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), ring finger protein 130 (RNF130) which does not have a 3D structure, and repurposed proton pump inhibitors (PPIs) for ADP-ribosylation factor 1 (ARF1). 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 AlphaFold[6] also relies heavily on the idea of end-to-end differentiability. Therefore, this new paradigm is appealing because it performs the entire learning process in a self-consistent and data-efficient manner, avoiding the error accumulation of complex pipelines.

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

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. Additionally, we identified ADP-ribosylation factor 1 (ARF1) as a new target for proton pump inhibitors (PPIs). 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 the other three external datasets: GPCR set, Kinase set, and a large external set. TransformerCPI2.0 also showed the greatest generalization ability among all the models (Extended Data Fig. 1d~i and Supplementary Table 1~3). Furthermore, we designed a time-split test set named the chEMBL27 dataset to predict new data that were deposited online after the training set, and TransformerCPI2.0 still outperformed the baseline models (Extended Data Fig. 1j, k). Since our training set was generated from chEMBL23, this test suggested that our model can learn from past knowledge and generalize to future data.

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

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

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

199 than TransformerCPI, GraphDTA, GCN and random guessing (Fig. 2f and
200 Supplementary Table 8). In addition, we investigated the performance of
201 TransformerCPI2.0 and baseline models on a subset where substitution of −CH3 by
202 −CF3 could increase or decrease the biological activity by at least three orders of
203 magnitude. TransformerCPI2.0 still outperformed the other models, and the accuracy
204 on this subset was higher than that on the whole dataset (Fig. 2g and Supplementary
205 Table 9). The predictive capacity of this test is much more challenging than that of the
206 whole dataset because the drastic biological activity change at this range involves
207 turning an active compound into an inactive one or vice versa.

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

216

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

218 SPOP, functioning as an adaptor of cullin3-RING ubiquitin ligase, mediates substrate
219 protein recognition and ubiquitination[40,41]. A previous study validated SPOP as an
220 attractive target for the treatment of clear-cell renal cell carcinoma (ccRCC), but it is a

221 challenging target in terms of protein–protein interactions[42]. SPOP is overexpressed and
222 misallocated in the cytoplasm of ccRCC cells, which induces proliferation and
223 promotes kidney tumorigenesis[43]. Phosphatase and tensin homologue (PTEN) and dual
224 specificity phosphatase 7 (DUSP7) are two substrates of SPOP[43]. PTEN acts as a
225 negative regulator of the phosphoinositide 3-kinase/AKT pathway, and DUSP7
226 dephosphorylates extracellular signal-regulated kinase (ERK)[44]. Accumulation of
227 cytoplasmic SPOP in ccRCC cells decreases cellular PTEN and DUSP7 by mediating
228 the degradation of these two cytoplasmic proteins, leading to an increase in
229 phosphorylated AKT and ERK and thus promoting the proliferation of ccRCC cells[43].

230 Here, virtual screening with TransformerCPI2.0 was performed to discover new
231 scaffold compounds that directly target SPOP (Fig. 3a, Supplementary Table 10) to test
232 the feasibility of the sequence-to-drug paradigm on challenging targets. Four
233 compounds were identified as initial hits by a fluorescence polarization (FP) assay (hit
234 rate ~5%), and 221C7 was the most active compound with an IC_{50} of 4.30 μM (Fig. 3b,
235 c, Extended Data Fig. 3a, b). To further confirm that 221C7 disrupts SPOP-substrate
236 interactions, an *in vitro* pull-down assay was performed. The results revealed that
237 compound 221C7 dose-dependently reduced PTEN protein binding to the SPOP MATH
238 domain ($SPOP^{MATH}$) (Fig. 3d). Then, a nuclear magnetic resonance (NMR) experiment
239 was conducted (Fig. 3e), and the result indicated direct binding between $SPOP^{MATH}$ and
240 221C7. These results verified that 221C7 disrupts SPOP-substrate interactions by
241 directly binding to $SPOP^{MATH}$. Drug resistance mutation analysis was applied to
242 interpret the prediction of TransformerCPI2.0, and important residues were mapped to

243 the structure of full-length SPOP (PDB: 3HQI, Fig. 3f). Important residues mostly
244 located on the MATH domain, and we then docked 221C7 to the predicted binding site
245 (Fig. 3g).

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

259 We conducted a coimmunoprecipitation and *in vivo* ubiquitination experiment at
260 the cellular level, and the results showed that 230D7 significantly disrupted the binding
261 of PTEN and DUSP7 to SPOP in a dose-dependent manner (Fig. 4c, d), leading to
262 decreases in PTEN and DUSP7 ubiquitination (Fig. 4e, f). Due to the inhibition of
263 PTEN and DUSP7 ubiquitination under 230D7 treatment, accumulation of cellular
264 PTEN and DUSP7 proteins was observed in 786-O cells treated with 230D7, causing

265 decreases in phosphorylated AKT and ERK (Fig. 4g, h). Next, we tested the cell
266 proliferation of three ccRCC cell lines (786-O, Caki-2, OS-CR-2) and two non-ccRCC
267 cell lines (4T-1, MDA-MB-231) in the presence of 230D7 (Fig. 4i). 230D7 specifically
268 inhibited the growth of ccRCC cell lines with an IC_{50} of approximately 20 μM
269 compared with non-ccRCC cell lines. To determine if 230D7 is suitable for *in vivo*
270 studies, we investigated the pharmacokinetics and acute toxicity profile of 230D7.
271 230D7 can be efficiently absorbed into the blood circulation following intraperitoneal
272 injection and has low acute toxicity (Extended Data Fig. 4d~h). A dose-dependent
273 reduction in 786-O tumour growth rate could be observed in NSG mice treated with
274 230D7 (Fig. 4j), revealing significant anti-ccRCC therapeutic effect of 230D7 *in vivo*.
275 No body weight loss in NSG mice was statistically noted during the entire
276 pharmacodynamics study of 230D7 (Extended Data Fig. 4i). Finally, we checked the
277 effect of 230D7 on oncogenic SPOP signalling in ccRCC xenograft tumours. As
278 expected, PTEN and DUSP7 are elevated in 230D7 treated groups, and p-AKT and p-
279 ERK levels are decreased (Fig. 4k). Moreover, we confirmed the high selectivity of
280 230D7 which does not target kinases (Extended Data Fig. 5). In conclusion, we
281 successfully discovered novel small-molecule inhibitors that target SPOP through a
282 sequence-to-drug paradigm, among which 230D7 showed therapeutic potential for
283 blocking SPOP activity to treat ccRCC.

284 After discovering novel inhibitors for SPOP, we applied this paradigm to discover
285 hits for a more challenging target RNF130 whose crystal structure is unknown. RNF130
286 is an E3 ubiquitin-protein ligase without structural information, and no chemical

287 binders have been reported. Our recent study revealed that RNF130 plays an important
288 role in autoimmune inflammation, suggesting its inhibition could be of potential
289 therapeutic value (unpublished data, under review). We utilized TransformerCPI2.0 to
290 screen compounds that directly bind to RNF130 (Extended Fig. 6a, Supplementary
291 Table 11) and discovered iRNF130-63 to be a binder of RNF130 (Extended Fig. 6b~d).
292 The direct binding between iRNF130-63 and RNF130 protein was confirmed through
293 surface plasmon resonance (SPR), and this binding exhibited a fast-on, fast-off kinetic
294 pattern with a K_D of 9.36 μM (Extended Fig. 6c). We also performed a cellular thermal
295 shift assay (CETSA), and the results supported that iRNF130-63 directly binds to and
296 thermally stabilizes the RNF130 proteins (Extended Fig. 6d). The success of
297 discovering novel hits for SPOP and RNF130 demonstrated that the sequence-to-drug
298 paradigm is practicable for virtual screening with encouraging prospects.
299
300 **Repositioning proton pump inhibitors as anticancer drugs by targeting ARF1.**
301 Benefiting from the end-to-end nature, the sequence-to-drug workflow can be inversely
302 used to enable drug target identification or drug repurposing. This means that we can
303 perform proteome-wide target screening, as only protein sequence information is
304 required except for a given drug molecule as the model input. Here, we select proton
305 pump inhibitors (PPIs) as a drug repurposing case study. To date, preclinical data and
306 clinical data support the use of PPIs in cancer treatment[46], but few new targets have
307 been found. TransformerCPI2.0 was applied to score 2,204 human proteins from the
308 DrugBank database[47] against four classical PPIs (rabeprazole, lansoprazole,

309 omeprazole and pantoprazole, Fig. 5a~b, Supplementary Table 12), and the results were
310 sorted by predicted interaction probability. After analysing the top 20 proteins, ARF1
311 attracted our attention due to its oncogenic effect on cancer stem cells (CSCs) by the
312 lipolysis pathway[48,49]. ARF1 is a small G protein and belongs to the RAS superfamily,
313 which cycles between an active GTP-bound and an inactive GDP-bound
314 conformation[50]. Recent studies have shown that the lipid metabolism regulated by
315 ARF1 selectively maintains cancer stem cells (CSCs), and ARF1 inhibition or
316 knockdown in CSCs leads to the accumulation of lipid droplets, further leading to
317 metabolic stress, which can not only selectively kill CSCs but also stimulate an
318 anticancer immune response and achieve lasting therapeutic effects[48,49]. Inhibition of
319 ARF1 activity is a promising direction for cancer immunotherapy, therefore, we
320 selected ARF1 for investigation.

321 PTS assay revealed dose-dependent T_m shifts (Fig. 5c, Extended Data Fig. 7b~d),
322 indicating that PPIs could directly bind to wild-type ARF1 ($ARF1^{WT}$) and destabilize
323 the protein. Drug resistance mutation analysis was then applied to interpret the
324 prediction of TransformerCPI2.0. The results indicated that amino acids 150~165, a
325 range that contains a cysteine (C159), contributed greatly to compound protein binding
326 (Fig. 5d, Extended Data Fig. 7a). Given that PPIs covalently bind to the cysteine of
327 H^+/K^+-ATPase[51], we conducted different assays to determine whether PPIs covalently
328 bind to ARF1. Therefore, another two PTS assays were conducted: 1) PPIs with
329 $ARF1^{WT}$ and the reducing agent dithiothreitol (DTT), which can break disulfide bonds;
330 and 2) PPIs with $ARF1^{C159A}$ where C159 was mutated to alanine. No T_m shifts were

331 observed in neither two assays (Fig. 5e, Extended Data Fig. 7b~d). Mass spectrometry
332 (MS) further validated that PPIs can covalently bind to $ARF1^{WT}$ but cannot bind to
333 $ARF1^{C159A}$ (Fig. 5f, Extended Data Fig. 7e~f), and two-dimensional mass spectrometry
334 determined the covalent binding site is C159 (Extended Data Fig. 8a~d). Among the
335 four PPIs, rabeprazole had the greatest effect on the thermal stability of ARF1 (Fig. 5g),
336 thus we selected rabeprazole for further functional study. According to the covalent
337 binding site, we provided a possible docking pose of rabeprazole (Fig. 5h). Activation
338 of ARF1 requires the release of GDP followed by the binding of GTP, a process
339 catalysed by guanine nucleotide exchange factor (GEF)[50]. Therefore, we performed
340 GDP/MANT-GTP nucleotide exchange catalysed by ARNO (a type of GEF) and found
341 that rabeprazole suppressed the nucleotide exchange process in a concentration-
342 dependent manner (Fig. 5i), verifying its inhibition of ARF1 activity.

343 According to previous works[48,49], we first detected the inhibitory effect of
344 rabeprazole on the activity level of ARF1 in CT26 cells (colon carcinoma cells) using
345 a G-LISA assay. The results showed that rabeprazole effectively inhibited ARF1
346 activity in CT26 cells in a concentration-dependent manner (Fig. 6a). In addition,
347 significant accumulation of lipid droplets was observed in CT26 cells treated with
348 rabeprazole (Fig. 6b). To evaluate the antitumour effect of rabeprazole *in vivo*, we
349 established colon cancer transplanted tumour models by injecting CT26 cells into
350 BALB/c mice. Rabeprazole treatment significantly suppressed the tumour growth in
351 mice as measured by tumour volume (Fig. 6c). In order to verify that rabeprazole induce
352 an antitumour immune response, we analysed the immune cell subsets of colon cancer

353 transplanted tumours by fluorescence-activated cell sorting (FACS) and found a
354 significant increase in $CD3^+$ $CD8^+$ T cells, as well as a significant decrease in $CD3^+$
355 $CD8^+$ $PD1^+$ T cells, $CD3^+$ $CD8^+$ $TIM3^+$ T cells and $CD3^+$ $CD8^+$ $PD1^+$ $TIM3^+$ T cells (Fig.
356 6d). Additionally, upregulation of CD8 and downregulation of PD1 were detected by
357 immunohistochemical staining (Fig. 6e), confirming that an antitumour immune
358 response was stimulated by rabeprazole. All these data suggest that rabeprazole inhibits
359 the growth of colon cancer by inducing an antitumour immune response. In summary,
360 the success of repurposing PPIs to ARF1 demonstrated that the inverse application of
361 the sequence-to-drug paradigm for drug repositioning is also practicable with
362 encouraging prospects.
363

364 **Discussion**

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

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

376 To address this problem, we proposed a new sequence-to-drug paradigm that
377 discovers modulators directly from protein sequences, jumping over intermediate steps.
378 This proof-of-concept study analyses the sequence-to-drug paradigm in three aspects:
379 its competitive performance against 3D structure-based molecular docking, its learning
380 of binding knowledge, and drug design for challenging targets and drug repurposing.
381 Apart from methodology, the sequence-to-drug paradigm successfully discovered novel
382 inhibitors for SPOP, among which 230D7 showed potential to treat ccRCC, and
383 discovered the first binder of RNF130. Given that SPOP is an adaptor of E3 ligase and
384 RNF130 is an E3 ligase, the hits we found have the potential to serve as novel warheads
385 for proteolysis-targeting chimaeras (PROTACs). PROTACs have been successfully
386 developed for harnessing the ubiquitin–proteasome system to degrade a protein of
387 interest, receiving tremendous attention as a new and exciting class of therapeutic
388 agents that promise to significantly impact drug discovery[52]. Due to the limited
389 structural knowledge of E3 ligases and their substrate recognition and regulation of
390 ubiquitination, the hits we discovered may promote the development of PROTACs.
391 Furthermore, through an inverse application of the sequence-to-drug paradigm, the
392 FDA-approved drug rabeprazole showed promise for expanding its indications to colon
393 cancer treatment by regulating lipid metabolism and inducing an antitumour immune
394 response. All these results demonstrate that the sequence-to-drug paradigm is a
395 promising direction for drug design.

396 Recently, due to the rapid growth of virtual libraries, such as DrugSpaceX[53] and

397 synthesis (REAL) combinatorial libraries[54] which cover spaces of 100 million to
398 multiple billions of chemicals, there is a high demand for developing computationally
399 efficient virtual screening approaches. The sequence-to-drug paradigm can be
400 combined with these large virtual libraries to rapidly discover new active scaffolds from
401 the unexplored chemical space. On the one hand, the explosion in available genomic
402 sequencing techniques and annotation techniques has revolutionized bioinformatics[55],
403 and the sequence-to-drug paradigm will likewise evolve rapidly by incorporating more
404 abundant multiple sequence alignment or functional annotation information. On the
405 other hand, we may perform more comprehensive proteome-wide virtual screening,
406 accelerating the discovery of new drug hits for novel but challenging biological targets.
407 Overall, our findings provide a proof of concept for the sequence-to-drug paradigm,
408 which we believe will become an essential component of future rational drug design
409 pipelines.
410

411 **References**

412 1 Gorgulla, C. *et al.* An open-source drug discovery platform enables ultra-large
413 virtual screens. *Nature* **580**, 663-668, doi:10.1038/s41586-020-2117-z (2020).
414 2 Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes.
415 *Nature* **566**, 224-229, doi:10.1038/s41586-019-0917-9 (2019).
416 3 Sadybekov, A. A. *et al.* Synthon-based ligand discovery in virtual libraries of
417 over 11 billion compounds. *Nature*, doi:10.1038/s41586-021-04220-9 (2021).
418 4 Zheng, M. *et al.* Computational chemical biology and drug design: Facilitating

419 protein structure, function, and modulation studies. *Med Res Rev* **38**, 914-950,
420 doi:10.1002/med.21483 (2018).
421 5 Zheng, M. *et al.* Computational methods for drug design and discovery: focus
422 on China. *Trends in Pharmacological Sciences* **34**, 549-559,
423 doi:https://doi.org/10.1016/j.tips.2013.08.004 (2013).
424 6 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold.
425 *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).
426 7 Baek, M. *et al.* Accurate prediction of protein structures and interactions using
427 a three-track neural network. *Science* **373**, 871-876,
428 doi:doi:10.1126/science.abj8754 (2021).
429 8 Tong, A. B. *et al.* Could AlphaFold revolutionize chemical therapeutics? *Nature*
430 *Structural & Molecular Biology* **28**, 771-772, doi:10.1038/s41594-021-00670-
431 x (2021).
432 9 Mullard, A. What does AlphaFold mean for drug discovery? *Nat Rev Drug*
433 *Discov* **20**, 725-727, doi:10.1038/d41573-021-00161-0 (2021).
434 10 Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the
435 human proteome. *Nature* **596**, 590-596, doi:10.1038/s41586-021-03828-1
436 (2021).
437 11 Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in
438 virtual screening for drug discovery: methods and applications. *Nature Reviews*
439 *Drug Discovery* **3**, 935-949, doi:10.1038/nrd1549 (2004).
440 12 Ni, D., Lu, S. & Zhang, J. Emerging roles of allosteric modulators in the

441 regulation of protein-protein interactions (PPIs): A new paradigm for PPI drug
442 discovery. *Medicinal Research Reviews* **39**, 2314-2342,
443 doi:https://doi.org/10.1002/med.21585 (2019).
444 13 Greener, J. G. & Sternberg, M. J. E. Structure-based prediction of protein
445 allostery. *Current Opinion in Structural Biology* **50**, 1-8,
446 doi:https://doi.org/10.1016/j.sbi.2017.10.002 (2018).
447 14 Stank, A., Kokh, D. B., Fuller, J. C. & Wade, R. C. Protein Binding Pocket
448 Dynamics. *Acc Chem Res* **49**, 809-815, doi:10.1021/acs.accounts.5b00516
449 (2016).
450 15 Teague, S. J. Implications of protein flexibility for drug discovery. *Nature*
451 *Reviews Drug Discovery* **2**, 527-541, doi:10.1038/nrd1129 (2003).
452 16 Zhu, T. *et al.* Hit identification and optimization in virtual screening: practical
453 recommendations based on a critical literature analysis. *Journal of medicinal*
454 *chemistry* **56**, 6560-6572, doi:10.1021/jm301916b (2013).
455 17 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444,
456 doi:10.1038/nature14539 (2015).
457 18 AlQuraishi, M. End-to-End Differentiable Learning of Protein Structure. *Cell*
458 *Syst* **8**, 292-301 e293, doi:10.1016/j.cels.2019.03.006 (2019).
459 19 Tsubaki, M., Tomii, K. & Sese, J. Compound-protein interaction prediction with
460 end-to-end learning of neural networks for graphs and sequences.
461 *Bioinformatics* **35**, 309-318, doi:10.1093/bioinformatics/bty535 (2019).
462 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 of Physical Chemistry Letters* **12**, 4247-4261, doi:10.1021/acs.jpclett.1c00867 (2021).

485 27 Kim, Q., Ko, J.-H., Kim, S., Park, N. & Jhe, W. Bayesian neural network with
486 pretrained protein embedding enhances prediction accuracy of drug-protein
487 interaction. *Bioinformatics* **37**, 3428-3435, doi:10.1093/bioinformatics/btab346
488 (2021).
489 28 Cai, T. *et al.* MSA-Regularized Protein Sequence Transformer toward
490 Predicting Genome-Wide Chemical-Protein Interactions: Application to
491 GPCRome Deorphanization. *Journal of Chemical Information and Modeling*
492 **61**, 1570-1582, doi:10.1021/acs.jcim.0c01285 (2021).
493 29 Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of Useful
494 Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better
495 Benchmarking. *Journal of Medicinal Chemistry* **55**, 6582-6594,
496 doi:10.1021/jm300687e (2012).
497 30 Bauer, M. R., Ibrahim, T. M., Vogel, S. M. & Boeckler, F. M. Evaluation and
498 Optimization of Virtual Screening Workflows with DEKOIS 2.0 – A Public
499 Library of Challenging Docking Benchmark Sets. *Journal of Chemical*
500 *Information and Modeling* **53**, 1447-1462, doi:10.1021/ci400115b (2013).
501 31 Cunningham Brian, C. & Wells James, A. High-Resolution Epitope Mapping of
502 hGH-Receptor Interactions by Alanine-Scanning Mutagenesis. *Science* **244**,
503 1081-1085, doi:10.1126/science.2471267 (1989).
504 32 Cote, B. *et al.* Discovery of MK-1439, an orally bioavailable non-nucleoside
505 reverse transcriptase inhibitor potent against a wide range of resistant mutant
506 HIV viruses. *Bioorg Med Chem Lett* **24**, 917-922,

507 doi:10.1016/j.bmcl.2013.12.070 (2014).

508 33 Wensing, A. M. *et al.* 2019 update of the drug resistance mutations in HIV-1.
509 *Top Antivir Med* **27**, 111-121 (2019).

510 34 Khan, N. H. *et al.* HIV Drug Resistance Mutations in Patients with HIV and
511 HIV-TB Coinfection After Failure of First-Line Therapy: A Prevalence Study in
512 a Resource-Limited Setting. *J Int Assoc Provid AIDS Care* **18**,
513 2325958219849061, doi:10.1177/2325958219849061 (2019).

514 35 Lai, M.-T. *et al.* In vitro characterization of MK-1439, a novel HIV-1
515 nonnucleoside reverse transcriptase inhibitor. *Antimicrob Agents Chemother* **58**,
516 1652-1663, doi:10.1128/AAC.02403-13 (2014).

517 36 Smith, S. J. *et al.* Rilpivirine and Doravirine Have Complementary Efficacies
518 Against NNRTI-Resistant HIV-1 Mutants. *J Acquir Immune Defic Syndr* **72**,
519 485-491, doi:10.1097/QAI.0000000000001031 (2016).

520 37 Stumpfe, D., Hu, Y., Dimova, D. & Bajorath, J. r. Recent progress in
521 understanding activity cliffs and their utility in medicinal chemistry:
522 miniperspective. *Journal of medicinal chemistry* **57**, 18-28 (2014).

523 38 Bajorath, J. Duality of activity cliffs in drug discovery. *Expert opinion on drug*
524 *discovery* **14**, 517-520 (2019).

525 39 Abula, A. *et al.* Substitution Effect of the Trifluoromethyl Group on the
526 Bioactivity in Medicinal Chemistry: Statistical Analysis and Energy
527 Calculations. *Journal of Chemical Information and Modeling* **60**, 6242-6250,
528 doi:10.1021/acs.jcim.0c00898 (2020).

529 40 Zhuang, M. *et al.* Structures of SPOP-Substrate Complexes: Insights into
530 Molecular Architectures of BTB-Cul3 Ubiquitin Ligases. *Molecular Cell* **36**,
531 39-50, doi:https://doi.org/10.1016/j.molcel.2009.09.022 (2009).
532 41 Xu, L. *et al.* BTB proteins are substrate-specific adaptors in an SCF-like
533 modular ubiquitin ligase containing CUL-3. *Nature* **425**, 316-321,
534 doi:10.1038/nature01985 (2003).
535 42 Guo, Z.-Q. *et al.* Small-Molecule Targeting of E3 Ligase Adaptor SPOP in
536 Kidney Cancer. *Cancer Cell* **30**, 474-484,
537 doi:https://doi.org/10.1016/j.ccell.2016.08.003 (2016).
538 43 Li, G. *et al.* SPOP Promotes Tumorigenesis by Acting as a Key Regulatory Hub
539 in Kidney Cancer. *Cancer Cell* **25**, 455-468,
540 doi:https://doi.org/10.1016/j.ccr.2014.02.007 (2014).
541 44 Chappell, J., Sun, Y., Singh, A. & Dalton, S. MYC/MAX control ERK signaling
542 and pluripotency by regulation of dual-specificity phosphatases 2 and 7. *Genes*
543 *Dev* **27**, 725-733, doi:10.1101/gad.211300.112 (2013).
544 45 Palm, K., Stenberg, P., Luthman, K. & Artursson, P. Polar Molecular Surface
545 Properties Predict the Intestinal Absorption of Drugs in Humans.
546 *Pharmaceutical Research* **14**, 568-571, doi:10.1023/A:1012188625088 (1997).
547 46 Spugnini, E. & Fais, S. Proton pump inhibition and cancer therapeutics: A
548 specific tumor targeting or it is a phenomenon secondary to a systemic buffering?
549 *Seminars in Cancer Biology* **43**, 111-118,
550 doi:https://doi.org/10.1016/j.semcancer.2017.01.003 (2017).

551 47 Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database
552 for 2018. *Nucleic Acids Res* **46**, D1074-D1082, doi:10.1093/nar/gkx1037
553 (2018).
554 48 Singh, S. R. *et al.* The lipolysis pathway sustains normal and transformed stem
555 cells in adult Drosophila. *Nature* **538**, 109-113, doi:10.1038/nature19788 (2016).
556 49 Wang, G. *et al.* Arf1-mediated lipid metabolism sustains cancer cells and its
557 ablation induces anti-tumor immune responses in mice. *Nature*
558 *Communications* **11**, 220, doi:10.1038/s41467-019-14046-9 (2020).
559 50 D'Souza-Schorey, C. & Chavrier, P. ARF proteins: roles in membrane traffic and
560 beyond. *Nature Reviews Molecular Cell Biology* **7**, 347-358,
561 doi:10.1038/nrm1910 (2006).
562 51 Olbe, L., Carlsson, E. & Lindberg, P. A proton-pump inhibitor expedition: the
563 case histories of omeprazole and esomeprazole. *Nature Reviews Drug*
564 *Discovery* **2**, 132-139, doi:10.1038/nrd1010 (2003).
565 52 Békés, M., Langley, D. R. & Crews, C. M. PROTAC targeted protein degraders:
566 the past is prologue. *Nature Reviews Drug Discovery*, doi:10.1038/s41573-021-
567 00371-6 (2022).
568 53 Yang, T. *et al.* DrugSpaceX: a large screenable and synthetically tractable
569 database extending drug space. *Nucleic Acids Res* **49**, D1170-D1178,
570 doi:10.1093/nar/gkaa920 (2021).
571 54 Grygorenko, O. O. *et al.* Generating Multibillion Chemical Space of Readily
572 Accessible Screening Compounds. *iScience* **23**, doi:10.1016/j.isci.2020.101681

(2020).

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

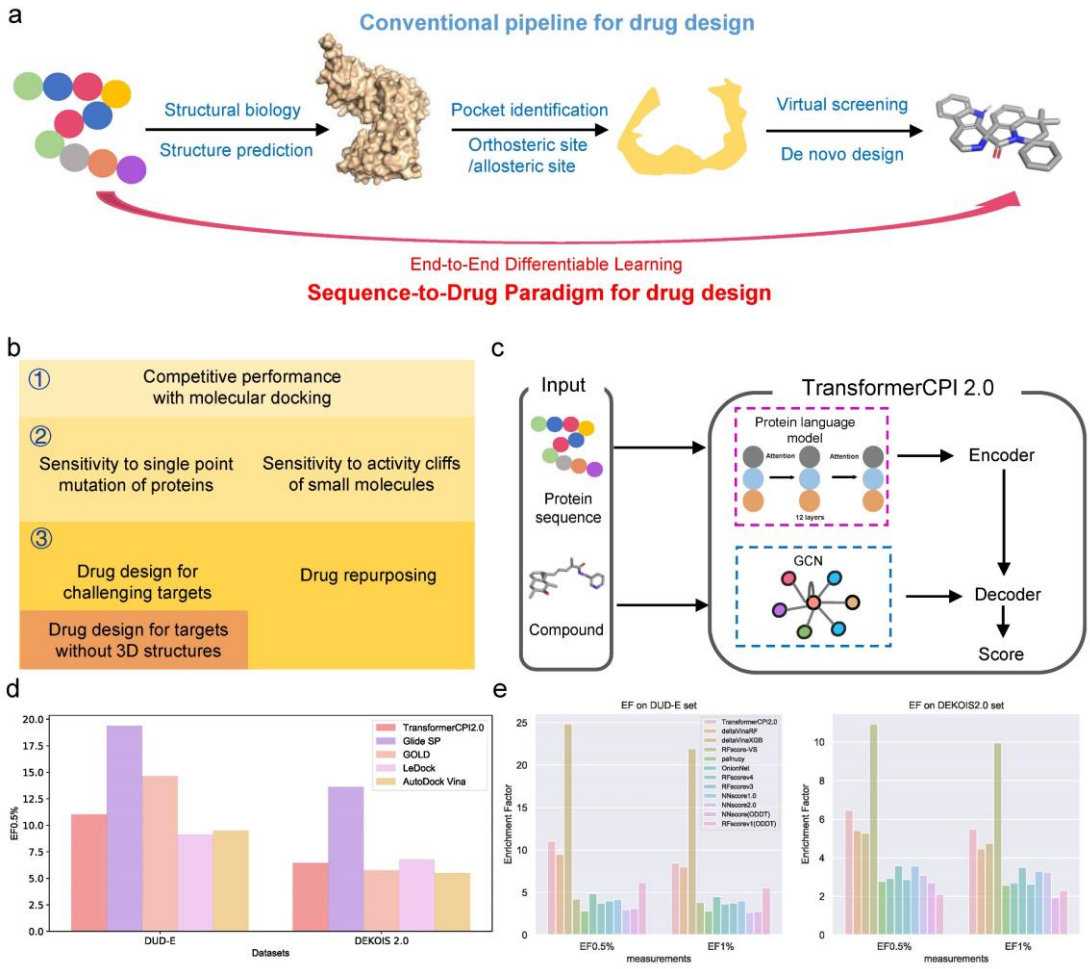


Fig. 1 | Competitive performance compared with 3D structure-based approaches. a, The conventional pipeline for target-based drug design and the sequence-to-drug paradigm. **b**, Three stages of the proof of concept of the sequence-to-drug paradigm, with each stage is labelled by different colours. **c**, 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.

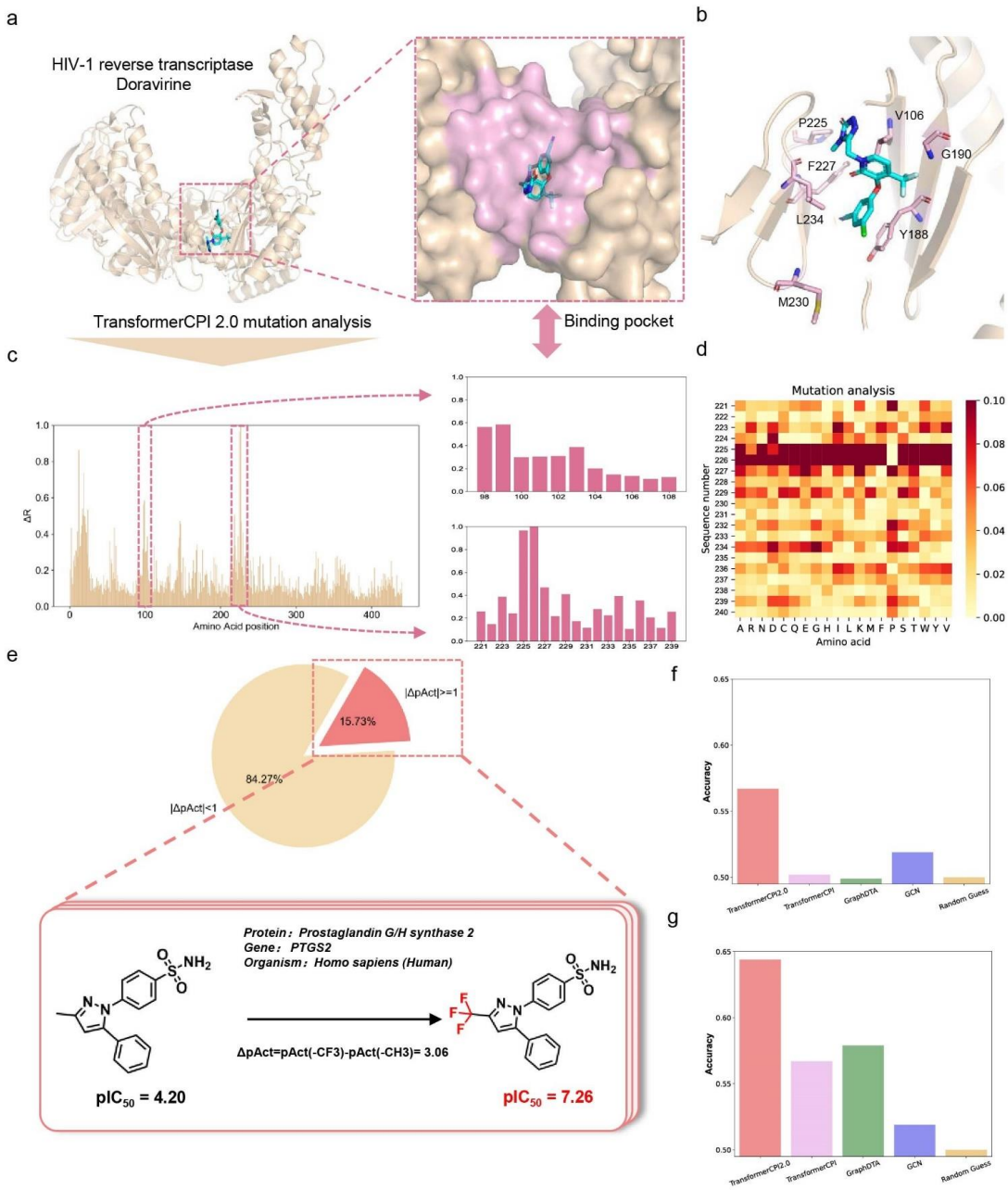


Fig. 2 | Drug resistance mutation analysis and substitution effect analysis of the trifluoromethyl group. a, The cocrystal structure of HIV-1 reverse transcriptase and doravirine (PDB: 4NCG). The binding pocket of doravirine is highlighted in pink. **b**, The binding mode of doravirine. The residues 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.

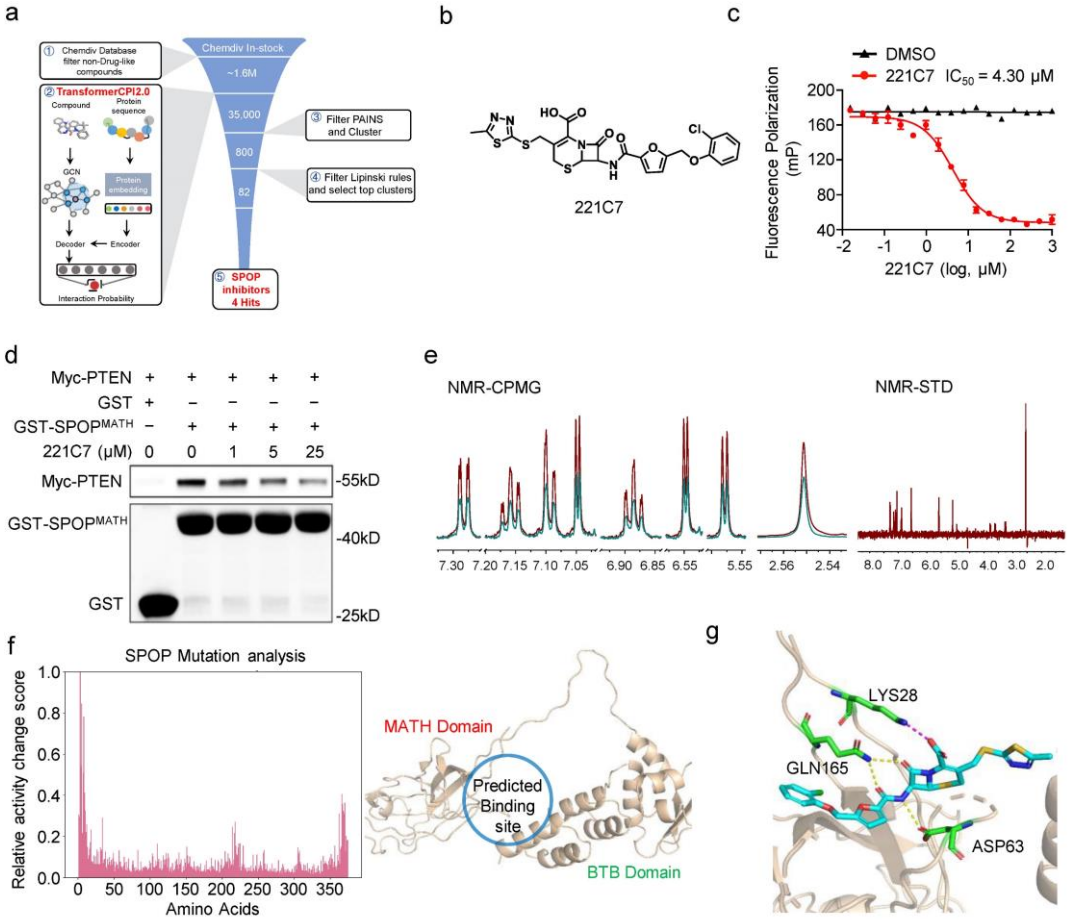


Fig. 3 | Discovering a novel scaffold hit of SPOP. a, Scheme of virtual screening protocol for small-molecule inhibitors of SPOP. **b**, Chemical structure of 221C7. **c**, 221C7 competitively inhibits puc_SBC1 peptide binding to SPOP^{MATH}, as measured by the FP assay. **d**, 221C7 disrupts protein 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. **g**, The docking pose of 221C7 binding to SPOP (PDB: 3HQI).

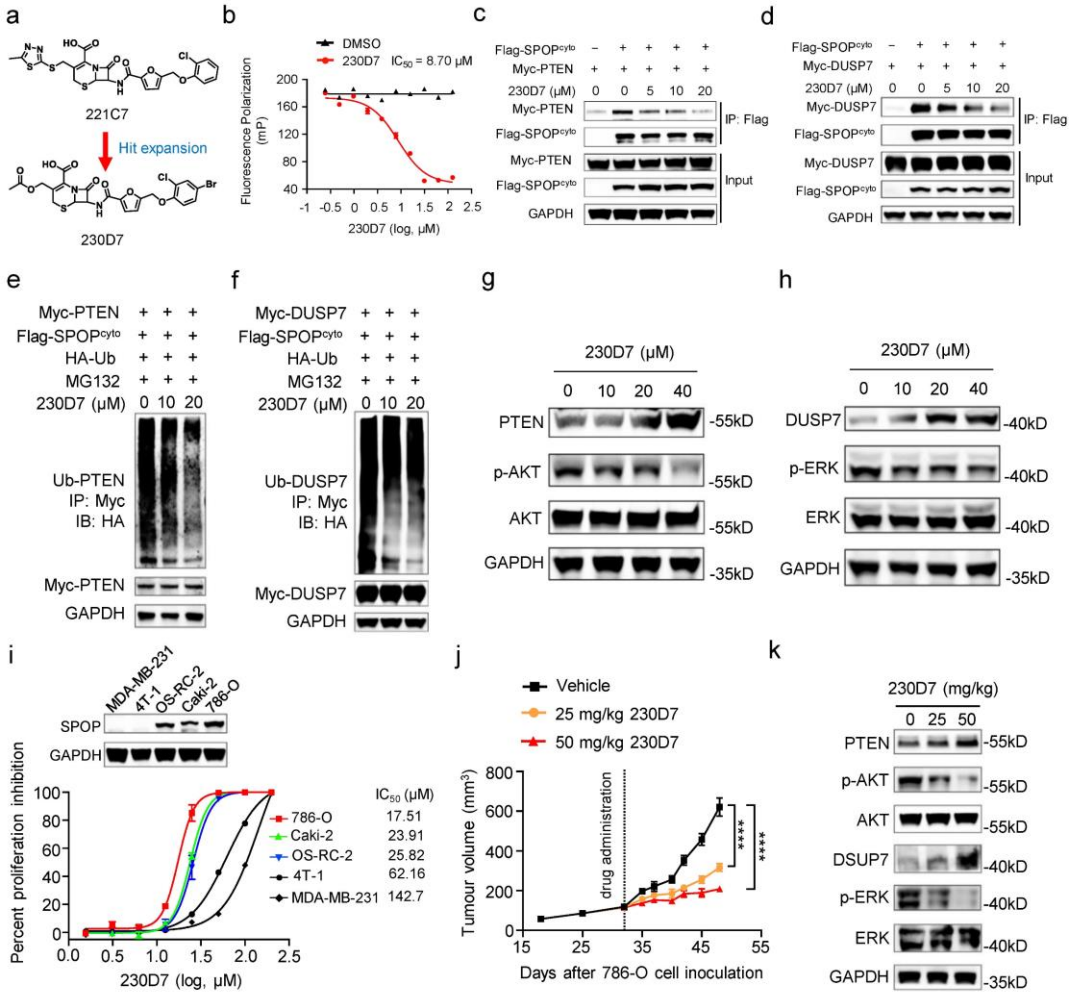
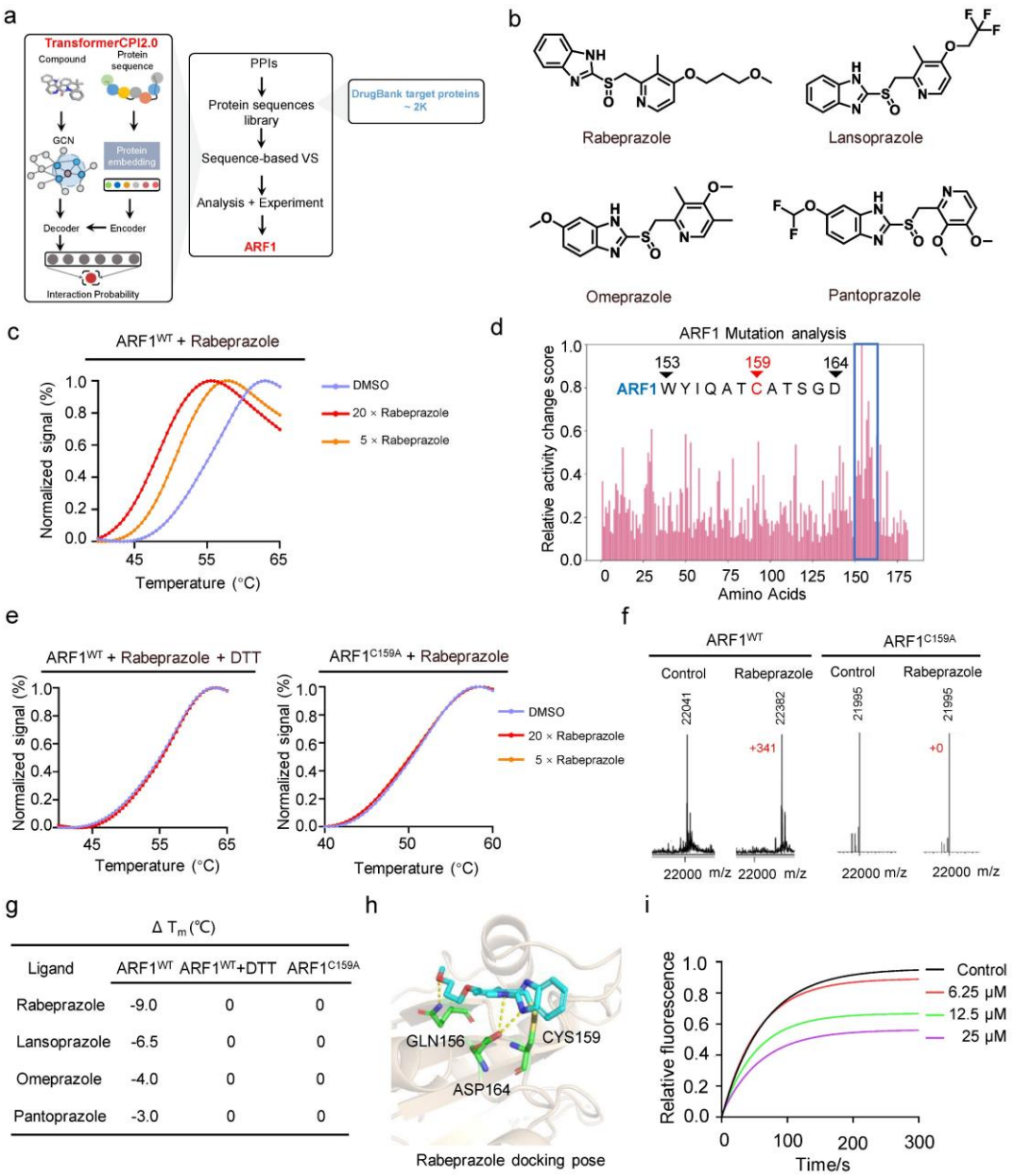


Fig. 4 | 230D7 showed therapeutic potential for blocking oncogenic SPOP activity to treat ccRCC. a, Hit expansion of 221C7 to 230D7. **b**, 230D7 competitively inhibits puc_SBC1 peptide binding to $SPOP^{MATH}$, as measured by the FP assay. **c~d**, SPOP-PTEN and SPOP-DUSP7protein interactions are inhibited 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.



Ligand	ARF1^{WT}	ARF1^{WT}+DTT	ARF1^{C159A}
Rabeprazole	-9.0	0	0
Lansoprazole	-6.5	0	0
Omeprazole	-4.0	0	0
Pantoprazole	-3.0	0	0

Fig. 5 | Identifying ARF1 as the new target of PPIs. a, Scheme of target identification protocol for PPIs. **b**, Chemical structures of four PPIs, rabeprazole, lansoprazole, omeprazole and pantoprazole. **c**, Effect of rabeprazole on the thermal stability of ARF1^{WT} in the PTS assay. **d**, Drug resistance mutation analysis interprets the prediction of omeprazole, and there is a cysteine (C159, the only cysteine residue of ARF1) residue among the important residues. **e**, Effects of rabeprazole on the thermal stability of ARF1^{WT} (containing DTT) and ARF1^{C159A} in the PTS assay. **f**, The protein molecular weights of ARF1^{WT} and ARF1^{C159A} in the presence or absence of rabeprazole were determined by mass spectrometer. **g**, Summary of PTS assay results. **h**, Potential docking pose of rabeprazole and ARF1 (PDB: 1HUR). **i**, GDP/MANT-GTP nucleotide exchange of ARF1 treated with rabeprazole.

581
582

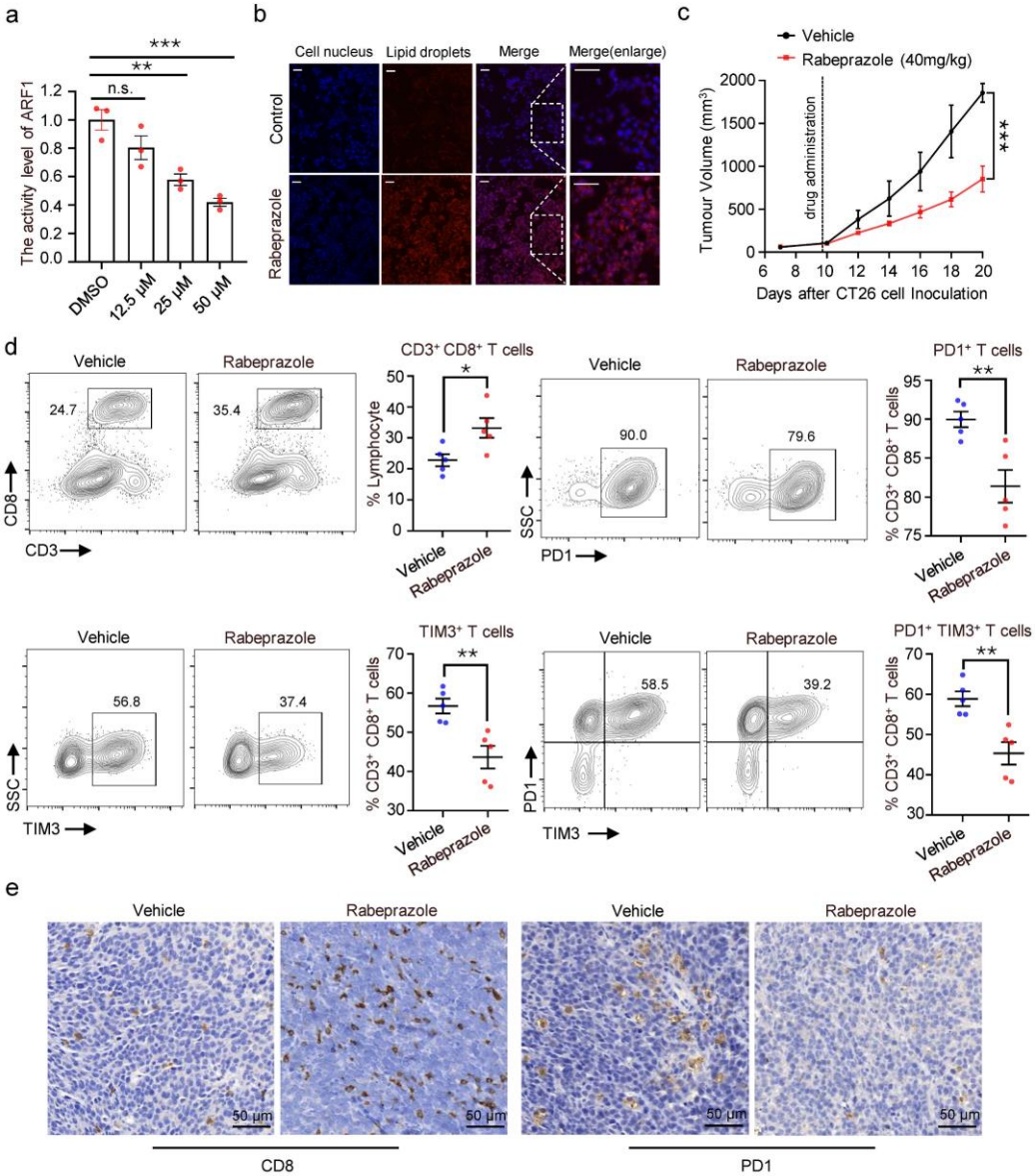


Fig. 6 | Rabeprazole induces antitumour immune response through lipid metabolism. **a**, The activity level of ARF1 in CT26 cells treated with rabeprazole for 48h was measured by using G-LISA assay. **b**, Fluorescent images of CT26 cells stained with DAPI (for nucleus) or Nile red (for lipid droplets) after treatment with rabeprazole (20 μM) or DMSO. Scale bars: 100 μm. **c**, *In vivo* efficacy of rabeprazole in CT26 transplanted tumour model in BALB/c mice. Mice were administrated rabeprazole 40 mg/kg daily for 10 days by intraperitoneal dosing. **d**, Impact of rabeprazole delivery on immune cell subsets in CT26 transplanted tumour model, assessed by flow cytometry analysis (a, c-d, Error bars represent mean ± SEM, with n = 5 for each group. P values were evaluated using 2-tailed unpaired t-test. *P < 0.05, **P < 0.01, ***P < 0.001). **e,** Immunohistochemical staining for cell surface markers (CD8, PD1) of tumour tissues in the indicated groups.

583
584

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[56] 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[57] and TAPE[58]. 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)[59] 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.

651 **Training details.** The TransformerCPI2.0 model was trained by the RAdam[60] optimizer
652 with a learning rate of 1e-5 and a weight decay of 1e-3. The batch size of 1 was selected
653 to ensure that the longest protein sequence fit into the GPU memory. We employed the
654 gradient accumulation technique to expand the actual batch size to 64. Training was
655 performed on one NVIDIA Tesla V100 (16G) GPU. The TransformerCPI2.0 model was
656 trained for ~50 epochs or ~1.5 weeks of wall clock time.

657 **ChEMBL dataset construction.** Inheriting our previous work, we followed two rules
658 to construct a dataset: (i) CPI data was collected from an experimentally validated
659 database and (ii) each ligand should exist in both positive and negative classes. To build
660 a universal deep learning model for all types of proteins, we selected the chEMBL_23[61]
661 database to construct a universal dataset to train TransformerCPI2.0.

662 **Data cleaning.** The ChEMBL_23 database was released on 1 May 2017 and contains
663 2,101,843 compound records, 1,735,442 compounds, 14,675,320 activities, 1,302,147
664 assays, 11,538 targets and 67,722 source documents. We downloaded the whole
665 database and cleaned the data using the following procedure:

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

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

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

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

673 early stage of drug discovery, only hit compounds whose IC_{50}, K_i or EC_{50} are at the μM
674 or even nM level will be further optimized. Additionally, public data are prone to a
675 certain experimental error, i.e., on average 0.5 log units for IC50 data[62,63]. The threshold
676 of 6 considers those CPI pairs whose IC50, Ki or EC50 are smaller than 1 μM as
677 positive samples, which causes the models to select CPI pairs with high activity. To
678 decrease the risk of experimental error in public data, a stricter threshold of 6.5 was
679 used in a previous work[64]. Therefore, we selected the threshold of 6.5 to define positive
680 data and negative data. Data whose atom number was more than 60 or whose protein
681 sequence length exceeded 4000 were filtered out, guaranteeing that all the data fit into
682 GPU memory. Finally, we constructed a ChEMBL dataset including 3,348 proteins,
683 69,616 compounds, 117,513 positive CPIs, 134,611 negative CPIs and 252,124 samples
684 in total.

685 **Dataset split and label reversal experiment.** We selected ligands that exist in both the
686 positive and negative classes to make the compound distribution in positive samples
687 and negative samples exactly the same, trying to eliminate the potential ligand bias as
688 much as possible. Consequently, we used a label reversal experiment to split the
689 ChEMBL dataset. For the ChEMBL dataset, we randomly selected 2,941 ligands and
690 pooled all the negative CPI samples containing these ligands into the test set.
691 Additionally, we selected another 2,900 ligands and pooled all their associated positive
692 samples into the test set. The remaining datasets were split randomly into a training set
693 and a validation set at a ratio of 10:1. The validation set was used to determine the
694 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

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

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

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

760 $$\Delta pAct = pAct(-CF3) - pAct(-CH3)\,.$$

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

783 positive molecules among compound libraries.

784 **Docking dataset.** We used two benchmark datasets, DUD-E and DEKOIS 2.0, to
785 compare TransformerCPI2.0 and other docking tools. DUD-E is a widely used
786 benchmark dataset for docking tool and scoring function comparisons. DUD-E includes
787 a total of 102 proteins with 22886 clustered ligands drawn from ChEMBL, each with
788 50 property-matched decoys drawn from ZINC. DEKOIS2.0 is often used as a
789 complementary benchmark for further validation. DEKOIS2.0 contains 81 structurally
790 diverse targets. Each target has 40 active compounds extracted from BindingDB and
791 1200 decoys generated from ZINC. DUD-E and DEKOIS 2.0 both serve as external
792 test sets for TransformerCPI2.0, which means that TransformerCPI2.0 was not trained
793 or fine-tuned on these datasets.

794 **Docking tools.** Four docking programs, Glide SP (version 8.2)[65], GOLD (version
795 2019)[66] with Piecewise Linear Potential (CHEMPLP) SF, LeDock (version 1.0)[67] and
796 AutoDock Vina[68], showed high docking power and computational efficiency and thus
797 were chosen as the baseline models. We focused on the enrichment factor (EF0.5%,
798 EF1%, EF5%) to evaluate different models, and the results of Glide SP, GOLD and
799 LeDock on the two benchmark sets were extracted from Chao *et al.*[69]. In addition to the
800 above four docking tools, many machine learning scoring functions have been tested
801 on DUD-E and DEKOIS 2.0[69]; thus, we compared these machine learning scoring
802 functions. We used the whole protein sequence as the default input of
803 TransformerCPI2.0 and calculated the average EF0.5%, EF1% and EF5% of DUD-E
804 and DEKOIS 2.0.

805 **Virtual screening of SPOP.** First, after filtering non-drug-like compounds from the
806 ChemDiv Library (San Diego, CA, USA), which contains approximately 1.6 million
807 in-stock compounds, TransformerCPI2.0 was applied to score the compounds, and the
808 top 35,000 molecules (~top 2%, ensuring compound diversity) were selected by
809 screening. Second, we filtered panassay interference compounds (PAINS) and clustered
810 these molecules automatically based on their extended-connectivity fingerprints
811 (ECFP), obtaining approximately 800 clusters. Third, we filtered these compounds by
812 the Lipinski rules and manually selected representative compounds from each cluster.
813 Finally, a total of 82 candidates were purchased for further experimental evaluation.
814 **Virtual screening of RNF130.** First, after filtering non-drug-like compounds from the
815 Chemspace Library (Monmouth Junction, NJ 08852, USA), which contains
816 approximately 2 million in-stock compounds, TransformerCPI2.0 was applied to score
817 the compounds, and the top 10,000 molecules (~top 0.5%, ensuring compound diversity)
818 were selected. Second, we filtered panassay interference compounds (PAINS) and
819 clustered these molecules automatically based on their extended-connectivity
820 fingerprints (ECFP), obtaining approximately 200 clusters. Third, we filtered these
821 compounds by the Lipinski rules and manually selected representative compounds from
822 each cluster. Finally, a total of 87 candidates were purchased for further experimental
823 evaluation.
824 **Target identification of PPIs.** We collected potential proteins from the DrugBank
825 database and selected proteins that already have active modulators. Then,
826 TransformerCPI2.0 was applied to score proteins against four classical PPIs

827 (omeprazole, rabeprazole, lansoprazole and pantoprazole), and the results were sorted
828 by predicted interaction probability. Next, we analysed the top 20 proteins by evaluating
829 their novelty, importance and feasibility, and finally chose ARF1 for experimental
830 validation.

831 **Compounds.** SPOP inhibitors were purchased from ChemDiv Library (San Diego, CA,
832 USA): 221C7, Y502-3210; 231A10, 5282-0816; 231D8, 8017-3040. 230D7 was
833 synthesised in our laboratory. RNF130 inhibitor was purchased from Chemspace
834 Library (Monmouth Junction, NJ 08852, USA): iRNF130-63, CSC138461036. PPIs
835 were purchased from MedChemExpress (Monmouth Junction, NJ, USA): rabeprazole,
836 HY-B0656; lansoprazole, HY-13662; Omeprazole, HY-B0113; pantoprazole, HY-
837 17507.

838 **Plasmid construction.** Wild-type, truncated or mutant versions of the human proteins
839 were used in this study: SPOP (UniProt accession number: O43791-1), PTEN (UniProt
840 accession number: P60484-1), DUSP7 (UniProt accession number: Q16829-1),
841 RNF130 (UniProt accession number: Q86XS8-1), ARF1 (UniProt accession number:
842 P84077-1) and ARNO (UniProt accession number: Q99418-1). For plasmid
843 construction, $SPOP^{WT}$ and $SPOP^{cyto}$ (residues 1-366) were subcloned into the pcDNA
844 3.1 vector with a Flag-tag, $SPOP^{MATH}$ (residues 28-166) and $ARNO^{Sec7}$ (residues 50-
845 250) was subcloned into the pGEX 6p-1 vector with a GST-tag, PTEN and DUSP7 were
846 subcloned into the pcDNA 3.1 vector with a Myc-tag, RNF130 (residues 1-304) was
847 subcloned into the pcDNA3.1 vector with a C-terminal His_8-Flag-tag. N-terminally
848 truncated human Δ17ARF1 and C159A-mutant Δ17ARF1 were subcloned into pProEX

849 HTb with a His_6-tag. All of the above plasmids were synthesized by Sangon Biotech
850 (Shanghai) Co., Ltd. pCMV-HA-Ub plasmid (CAT#. kl-zl-0513) was purchased from
851 Shanghai Kelei Biological Technology Co., Ltd.

852 **Recombinant protein expression and purification.** For expression of $SPOP^{MATH}$,
853 GST-tagged $SPOP^{MATH}$ plasmid was transformed into BL21-CodonPlus (DE3)-RIPL
854 Cells (CAT#. EC1007, Shanghai Weidi Biotechnology Co., Ltd.), and then the cells
855 were grown in lysogeny broth (LB) medium and induced by isopropyl β-D-1-
856 thiogalactopyranoside (IPTG) at a final concentration of 0.5 mM at 16 °C overnight.
857 Cells were harvested and lysed in solution (20 mM HEPES pH 7.4, 200 mM NaCl, 1
858 mM dithiothreitol) by sonication and then centrifuged at 16,000 rpm for 1 hour at 4 °C.
859 The supernatants were filtered by 0.22 μM syringe filters and purified on GST Trap
860 columns (GE Healthcare) by elution with 10 mM glutathione. The eluted components
861 were loaded onto desalting columns (GE Healthcare) to remove glutathione and
862 incubated with PreScission Protease for 6-8 hours at 4 °C. The components were
863 reloaded onto GST Trap columns to remove the GST tags and further purified by a
864 Superdex 75 10/300 GL column. The purified $SPOP^{MATH}$ protein was concentrated and
865 stored in buffer (20 mM HEPES pH 7.4, 200 mM NaCl) at -80 °C.

866 RNF130 protein was expressed in Expi-293F (Invitrogen) using Expifectamine
867 transfection reagent according to the manufacturer's instructions. Cells were collected
868 3 days after transfection. Proteins were first captured by Ni^{2+}-Sepharose 6 Fast Flow
869 resin (GE healthcare) and then further purified by gel filtration chromatography with a
870 Superdex S200 column (GE Healthcare). The purified protein was concentrated and

871 stored in buffer (20 mM HEPES pH 7.5, 150 mM NaCl and 1 mM TCEP) at -80 °C.
872 For expression of Δ17ARF1 and Δ17ARF1^{C159A}, the His-tagged recombinant plasmid
873 was transformed into BL21-CodonPlus (DE3)-RIPL cells, and then the cells were
874 grown in LB medium and induced by IPTG at a final concentration of 0.1 mM at 25 °C
875 for 6 hours. The cells were harvested and lysed in solution (20 mM Tris, 100 mM NaCl,
876 5 mM $MgCl_2$, 10 mM imidazole, pH 8.0) by sonication and then centrifuged at 16,000
877 rpm for 1 hour at 4 °C. The supernatants were filtered by 0.22 μM syringe filters and
878 purified on HiTrap column (GE Healthcare) by elution with 300 mM imidazole. Protein
879 sample was then purified by a Superdex 75 10/300 GL column. Finally, the purified
880 Δ17ARF1 and Δ17ARF1^{C159A} proteins were concentrated and stored in buffer (20 mM
881 Tris, 100 mM NaCl, pH 8.0) at -80 °C.
882 The purification process ARNO^{Sec7} protein was the same as that of the SPOP^{MATH}
883 protein except the cells were induced by 0.1 mM IPTG at 37 °C for 3 hours.
884 **Fluorescence polarization (FP).** Fluorescence polarization experiments were
885 conducted in a 384-well black plate (Corning, 3575) using a 42 μL reaction system.
886 FITC-labelled SPOP substrate puc_SBC1 (FITC-LACDEVTSTTSSSTA)[70,71]
887 (synthesized by GL Biochem (Shanghai) Ltd) was used for the probe. Then, 20 μL of
888 reaction buffer (20 mM HEPES, pH 7.4) containing 200 nM SPOP^{MATH} protein was
889 incubated with 2 μL of compound for 30 minutes at room temperature, and 20 μL of
890 reaction buffer containing 200 nM probe was added. Fluorescence polarization (mP)
891 signals were measured by a fluorescence mode (excitation filter 480 nm, emission filter
892 535 nm) in Spark microplate reader (Tecan).

893 **Protein thermal shift (PTS).** The Bio–Rad CFX96 RealTime PCR Detection System
894 was utilized to monitor the thermal stability of ARF1 and $SPOP^{MATH}$ protein. PTS
895 experiments were performed in a 96-well PCR plate (DN Biotech (Hong Kong) Co.,
896 Ltd.) with a 20 μL reaction system. A total of 20 μL of reaction buffer containing protein
897 (5 μM for $SPOP^{MATH}$ or 2.5 μM for ARF1), 5 × SYPRO Orange Protein Gel Stain
898 (Sigma, S5692) and indicated concentration of compound. The signals of all reaction
899 systems were continuously monitored and recorded from 25 °C to 90 °C for
900 approximately 45 minutes. The T_m values of $SPOP^{MATH}$ and ARF1 were measured using
901 CFX manager software version 3.1.
902 **Nuclear magnetic resonance (NMR).** NMR spectroscopy experiments were
903 performed using a 600 MHz spectrometer (AVANCE III, Bruker) to validate protein–
904 ligand interactions. In Carr-Purcell-Meiboom-Gill (CPMG) and saturation transfer
905 difference (STD) NMR experiments, compound was dissolved to a final concentration
906 of 200 μM in a solution of PBS formed with D_2O containing 5 μM $SPOP^{MATH}$ protein
907 and 5% DMSO-d^6.
908 **Mass spectrometry analysis.** The experiment was performed on the mass spectrometry
909 service platform of Shanghai Institute of Materia Medica, Chinese Academy of
910 Sciences. The protein (100 μM) was incubated with compounds (1 mM) or solvent
911 control overnight at 4 °C, and then the protein molecular weights were determined by
912 Q Exactive (Thermo) and 6545 XT (Aglient) mass spectrometer. For compound binding
913 site identification, the proteins were digested with trypsin (10 ng/μL) at 37 °C for 17 h.
914 The next day, after centrifugation, the supernatant was lyophilized, desalted, and

915 lyophilized again, followed by the addition of 0.1% FA solution to dissolve peptide
916 lyophilized powder. After centrifugation, the supernatant was detected by mass
917 spectrometry (Q-Exactive). The MS data was analysed via software MaxQuant (version
918 1.6.5.0). The false discovery rate (FDR) for peptides and proteins was controlled <1%
919 by Andromeda search engine.

920 **Surface plasmon resonance (SPR) for RNF130.** The SPR binding assay was
921 performed using a Biacore T200 instrument (GE Healthcare). The purified RNF130
922 protein was covalently immobilized onto a CM5 sensor chip (Cytiva) by a standard
923 amine-coupling procedure in 10 mM sodium acetate (pH 4.5) with running buffer HBS
924 (50 mM HEPES pH 7.4, 150 mM NaCl). iRNF130-63 was serially diluted and injected
925 onto the sensor chip at a flow rate of 30 μL/min for 120 s (contact phase), followed by
926 120 s of buffer flow (dissociation phase). The equilibrium dissociation constant (K_D)
927 value was derived using Biacore T200 Evaluation software (version 1.0, GE
928 Healthcare).

929 **Guanine nucleotide exchange assay for ARF1**. First, with the participation of EDTA
930 (a metal chelating agent capable of chelating magnesium ions, which are critical for the
931 binding of GDP/GTP to ARF1), GDP was loaded onto the ARF1 protein by incubating
932 ARF1 with a 20-fold molar concentration of GDP. Excess magnesium chloride was
933 used to terminate the loading reaction, followed by the removal of excess GDP by a
934 NAP-5 column to produce $ARF1^{GDP}$ protein. Next, $ARF1^{GDP}$ protein (20 μM) was
935 mixed with compounds and Mant-GTP (10 μM) in reaction buffer and incubated in the
936 dark for 15 minutes. Exchange reactions were initiated by the injection of $ARNO^{Sec7}$ (1

937 μM).

938 **Cell lines.** 293T, MDA-MB-231, 4T-1 and CT26 cells were obtained from the
939 American Type Culture Collection (ATCC). OS-RC-2 and Caki-2 cells were purchased
940 from the National Biomedical Laboratory Cell Resource Bank. 786-O cells were kindly
941 provided by the Cell Bank/Stem Cell Bank, Chinese Academy of Sciences. 293T and
942 MDA-MB-231 cells were cultured in DMEM medium (BasalMedia, L110KJ)
943 supplemented with 10% foetal bovine serum (FBS, Gibco, 10099141C) and 1%
944 Penicillin-Streptomycin (PS, Gibco, 2321118). 786-O, OS-RC-2, 4T-1 and CT26 cells
945 were cultured in RPMI-1640 medium (BasalMedia, L210KJ) supplemented with 10%
946 FBS and 1% PS. Caki-2 cells were cultured in McCoy's 5A medium (Gibco, 2193071)
947 supplemented with 10% FBS and 1% PS. All cells were incubated at 37 °C under a 5%
948 (v/v) CO_2 atmosphere.

949 **G-LISA.** A Cytoskeleton ARF1 G-LISA Activation Assay Kit (BK132) was used to
950 detect cellular ARF1-GTP levels according to the manufacturer's instructions.

951 **Western blot.** Total proteins from cells were lysed in RIPA lysis buffer (Beyotime,
952 P0013C) containing phosphatase inhibitor (Bimake, B15001) and protease inhibitor
953 (Bimake, B14001) on ice. Cell lysates were centrifuged at 13,000 × g for 15 minutes at
954 4 °C. The BCA protein assay kit (Thermo Scientific, 23225) was used to quantify the
955 protein concentration. Equal amounts of total proteins were separated by 10% SDS–
956 PAGE and then transferred onto nitrocellulose membranes. The membranes were
957 blocked with 5% skim milk in TBST for 1 hour at room temperature and then incubated
958 with primary antibodies overnight at 4 °C. Then membranes were incubated with HRP-

959 conjugated anti-rabbit antibody (secondary antibody, Promega, W4011) for 1 hour at
960 room temperature. Finally, the immune complexes were detected with an ECL kit
961 (Meilun, MA0186) and visualized as well as quantified using GenGnome XRQ NPC.
962 The following primary antibodies were used: anti-Flag (Cell Signaling Technology,
963 14793), anti-myc (Cell Signaling Technology, 2278), anti-GAPDH (Cell Signaling
964 Technology, 5174), anti-PTEN (Cell Signaling Technology, 9559), anti-HA (Cell
965 Signaling Technology, 3724), anti-p-ERK$^{T202/Y204}$ (Cell Signaling Technology, 4376),
966 anti-p-AKT^{Thr308} (Cell Signaling Technology, 4056), anti-ERK (Cell Signaling
967 Technology, 9102), anti-AKT (Cell Signaling Technology, 9272), anti-DUSP7
968 (ABGENT, AP8450a), anti-SPOP (Abcam, ab192233), anti-GST (Absin, abs830010)
969 and anti-β-Tubulin (Cell Signaling Technology, 15115).
970 ***In vitro* GST pull-down.** The plasmids (Myc-PTEN or Myc-DUSP7) were transiently
971 transfected into 293T cells. After transfection for 24 hours, the cells were harvested and
972 lysed in cell lysis buffer for Western and IP (Beyotime, P0013) containing protease
973 inhibitor on ice. GST or GST-SPOP^{MATH} proteins bound to GST magnetic beads
974 (GenScript, L00327) were incubated with the cell lysates (Myc-PTEN or Myc-DUSP7)
975 in the presence of different doses of compound for 2 hours at room temperature. The
976 beads were washed 3 times with PBST, and the precipitated proteins were eluted with
977 1 × SDS loading buffer (Beyotime, L00327) at 100 °C for 5 minutes and analysed by
978 Western Blot.
979 **Cellular thermal shift assay (CETSA).** 293T cells transfected with Flag-RNF130 for
980 48 hours were collected and lysed in 20 mM Tris pH 7.5, 150 mM NaCl and 1% Triton

981 X-100. Then, 50 μM iRNF130-63 or DMSO was added to the supernatant and
982 incubated at 25 °C for 30 minutes. After denaturing at various temperatures for 3
983 minutes on a temperature gradient PCR instrument (Eppendorf), the samples were
984 centrifuged at 20000 × g for 30 minutes at 4 °C, and the supernatants were analysed by
985 Western blot.

986 **Coimmunoprecipitation (Co-IP).** The plasmids (Flag-SPOP^{cyto}, Myc-PTEN or Myc-
987 DUSP7) were transiently cotransfected into 293T cells. After transfection for 24 hours,
988 293T cells were treated with different doses of compound for another 24 hours. The
989 cells were harvested and lysed in cell lysis buffer for Western and IP containing protease
990 inhibitor on ice. Approximately 80% of the total lysates were immunoprecipitated with
991 anti-Flag-conjugated magnetic beads (Bimake, B26102) for 2 hours at room
992 temperature, and other lysates were used as input. The magnetic beads were then
993 washed 3 times with PBST, and the immunoprecipitated proteins were eluted with 1 ×
994 SDS loading buffer at 100 °C for 5 minutes. The IP and lysate samples were analysed
995 by Western Blot.

996 ***In vivo* ubiquitination.** The plasmids (Myc-PTEN or Myc-DUSP7, Flag-SPOP^{cyto},
997 HA-Ub) were transiently cotransfected into 293T cells. After transfection for 24 hours,
998 293T cells were treated with different doses of compound for 24 hours. The cells were
999 then treated with 10 μM protease inhibitor MG132 (MedChemExpress (Monmouth
1000 Junction, NJ, USA), HY-13259) for another 4 hours before harvesting. Next, the cells
1001 were lysed in cell lysis buffer for Western and IP containing protease inhibitor on ice.
1002 Approximately 80% of the total lysates were immunoprecipitated with anti-Myc-

1003 conjugated magnetic beads (Bimake, B26302) for 2 hours at room temperature, and the
1004 other lysates were used as input. The magnetic beads were then washed 3 times with
1005 PBST, and the immunoprecipitated proteins were eluted with 1 × SDS loading buffer
1006 at 100 °C for 5 minutes. The ubiquitination levels were detected using a Western Blot
1007 assay.

1008 **Cell proliferation.** Cells were seeded in 96-well plates and incubated with serially
1009 diluted compounds for 72 hours. Cell viability was determined using the CellTiter-
1010 Glo® Luminescent Cell Viability Assay kit (Promega, G7573) following the
1011 manufacturer's instructions. IC_{50} values were determined by nonlinear regression
1012 (curve fit) using a variable slope (four parameters) in GraphPad Prism (9.0).

1013 **Animals.** All procedures performed on animals were in accordance with regulations
1014 and established guidelines and were reviewed and approved by the Institutional Animal
1015 Care and Use Committee at the Shanghai Institute of Materia Medica, Chinese
1016 Academy of Sciences (IACUC Issue NO. 2022-01-JHL-27 for NSG mice; IACUC
1017 Issue NO. 2021-03-JHL-22 for BALB/c mice). NSG mice were obtained from
1018 Shanghai Model Organisms Center, Inc; BALB/c mice were obtained from Beijing
1019 Huafukang Biotechnology Co. Ltd (Beijing, China). Six- to eight-week-old mice were
1020 used for the studies and were maintained with free access to pellet food and water in
1021 plastic cages ate 21 ± 2℃ and kept on a 12 hours light/dark cycle.

1022 **Pharmacokinetics.** The pharmacokinetic profiles of compound 230D7 were
1023 determined in male BALB/c mice. The test compound 230D7 was dissolved in solution
1024 containing DMSO, PEG400, PBS (5/5/90, v/v) and administered via intraperitoneal

1025 administration (i.p.) at 10 mg/kg. Serial blood samples were collected at 0.25, 0.5, 1, 2,
1026 4, 8, 24 hours after dosing and centrifuged to obtain the plasma fraction. A 10 μL aliquot
1027 of plasma was deproteinized with 100 μL acetonitrile/methanol (1/1, v/v) containing
1028 internal standard. After centrifugation, the supernatant was diluted with a certain
1029 proportion of acetonitrile/water (1/1, v/v), mixed and centrifuged at 4,000 rpm for 10
1030 minutes. Finally, the aliquots of the diluted supernatant were injected into LC-MS/MS
1031 system.

1032 **Acute toxicity.** BALB/c mice were used to evaluate the toxicity of compound 230D7.
1033 The mice were randomly divided into 3 groups (n = 3) and treated with different doses
1034 of compound 230D7 (0, 50, 100 mg/kg) by intraperitoneal administration daily for a
1035 week. The body weights of mice were measured every day and the significant organs
1036 (heart, kidney, lung, liver and spleen) were harvested, weighted and used for
1037 histological analysis at the last day.

1038 **H&E staining.** For histological analysis of BALB/c mice in 230D7-treated or vehicle
1039 control groups, staining for H&E staining were performed using standard histological
1040 techniques. Isolated organ tissues were fixed in neutral paraformaldehyde for 24 h and
1041 embedded in paraffin wax. H&E staining of these tissue examples were carried out by
1042 Servicebio company.

1043 **786-O cells xenograft tumour growth.** NSG mice were used to evaluate the
1044 pharmacodynamics of 230D7. The 786-O cells xenograft tumour model was established
1045 by the subcutaneous injection of 786-O cells (1×10^7) into the NSG mice. When the
1046 tumour reached the volume of approximately 100 mm^3, the mice were randomly

1047 divided into three groups (n = 7) and intraperitoneally treated with different dosages of
1048 230D7 (0, 25, 50 mg/kg, 230D7 was synthesised in our laboratory) in solution
1049 containing DMSO, PBS (5/90, v/v) once a day for 16 days. Body weight and tumour
1050 size were measured every two or three days, and tumour volume was calculated using
1051 the formula: $V = (L \times W^2)/2$ (L, length; W, width). At the end of the experiment, the
1052 mice were euthanized and the tumours were harvested for Western Blot and other
1053 studies.

1054 **CT26 cells transplanted tumour growth.** BALB/c mice were used to evaluate the
1055 pharmacodynamics of rabeprazole (purchased from MedChemExpress, HY-B0656).
1056 The CT26 cells transplanted tumour model was established by the subcutaneous
1057 injection of CT26 cells (2.5×10^5) into mice. When the tumour reached the volume of
1058 approximately 100 mm^3, the mice were randomly divided into two groups (n = 5) and
1059 intraperitoneally treated with rabeprazole (0, 40 mg/kg) in solution containing DMSO,
1060 PEG300, PBS, (1/10/89, v/v/v) once a day for 10 days. The tumour size was recorded
1061 using callipers, and tumour volume was calculated using the formula: $V = (L \times R^2)/2$.
1062 At the end of the experiment, the mice were euthanized and the tumours were harvested
1063 for immunohistochemistry and fluorescence-activated cell sorting.

1064 **Nile red staining.** CT26 cells were seeded into Lab-Tek™ II Chamber Slide systems
1065 (Thermo) and incubated with rabeprazole or vehicle. Then, the cells were washed with
1066 PBS for 15 minutes, fixed with Immunol Staining Fix Solution (Beyotime, P0098) for
1067 30 minutes and washed with PBS again, followed by treatment with Immunostaining
1068 Permeabilization Buffer with Triton X-100 (Beyotime, P0098) for 30 minutes and

1069 washing with PBS again. To stain the lipid droplets, the cells were incubated with Nile
1070 Red (2 μM) in the dark for 10-30 minutes and then washed with PBS before the nuclei
1071 were stained with Antifade Mounting Medium with DAPI (Beyotime, P0131).
1072 Fluorescence images were captured using an OLYMPUS IX73 fluorescence
1073 microscope.

1074 **Immunohistochemistry (IHC)**. The isolated CT26 tumour tissue was fixed with
1075 neutral paraformaldehyde, and subsequent staining of cell surface markers was
1076 performed by Servicebio Company.

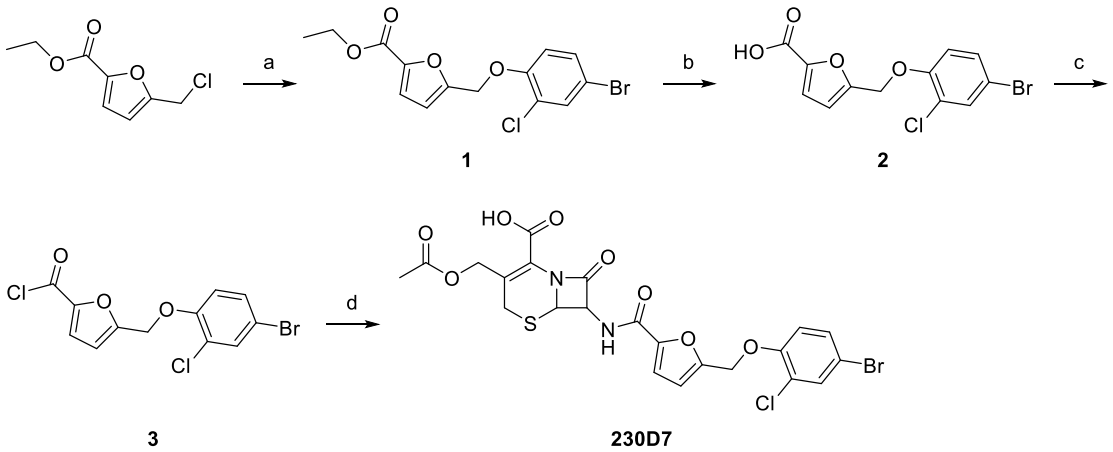
1077 **Fluorescence-activated cell sorting (FACS)**: We analyzed the infiltration of immune
1078 cell subsets in CT26 transplanted tumour tissue by fluorescence-activated cell sorting
1079 (FACS) analysis. After the mice were euthanized, the tumour tissues were stripped and
1080 cut into pieces, then digested at 37 ℃ for 60 minutes with tumour tissue digestive buffer
1081 (0.1% collagenase, 0.001% hyaluronidase, 0.002% DNA enzyme, 120 μM $CaCl_2$ and
1082 120 μM $MgCl_2$ in RPMI1640 medium). The digested tumour tissues were filtered with
1083 200 mesh gauze, followed by the lysis of red blood cells with ammonium chloride
1084 solution, and filtered again to obtain single-cell suspensions in PBS. For discriminating
1085 the living and dead cells, single-cell suspensions were stained on ice with Fixable
1086 Viablity Stain 700 (BD Horizan, 564997) for 10 min and then terminated with cell
1087 staining buffer (PBS containing 2% FBS). Fc receptors on the cell surface are blocked
1088 by anti-Mouse CD16/32 antibody (10 minutes on ice). Then, appropriately conjugated
1089 fluorescent primary antibodies were added to stain cell surface markers. Finally, cells
1090 were suspended with cell staining buffer for flow cytometry analysis using Beckman

CytoFelx. The data were analyzed by Flowjo sofrware.

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 $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

Synthesis of 230D7. Unless otherwise specified, all reagents and solvents were purchased from commercial sources and used without further purification. ^{1}HNMR spectra was recorded on Mercury-600 spectrometers at room temperature. Chemical shifts are referenced to the residual solvent peak and reported in ppm (δ scale), and all coupling constant (*J*) values are given in Hz. ESI-LRMS data were measured on Thermo Exactive Orbitrap plus spectrometer. Flash column chromatography was performed on Flash 300 Isolera one. ^{1}H NMR and ESI-LRMS data were reported on Extended Data Fig. 9.

Scheme I. Synthetic Route to Compound **230D7**



Reagents and conditions: (a) 4-bromo-2-chlorophenol, K_2CO_3, DMF, 60 °C, 5 h; (b) NaOH, MeOH, H_2O, 50 °C, 2 h; (c) $SOCl_2$, r.f., 2 h; (d) β-lactam amine, $NaHCO_3$, acetone, H_2O, r.t., 4 h.

*ethyl 5-((4-bromo-2-chlorophenoxy)methyl)furan-2-carboxylate (**1**)*

To a solution of 4-bromo-2-chlorophenol (1.0 g, 4.9 mmol) and K_2CO_3 (1.1 g, 8.0 mmol) was added 7 mL of DMF at 60 °C for 5 h. After reaction completion, H_2O (50 mL) was added to the mixture followed by extraction with EtOAc (3 × 50 mL). The organic phases were collected and evaporated. The crude product was purified by column chromatography and eluted with 20% ethyl acetate in petroleum ether as a solvent to obtain a white solid product **1** (1.2 g) in 63% yield.

*5-((4-bromo-2-chlorophenoxy)methyl)furan-2-carboxylic acid (**2**)*

To a solution of **1** (1.2 g, 3.4 mmol) and NaOH (0.48 g, 10.2 mmol) was added 15 mL of MeOH and 15 mL of H_2O at 50 °C for 5 h. After reaction completion, MeOH was removed under reduced pressure. In the reaction mixture, 1.0 M aqueous HCl (5.0 mL) was added and pH value was adjusted to 4 to precipitate 1.09 g product **2** without further purification in 98% yield.

*5-((4-bromo-2-chlorophenoxy)methyl)furan-2-carbonyl chloride (**3**)*

To a solution of **2** (0.20g, 0.61 mmol) in $SOCl_2$ (3.0 mL) was stirred and heated at reflux for 2 h. After reaction completion, $SOCl_2$ was removed under reduced pressure to obtain 0.21 g desired product **3** without further purification for next step in 100% yield.

*3-(acetoxymethyl)-7-(5-((4-bromo-2-chlorophenoxy)methyl)furan-2-carboxamido)-8-oxo-5-thia-1-azabicyclo[4.2.0]oct-2-ene-2-carboxylic acid (**230D7**)*

3-(acetoxymethyl)-7-amino-8-oxo-5-thia-1-azabicyclo[4.2.0]oct-2-ene-2-carboxylic acid (0.21 g, 0.76mmol) was dissolved in 20 mL sat. aqueous NaHCO3 and 7 mL acetone at 0 °C. To the stirring solution, **3** (0.32 g, 0.91 mmol) in 5 mL acetone was

added to reaction mixture dropwise in 15 min. After ice water bath was removed, reaction was allowed to stir at room temperature for another 4 h. After reaction completion, pH value was adjusted to 4 with 1.0 M aqueous HCl. Subsequently, the mixture was filtered to collect precipitate as crude product. The crude product was purified by column chromatography and eluted with 5% MeOH in DCM as a solvent to obtain 0.30 g white solid product (**230D7**) in 68% yield. ^{1}H NMR (600 MHz, DMSO-d_6) δ 13.71 (s, 1H), 9.35 (d, J = 8.0 Hz, 1H), 7.72 (d, J = 3.0 Hz, 1H), 7.45 (dd, J = 9.0, 3.0 Hz, 1H), 7.38 (d, J = 3.0 Hz, 1H), 7.33 (d, J = 9.0 Hz, 1H), 6.78 (d, J = 3.0 Hz, 1H), 5.82 (dd, J = 8.0, 5.0Hz, 1H), 5.24 (s, 2H), 5.18 (d, J = 5.0 Hz, 1H), 5.00 (d, J = 13.0 Hz, 1H), 4.71 (d, J = 13.0 Hz, 1H), 3.65 (d, J = 18.0 Hz, 1H), 3.51 (d, J = 18.0 Hz, 1H), 2.04 (s, 3H), ESI-MS: calcd for $C_{22}H_{18}BrClN_2O_8S$ $[M-H]^-$: 584.9, found: 584.9.

Data availability

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

Code availability

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

References

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

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

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

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

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

61 Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res* **45**, D945-D954, doi:10.1093/nar/gkw1074 (2017).

62 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).

1175 63 Kramer, C., Kalliokoski, T., Gedeck, P. & Vulpetti, A. The Experimental
1176 Uncertainty of Heterogeneous Public Ki Data. *Journal of Medicinal Chemistry*
1177 **55**, 5165-5173, doi:10.1021/jm300131x (2012).
1178 64 Lenselink, E. B. *et al.* Beyond the hype: deep neural networks outperform
1179 established methods using a ChEMBL bioactivity benchmark set. *Journal of*
1180 *Cheminformatics* **9**, 45, doi:10.1186/s13321-017-0232-0 (2017).
1181 65 Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and
1182 Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal*
1183 *Chemistry* **47**, 1739-1749, doi:10.1021/jm0306430 (2004).
1184 66 Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and
1185 validation of a genetic algorithm for flexible docking11Edited by F. E. Cohen.
1186 *Journal of Molecular Biology* **267**, 727-748,
1187 doi:https://doi.org/10.1006/jmbi.1996.0897 (1997).
1188 67 Zhang, N. & Zhao, H. Enriching screening libraries with bioactive fragment
1189 space. *Bioorg Med Chem Lett* **26**, 3594-3597, doi:10.1016/j.bmcl.2016.06.013
1190 (2016).
1191 68 Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of
1192 docking with a new scoring function, efficient optimization, and multithreading.
1193 *J Comput Chem* **31**, 455-461, doi:10.1002/jcc.21334 (2010).
1194 69 Shen, C. *et al.* Beware of the generic machine learning-based scoring functions
1195 in structure-based virtual screening. *Brief Bioinform*, doi:10.1093/bib/bbaa070
1196 (2020).

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

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

Competing interests

The Shanghai Institute of Materia Medica has applied for Chinese patents that covers SPOP compounds.

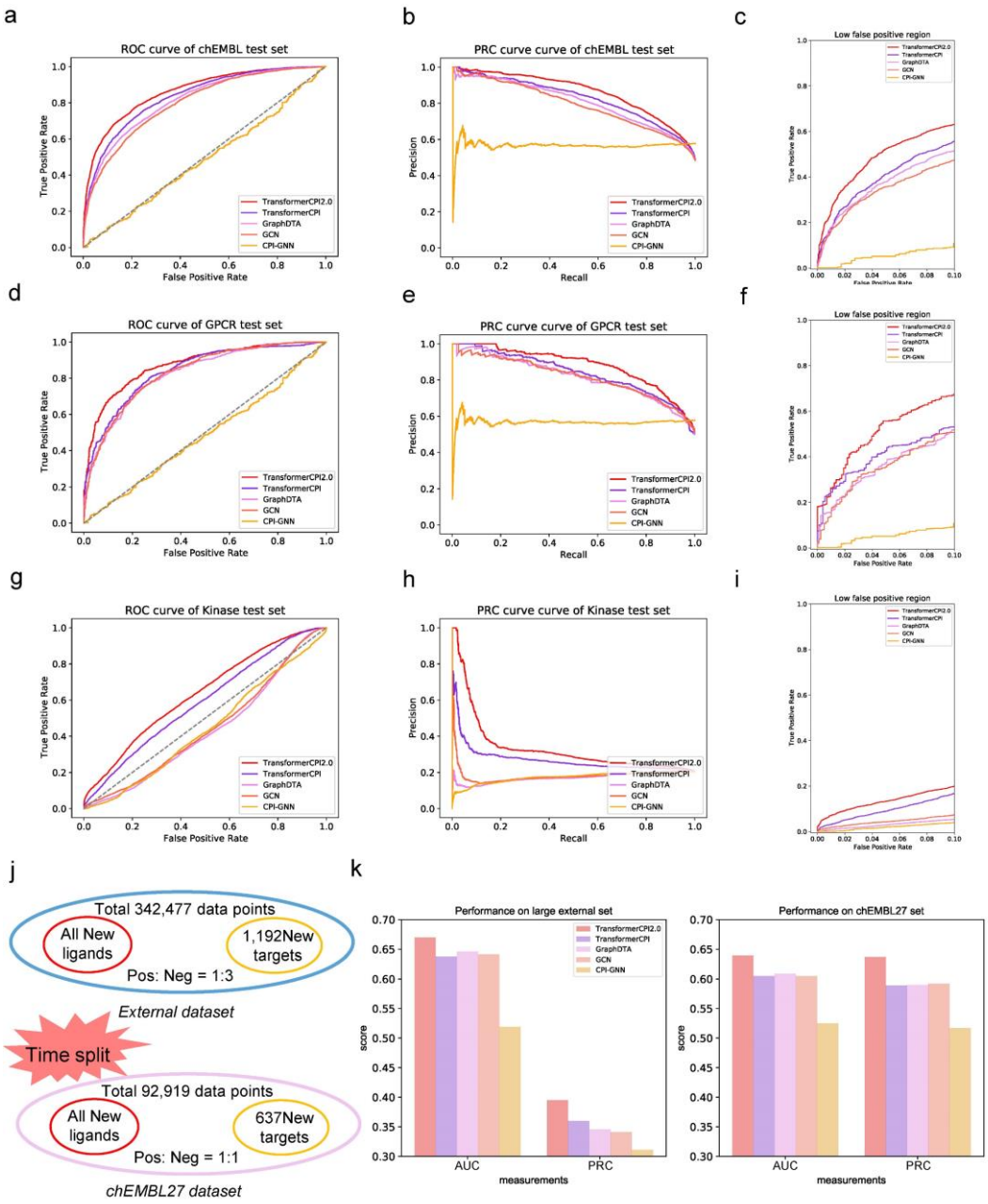
Acknowledgements

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

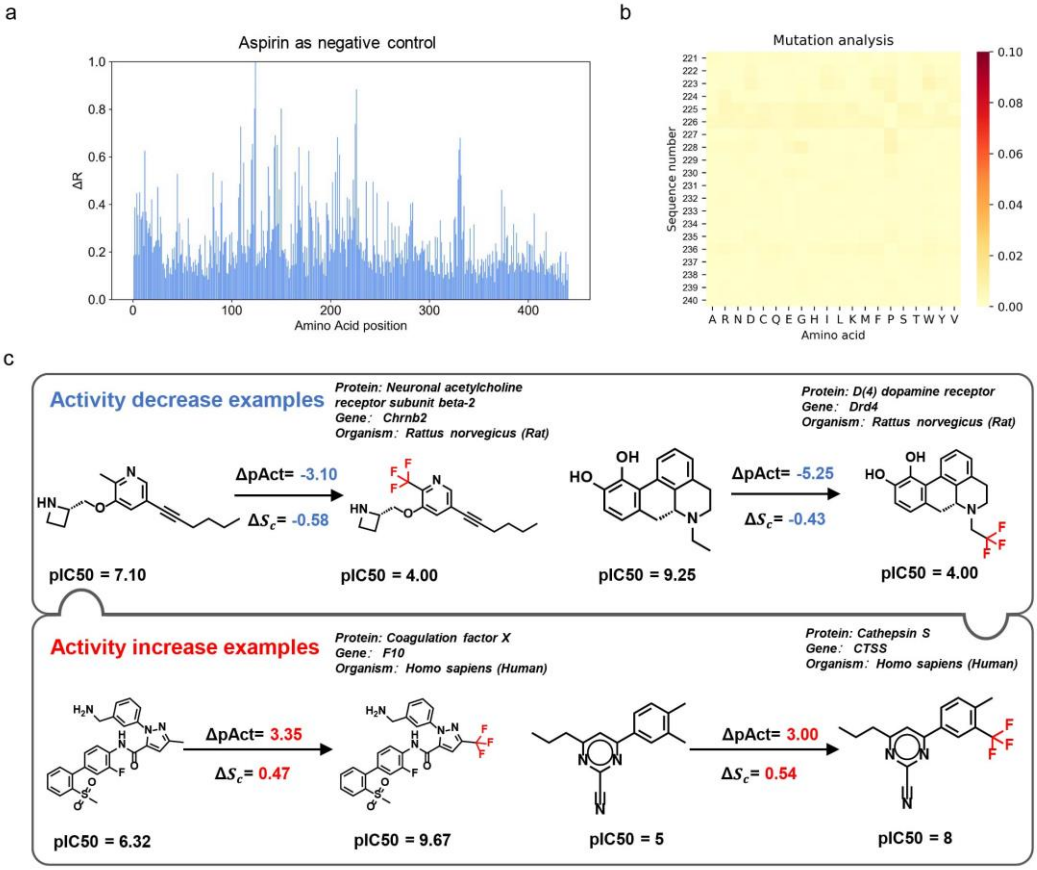
Author Contributions

L.C., Z.F., J.C. and R.Y. designed and performed the experiments, prepared the figures, and wrote the manuscript; L.C. designed TransformerCPI2.0 and conducted computational work; T.Y. conducted computational analysis; Z.F., C.Z., Z.C., R.C. and R.Y. contributed to the biological experiments on SPOP inhibitors; H.G., C.Z. and X.H.

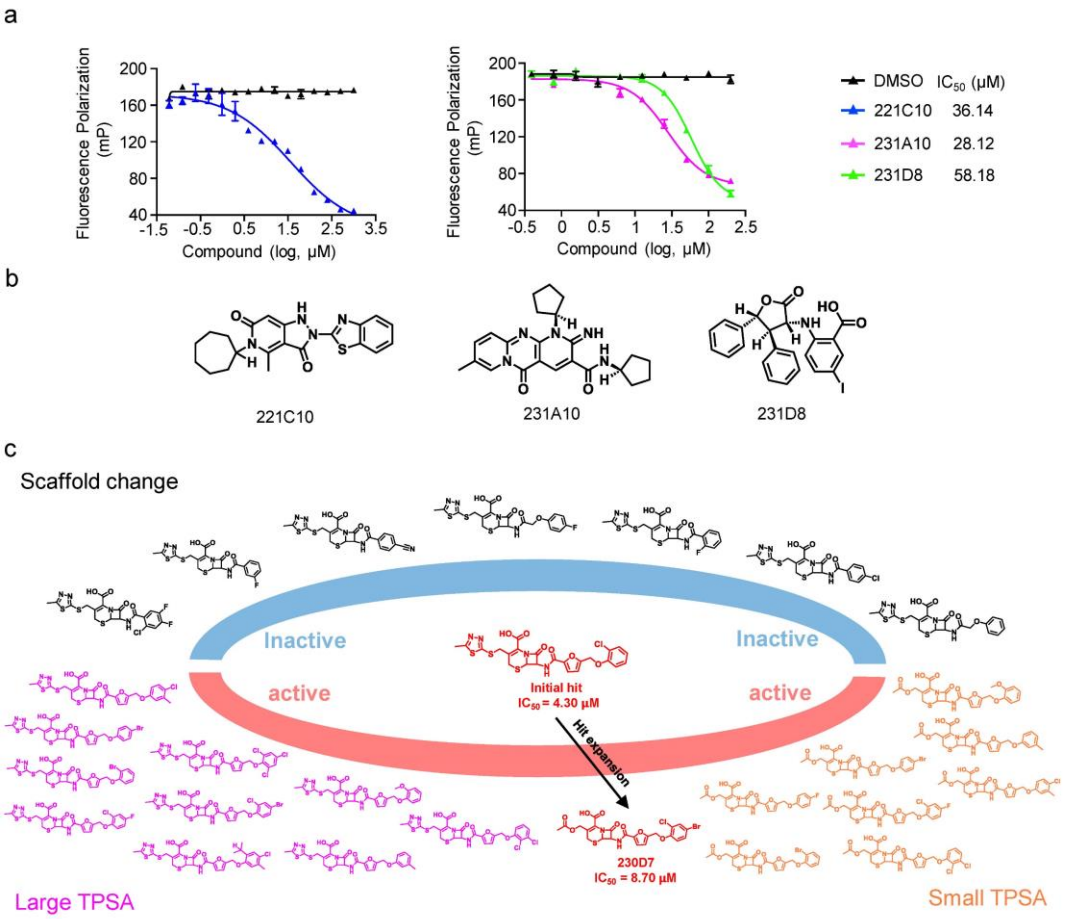
contributed to the synthesis of SPOP inhibitors; Y.D. and N.Z. contributed to nuclear magnetic resonance experiments on SPOP inhibitors; R.Y. contributed to the biological experiments on RNF130; and J.C., Y.Z., K.Z. and R.Y. contributed to the biological experiments on ARF1 inhibitors. M.Z., S.Z. and H.J. initiated, designed and supervised this study.



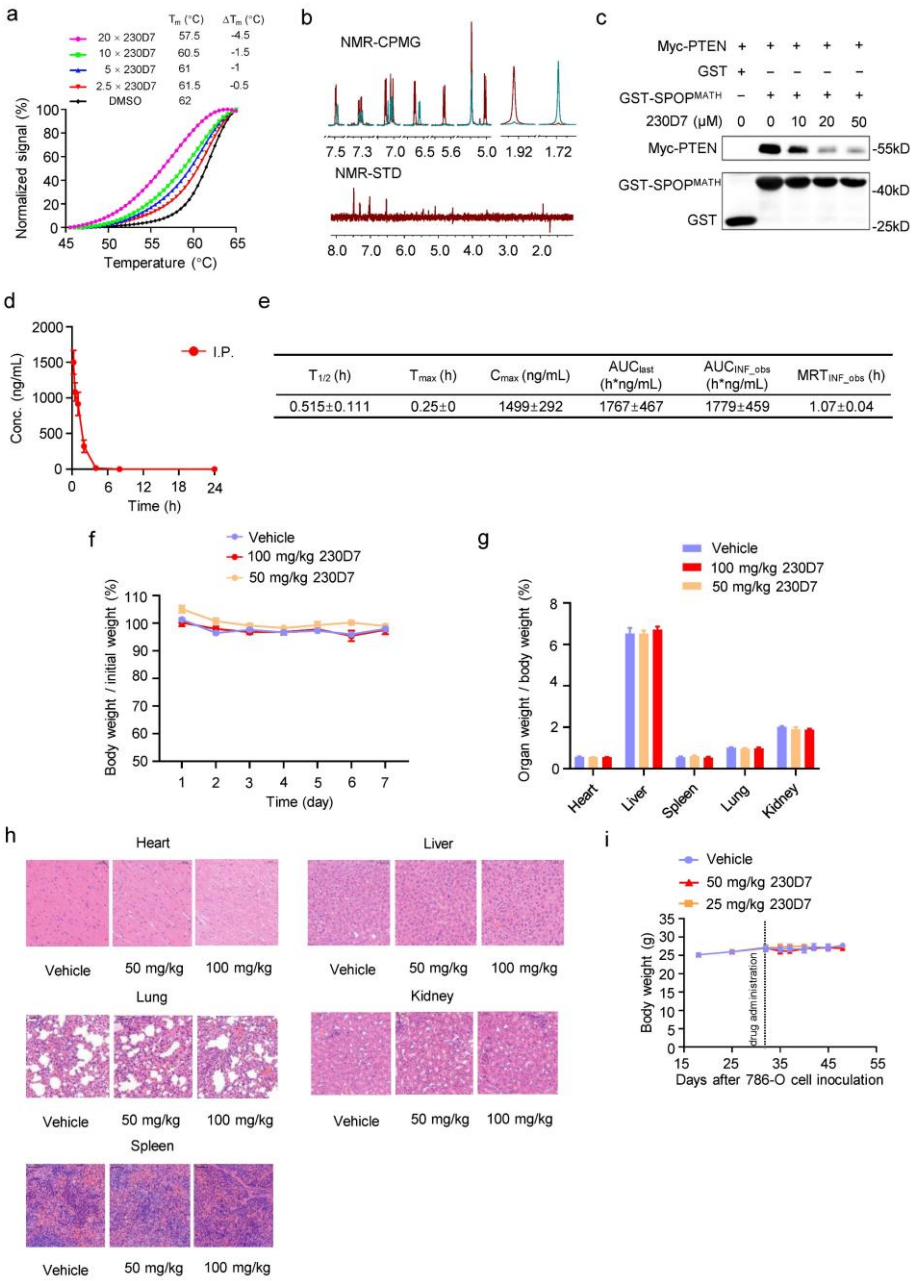
Extended Data Fig. 1 | **TransformerCPI 2.0 predicts compound–protein interaction without protein structure. a,** AUC curves of TransformerCPI2.0 and baseline models on the chEMBL set. **b**, PRC curves of TransformerCPI2.0 and baseline models on the chEMBL set. **c**, ROC curves in low-false-positive 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.



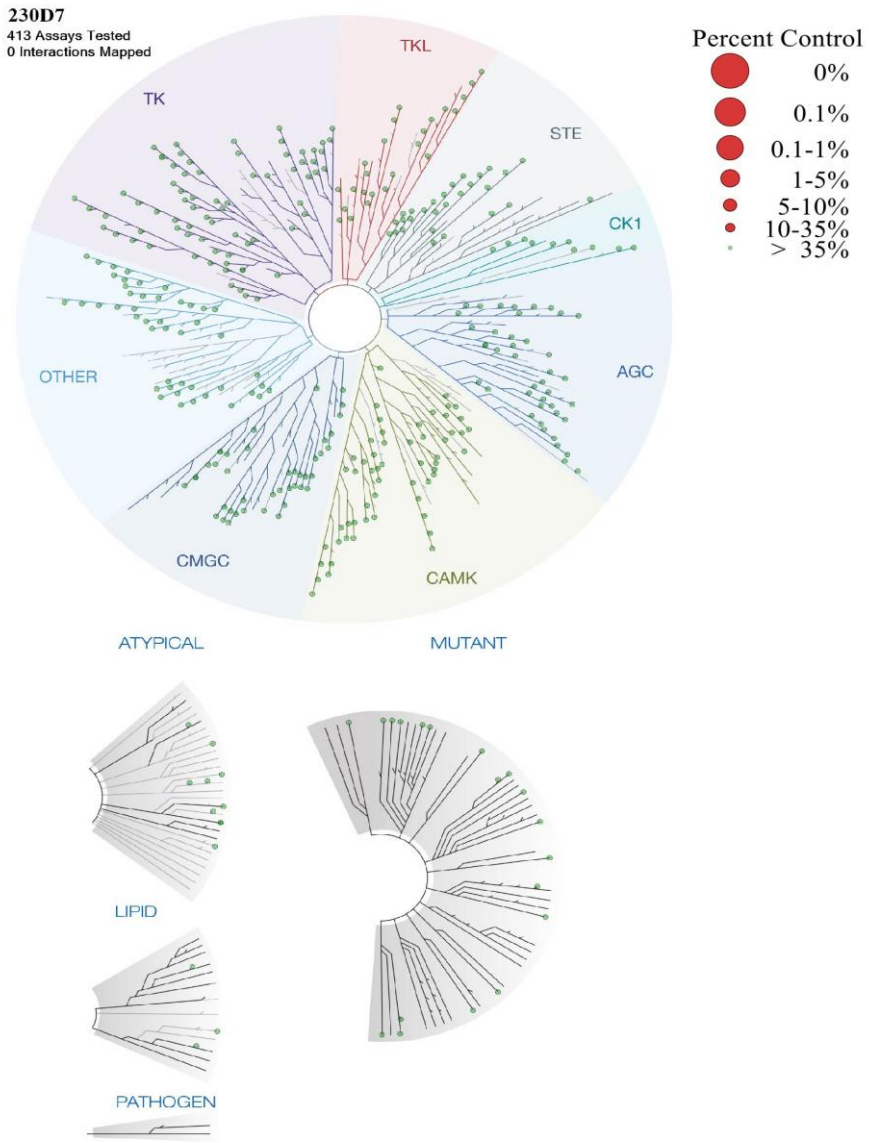
Extended Data Fig. 2 | Supplementary data from drug resistance mutation analysis and substitution effect analysis of the trifluoromethyl group. a, Relative activity change score (ΔR) calculated by TransformerCPI 2.0 for each amino acid position. The pink boxes mark the high ΔR regions, which are plotted 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. 3. **a**, The FP assay of three initial hits predicted by TransformerCPI2.0. **b**, Chemical structures of the three initial hits predicted by TransformerCPI2.0. **c**, A similarity search of 221C7 was conducted, and 26 compounds were purchased, 19 of which were active in the FP assay.

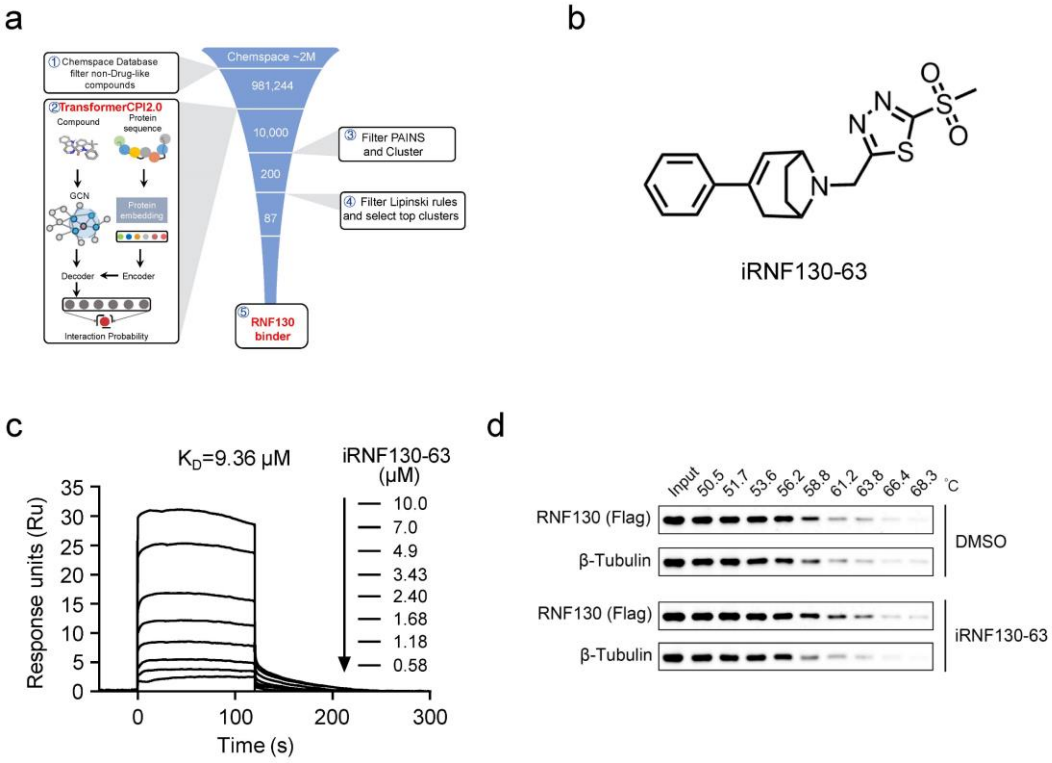


Extended Data Fig. 4 | Supplementary data for Fig. 4. a, Thermostability of SPOP^{MATH} treated with different concentrations of 230D7. The thermal stability of SPOP^{MATH} was quantified by the ΔT_m. **b**, NMR measurement of direct binding between 230D7 and SPOP^{MATH}. CPMG NMR spectra for 230D7 (red), 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.

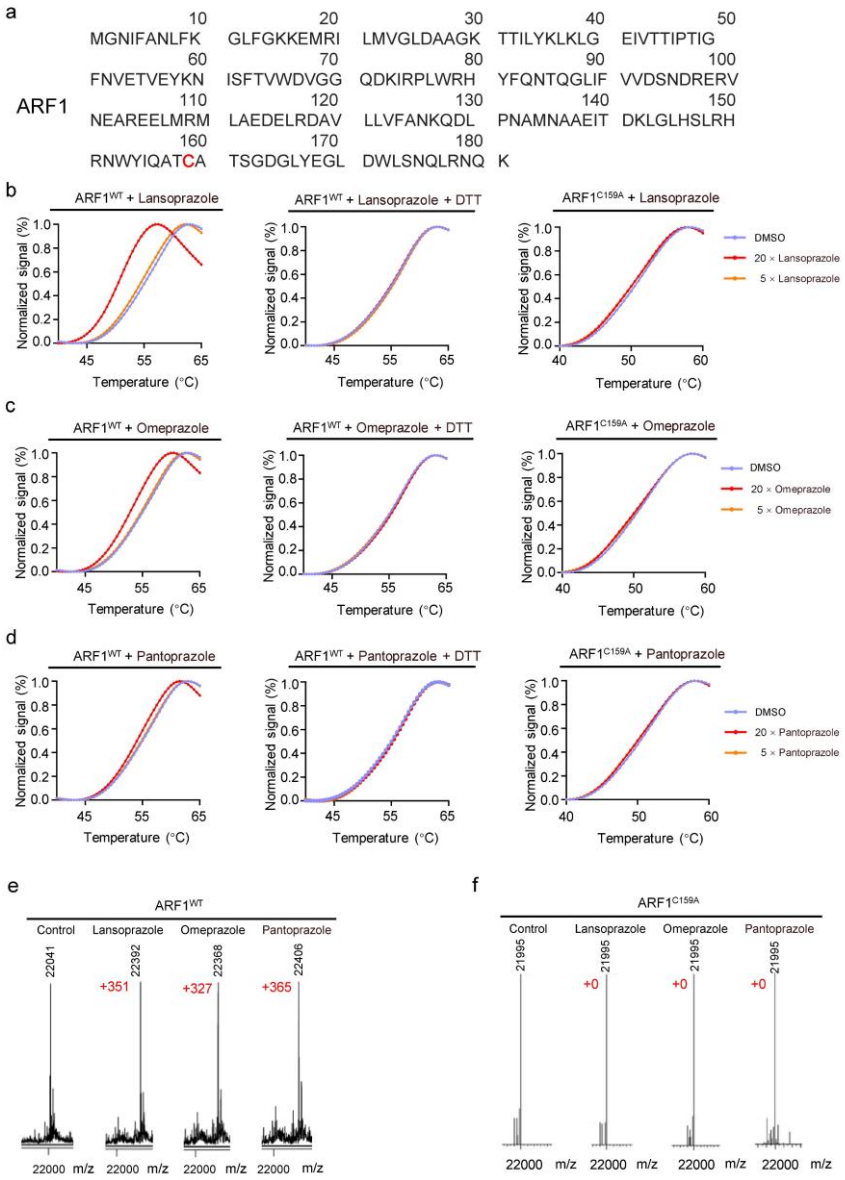


Extended Data Fig. 5 | Kinome profiling of 230D7. 230D7 (10 μM) was submitted for a KinaseProfiler (eurofins) to quantify interactions with 413 human wild-type/mutant kinases. The results are displayed as a TREESPOT interaction map. Image generated using TREEspot™ Software Tool and reprinted with permission from KINOMEscan®, a division of DiscoveRx Corporation, © DISCOVERX CORPORATION 2010.

1231

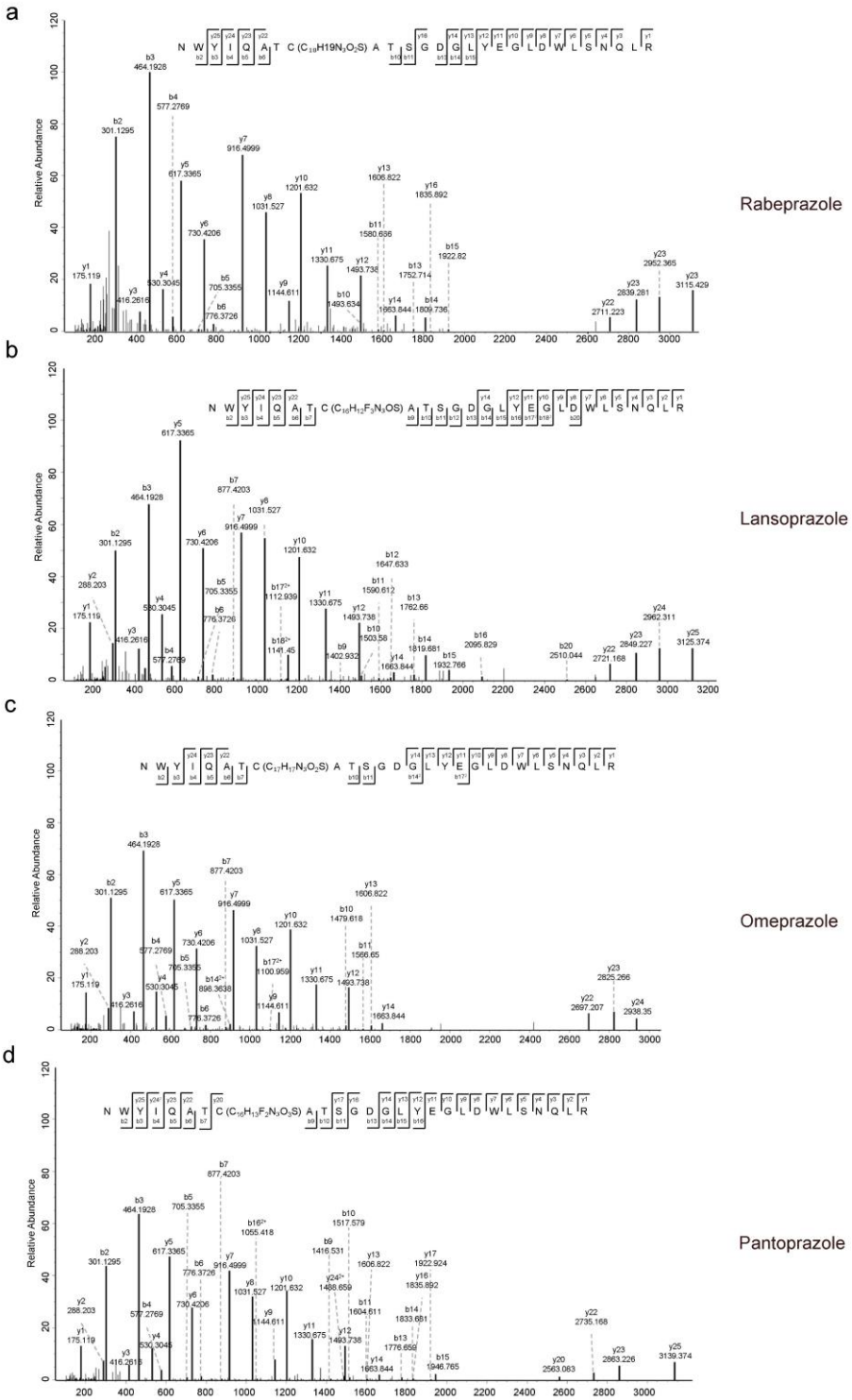


Extended data Fig. 6 | Discovering the chemical binder of RNF130. a, The virtual screening procedure of RNF130. **b**, Chemical structure of iRNF130-63. **c**, Surface plasmon resonance analysis examining the direct binding affinities of iRNF130-63 to RNF130. Graphs of equilibrium unit responses versus 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.



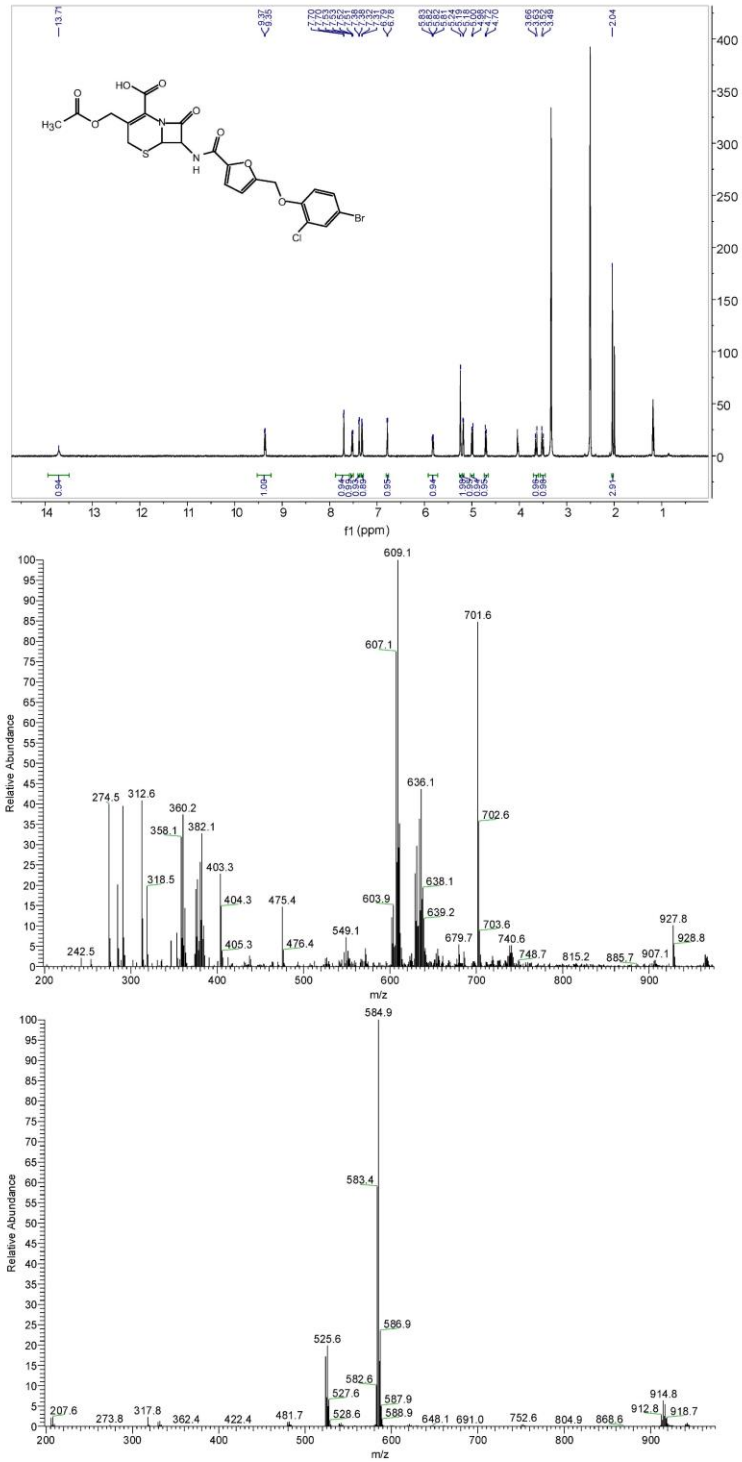
Extended Data Fig. 7 | Supplementary data for Fig 5. a, Amino acid sequence of ARF1, and C159 (marked red) is the only cysteine residue. **b**, PTS assay of lansoprazole+ ARF1^{WT} (left panel); PTS assay of lansoprazole+ ARF1^{WT}+ DTT (middle panel); PTS assay of ARF1^{C159A}+ lansoprazole (right panel). **c**, PTS assay of omeprazole+ ARF1^{WT} (left panel); PTS assay of omeprazole+ ARF1^{WT}+ DTT (middle panel); PTS assay of ARF1^{C159A} + omeprazole (right panel). **d**, PTS assay of pantoprazole+ ARF1^{WT} (left panel); PTS assay of pantoprazole+ ARF1^{WT}+ DTT (middle panel); PTS assay of ARF1^{C159A}+ pantoprazole (right panel). **e**, Deconvoluted electrospray ionization mass spectra of ARF1^{WT} in the presence of lansoprazole or omeprazole or pantoprazole. **f**, Deconvoluted electrospray ionization mass spectra of ARF1^{C159A} in the presence of lansoprazole or omeprazole or pantoprazole.

1234
1235



Extended Data Fig. 8 | Two-dimensional mass spectra of rabeprazole, lansoprazole, omeprazole, and pantoprazole. a~d, Q-Exactive tandem mass spectra results showed the modified peptide of ARF1, demonstrating that ARF1 was covalently modified by, rabeprazole (a), lansoprazole (b), omeprazole (c) and pantoprazole (d) at cysteine 159.

1236
1237



Extended Data Fig. 9 | ^{1}H NMR and ESI-LRMS spectra of 230D7.

1238