice were used to evaluate the toxicity of compound 230D7.
742 The mice were randomly divided into 3 groups (n = 3) and treated with different doses
743 of compound 230D7 (0, 50, 100 mg/kg) by intraperitoneal administration daily for a

week. The body weights of mice were measured every day and the significant organs (heart, kidney, lung, liver and spleen) were harvested, weighted and used for histological analysis at the last day.

H&E staining. For histological analysis of BALB/c mice in 230D7-treated or vehicle control groups, staining for H&E staining were performed using standard histological techniques. Isolated organ tissues were fixed in neutral paraformaldehyde for 24 h and embedded in paraffin wax. H&E staining of these tissue examples were carried out by Servicebio company.

786-O cells xenograft tumour growth. NSG mice were used to evaluate the pharmacodynamics of 230D7. The 786-O cells xenograft tumour model was established by the subcutaneous injection of 786-O cells (1×10^7) into the NSG mice. When the tumour reached the volume of approximately 100 mm^3, the mice were randomly divided into three groups (n = 7) and intraperitoneally treated with different dosages of 230D7 (0, 25, 50 mg/kg, 230D7 was synthesised in our laboratory) in solution containing DMSO, PBS (5/90, v/v) once a day for 16 days. Body weight and tumour size were measured every two or three days, and tumour volume was calculated using the formula: $V = (L \times W^2)/2$ (L, length; W, width). At the end of the experiment, the mice were euthanized and the tumours were harvested for Western Blot and other studies.

Statistical analysis. GraphPad Prism 9.0 software was used to perform statistical analysis. Differences of quantitative data between groups were calculated using 2-tailed unpaired t-test. The significance level was set as $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$,

766 ****$P < 0.0001$.

767

768 **Data availability**

769 ChEMBL database is available at https://www.ebi.ac.uk/chembl/, DUDE dataset is
770 available at http://dude.docking.org/, DEKOIS2.0 dataset is available at
771 http://www.pharmchem.uni-tuebingen.de/dekois/, and DrugBank dataset is available at
772 https://go.drugbank.com/. The commercial Chemspace Library is available at
773 https://chem-space.com/, and ChemDiv Library is available at
774 https://www.chemdiv.com/. All data are available in the main text or the supplementary
775 materials.

776

777 **Code availability**

778 Source code of TransformerCPI is available at https://github.com/lifanchen-
779 simm/transformerCPI. The details of upgrading TransformerCPI to TransformerCPI2.0
780 are provided in the Method section.

781

782 ## References

783 1 Gorgulla, C. *et al.* An open-source drug discovery platform enables ultra-large
784 virtual screens. *Nature* **580**, 663-668, doi:10.1038/s41586-020-2117-z (2020).

785 2 Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes.
786 *Nature* **566**, 224-229, doi:10.1038/s41586-019-0917-9 (2019).

787 3 Sadybekov, A. A. *et al.* Synthon-based ligand discovery in virtual libraries of

over 11 billion compounds. *Nature*, doi:10.1038/s41586-021-04220-9 (2021).

4 Zheng, M. *et al.* Computational chemical biology and drug design: Facilitating protein structure, function, and modulation studies. *Med Res Rev* **38**, 914-950, doi:10.1002/med.21483 (2018).

5 Zheng, M. *et al.* Computational methods for drug design and discovery: focus on China. *Trends in Pharmacological Sciences* **34**, 549-559, doi:https://doi.org/10.1016/j.tips.2013.08.004 (2013).

6 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).

7 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876, doi:doi:10.1126/science.abj8754 (2021).

8 Tong, A. B. *et al.* Could AlphaFold revolutionize chemical therapeutics? *Nature Structural & Molecular Biology* **28**, 771-772, doi:10.1038/s41594-021-00670-x (2021).

9 Mullard, A. What does AlphaFold mean for drug discovery? *Nat Rev Drug Discov* **20**, 725-727, doi:10.1038/d41573-021-00161-0 (2021).

10 Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590-596, doi:10.1038/s41586-021-03828-1 (2021).

11 Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews*

810 *Drug Discovery* **3**, 935-949, doi:10.1038/nrd1549 (2004).

811 12 Ni, D., Lu, S. & Zhang, J. Emerging roles of allosteric modulators in the
812 regulation of protein-protein interactions (PPIs): A new paradigm for PPI drug
813 discovery. *Medicinal Research Reviews* **39**, 2314-2342,
814 doi:https://doi.org/10.1002/med.21585 (2019).

815 13 Greener, J. G. & Sternberg, M. J. E. Structure-based prediction of protein
816 allostery. *Current Opinion in Structural Biology* **50**, 1-8,
817 doi:https://doi.org/10.1016/j.sbi.2017.10.002 (2018).

818 14 Stank, A., Kokh, D. B., Fuller, J. C. & Wade, R. C. Protein Binding Pocket
819 Dynamics. *Acc Chem Res* **49**, 809-815, doi:10.1021/acs.accounts.5b00516
820 (2016).

821 15 Teague, S. J. Implications of protein flexibility for drug discovery. *Nature*
822 *Reviews Drug Discovery* **2**, 527-541, doi:10.1038/nrd1129 (2003).

823 16 Zhu, T. *et al.* Hit identification and optimization in virtual screening: practical
824 recommendations based on a critical literature analysis. *Journal of medicinal*
825 *chemistry* **56**, 6560-6572, doi:10.1021/jm301916b (2013).

826 17 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444,
827 doi:10.1038/nature14539 (2015).

828 18 AlQuraishi, M. End-to-End Differentiable Learning of Protein Structure. *Cell*
829 *Syst* **8**, 292-301 e293, doi:10.1016/j.cels.2019.03.006 (2019).

830 19 Tsubaki, M., Tomii, K. & Sese, J. Compound-protein interaction prediction with
831 end-to-end learning of neural networks for graphs and sequences.

Bioinformatics **35**, 309-318, doi:10.1093/bioinformatics/bty535 (2019).

20 Chen, L. *et al.* TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* **36**, 4406-4414, doi:10.1093/bioinformatics/btaa524 (2020).

21 Nguyen, T. *et al.* GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, doi:10.1093/bioinformatics/btaa921 (2020).

22 Li, S. *et al.* MONN: A Multi-objective Neural Network for Predicting Compound-Protein Interactions and Affinities. *Cell Systems* **10**, 308-322.e311, doi:10.1016/j.cels.2020.03.002 (2020).

23 Ozturk, H., Ozgur, A. & Ozkirimli, E. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics* **34**, i821-i829, doi:10.1093/bioinformatics/bty593 (2018).

24 Karimi, M., Wu, D., Wang, Z. & Shen, Y. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **35**, 3329-3338, doi:10.1093/bioinformatics/btz111 (2019).

25 Zhao, Q., Zhao, H., Zheng, K. & Wang, J. HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics*, doi:10.1093/bioinformatics/btab715 (2021).

26 Yang, Z., Zhong, W., Zhao, L. & Chen, C. Y.-C. ML-DTI: Mutual Learning Mechanism for Interpretable Drug–Target Interaction Prediction. *The Journal*

854 *of Physical Chemistry Letters* **12**, 4247-4261, doi:10.1021/acs.jpclett.1c00867
855 (2021).
856 27 Kim, Q., Ko, J.-H., Kim, S., Park, N. & Jhe, W. Bayesian neural network with
857 pretrained protein embedding enhances prediction accuracy of drug-protein
858 interaction. *Bioinformatics* **37**, 3428-3435, doi:10.1093/bioinformatics/btab346
859 (2021).
860 28 Cai, T. *et al.* MSA-Regularized Protein Sequence Transformer toward
861 Predicting Genome-Wide Chemical-Protein Interactions: Application to
862 GPCRome Deorphanization. *Journal of Chemical Information and Modeling*
863 **61**, 1570-1582, doi:10.1021/acs.jcim.0c01285 (2021).
864 29 Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of Useful
865 Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better
866 Benchmarking. *Journal of Medicinal Chemistry* **55**, 6582-6594,
867 doi:10.1021/jm300687e (2012).
868 30 Bauer, M. R., Ibrahim, T. M., Vogel, S. M. & Boeckler, F. M. Evaluation and
869 Optimization of Virtual Screening Workflows with DEKOIS 2.0 – A Public
870 Library of Challenging Docking Benchmark Sets. *Journal of Chemical*
871 *Information and Modeling* **53**, 1447-1462, doi:10.1021/ci400115b (2013).
872 31 Cunningham Brian, C. & Wells James, A. High-Resolution Epitope Mapping of
873 hGH-Receptor Interactions by Alanine-Scanning Mutagenesis. *Science* **244**,
874 1081-1085, doi:10.1126/science.2471267 (1989).
875 32 Cote, B. *et al.* Discovery of MK-1439, an orally bioavailable non-nucleoside

876 reverse transcriptase inhibitor potent against a wide range of resistant mutant
877 HIV viruses. *Bioorg Med Chem Lett* **24**, 917-922,
878 doi:10.1016/j.bmcl.2013.12.070 (2014).
879 33 Wensing, A. M. *et al.* 2019 update of the drug resistance mutations in HIV-1.
880 *Top Antivir Med* **27**, 111-121 (2019).
881 34 Khan, N. H. *et al.* HIV Drug Resistance Mutations in Patients with HIV and
882 HIV-TB Coinfection After Failure of First-Line Therapy: A Prevalence Study in
883 a Resource-Limited Setting. *J Int Assoc Provid AIDS Care* **18**,
884 2325958219849061, doi:10.1177/2325958219849061 (2019).
885 35 Lai, M.-T. *et al.* In vitro characterization of MK-1439, a novel HIV-1
886 nonnucleoside reverse transcriptase inhibitor. *Antimicrob Agents Chemother* **58**,
887 1652-1663, doi:10.1128/AAC.02403-13 (2014).
888 36 Smith, S. J. *et al.* Rilpivirine and Doravirine Have Complementary Efficacies
889 Against NNRTI-Resistant HIV-1 Mutants. *J Acquir Immune Defic Syndr* **72**,
890 485-491, doi:10.1097/QAI.0000000000001031 (2016).
891 37 Stumpfe, D., Hu, Y., Dimova, D. & Bajorath, J. r. Recent progress in
892 understanding activity cliffs and their utility in medicinal chemistry:
893 miniperspective. *Journal of medicinal chemistry* **57**, 18-28 (2014).
894 38 Bajorath, J. Duality of activity cliffs in drug discovery. *Expert opinion on drug*
895 *discovery* **14**, 517-520 (2019).
896 39 Abula, A. *et al.* Substitution Effect of the Trifluoromethyl Group on the
897 Bioactivity in Medicinal Chemistry: Statistical Analysis and Energy

898 Calculations. *Journal of Chemical Information and Modeling* **60**, 6242-6250,
899 doi:10.1021/acs.jcim.0c00898 (2020).
900 40 Zhuang, M. *et al.* Structures of SPOP-Substrate Complexes: Insights into
901 Molecular Architectures of BTB-Cul3 Ubiquitin Ligases. *Molecular Cell* **36**,
902 39-50, doi:https://doi.org/10.1016/j.molcel.2009.09.022 (2009).
903 41 Xu, L. *et al.* BTB proteins are substrate-specific adaptors in an SCF-like
904 modular ubiquitin ligase containing CUL-3. *Nature* **425**, 316-321,
905 doi:10.1038/nature01985 (2003).
906 42 Guo, Z.-Q. *et al.* Small-Molecule Targeting of E3 Ligase Adaptor SPOP in
907 Kidney Cancer. *Cancer Cell* **30**, 474-484,
908 doi:https://doi.org/10.1016/j.ccell.2016.08.003 (2016).
909 43 Li, G. *et al.* SPOP Promotes Tumorigenesis by Acting as a Key Regulatory Hub
910 in Kidney Cancer. *Cancer Cell* **25**, 455-468,
911 doi:https://doi.org/10.1016/j.ccr.2014.02.007 (2014).
912 44 Chappell, J., Sun, Y., Singh, A. & Dalton, S. MYC/MAX control ERK signaling
913 and pluripotency by regulation of dual-specificity phosphatases 2 and 7. *Genes*
914 *Dev* **27**, 725-733, doi:10.1101/gad.211300.112 (2013).
915 45 Palm, K., Stenberg, P., Luthman, K. & Artursson, P. Polar Molecular Surface
916 Properties Predict the Intestinal Absorption of Drugs in Humans.
917 *Pharmaceutical Research* **14**, 568-571, doi:10.1023/A:1012188625088 (1997).
918 46 Békés, M., Langley, D. R. & Crews, C. M. PROTAC targeted protein degraders:
919 the past is prologue. *Nature Reviews Drug Discovery*, doi:10.1038/s41573-021-

920 00371-6 (2022).

921 47 Yang, T. *et al.* DrugSpaceX: a large screenable and synthetically tractable
922 database extending drug space. *Nucleic Acids Res* **49**, D1170-D1178,
923 doi:10.1093/nar/gkaa920 (2021).

924 48 Grygorenko, O. O. *et al.* Generating Multibillion Chemical Space of Readily
925 Accessible Screening Compounds. *iScience* **23**, doi:10.1016/j.isci.2020.101681
926 (2020).

927 49 Bileschi, M. L. *et al.* Using deep learning to annotate the protein universe.
928 *Nature Biotechnology*, doi:10.1038/s41587-021-01179-w (2022).

929 50 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep
930 Bidirectional Transformers for Language Understanding. *arXiv e-prints*,
931 arXiv:1810.04805 (2018).

932 51 Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified
933 rational protein engineering with sequence-based deep representation learning.
934 *Nat Methods* **16**, 1315-1322, doi:10.1038/s41592-019-0598-1 (2019).

935 52 Bepler, T. & Berger, B. *Learning protein sequence embeddings using*
936 *information from structure*. (2019).

937 53 Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph
938 Convolutional Networks. *arXiv e-prints*, arXiv:1609.02907 (2016).

939 54 Liu, L. *et al.* On the Variance of the Adaptive Learning Rate and Beyond. *arXiv*
940 *e-prints*, arXiv:1908.03265 (2019).

941 55 Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res* **45**, D945-

D954, doi:10.1093/nar/gkw1074 (2017).

56 Papadatos, G., Gaulton, A., Hersey, A. & Overington, J. P. Activity, assay and target data curation and quality in the ChEMBL database. *Journal of Computer-Aided Molecular Design* **29**, 885-896, doi:10.1007/s10822-015-9860-5 (2015).

57 Kramer, C., Kalliokoski, T., Gedeck, P. & Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public Ki Data. *Journal of Medicinal Chemistry* **55**, 5165-5173, doi:10.1021/jm300131x (2012).

58 Lenselink, E. B. *et al.* Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of Cheminformatics* **9**, 45, doi:10.1186/s13321-017-0232-0 (2017).

59 Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry* **47**, 1739-1749, doi:10.1021/jm0306430 (2004).

60 Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking11Edited by F. E. Cohen. *Journal of Molecular Biology* **267**, 727-748, doi:https://doi.org/10.1006/jmbi.1996.0897 (1997).

61 Zhang, N. & Zhao, H. Enriching screening libraries with bioactive fragment space. *Bioorg Med Chem Lett* **26**, 3594-3597, doi:10.1016/j.bmcl.2016.06.013 (2016).

62 Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.

964 *J Comput Chem* **31**, 455-461, doi:10.1002/jcc.21334 (2010).

965 63 Shen, C. *et al.* Beware of the generic machine learning-based scoring functions
966 in structure-based virtual screening. *Brief Bioinform*, doi:10.1093/bib/bbaa070
967 (2020).

968 64 Zhuang, M. *et al.* Structures of SPOP-substrate complexes: insights into
969 molecular architectures of BTB-Cul3 ubiquitin ligases. *Mol Cell* **36**, 39-50,
970 doi:10.1016/j.molcel.2009.09.022 (2009).

971 65 Guo, Z. Q. *et al.* Small-Molecule Targeting of E3 Ligase Adaptor SPOP in
972 Kidney Cancer. *Cancer Cell* **30**, 474-484, doi:10.1016/j.ccell.2016.08.003
973 (2016).

974

975 **Competing interests**

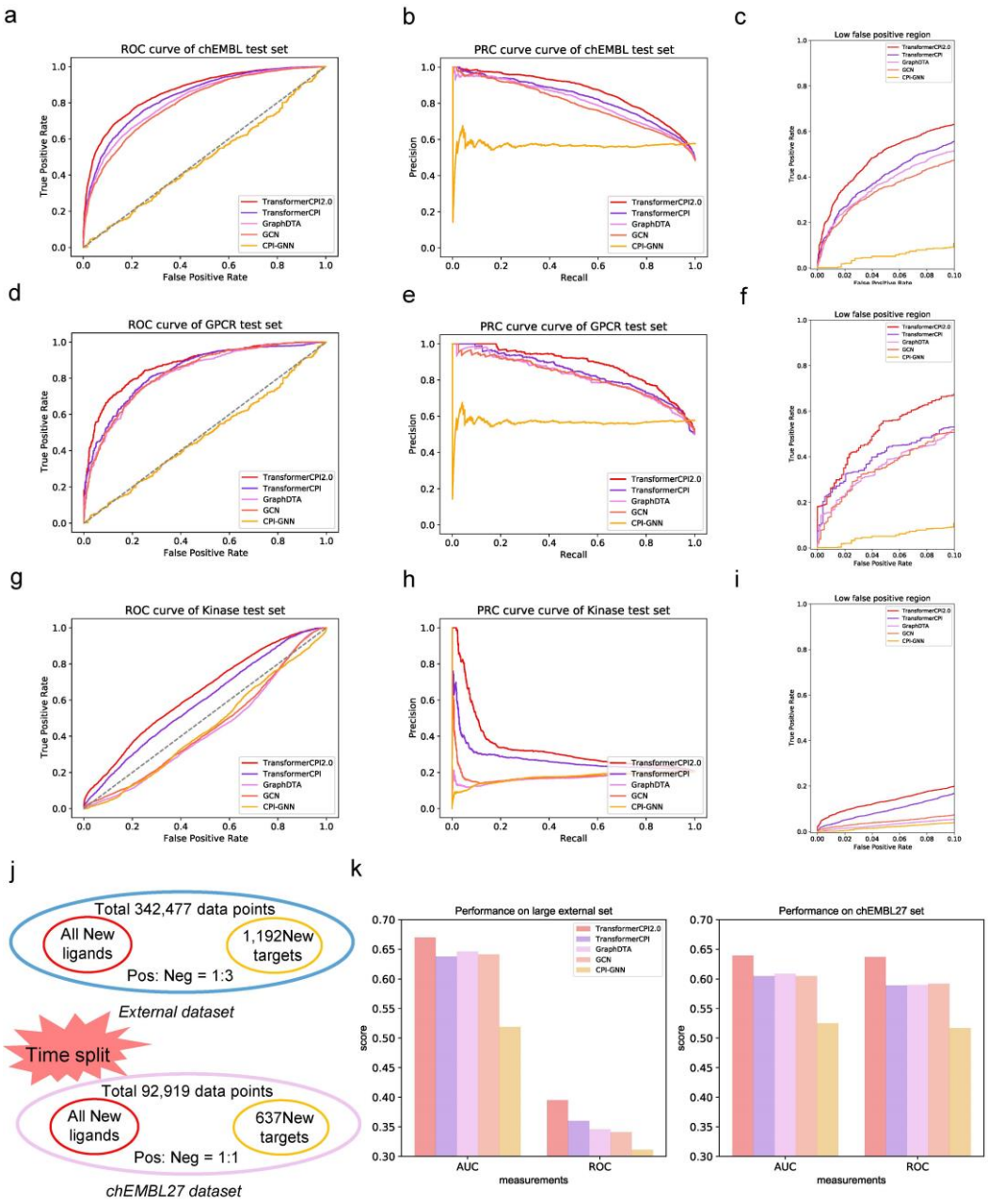
976 The authors declare that they have no competing interests.

977

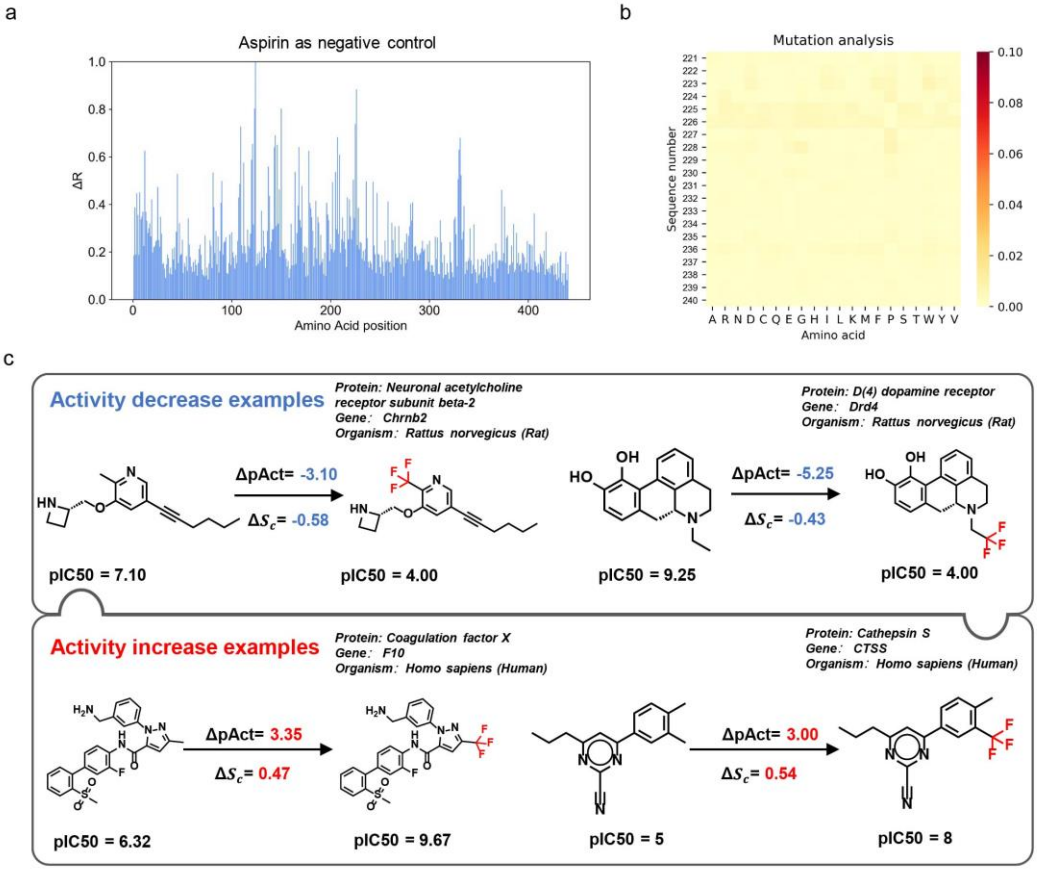
978 **Acknowledgements**

979 We gratefully acknowledge financial support from Lingang Laboratory (LG202102-01-
980 02 to M.Z. and LG-QS-202204-01 to S.Z.), the National Natural Science Foundation
981 of China (81903639 to S.Z.), the Shanghai Municipal Science and Technology Major
982 Project and the Natural Science Foundation of Shanghai (22ZR1474300 to S.Z.).

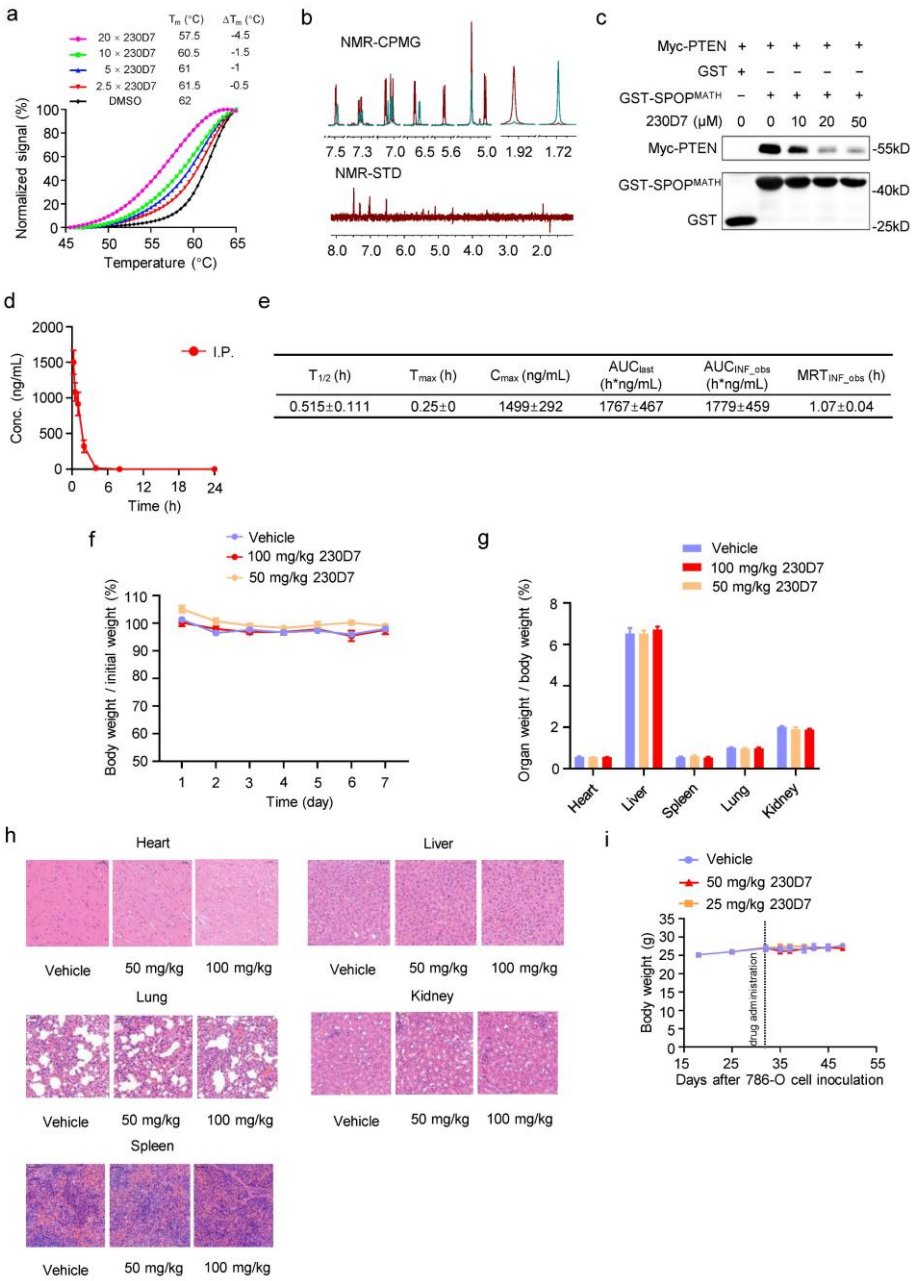
983



Extended Data Fig. 1 | **TransformerCPI 2.0 predicts compound–protein interaction without protein structure. a,** AUC curves of TransformerCPI2.0 and baseline models on the chEMBL set. **b**, PRC curves of TransformerCPI2.0 and baseline models on the chEMBL set. **c**, ROC curves in low-false-positive region. **d,** AUC curves of TransformerCPI2.0 and baseline models on the GPCR set. **e**, PRC curves of TransformerCPI2.0 and baseline models on the GPCR set. **f**, ROC curves in a low-false-positive region. **g**, AUC curves of TransformerCPI2.0 and baseline models on the Kinase set. **h**, PRC curves of TransformerCPI2.0 and baseline models on the Kinase set. **i,** ROC curves in a low-false-positive region. **j**, The arrangement of the external dataset and chEMBL27 dataset. **k**, The performance of TransformerCPI2.0 and baseline models on the external set and chEMBL27 set.

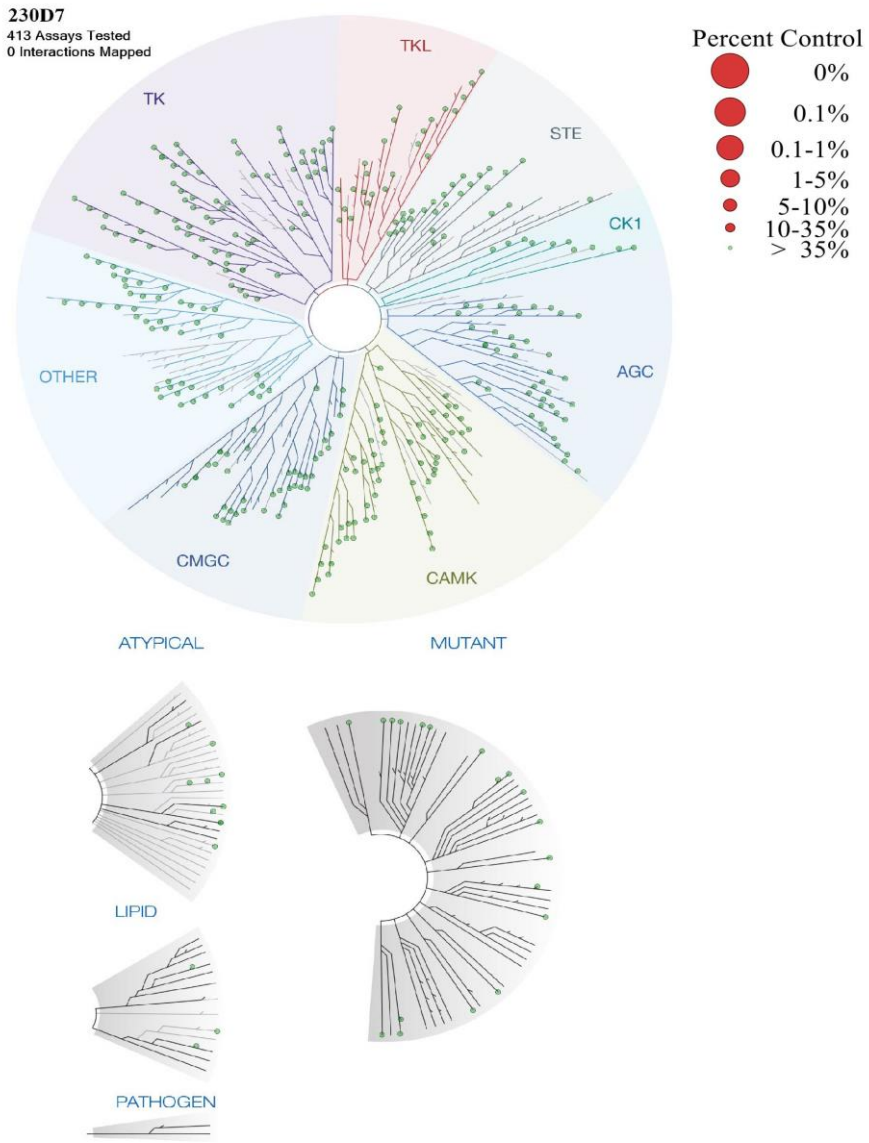


Extended Data Fig 2. | Supplementary data from drug resistance mutation analysis and substitution effect analysis of the trifluoromethyl group. a, Relative activity change score (ΔR) calculated by TransformerCPI 2.0 for each amino acid position. The pink boxes mark the high ΔR regions, which are plotted in detail on the right. **b**, The heatmap plots the activity change score of positions 221 to 240, where each position is mutated to the 20 different amino acids (including itself). The darker colour represents the higher activity change score caused by the mutation. **c**, Additional two activity decrease examples and activity increase examples are shown. The predictions of TransformerCPI2.0 are consistent with the ground truth for proteins and compounds not present in the training set.



Extended Data Fig 3. | Supplementary data for Fig. 4. a, Thermostability of SPOP^{MATH} treated with different concentrations of 230D7. The thermal stability of SPOP^{MATH} was quantified by the ΔT_m. **b**, NMR measurement of direct binding between 230D7 and SPOP^{MATH}. CPMG NMR spectra for 230D7 (red), 230D7 in the presence of 5 μM SPOP^{MATH} (green). The STD spectrum for 230D7 is recorded in the presence of 5 μM SPOP^{MATH}. **c**, 230D7 disrupts protein binding between SPOP^{MATH} and PTEN, as measured by *in vitro* pull-down assay. **d**, Concentrations (ng/mL) of 230D7 in BALB/c mice plasma after i.p. administration of 10 mg/kg 230D7. **e**, The Pharmacokinetic parameters of 230D7 after an intraperitoneal dose in BALB/c mice. **f**, The body weight of BALB/c mice treated with different dosages of 230D7 daily for 7 days (n=3 mice each group). **g**, The weight of different organs (heart, liver, spleen, lung, and kidney) of BALB/c mice treated with different dosages of 230D7 daily for 7 days (n=3 mice each group). **h**, Representative histological morphology of H&E-stained tissue sections of BALB/c mice in 230D7-treated or vehicle control groups. **i**, The body weight of NSG mice were measured during the entire pharmacodynamics study of 230D7.

987
988



Extended Data Fig 4. | Kinome profiling of 230D7. 230D7 (10 μM) was submitted for a KinaseProfiler (eurofins) to quantify interactions with 413 human wild-type/mutant kinases. The results are displayed as a TREESPOT interaction map. Image generated using TREEspot™ Software Tool and reprinted with permission from KINOMEscan®, a division of DiscoveRx Corporation, © DISCOVERX CORPORATION 2010.