



Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com



ORIGINAL ARTICLE

Deep learning-based discovery of tetrahydrocarbazoles as broad-spectrum antitumor agents and click-activated strategy for targeted cancer therapy



Xue Liu^{a,†}, Yalan Lu^{b,c,†}, Qichen Chen^{d,†}, Minjian Yang^e, Shize Li^a, Hanyu Sun^a, Xiangying Liu^a, Jingjie Yan^a, Liangning Li^a, Nan Xiang^{a,f}, Yan Lu^c, Qi Geng^c, Yiqiao Deng^g, Baolian Wang^a, Jing Jin^{a,h,*}, Hong Zhao^g, Xiandao Pan^a, Ahmed Al-Harrasiⁱ, Tingting Du^{a,h,*}, Wei Song^{c,*}, Xiaojian Wang^{a,*}

^aState Key Laboratory of Bioactive Substances and Functions of Natural Medicines, Institute of Materia Medica, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing 100050, China

^bKey Laboratory of Human Disease Comparative Medicine, Chinese Ministry of Health, Beijing Key Laboratory for Animal Models of Emerging and Remerging Infectious Diseases, Institute of Laboratory Animal Science, Chinese Academy of Medical Sciences and Comparative Medicine Center, Peking Union Medical College, Beijing 100021, China

^cDepartment of Biochemistry and Molecular Biology, State Key Laboratory of Medical Molecular Biology, Institute of Basic Medical Sciences Chinese Academy of Medical Sciences, School of Basic Medicine Peking Union Medical College, Beijing 100005, China

^dDepartment of Colorectal Surgery, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, China

^ePeking University Health Science Center-Stonewise Joint Laboratory of Biomedical AI and Data Technologies, Beijing 100080, China

^fKey Laboratory of Molecular Pharmacology and Drug Evaluation (Yantai University), Ministry of Education, Collaborative Innovation Center of Advanced Drug Delivery System and Biotech Drugs in Universities of Shandong, Yantai University, Yantai 264005, China

^gDepartment of Hepatobiliary Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China

^hBeijing Key Laboratory of Key Technologies for Preclinical Research and Development of Innovative Drugs in Pharmacokinetics and Pharmacodynamics, Beijing 100050, China

ⁱNatural and Medical Sciences Research Center, University of Nizwa, Nizwa 616, Oman

*Corresponding authors.

E-mail addresses: rebeccagold@imm.ac.cn (Jing Jin), ninadu@imm.ac.cn (Tingting Du), songwei@ibms.pumc.edu.cn (Wei Song), wangxiaojian@imm.ac.cn (Xiaojian Wang).

[†]These authors made equal contributions to this work.

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

<https://doi.org/10.1016/j.apsb.2025.10.005>

2211-3835 © 2025 The Authors. Published by Elsevier B.V. on behalf of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Received 4 March 2025; received in revised form 16 May 2025; accepted 30 June 2025

KEY WORDS

Deep learning;
Phenotypic screening;
Tetrahydrocarbazoles;
Drug delivery;
Click-activated prodrug;
Antitumor;
Drug discovery;
p53

Abstract Phenotypic screening has played an important role in discovering innovative small-molecule drugs and clinical candidates with unique molecular mechanisms of action. However, conducting cell-based high-throughput screening from vast compound libraries is extremely time-consuming and expensive. Fortunately, deep learning has provided a new paradigm for identifying compounds with specific phenotypic properties. Herein, we developed a data-driven classification-generation cascade model to discover new chemotype antitumor drugs. Through wet-lab validation, WJ0976 and WJ0909 were identified as tetrahydrocarbazole derivatives and displayed potent broad-spectrum antitumor activity as well as growth inhibitory properties against multidrug-resistant cancer cells. Furthermore, the *R*-(–)-WJ0909 (WJ0909B), demonstrated optimal antitumor efficacy *in vitro* and *ex vivo* patient-derived organoids (PDOs). Further investigations revealed that WJ0909B upregulates p53 expression and cause mitochondria-dependent endogenous apoptosis. Moreover, WJ0909B and the click-activated prodrug WJ0909B-TCO potently inhibited tumor growth in cell-derived xenograft models. This research highlights the significant potential of deep learning-guided approach to phenotypic drug discovery for anticancer drugs and the strategy of click-activated prodrug for targeted cancer therapy.

© 2025 The Authors. Published by Elsevier B.V. on behalf of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Phenotypic screening, as an alternative to target-focused approaches, has re-emerged and occupied an important position in the discovery of first-in-class small-molecule drugs since 2011^{1,2}. It is considered to have significant advantages in identifying drug leads and clinical candidates with new molecular mechanisms³. In addition, the recent rapid advances in phenotypic screening technologies, including induced pluripotent stem (iPS)-cell technologies, CRISPR-Cas gene-editing tools, organoids, and imaging assay technologies, have enabled the rapid discovery of drug candidates with specific pharmacological properties⁴. However, the screening of the vast number of compounds in these libraries is extremely time-consuming and expensive^{5,6}. This raises the critical question of how to best use this prospective approach to identify new drugs through phenotypic screening⁷.

Fortunately, machine learning-driven classifiers and generative deep-learning (GDL) models have emerged as promising approaches for drug discovery⁸. These approaches have increased the probability of identifying new drug chemotypes with specific phenotypic properties^{9–11}. For example, machine learning-driven classifiers have been employed to analyze a large number of compounds for phenotypic drug discovery, successfully contributing to the identification of antibiotics^{12–14} and antiplasmodials¹⁵. Meanwhile, GDL models can learn probability distributions of molecules and sample molecules from the corresponding chemical space¹⁶, resulting in the discovery of numerous drug candidates with entirely new chemical structures¹⁴, such as the RIPK1 inhibitor RI-962¹⁷ and novel macrocyclic JAK2 inhibitor¹⁸. These alongside other pivotal studies illustrate the importance of deep-learning approaches toward the identification of structurally and functionally new drug candidates^{19,20}.

Given that cancer is composed of a very large number of molecularly and phenotypically distinct diseases^{21,22}, phenotypic screening approach remains a crucial strategy for identifying, validating, and developing new anticancer drugs with optimal molecular action mechanisms²³, especially for chemo-refractory and drug-resistant tumors^{24,25} where improved therapeutic options are urgently needed^{26,27}.

To the best of our knowledge, this study was the first to establish a new cascade model by integrating machine learning-driven classifiers and GDL models for the discovery of desirable antitumor drugs through phenotypic data. Firstly, the machine learning-driven classifiers were used to identify hit compounds from in-house library. These hit compounds were subsequently subjected to lead optimization using GDL models. Through the integrated model prediction and wet-lab validation, tetrahydrocarbazole derivatives WJ0976 and WJ0909 were identified as potent antitumor drugs with broad-spectrum activity, exhibiting half-maximal inhibitory concentration (IC₅₀) values of 0.2–45 nmol/L against various cancer cell lines and demonstrating growth inhibitory properties against drug-resistance cells induced by paclitaxel, cisplatin, or doxorubicin. Furthermore, WJ0909B, as the *R* enantiomer of WJ0909, exhibited optimal efficacy *in vitro* and showed significant therapeutic effect *ex vivo* PDOs. Further investigations revealed that WJ0909B induced mitochondrial dysfunction and endogenous apoptosis through mechanisms involving the p53 signaling pathway. Notably, WJ0909B demonstrated potential therapeutic properties in xenograft models, and the click-activated prodrug WJ0909B-TCO exhibited tumor growth inhibition with minimal toxicity. This work underscores the use of deep learning in phenotypic data-driven antitumor drug discovery, introduces tetrahydrocarbazole derivatives as promising broad-spectrum antitumor agents, and

leverages the click-activated prodrug strategy to precisely control drug activation in tumor sites.

2. Results and discussion

2.1. Deep learning-guided discovery of the broad-spectrum antitumor drugs WJ0976 and WJ0909

In order to establish a new paradigm for phenotypic data-driven antitumor drug identification by deep learning, this study proposed a cascade of fingerprint-based classifier and GDL model¹⁹ to automate the process of lead compound screening and optimization. Specifically, as shown in Fig. 1A, compounds with antiproliferative activities were collected from databases and literature, and composing a unified dataset for subsequent training of classification and generation model. The fingerprint-based classifier was then exploited to identify hit compounds from the database, following which the selected hit compounds were optimized *via* the GDL model to afford candidates with higher activity. This approach enabled the successful discovery of tetrahydrocarbazoles broad-spectrum antitumor drugs, WJ0976 and WJ0909.

Initially, molecules with antiproliferative activities were collected from databases and literature to yield a unified dataset. This would enable the identification of the chemical substructures associated with antitumor activity by binary classification model and the generation of optimized compounds by GDL model. Then 1112 molecules were collected from the Cancer Drug Resistance (CancerDR) Database²⁸, the L5500 Database (a commercial compound library from TargetMol), and self-collected literature (Supporting Information Fig. S1). These molecules were divided into 399 active and 713 inactive molecules based on an IC₅₀ value of 1 $\mu\text{mol/L}$ as the hit cutoff. Following binarization, they were used to train the binary classification model, enabling the identification of compounds with antitumor activity. When training the generative model, we included only 399 active molecules in the training set.

Due to the limited number of broad-spectrum anti-tumor molecules collected, we selected machine learning methods that more suitable for small datasets and a commonly used molecular representation method, ECFP4 fingerprint, for constructing the eXtreme Gradient Boosting (XGBoost) classification model^{29,30}, followed by the use of the GDL model to optimize antitumor compounds. As shown in Fig. 1B, the classifier uses multiple trees for decision, result of each tree is the difference between the target value and the prediction results of all previous trees, and the final accurate prediction result is obtained by accumulating all these results. The architecture of the GDL model, illustrated in Fig. 1C—is based on the BART framework with a bidirectional encoder and an autoregressive decoder. Meanwhile, the GDL model also involved advanced attention mechanisms. As shown in Fig. 1D, each row in the input information H represents a corresponding input vector. Additionally, three matrices include W_q , W_k , and W_v , which are responsible for transforming the input information H into the corresponding query space Q , key space K , and value space V . After obtaining the representations of the input information in different spaces K , Q , and V , the attention output vector for each position is computed, ultimately yielding the reliable context output.

Fingerprint-based classifier was trained afterwards. The datasets were randomly split into training and test sets at a 4:1 ratio.

The resultant model achieved a receiver operating characteristic curve-area under the curve (ROC-AUC) of 0.921 on the test data and 0.988 on the test data (Fig. 2), indicating a strong predictive performance for the antitumor activities of compounds across a wide range. For comparison, we also attempted the models using Random Forest (RF) and Support Vector Machine (SVM) methods on the same dataset. The ROC-AUC scores of RF and SVM model on the test set were 0.901 and 0.874, respectively, both of which performed noticeably worse than the XGBoost method (Supporting Information Figs. S2 and S3).

To validating the genuine accuracy of the model, the binary classification model was applied to screen the antitumor agents from L4010 Database (a commercial compound dataset containing 20,952 compounds from TargetMol) (Fig. 2A and B). Firstly, in order to meet the requirement that the selected molecule can be further optimized as a hit as a structural fragment with antitumor potential, we first selected molecules with molecular weights ranging from 200 to 500, and obtained 16,516 molecules. Then these molecules were assigned with prediction scores between 0 and 1 by our model (Supporting Information Fig. S4). A higher predicted score indicates that the compound is more likely to possess antitumor activity. Next, we curated the top 1‰ molecules that were most strongly predicted to display antitumor properties by our model and empirically tested these for growth inhibition. We observed that 4 predicted molecules displayed growth inhibition against HeLa and HepG2 with IC₅₀ values less than 10 $\mu\text{mol/L}$ (Supporting Information Table S1), indicating that our classification model which driven by phenotypic data has acceptable accuracy.

Subsequently, to identify structurally novel scaffolds, an in-house library was utilized. Using the same screening process, three molecules ranked in the top 1‰ of the in-house library were selected for wet-lab evaluation (Fig. 2C and D, Supporting Information Table S2). Notably, two of the three compounds exhibited an IC₅₀ of <10 $\mu\text{mol/L}$. Compound **19**, containing a tetrahydrocarbazole scaffold, exhibited the most potent antitumor activity against HeLa and HepG2 cells with IC₅₀ values of 0.147 and 0.198 $\mu\text{mol/L}$, respectively, and was selected for further optimization.

Local Interpretable Model-agnostic Explanations (LIME) analysis was employed to elucidate the critical structural determinants contributing to the antitumor activity of compound **19**. The interpretable model highlighted the carboxamide ($-\text{CONH}_2$) and 3-amino ($-\text{NH}_2$) moieties as the most influential pharmacophoric features (Supporting Information Fig. S5). These functional groups participate in key hydrogen-bonding interactions with target proteins and are essential for maintaining sub-nanomolar potency, their removal or modification may lead to a decrease in activity.

GDL models have advanced the exploration of chemical spaces, especially in *de novo* molecular design. In this study, we utilized the BART-based GDL model to optimize compound **19** (Fig. 2). By using 399 active compounds summarized above, transfer learning was performed on the BART-based GDL model, which had been pre-trained on 50 million molecules from the ZINC^{31,32} database and 1 million molecules from the ChEMBL database³³ in our previous work. As shown in Supporting Information Fig. S6, the trained model, which ran for 750 epochs and achieved minimum loss, was then evaluated. We assessed the model's capacity to generate molecules with antitumor activity by inputting 400 random fragments into the GDL model for molecular generation³⁴. Principal Components Analysis (PCA) was used to visualize the chemical space and showed distinct differences in the distribution of molecules with and without antitumor activity

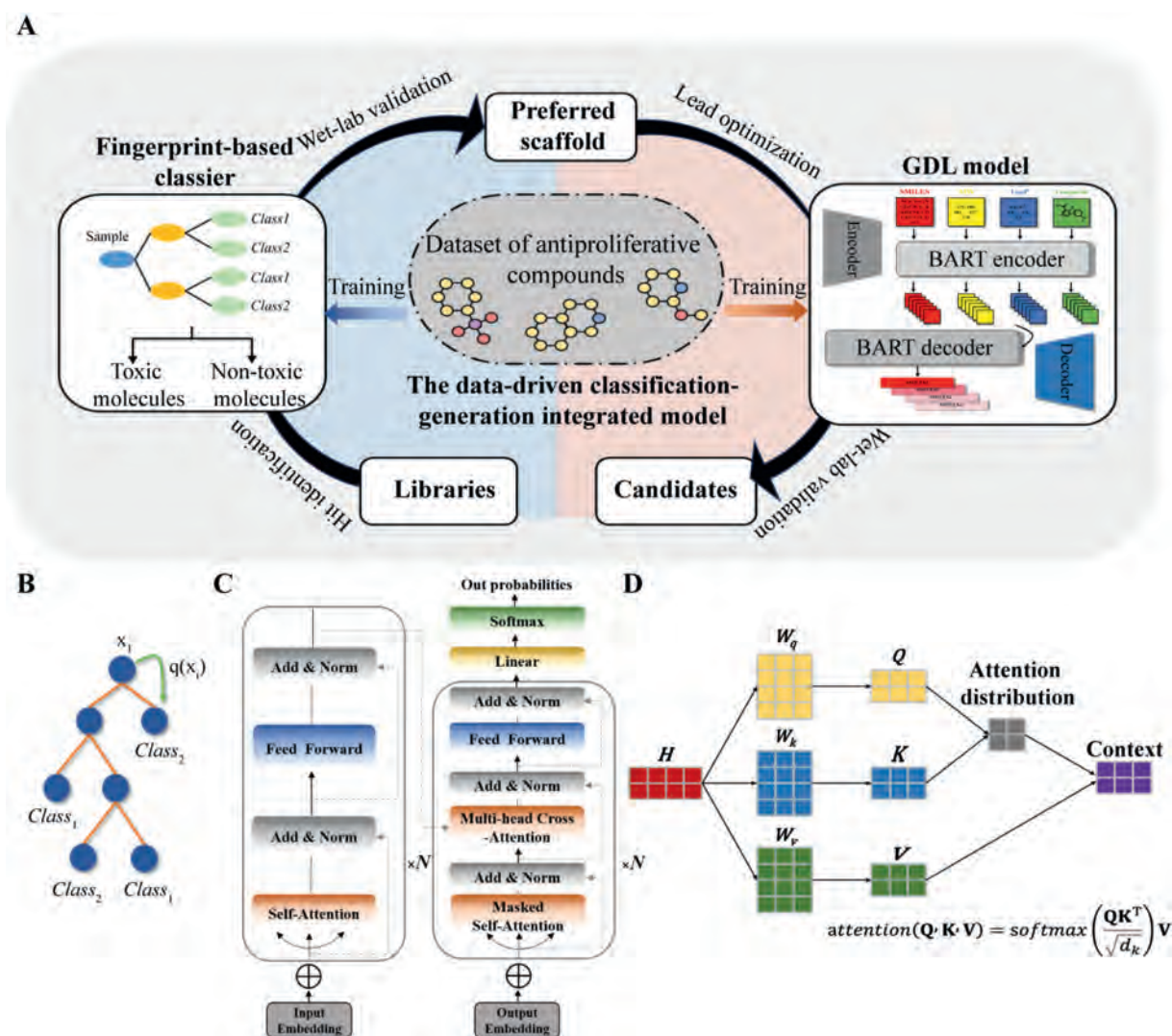


Figure 1 Pipeline of the discovery of broad-spectrum antitumor drugs by unified data-driven classification-generation integrated model and the architectures of models. (A) Workflow of the identification of desirable antitumor drugs by fingerprint-based classifiers and GDL model. (B) The architectures of the fingerprint-based classifier. (C) The architectures of the BART-based GDL model. (D) Illustration of attention mechanism in BART-based GDL model.

(Fig. 2E). The generated molecules demonstrate distribution characteristics that bear greater resemblance to those of antitumor compounds. Meanwhile, the generated molecules displayed similarities to antitumor compounds, particularly in their physico-chemical properties, such as molecular weight and $\log P$ distribution (Fig. 2F and G). The model exhibited excellent generative ability, with 97.8% validity in generating molecules and 94.88% of the generated molecules featuring entirely novel structures. Despite the GDL model demonstrating a uniqueness of only 39.66%, the internal diversity of the generated unique molecules was high (0.8854). Among these molecules, 62.41% have a Tanimoto coefficient (Tc) greater than 0.6 which were similar to them in the training set, and 7.52% with Tc values less than 0.4, indicating relatively novel structures (Supporting Information Table S3). The molecular generation model, following transfer training, demonstrated excellent molecular generation capability and could generate novel molecular structures while meeting the characteristics of antitumor drugs.

The preliminary testing of the screened hit compound **19** suggested that the 3-amino-2,3,4,9-tetrahydro-1*H*-carbazole-8-carboxamide scaffold had potential antitumor activity. Based on the synthetic feasibility, the BART-based GDL model was utilized to perform fixed-point structural modifications to the scaffold for two rounds of optimization (Fig. 2H). We first defined substituents on the benzene ring as fragment 1 and inputted it into the GDL model without specific properties. The model generated 200 molecules with a notable inclination toward substituted phenyl groups. Among the top 50 molecules, 4 compounds (**20–23**) with substituted phenyl groups and relatively simple synthetic pathways were selected for chemical synthesis and biological evaluation.

The selected compounds were synthesized, with the synthesis process of compound **21** (WJ0976) detailed in Supporting Information Fig. S7. Briefly, commercially available 2-amino-4-bromobenzoic acid was initially converted to **2a** through diazotization, followed by Fischer indole synthesis, yielding the key intermediate **3a**. Subsequently, the carboxyl group of **3a** was

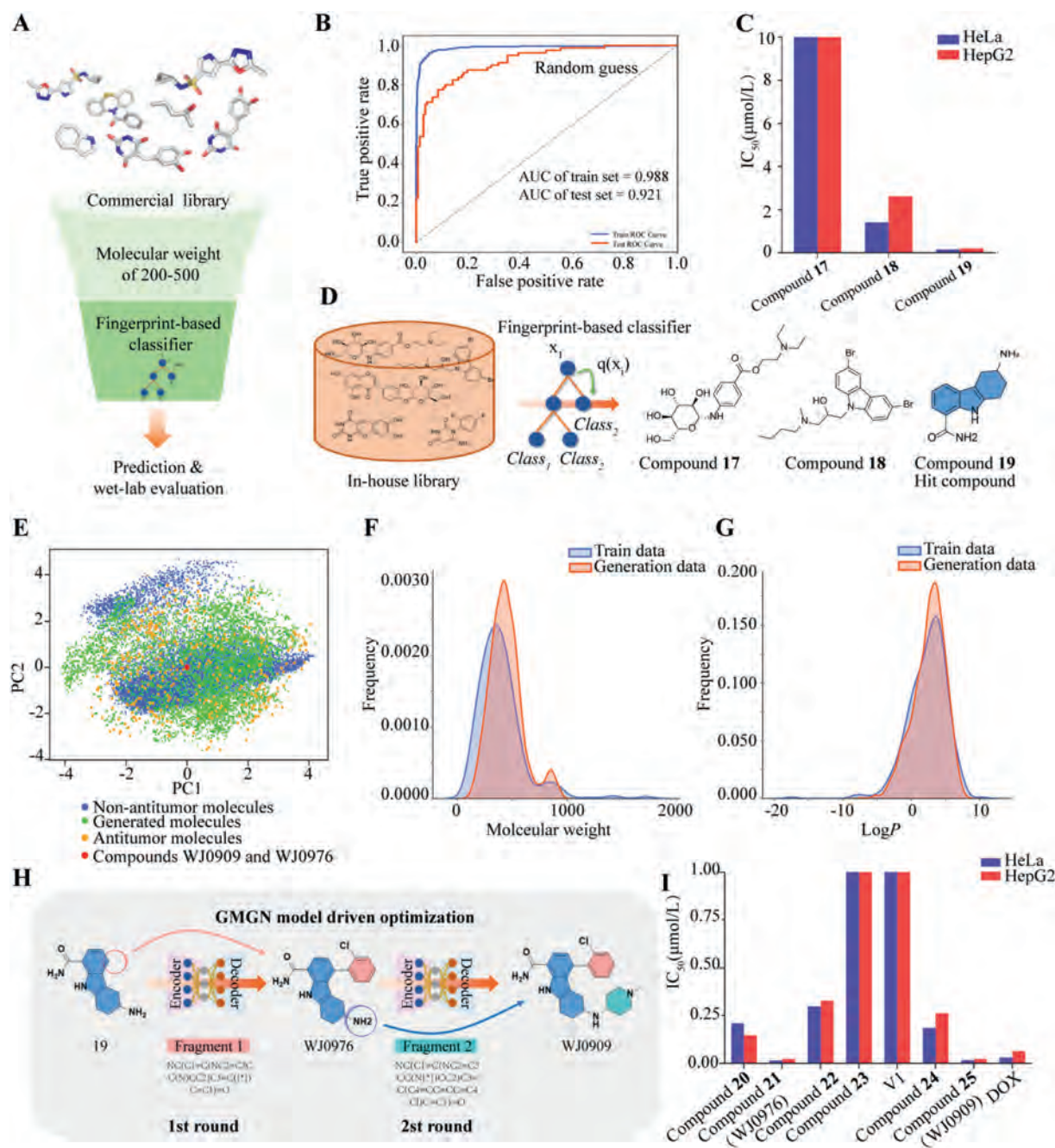


Figure 2 Phenotypic data-driven antitumor drug discovery with deep learning. (A) Workflow for screening potential antineoplastic drugs in commercial library with fingerprint-based classifiers. (B) ROC-AUC plot of the fingerprint-based classifiers on training and test datasets. (C) Antiproliferative efficacy of compounds **17**–**19** in HeLa and HepG2 cells ($n = 3$). (D) Workflow for screening potential antitumor drugs in an in-house library with fingerprint-based classifiers. (E) Distribution of non-antitumor molecules, antitumor molecules, generated molecules, and optimized compounds (WJ0976 and WJ0976) represented by PCA plots. (F, G) Molecule weight and water-octanol partition coefficient (LogP) of the training and generated data are represented by kernel density estimation. (H) Two rounds of GDL-model-based structural optimization of compound **1**. (I) Antiproliferative efficacy of compounds **20**–**26**, **V1** and DOX in HeLa and HepG2 cells. Data are shown as mean $\pm$ SD ($n = 3$).

converted to carboxamide **5a**, with the amine moiety of **3a** being protected with *N*-Boc to produce **4a**. Compound **6a** was then synthesized from **5a** and 2-chlorophenylboronic acid using Suzuki conditions. Finally, WJ0976 was obtained by removing the *N*-Boc protecting group from **6a** under trifluoroacetic acid conditions. The syntheses of the other compounds followed a similar method and are provided in the Supporting Information

The antitumor activity of the obtained compounds was subsequently assessed (Fig. 2I, Supporting Information Table S4). WJ0976 exhibited considerable cytotoxic effects against HeLa ($IC_{50} = 0.015 \mu\text{mol/L}$) and HepG2 ($IC_{50} = 0.023 \mu\text{mol/L}$) cells, 10-fold greater than the cytotoxicity of hit compound **19** (Supporting Information Table S5). Compounds **20** and **22** exhibited similar potency to compound **19**, while compound **23** exhibited

lower activity. To visualize the similarity between WJ0976 and known broad-spectrum cytotoxic drugs, scaffold-like molecules were collected from literature and Tc values were calculated and averaged to obtain the final score. As illustrated in Table S5, the majority of collected compounds have average Tc scores higher than 0.5, indicating that similar structures have been reported. Compound **V1**, the most structurally similar to WJ0976, was reported as BTK inhibitors with moderate cell proliferation inhibitory activity. We further performed antitumor efficiency of **V1** against HeLa and HepG2 and exhibited an IC_{50} of $>10 \mu\text{mol/L}$, respectively. Notably, WJ0976 differed from compound **V1**, demonstrating the powerful ability of WJ0976 in tumor suppression. Based on the analysis of relative scaffold diversity, it can be noticed that 3-amino group were essential for WJ0976 to exert its broad-spectrum tumor proliferation inhibitory activity.

Then WJ0976 was fragmented at the amino site to define fragment 2 for the second round of optimization. Among the top 20 molecules, compounds **24** and **25** (WJ0909) were selected for chemical synthesis and bioactivity evaluation owing to their easy synthetic accessibility (Fig. 2H, Supporting Information Table S4). WJ0909 significantly inhibited cancer cell growth, with IC_{50} values of 0.017 and 0.024 $\mu\text{mol/L}$ in HeLa and HepG2 cells, respectively, while compound **24** was less potent.

Using this cascade model, the preliminary active compound **19** we initially identified using a fingerprint-based classification model. Subsequently, structure optimization was conducted based on the scaffold of compound **19** through generative modeling, resulting in the final compounds WJ0976 and WJ0909, thus completing a closed-loop screening and optimization process based on deep learning. In more complex scenarios, generated molecules can undergo wet-lab validation before being reintroduced into the screening model for iterative improvement. Overall, the findings demonstrate the ability of our model to identify and optimize hit/lead compounds with unreported scaffolds, highlighting the significant potential of deep learning-guided approach to phenotypic drug discovery for anticancer drugs.

2.2. WJ0909B exhibited broad-spectrum antitumor activity

Based on the significant antitumor efficacy of WJ0976 and WJ0909 against HeLa and HepG2 cell lines, we proceeded to assess their growth-inhibitory effects in 24 other cell lines, encompassing 14 solid tumors and 10 hematological malignancies (Fig. 3A). Both WJ0976 and WJ0909 demonstrated significant inhibition of tumor proliferation across various cancer cell lines with IC_{50} values of approximately 20 nmol/L, comparable or even better than that of doxorubicin (Fig. 3B).

Considering the chiral nature of WJ0976 and WJ0909, it piqued our interest to examine the impact of stereochemistry on antitumor activity. Two pairs of enantiomers: *S*-(+)-WJ0976 (WJ0976A), *R*-(-)-WJ0976 (WJ0976B), *S*-(+)-WJ0909 (WJ0909A), *R*-(-)-WJ0909 (WJ0909B) were obtained through the separation of key intermediates using tartaric acid as a chiral chemical building block. The absolute configurations of these enantiomers were determined using single-crystal X-ray diffraction (Supporting Information Figs. S8 and S9) and quantum chemical electronic circular dichroism (ECD) calculations (Supporting Information Fig. S10). The anticancer activities of these enantiomers were then evaluated against 25 tumor cell lines *in vitro*. As illustrated in Fig. 3A, WJ0976A, WJ0976B, and WJ0909B exhibited more potent tumor-suppressive activity than doxorubicin across most cancer cell lines, with IC_{50} values

ranging from 0.5 to 30 nmol/L, more than 30 times more active than WJ0909A (IC_{50} values 219–3101 nmol/L) (Supporting Information Table S6). Overall, it was evident that WJ0909B exhibited optimal antitumor activity *in vitro*.

Since drug resistance is a major obstacle in cancer treatment, both compounds WJ0976 and WJ0909 feature structural scaffolds that differ from those of known broad-spectrum antineoplastic agents. Our goal was to investigate whether they could address the drug resistance associated with conventional therapies. We evaluated the impact of WJ0976A, WJ0976B, and WJ0909B on drug-resistant cell lines that emerged following treatment with commonly used chemotherapeutic drugs such as doxorubicin, paclitaxel, and cisplatin. We observed significant antiproliferative effects, with IC_{50} values below 45 nmol/L, against MCF-7/PTX, A549/DDP, and K562/DOX cell lines (Fig. 3C–E, Supporting Information Table S7). Therefore, WJ0976A, WJ0976B, and WJ0909B effectively reversed multidrug resistance, indicating a potentially new treatment option for chemoresistant tumors.

To better mimic the therapeutic impact of WJ0909B in a more complex cellular environment, we employed five PDOs derived directly from patient colorectal tumor tissue. After treating the PDOs with varying concentrations of WJ0909B for 6 days, CTG reagent was used to assess their viability (Fig. 3F, Supporting Information Table S8). The results revealed that treatment with WJ0909B inhibited growth of PDOs from five different patients, with IC_{50} values of CC116 PDO: 3.1 nmol/L, CC141 PDO: 3.3 nmol/L, CC108 PDO: 20.0 nmol/L, CC87 PDO: 8.4 nmol/L, and CC59 PDO: 8.3 nmol/L. Brightfield microscopy images illustrated that treatment with WJ0909B compromised the structural integrity of the PDOs, which had initially displayed distinct edges and complete morphology, resulting in disintegration into single cells or cell populations (Fig. 3G). Furthermore, we utilized ImageJ software to conduct fitting calculations and statistical analyses on the PDO area (Fig. 3H), confirming the significant inhibitory effect of WJ0909B on colorectal cancer PDO growth.

2.3. WJ0909B caused mitochondria-dependent endogenous apoptosis via the p53 signaling pathway

While tetrahydrocarbazole scaffolds are known in drug discovery (e.g., as kinase inhibitors or antimicrobial agents), WJ0909B represents a novel chemotype with distinct structural features (3-amino-2,3,4,9-tetrahydro-1*H*-carbazole-8-carboxamide) and unparalleled potency (IC_{50} values of 0.2–45 nmol/L across 25 cancer cell lines), suggested that WJ0909B exerts its antitumor effects through a unique mechanism which distinct from those known tetrahydrocarbazole derivatives. We desired to explore the shared molecule mechanism in different cell lines. An integrated transcriptomic and proteomic analysis of WJ0909B-treated RKO and RBE cells revealed significant differences in genes (Supporting Information Fig. S11A–S11H) and proteins (Supporting Information Fig. S12). Subsequent transcriptome and proteome sequencing and KEGG pathway enrichment showed the p53 signaling pathway as the sole coregulated pathway that underwent significant changes following WJ0909B treatment (Fig. 4A–D, Fig. S11I and S11J), indicating that this pathway may mediate the biological function of WJ0909B.

We then proceeded to evaluate the impact of WJ0909B on the p53 signaling pathway. Western blot analysis revealed that WJ0909B upregulated p53 expression in a dose- and time-dependent manner in RKO cells (Fig. 4E, Supporting Information Fig. S13A). At the same time, the protein expression of p21, a

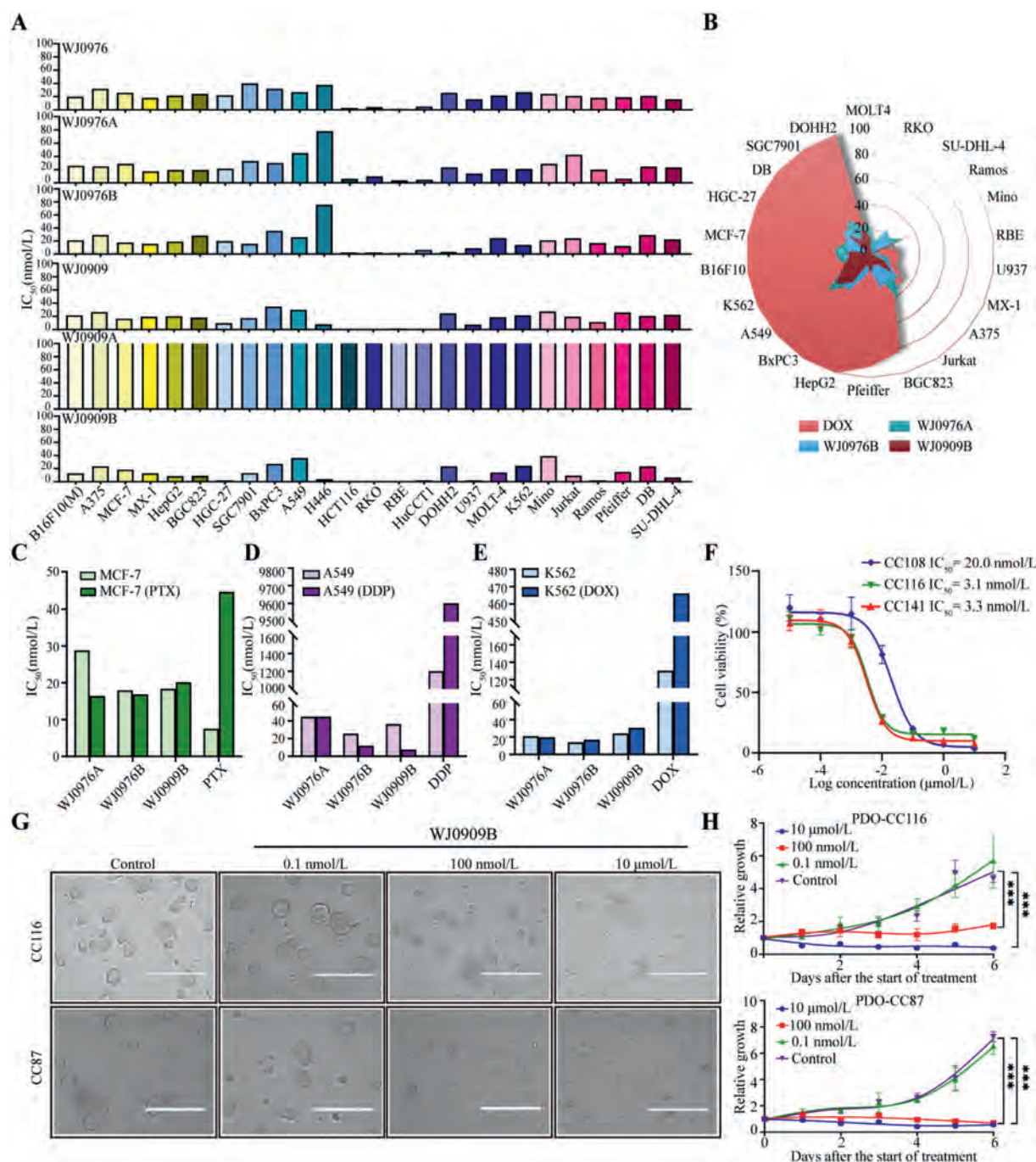


Figure 3 WJ0976A, WJ0976B, and WJ0909B exhibit broad-spectrum antitumor activity. (A, B) Cancer cells were incubated with various concentrations of compounds for 96 h ($n = 3$), and IC_{50} values were determined by MTT and CTG assay, the average IC_{50} values were shown. (C–E) Resistant human breast carcinoma MCF-7/paclitaxel (PTX), human lung adenocarcinoma A549/cisplatin (DDP), and human chronic myelogenous leukemia K562/doxorubicin (DOX) cell lines were treated with different concentrations of WJ0976A, WJ0976B, and WJ0909B for 96 h ($n = 3$) and IC_{50} values were determined by CellTiter-Glo assay and averaged. (F) Growth inhibition against patient-derived colorectal cancer organoids was assessed by CellTiter-Glo 3D cell viability assay after incubation with various concentrations of WJ0909B. The IC_{50} curves of three representative PDOs are shown ($n = 3$). (G) Brightfield microscopy images of patient-derived colorectal cancer organoids degradation after treatment with various concentrations of WJ0909B. Scale bar (white), 200 μ m. (H) The area of PDOs was calculated using ImageJ to build the growth curve. Data are shown as mean $\pm$ SD. * P < 0.05, ** P < 0.01, *** P < 0.001.

key player in the p53 signaling pathway, was markedly higher in WJ0909B-treated cells than in the untreated controls. These findings were further confirmed in the RBE, HCT116, and HuCCT1 cell lines (Fig. 4F, Fig. S13B and S13C). These results

suggested a correlation between heightened p53 expression and the biological function of WJ0909B, aligning with the outcomes of the transcriptome and proteome sequencing analyses. Subsequently, the effect of WJ0909B on apoptosis in both wild-type

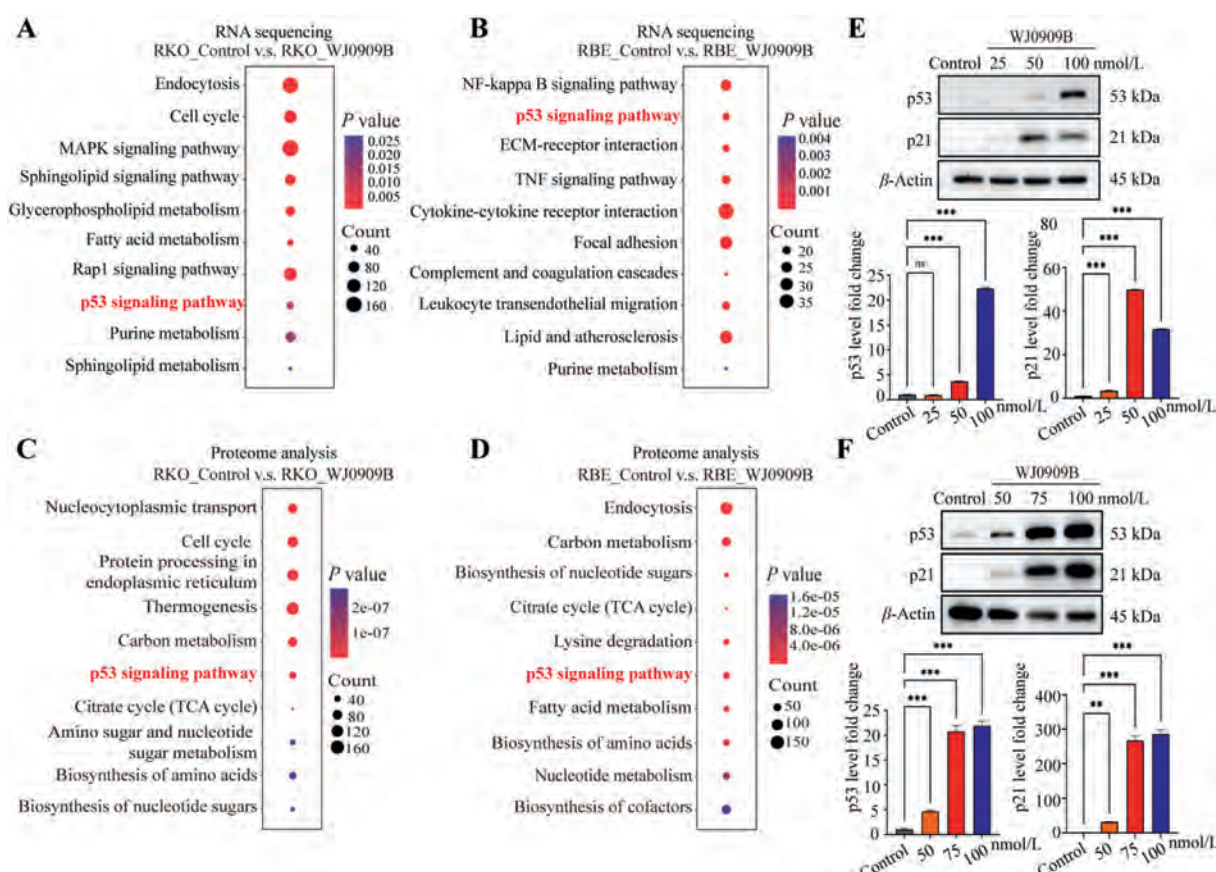


Figure 4 WJ0909B functions via p53 signaling. (A, B) KEGG enrichment analysis of differentially expressed genes (DEGs). The advanced bubble chart shows the KEGG enrichment of DEGs in signaling pathways. (C, D) KEGG enrichment analysis of differentially expressed proteins. The advanced bubble chart shows KEGG enrichment of differentially expressed proteins in signaling pathways. (E, F) Protein expression of p53 and p21 in RKO and RBE was quantified by Western blotting after treatment with varying concentrations of WJ0909B for 12 h. β -Actin was used as an internal control. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

(WT) and p53-KO HCT116 cells (Supporting Information Fig. S14A) was evaluated using two independent assays: (1) cell viability assay by CTG (Fig. S14B) and (2) Cell apoptosis assay by Annexin V-FITC/PI staining (Fig. S14C and S14D). Notably, while WJ0909B treatment induced significant apoptosis in WT HCT116 cells (33.18% increase in apoptotic cells at 50 nmol/L), this effect was markedly attenuated in p53-KO cells (only 14.08% increase in apoptosis under the same conditions). The stark contrast in apoptotic response between WT and p53-KO cells strongly supports our conclusion that WJ0909B-mediated apoptosis is predominantly dependent on functional p53 signaling pathway.

The upregulation of p53 is one of the main factors of cell apoptosis caused by mitochondrial damage^{35,36}. Gene ontology (GO) enrichment analysis of the transcriptome revealed that the differentially expressed genes were largely involved in mitochondria-related biological processes (Fig. 5A and B, Supporting Information Fig. S15A and S15B) and the intrinsic apoptotic signaling pathway (Fig. 5C–E, Fig. S15C and S15D). To assess the phenotypic effects of WJ0909B, we first used MitoTracker to examine its effects on mitochondria. Before treatment with WJ0909B, we observed that most mitochondria in RKO and RBE cells appeared as elongated tubules (Fig. 5F). By contrast, a significant increase in the content of fragmented or

intermediate mitochondria was observed in WJ0909B-treated cells. Additionally, transmission electron microscopy (TEM) was used to observe changes in mitochondrial morphology following WJ0909B treatment (Fig. 5G, Supporting Information Fig. S16). The mitochondria of RKO and RBE were readily visible owing to their characteristic cristae. However, with increasing concentration of WJ0909B, the mitochondria became smaller, with increased electron density, swelling of the mitochondrial ridges, and mitochondrial disorganization or even breakage, indicating a significant alteration in response to WJ0909B.

RKO and RBE cells were incubated with different concentrations of WJ0909B, followed by Annexin V-FITC-PI staining and flow cytometry analysis to assess the apoptotic effect of WJ0909B on tumor cells (Fig. 5H and I). The results showed that WJ0909B induced an overt apoptotic response in the cells in a dose-dependent manner. The same results were observed on the HuCCT1 and HCT116 cell lines (Supporting Information Fig. S17). These findings indicated that WJ0909B likely triggers mitochondria-dependent endogenous apoptosis via mechanisms involving the p53 signaling pathway. Compared to clinically evaluated p53 activators (e.g., nutlin-3, APR-246; IC₅₀ = 2–50 μ mol/L), WJ0909B exhibits superior potency (IC₅₀ = 0.5–30 nmol/L) and a distinctive dual mechanism involving both mitochondrial dysfunction induction and p53 pathway activation. These findings suggest WJ0909B

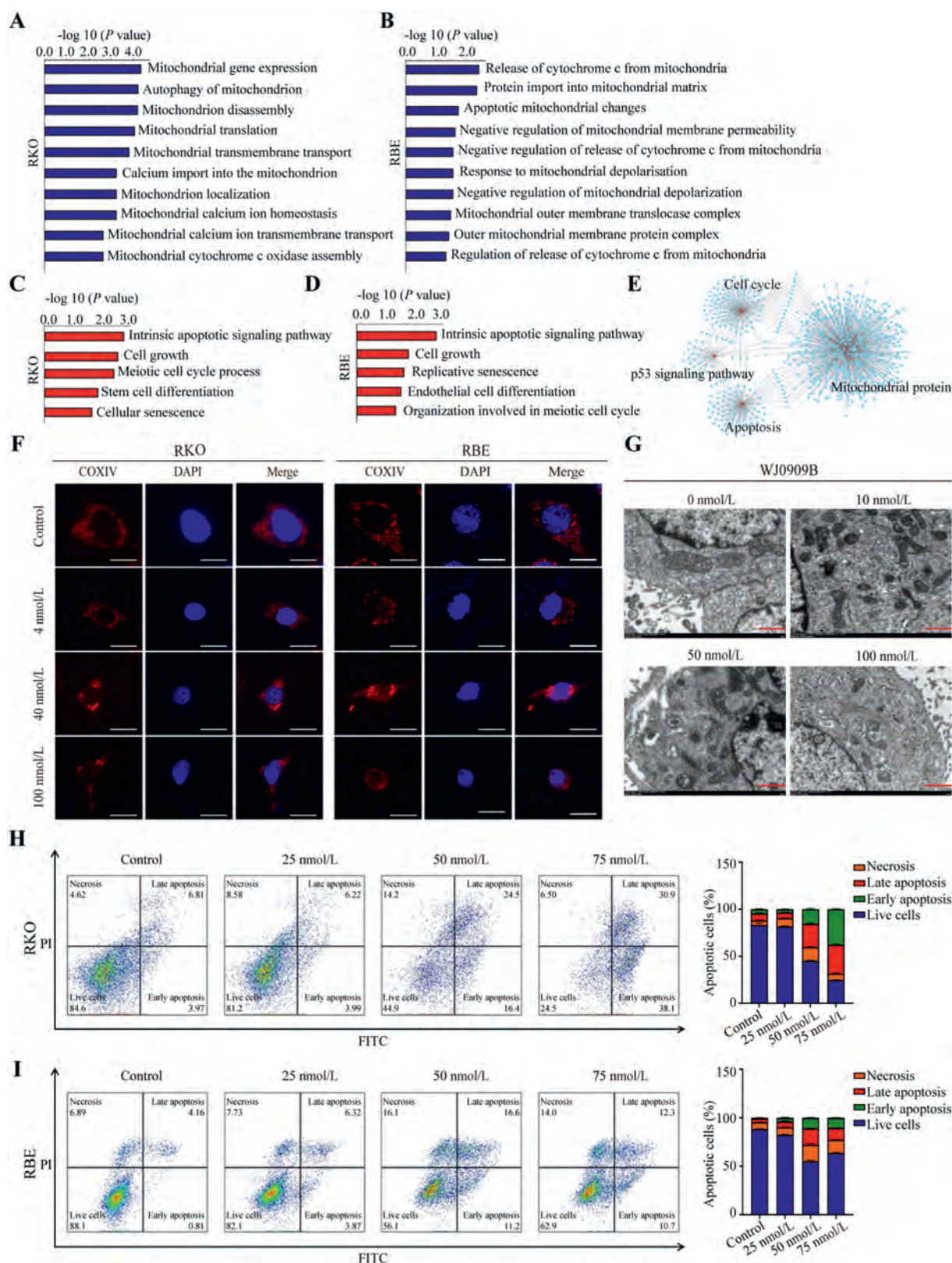


Figure 5 WJ0909B caused mitochondria-dependent endogenous apoptosis. (A–D) Enriched GO molecular function and GO biological processes of significantly upregulated and downregulated genes in RKO and RBE cells after WJ0909B treatment. (E) Integrated transcriptomic and proteomic analysis. (F) Mitochondrial mass of RKO and RBE cells visualized *via* confocal microscopy excited with COXIV, DAPI, and Merge after 48 h exposure to 0, 4, 40, and 100 nmol/L WJ0909B. Scale bar (white), 10 μ m. (G) Representative TEM images of RBE mitochondria after exposure to 0, 10, 50, and 100 nmol/L WJ0909B. Scale bar (red), 1.0 μ m. (H, I) WJ0909B induced apoptosis in RKO and RBE cell lines as visualized *via* the FCM detection of Annexin V-FITC/PI staining.

operates through a novel mechanism of action that warrants further investigation.

2.4. Pharmacokinetic characteristics of WJ0909B pharmacokinetics in BALB/c mice

To further investigate the druggability of WJ0909B, pharmacokinetic (PK) experiments were performed in BALB/c mice. When administered intravenously (i.v.) at a dose of 2 mg/kg and orally (p.o.) at a dose of 10 mg/kg, WJ0909B exhibited area under the curve (AUC_{0–t}) values of 246.45 and 436.76 h*ng/mL, respectively, indicating appropriate drug exposure. It demonstrated a half-life ($T_{1/2}$) of 11.96 h and a bioavailability of 35.4% after oral administration, affirming its status as an orally available antitumor drug (Table 1).

2.5. Potent antitumor efficacy of WJ0909B in mouse models

To comprehensively assess the *in vivo* antitumor efficacy of WJ0909B, it was administered orally (p.o.) or intravenously (i.v.) in RKO-bearing mice. A daily oral administration of 20 mg/kg WJ0909B significantly reduced tumor size in the treated groups, with 64% of the tumors regressing (Fig. 6A and B). This outcome was comparable with that caused by administering 10 mg/kg WJ0909B i.v. once every 3 days. To identify the pharmacodynamics (PD) markers, we conducted immunohistochemistry to analyze the xenograft tumors harvested from mice. The analysis revealed an increase in the number of p53-positive cells and a decrease in the number of Ki67-positive cells in the WJ0909B-treated tumors (Fig. 6C and D). These findings indicate that the *in vivo* antitumor activity of WJ0909B involves apoptosis and the p53 signaling pathway.

To further explore the therapeutic potential of WJ0909B on other types of tumors, NOD/SCID mice bearing H446 lung cancer were treated with vehicle, 5, 7.5, or 10 mg/kg WJ0909B by intravenous injection once every three days. Similarly, we observed that WJ0909B demonstrated dose-dependent antitumor activity and significantly inhibited the growth of H446 tumors when treated with 10 mg/kg WJ0909B, achieving an 85% tumor growth inhibition rate (Fig. 6E and F). These results highlight the potential antitumor activity of WJ0909B in preclinical applications.

2.6. WJ0909B-TCO combined with cRGDyk-Tz inhibited tumor growth with minimal toxicity

During the treatment, a significant loss in body weight was observed in all the WJ0909B-treated groups (Fig. 6B and F). We conducted histopathology analysis to evaluate the systemic toxicity of WJ0909B in RKO tumor-bearing mice. After 15 days of oral administration of 20 mg/kg/day WJ0909B, the major organs of the mice were stained with H&E. Supporting Information Fig. S18 and Fig. 6D shows that, apart from nuclear shrinkage in liver cells, no significant differences were observed in the other organs, suggesting that while WJ0909B effectively inhibited tumor growth, it also caused liver damage.

We further evaluated the effects of WJ0909B on peripheral blood mononuclear cells. Quantitative analysis revealed no significant selectivity for tumor cells and normal cells, as evidenced by nearly identical IC₅₀ values (2.99 nmol/L) (Supporting Information Fig. S19). This observation aligns with the design principles of our deep learning model, which prioritized broad cytotoxic phenotypes without mechanistic filters for tumor-specific pathways. Although this pan-cytotoxicity may initially seem like a disadvantage, the clinical success of targeted delivery systems indicates that tumor selectivity can be designed through spatial control rather than inherent molecular characteristics and we decided to explore a prodrug strategy to enable tumor-selective drug delivery and enhance therapeutic safety.

In our investigation of tetrahydrocarbazolamide-containing scaffold compounds, we found that acetylating the 3-amino group of WJ0976 (WJ1910) led to a significant loss of activity (Supporting Information Fig. S20A and S20B). This suggests that the amino group plays an essential role in antitumor activity. Shasqi's Click-Activated Prodrugs Against Cancer platform has been successfully used to activate cytotoxic drugs at the tumor site. For example, a click chemistry reaction between TCO and tetrazine (Tz) groups has been successfully applied in DOX^{37,38}. The prodrug's design leverages the biorthogonal TCO-Tz reaction, which has demonstrated safety and efficacy in humans, and is currently being evaluated in a phase 1/2a clinical trial^{39,40}. Unlike antibody-drug conjugates (ADCs), which are limited by antigen heterogeneity, the click-activated strategy is modular and adaptable to diverse tumor types. The use of cRGDyk (targeting $\alpha v \beta 3$ integrins) ensures applicability to aggressive cancers (e.g., glioblastoma, pancreatic cancer) while enabling local activation in

Table 1 Key pharmacokinetic characteristics of WJ0909B obtained in a preliminary pharmacokinetic assessment^a.

Parameters	WJ0909B	
	p.o. (10 mg/kg)	i.v. (2 mg/kg)
$T_{1/2}$ (h)	6.95 ± 1.88	11.96 ± 4.52
T_{max} (h)	2.67 ± 1.15	0.033
C_{max} (ng/mL)	98.07 ± 48.99	652.82 ± 60.10
AUC _(0–t) (h*ng/mL)	436.76 ± 141.81	246.45 ± 35.12
AUC _(0–∞) (h*ng/mL)	478.86 ± 143.63	284.60 ± 55.77
MRT _{0–t} (h)	6.52 ± 1.16	4.15 ± 0.34
MRT _(0–∞) (h)	8.89 ± 2.67	9.05 ± 3.12
V (mL/kg)	/	119,576.07 ± 35,919.20
CL (mL/H/kg)	/	7216.37 ± 1454.61
F (%)	35.4	/

^aData represent mean ± standard deviation; $n = 3$.

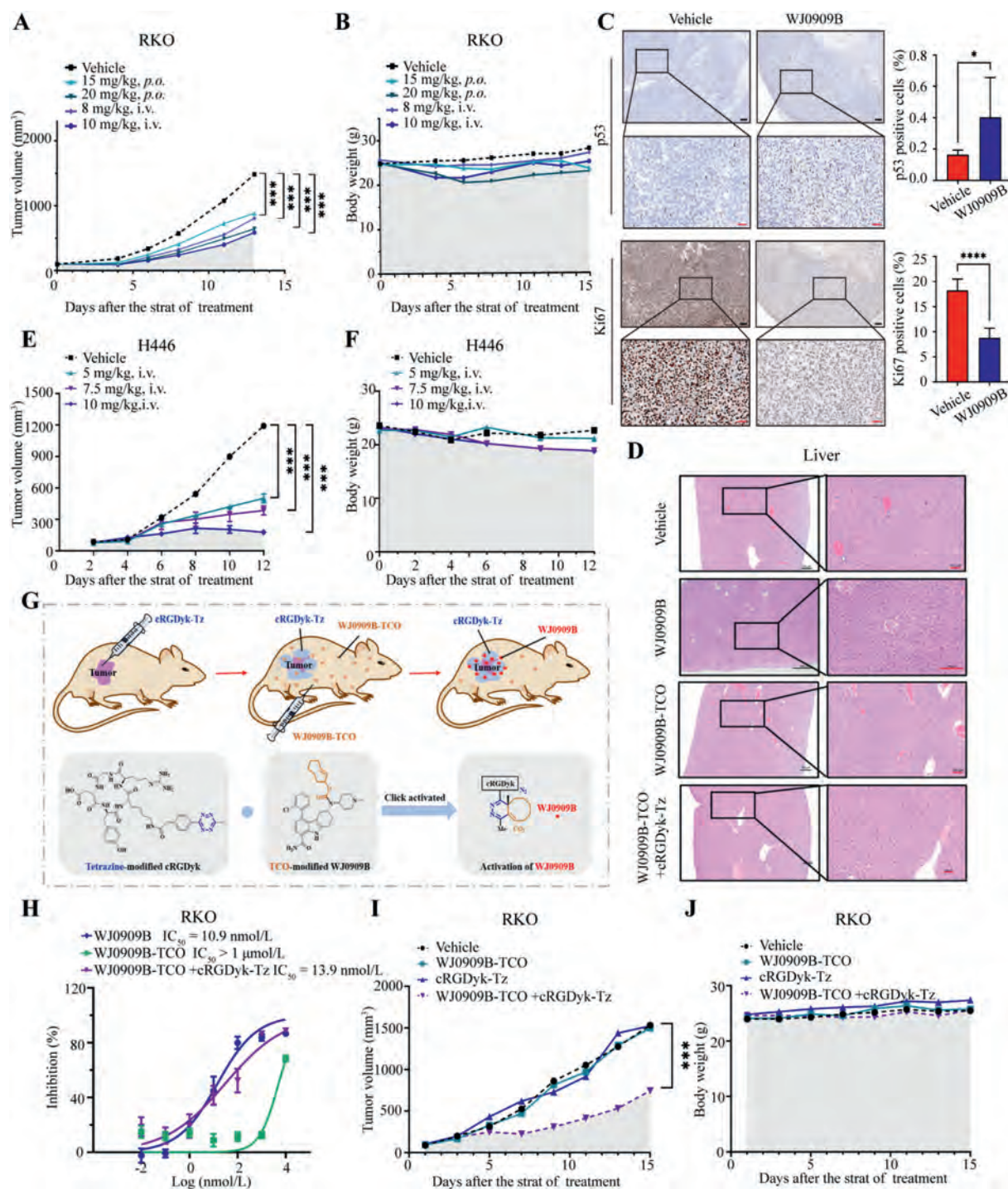


Figure 6 *In vivo* antitumor efficacy of WJ0909B and the click-activated strategy for targeted cancer therapy. (A, B) BALB/c nude mice bearing xenograft tumors of RKO colorectal cancer were treated with vehicle, two oral doses (15 and 20 mg/kg) of WJ0909B, or two intravenous doses (8 and 10 mg/kg) of WJ0909B. Tumor volume (A) and body weight (B) were measured. (C) The expression of Ki67 and p53 in RKO-derived tumor samples was analyzed by H&E. Representative images of tumors treated with vehicle or 20 mg/kg oral WJ0909B were shown. Scale bar (black), 100 μ m. Scale bar (red), 50 μ m. (D) Evaluation of systemic toxicities to the liver. H&E staining of major organs from BALB/c nude mice after treatment with vehicle, WJ0909B, WJ0909B-TCO or WJ0909B-TCO combined with cRGDyK-Tz. Scale bar (black) 500 μ m, (red) 100 μ m. (E and F) NOD/SCID mice bearing xenograft tumors of H446 lung cancer were treated with vehicle, 5, 7.5, or 10 mg/kg WJ0909B by intravenous injection. Tumor volume (E) and body weight (F) were measured. (G) Designs for click-activated prodrug WJ0909B-TCO combined with cRGDyK-Tz to achieve high tumor specificity. cRGDyK-Tz biopolymer is locally injected at the tumor site. The WJ0909B-TCO prodrug is infused systemically and captured by cRGDyK-Tz at the desired site *via* rapid click chemistry between Tz and TCO, followed by precisely controlled activation of WJ0909B at the tumor site. (H) IC_{50} values of WJ0909B, WJ0909B-TCO, and WJ0909B-TCO combined with cRGDyK-Tz in RKO were determined *in vitro*. (I, J) BALB/c nude mice were implanted with RKO cells in the left flank, then treated with vehicle, cRGDyK-Tz,

non-injected lesions. Based on the same retrofit strategy, we designed a combination of TCO-modified WJ0909B (WJ0909B-TCO) and a Tz-modified cRGDyk (cRGDyk-Tz) to achieve the local capture and activation of WJ0909B at the tumor site (Fig. 6G). The feasibility of this approach was verified by the observed IC_{50} values of WJ0909B, WJ0909B-TCO, and WJ0909B-TCO combined with cRGDyk-Tz in RKO cells. As expected, WJ0909B-TCO exhibited an IC_{50} value of 2.52 $\mu\text{mol/L}$, which was 68-fold less than that of WJ0909B (Fig. 6H), thus demonstrating the suitability of the TCO-modified prodrug strategy and a substantial improvement in its safety. Additionally, the combination of WJ0909B-TCO and cRGDyk-Tz inhibited cell proliferation in RKO cells to a similar extent as WJ0909B treatment alone ($IC_{50} = 0.027 \mu\text{mol/L}$), demonstrating the successful and complete activation of WJ0909B *via* the rapid click chemistry reaction between Tz and TCO moieties. In conclusion, WJ0909B-TCO can significantly reduce toxicity and, in combination with cRGDyk-Tz, can be successfully and completely activated to WJ0909B in tumor cells (Fig. S20C–S20E).

After completing the initial pharmacokinetic evaluation of WJ0909B-TCO (Supporting Information Table S9), an *in vivo* pharmacodynamic study involving RKO tumor-bearing mice was performed to evaluate whether the combination of WJ0909B-TCO and cRGDyk-Tz could maintain high antitumor activity with minimal toxicity. The mice were treated with vehicle, cRGDyk-Tz, WJ0909B-TCO, or a combination of WJ0909B-TCO with cRGDyk-Tz. In the combination group, mice were first peritumorally injected with cRGDyk-Tz, followed by an intraperitoneal injection of WJ0909B-TCO (10 mg/kg/dose WJ0909B eq.). As depicted in Fig. 6I and J, the combination treatment significantly slowed tumor growth compared with the other groups, with a tumor regression rate of 58% and no significant weight loss. To evaluate the stability of WJ0909B and click-activated strategy for target cancer therapy. Plasma/tumor pharmacokinetics of prodrug WJ0909B-TCO and its activated payload WJ0909B post-cRGDyk-Tz administration were tracked (Table S10). The results revealed that WJ0909B-TCO exhibited >90% integrity in plasma with minimal premature release. WJ0909B can be significantly detected in tumors of mice treated with WJ0909B-TCO combined with cRGDyk-Tz. The concentration of WJ0909B dispersed in plasma is relatively low, underscoring the prodrug WJ0909B-TCO has good stability and this prodrug strategy has good tumor specificity. Further histopathology analysis indicated that there was no significant damage to the liver in combination treatment groups (Fig. 6D, Supporting Information Fig. S21). These preliminary results confirm that the click-activated prodrug WJ0909B-TCO and cRGDyk-Tz could inhibit tumor growth with minimal toxicity.

3. Conclusions

Phenotypic screening plays a crucial role in the discovery of first-in-class small-molecule drugs. It is considered to have significant advantages in discovering drug leads and clinical candidates with novel scaffolds and new mechanisms of action. However, the screening of the vast number of compounds is extremely time-

consuming and expensive, which raises the critical question of how to efficiently conduct the phenotypic screening approach to discover new drugs.

Given the development of artificial intelligence, deep learning methods have become important tools for drug discovery. In this paper, a data-driven model has been built by integrating deep learning-driven classifiers and GDL models for the discovery of desirable antitumor drugs. Through wet-lab validation, WJ0976 and WJ0909, which were different from known tetrahydrocarbazole derivatives, were identified as potent antitumor drugs with broad-spectrum activity, exerted the half-maximal inhibitory concentration (IC_{50}) values of 0.2–45 nmol/L against various cancer cell lines. To our surprise, they exhibited growth inhibitory properties against multi-drug resistant tumor cells. Meanwhile, the *R*-(–)-WJ0909 (WJ0909B) exhibited optimal efficacy *in vitro* and showed significant therapeutic effect in *ex vivo* patient-derived cancer organoids (PDOs). Subsequent mechanisms investigation revealed that WJ0909B induced mitochondrial dysfunction and endogenous apoptosis by involving the p53 signaling pathway. Notably, WJ0909B demonstrated remarkable therapeutic properties in xenograft models, and the click-activated prodrug WJ0909B-TCO exhibited high tumor specificity with minimal toxicity.

4. Experimental

4.1. Data preparation

Molecules with antiproliferative activities were collected from databases and literature for the construction of training dataset. 1114 molecules were collected from Cancer Drug Resistance (CancerDR) Database, L5500 Database (a commercial compound library from TargetMol) and self-collected literatures. Using IC_{50} value of 1 $\mu\text{mol/L}$ as a hit cut-off, 1112 molecules were divided into 399 active molecules and 713 inactive molecules. After binarization, these molecules were used to train a binary classification model that predicts the probability of whether a compound may exhibit broad-spectrum antitumor activity. Datasets were randomly split into training and test sets with a ratio of 4:1 for the training of fingerprint-based molecular classifier, and all active molecules were selected for the building of the training set of GDL model.

The molecular formulas and molecular properties including molecular weight (MW), lipid-water distribution coefficient (LogP), drug-likeness (QED) and synthetic accessibility (SA) are calculated by RDKit (version 2021.03.1). RECAP and BRICS methods were used to cut the entire molecules into fragments.

4.2. Data representation in fingerprint-based molecular classifier

Molecules were represented by extended-connectivity fingerprint (ECFP), which is a vector with a fixed length, which initially uses unique identifiers to demonstrate structures around all heavy atoms of a molecule with a defined radius. ECFP4 was employed

WJ0909B-TCO, or WJ0909B with cRGDyk-Tz. The combined treatment group started when tumor volumes reached 100 mm³ with a daily peritumoral injection of 42 mg/kg cRGDyk-Tz, followed by a daily IP injection of 15 mg/kg WJ0909B-TCO. Tumor volume (I) and body weight (J) were measured. **P* < 0.05, ***P* < 0.01, ****P* < 0.001.

for its advantages in capturing molecular structural information and providing robust feature representations for predictive tasks of classification models. The appended number is the effective diameter of the largest feature and is equal to twice the number of iterations performed. Finally, 300-dimensional embeddings were generated for all compounds. ECFP4 was calculated using the RDKit (version 2021.03.1).

4.3. Model architecture of fingerprint-based molecular classifier

The model based on eXtreme Gradient Boosting, which is a scalable end-to-end tree boosting system, was default cited from our previous work and was available at GitHub (https://github.com/WJmodels/Fingerprint-based_Molecular_Classifier). This model is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. It was built on previous ideas in gradient boosting, which builds a sequential series of smaller trees where each tree corrects for the residuals in the predictions made by all the previous trees. The model is designed to perform a binary classification task, where each decision tree predicts based on the input ECFP4 fingerprint. The raw score is obtained by weighted summation of the leaf node weights from all trees. A sigmoid function is then applied to the raw score to calculate the probability of the molecule being positive. The loss function of the classifier can be represented by the predicted values and the true values are shown as Eq. (1):

$$L = \sum_{i=1}^n l(y_i, \hat{y}_i) \quad (1)$$

Here, n represents the number of samples, y_i represents actual value, and $\hat{y}_i$ represents predicted value.

The prediction accuracy of model is determined by both the bias and variance of the model. The loss function represents the model's bias. To reduce variance and prevent overfitting, a regularization term Ω should be added to the objective function, the metrics were defined using Eq. (2):

$$\text{Obj} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{i=1}^t \Omega(f_i) \quad (2)$$

The regularization term is the sum of the complexities of all t trees.

In eXtreme Gradient Boosting, each tree is added one by one. Thus, the Obj can be rewritten as Eq. (3):

$$\text{Obj} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_i(x_i)) + \Omega(f_i) + \text{Constant} \quad (3)$$

Here, $\hat{y}_i^{(t-1)}$ is the prediction value provided by the $(t-1)$ model, which is a known constant, while $f_i(x_i)$ is the prediction value of the new model to be added this time.

This objective function was then optimized by using the gradient descent method. Herein, we introduce some parameters we used to build classification models: eta (step size shrinkage) was set to 0.3; the max depth (maximum depth of a tree) was set to 4; the colsample_bytree (what fraction of descriptors would be examined for each tree) was set to 1; and the colsample_bylevel was set to 1. Gamma was set to 1; the objective was set to "binary: logistic". Given that the classifier is tasked with determining whether a molecule exhibits broad-spectrum antitumor activity, the model is designed to apply more stringent classification standard in order to minimize false positives. To this end, the self-collected class-imbalanced training set was employed. The

fingerprint-based molecular classifier was trained for 300 epochs, at which point the model achieved optimal performance on the test set before decline in accuracy was observed.

Random forest (RF) is an ensemble learning method that constructs multiple decision trees using a bagging approach. It employs bootstrap sampling to generate a random subset of the training data, and at each split, a random selection of features is used to build individual trees. This randomization helps mitigate overfitting and reduces the correlation between the trees, enhancing the model's generalization ability. For classification tasks, the final prediction is made through a majority voting mechanism, where the predicted class is determined by the most frequent class among the individual trees' predictions. In our methods, key parameters were systematically selected *via* Bayesian optimization using the Tree-structured Parzen Estimator (TPE) algorithm. The hyperparameter search space included: maximum tree depth: [5, 20], number of estimators: [50, 100], maximum features per split: [5, 30].

Support Vector Machine (SVM) is a widely recognized supervised machine learning technique, particularly effective in quantitative structure-activity relationship (QSAR) modeling. SVM excels in handling challenges such as small sample sizes, non-linearity, and high-dimensional spaces. The SVM classifier was optimized through a Bayesian hyperparameter search framework to enhance classification performance. Hyperparameters were systematically explored using the TPE algorithm *via* the Hyperopt library, with the following search space defined: Regularization parameter: Sampled from a logarithmic uniform distribution [0.001, 1000] to balance margin maximization and training error minimization. Kernel function: evaluated among radial basis function, polynomial, and sigmoid kernels to capture nonlinear decision boundaries. Kernel coefficient: optimized *via* a hybrid strategy, including automatic scaling ("auto") and continuous sampling [0.0001, 8], adapting to feature space complexity. A total of 50 optimization iterations were executed to minimize the negative mean accuracy, calculated as the average of training and test set accuracies.

4.4. Explainability of fingerprint-based molecular classifier

Local Interpretable Model-Agnostic Explanations (LIME) offers a transparent way to understand the decision-making process of a model by approximating the complex black-box model with a simple, interpretable surrogate model around the specific instance of interest. LIME constructs a linear surrogate model that mimics the behavior of the original model in the local neighborhood of the instance to be explained. This surrogate model is then utilized to derive feature importance scores, providing local explanations for the model's predictions. LIME was chosen for the explainability of the fingerprint-based molecular classifier, with the most important fragments in the model's scoring being identified based on molecular fingerprints. Among these, three most significant molecular fingerprints were selected for visualization.

4.5. Data representation in GDL model

Chemical structures and fragments were represented as SMILES strings by multiple SMILES strings rooted at different atoms in each molecule. These strings were tokenized through NLP strings, such as 'C', 'N', and '('. Further, every SMILES string was expanded with the leading token <SMILES> and lagging token </SMILES>. Similarly, the leading and lagging tokens for

fragment SMILES string expansion were $\langle \text{fragment} \rangle$ and $\langle / \text{fragment} \rangle$, respectively. Property ranges are numerical lists of property values. MW ranged within 0–1000 without decimal digits, QED within 0–1 with two decimal digits, $\log P$ within -4 to 7 with one decimal digit and SA within 0–10 with one decimal digit.

4.6. Model architecture of the BART-based GDL model

The model was default cited from our previous work and was available at GitHub (<https://github.com/WJmodels/CMGN>). The BART backbone in this work was composed of 6 encoder layers with 12 attention heads, 6 decoder layers with 12 attention heads and 768-dimensional hidden units.

The multi-head attention mechanism allows the encoder and decoder to focus on different tokens simultaneously, thus BART can successfully process long-range strings, and shows brilliant performance when generating SMILES strings. A multi-head attention unit is concatenated and projected into the final values. The attention layer takes three matrices as input: the matrix Q packed with a set of queries, the matrix K with keys, and the matrix V with values. The attention is computed as Eq. (4):

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

The goal of training for a specific source sequence is to reduce the difference between the predicted sequence and its target sequence, which is calculated using the cross-entropy loss function as Eq. (5):

$$\text{Loss} = - \sum_{i=1}^k y_i \log(x_i) \quad (5)$$

where k is the token number of the target sequence, and y_i and x_i are the actual values and predicted values at the corresponding position of the target sequence, respectively.

Models were trained using an AdamW optimizer with a batch size of 32. The learning rate is set to 5×10^{-5} , and weight decay is used for regularization, with a value of 1×10^{-4} . The training accuracy was computed using the bf16 precision format. The model was trained on a server equipped with two Intel Xeon Gold 5320 CPU 32-core 2.20 GHz processors using four NVIDIA A100 GPUs. When training the GDL model, we perform fine-tuning on a previously reported model CMGN. In previous work, a dataset consisting of fifty million molecules in SMILES format was curated from the ZINC database and used for pretraining over two epochs. Then the fine-tuning was performed on the pretraining model based on 399 self-collected antitumor molecules for 1000 epochs.

4.7. Evaluation metrics

Metrics were applied to evaluate the ability of the GDL model to generate molecules, as shown in Eqs. (6)–(9).

Validity: fraction of validly generated molecules. RDKit was used for validity checks.

$$\text{Valid} = \text{Chemically valid SMILES} / \text{Generated smiles} \quad (6)$$

Uniqueness: proportion of unique structures generated.

$$\text{Unique} = (\text{Non-duplicate; valid structures}) / \text{Valid structures} \quad (7)$$

Novelty: proportion of generated molecules not in the training set.

$$\text{Novelty} = \text{molecules not in training set} / \text{unique structures} \quad (8)$$

Internal diversity: measure of diversity of the generated molecules calculated as the average Tc in the set of generated molecules.

$$\text{Internal diversity} = 1 - 1/|\text{set}(G)|^2 \sum_{(a,b) \in \text{set}(G)} \text{TC}(m_a, m_b) \quad (9)$$

4.8. Chemical synthesis

The primary synthetic data are available in the Supporting Information

4.9. In vitro antitumor activity analysis

Cells were distributed into 96-well white round-bottom plates, with each well receiving 1000 cells in RPMI-1640 medium supplemented with 10% FBs. Following a 24-h incubation period, diluted compounds were introduced to the wells. After 96 h, a Cell Titer-Glo luminescent cell viability assay (Promega, Madison, WI, USA) was performed to assess cell viability. The luminescent readings were normalized as percentages relative to untreated cells, and the IC₅₀ values for each compound were determined.

4.10. Culture of human intestinal organoids

Patient consent and approval were obtained from the Institutional Research Ethics Committee for research purposes (NCC2019C-016). Advanced Dulbecco's modified Eagle's medium/F12 was supplemented with penicillin/streptomycin, 10 mmol/L HEPES, 2 mmol/L GlutaMAX, $1 \times \text{B27}$ (Life Technologies), 10 nmol/L gastrin I (Sigma) and 1 mmol/L N-acetylcysteine (Sigma). The following niche factors were used: 50 ng/mL mouse recombinant EGF, 100 ng/mL mouse recombinant noggin (Peprotech), 10% R-spondin-1 conditioned medium (kindly provided by C. Kuo, Stanford University) 34, 50% Wnt-3A conditioned medium 15, 500 nmol/L A83-01 (Tocris) and 10 $\mu\text{mol/L}$ SB202190 (Sigma). To select mutant organoids, the following reagents were used: 10 ng/mL human recombinant TGF- β (R&D), 100 ng/mL human recombinant BMP4 (Peprotech), 1 $\mu\text{mol/L}$ EGFR inhibitor (EMD Millipore), 50 nmol/L PD325901 (MEK inhibitor) (Sigma) or ($\pm$)-nutlin-3 (Cayman Chemical). For the organoid formation assay and before genome editing, organoids were dissociated into single cells with TrypLE Express (Life Technologies) and 1000 cells were cultured in a 48-well plate under the above culture conditions for 10 days. To prevent anoikis, 10 $\mu\text{mol/L}$ Y-27632 was included in the culture medium for the first 2 days (ref. 3). The growth of organoids was estimated by area of intact organoids using brightfield microscopy images.

4.11. Immunofluorescence

RKO and RBE cells were grown in a 4-well Polystyrene chamber piece (Biologix Group Limited, 07-2104) overnight, and then treated with a series of concentrations (0, 4, 40, 100 nmol/L) of WJ0909B for 48 h. Cells were then fixed with 4% poly-formaldehyde for 10 min at room temperature (RT), followed by

permeabilized with 0.1% Triton X-100 (in PBS) for 10 min. After permeabilization, cells were blocked with 5% BSA in PBS for 1 h at RT. Cells were further incubated with antibody against COX IV (CST, 4850) at a ratio of 1:100 overnight at 4 °C. After washing with PBS, cells were incubated with Alexa Fluor™ Plus 594 anti-rabbit antibody (Invitrogen, A32740) at a ratio of 1:200 for 1 h at RT. The cells were later counterstained with Fluoroshield Mounting Medium with DAPI (Abcam, ab104139). Images were obtained with fluorescence microscope GE Deltavision OMX SR. Images were quantitatively assessed using ImageJ software with seven visual fields.

4.12. Electron microscopy

RKO and RBE cells were inoculated into 10 cm dishes overnight, and then treated with a series of concentrations of WJ0909B (RKO cells [0, 2, 40, 100 nmol/L] and RBE cells [0, 10, 50, 100 nmol/L] for 48 h. The collected cells were fixed with Glutaraldehyde Fixative (2.5%, for electron microscopy) (Bio-sharp, BL911A) at 4 °C for another 48 h. For the electron microscopy in WJ0909B treated RKO and RBE tumor cells, cells were fixed in 0.1 mol/L phosphate buffered PB (PH = 7.4), containing 1% osmium acid (Ted Pella Inc), at RT for 2 h. After fixation, the samples were stained with 812 aqueous overnight at 37 °C. Sample blocks were cut from the dishes with a diamond knife using an ultramicrotome (Leica UC7), and the sections were placed on bare copper TEM grids. Ultrathin sections were stained with 2.6% lead citrate solution and imaged using a MegaView 3 digital camera and iTEM software with a HITACHI electron microscope.

4.13. Western blotting

For all Western blots, the cell lysate was prepared in 1 × SDS loading buffer. Total protein was applied to SDS-PAGE and transferred to polyvinylidene fluoride (Merck Millipore). The membranes were blocked in 5% nonfat milk or 5% BSA for 1 h at room temperature before incubated with a primary antibody overnight at 4 °C. Membranes were incubated with a primary antibody against p53 (Protein Tech, PTM-5084), p21 (CST, 2947), β -actin (ZSGB-BIO, TA-09). Horseradish peroxidase (HRP)-linked anti-rabbit IgG (ZSGB-BIO, ZB-2301) or HRP-linked anti-mouse IgG (ZSGB-BIO, ZB-2305) was used as a secondary antibody, corresponding to different proteins. Protein bands were detected using a super sensitive ECL luminescence reagent (LABLEAD, E1070).

4.14. Apoptosis assay

An Annexin V/PI (Multi Sciences, Hangzhou, China) assay was used to identify apoptotic and necrotic cells following incubation with WJ0909B (0–100 nmol/L) for 24 h. Following the manufacturer's instructions, live cells were harvested and resuspended in 1 × Binding Buffer. Then, stain (PI and FITC-conjugated annexin V in the amount specified by the instructions) it for 15 min in the dark at room temperature, before analysis using a flow cytometer (Guava easy Cyte 12HT, LUMINEX, Texas, USA). In the resulting dot plots, the X-axis represents the intensity of green fluorescence (Annexin V-FITC) and the Y-axis represents the intensity of red fluorescence (PI).

4.15. Assessment of pharmacokinetic properties

The PK properties of compounds were examined in male BALB/c mice ($n = 3$ per group, weight: 20–25 g). Compounds were dissolved in saline with 10% (v/v) DMSO. The animals were administered with a single dose of 2 mg/kg intravenous injection (i.v.) and 10 mg/kg oral gavage (p.o.). Blood samples were collected at 2, 5, 15, and 30 min, 1, 2, 3, 4, 6, 8, 12 and 24 h, and centrifuged to isolate plasma. Subsequently, the plasma compound concentrations were determined by LC–MS (HPLC–QE Orbitrap MS), and the PK parameters were calculated using Phoenix WinNonlin 7.0.

4.16. Animal studies

The protocol for animal experiments was approved by the Ethics Committee for Animal Experiments of the Institute of Materia Medica, Chinese Academy of Medical Sciences & Peking Union Medical College. BALB/c and NOD/SCID mice were purchased from Vital River (Beijing Vital River Laboratory Animal Technology Co., Ltd.). All mice were bred under standard conditions and used at the age of 6–8 weeks when the body weight was ~25 g. WJ0909B or vehicle was dosed by i.v. or p.o.. Tumor and body weight measurements were performed three times per week using calipers and a weighing scale, respectively. Mice were euthanized when tumor volume approached ~2000 mm³ or tumors ulcerated. When necessary, tumor samples were collected at specific time points and frozen for Pharmacodynamics and toxicology studies. Ki67 and p53 were measured by HE.

4.17. Statistical analysis

Data on Figures represent mean $\pm$ standard deviation (SD). Unless otherwise noted, the differences between two groups were analyzed by unpaired Student's *t*-test, and differences with *P* value < 0.05 were considered significant.

Acknowledgments

This work was supported by CAMS Innovation Fund for Medical Sciences, China; (No. 2021-I2M-1-028 and No. 2021-I2M-1-054, China), the National Natural Science Foundation of China, China; (No. 82303782, China), the China Postdoctoral Science Foundation, China; (2024M763807, China) and the 2024 China Industrial Technology Infrastructure Public Service Platform Project, China; (GN2024-31-4700). The computing resources were supported by Biomedical High Performance Computing Platform, Chinese Academy of Medical Sciences, China. We would like to sincerely thank the IMM Compound Library for the assistance with screening.

Author contributions

Xiaojuan Wang, Tingting Du, Wei Song and Hong Zhao designed this project. Xue Liu and Yalan Lu assisted in designing this project. Xue Liu, Xiangying Liu and Jingjie Yan performed the chemical synthesis. Hanyu Sun, Minjian Yang, Liangning Li and Ahmed Al-Harrasi assisted in building deep learning models. Qichen chen, Yalan Lu, Shize Li, Yiqiao Deng, Yan Lu and Nan Xiang performed *in vitro* and *in vivo* assays. Xiandao Pan and Jing Jin were involved in the experimental design. Qi Geng analyzed

for the proteomics and transcriptome sequencing data. Baolian Wang performed *in vivo* pharmacokinetic experiments. Xue Liu, Xiangying Liu, Hanyu Sun and Xiaojian Wang contributed to the writing, review and editing of the manuscript. All authors have given approval to the final version of the manuscript.

Data availability

Crystallographic data of compounds WJ0976B and WJ0909B have been deposited to the Cambridge Crystallographic Data Centre (CCDC) with depository numbers CCDC-2363192 (WJ0909B) and CCDC-2363193 (WJ0976B). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium with the dataset identifier PXD042710. RNA sequencing data is available at the NCBI Sequence Read Archive under accession PRJNA1126411. All other data are available upon reasonable request.

Code availability

The codes for classification and generating models are available from the GitHub repository: https://github.com/WJmodels/Fingerprint-based_Molecular_Classifier and <https://github.com/WJmodels/CMGN>.

Conflicts of interest

The authors declare no competing financial interest.

Appendix A. Supporting information

Supporting information to this article can be found online at <https://doi.org/10.1016/j.apsb.2025.10.005>.

References

- Costello JC, Stolovitzky G. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. *Clin Pharmacol Ther* 2013;**93**:396–8.
- Swinney DC, Anthony J. How were new medicines discovered?. *Nat Rev Drug Discov* 2011;**10**:507–19.
- Sadri A. Is target-based drug discovery efficient? Discovery and “off-target” mechanisms of all drugs. *J Med Chem* 2023;**66**:12651–77.
- Moffat JG, Vincent F, Lee JA, Eder J, Prunotto M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat Rev Drug Discov* 2017;**16**:531–43.
- Qin T, Gao XF, Lei L, Feng J, Zhang WX, Hu YH, et al. Machine learning- and structure-based discovery of a novel chemotype as FXR agonists for potential treatment of nonalcoholic fatty liver disease. *Eur J Med Chem* 2023;**252**:115307.
- Brown DG, May Dracka TL, Gagnon MM, Tommasi R. Trends and exceptions of physical properties on antibacterial activity for gram-positive and gram-negative pathogens. *J Med Chem* 2014;**57**:10144–61.
- Vincent F, Nueda A, Lee J, Schenone M, Prunotto M, Mercola M. Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nat Rev Drug Discov* 2022;**21**:899–914.
- Issa NT, Stathias V, Schürer S, Dakshanamurthy S. Machine and deep learning approaches for cancer drug repurposing. *Semin Cancer Biol* 2021;**68**:132–42.
- Ahmad S, Xu J, Feng JA, Hutchinson A, Zeng H, Ghiabi P, et al. Discovery of a first-in-class small-molecule ligand for WDR91 using DNA-encoded chemical library selection followed by machine learning. *J Med Chem* 2023;**66**:16051–61.
- Smer-Barreto V, Quintanilla A, Elliott RJR, Dawson JC, Sun J, Campa VM, et al. Discovery of senolytics using machine learning. *Nat Commun* 2023;**14**:3445.
- Wang P, Ho D. Deep learning and drug discovery for healthy aging. *ACS Cent Sci* 2023;**9**:1860–3.
- Liu G, Catacutan DB, Rathod K, Swanson K, Jin W, Mohammed JC, et al. Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*. *Nat Chem Biol* 2023;**19**:1342–50.
- Stokes JM, Yang K, Swanson K, Jin W, Cubillos Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. *Cell* 2020;**180**:688–702.
- Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, Omori S, et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature* 2023;**626**:177–85.
- Keshavarzi Arshadi A, Salem M, Collins J, Yuan JS, Chakrabarti D. DeepMalaria: artificial intelligence driven discovery of potent antiparasmodials. *Front Pharmacol* 2019;**10**:1526.
- Powers AS, Yu HH, Suriana P, Koodli RV, Lu T, Paggi JM, et al. Geometric deep learning for structure-based ligand design. *ACS Cent Sci* 2023;**9**:2257–67.
- Li YS, Zhang LT, Wang YF, Zou J, Yang RC, Luo XL, et al. Generative deep learning enables the discovery of a potent and selective RIPK1 inhibitor. *Nat Commun* 2022;**13**:6891.
- Diao YY, Liu DD, Ge H, Zhang RR, Jiang KX, Bao RH, et al. Macrocyclization of linear molecules by deep learning to facilitate macrocyclic drug candidates discovery. *Nat Commun* 2023;**14**:4552.
- Yang MJ, Sun HY, Liu X, Xue X, Deng YF, Wang XJ. CMGN: a conditional molecular generation net to design target-specific molecules with desired properties. *Briefings Bioinf* 2023;**24**:bbad185.
- Boldini D, Friedrich L, Kuhn D, Sieber SA. Machine learning assisted hit prioritization for high throughput screening in drug discovery. *ACS Cent Sci* 2024;**10**:823–32.
- Cox TR. The matrix in cancer. *Nat Rev Cancer* 2021;**21**:217–38.
- Hoelder S, Clarke PA, Workman P. Discovery of small molecule cancer drugs: successes, challenges and opportunities. *Mol Oncol* 2012;**6**:155–76.
- Moffat JG, Rudolph J, Bailey D. Phenotypic screening in cancer drug discovery — past, present and future. *Nat Rev Drug Discov* 2014;**13**:588–602.
- Agnello S, Brand M, Chellat MF, Gazzola S, Riedl RA. Structural view on medicinal chemistry strategies against drug resistance. *Angew Chem Int Ed Engl* 2019;**58**:3300–45.
- Cabanos HF, Hata AN. Emerging insights into targeted therapy-tolerant persister cells in cancer. *Cancers* 2021;**13**:2666.
- Scott EC, Baines AC, Gong Y, Moore R, Pamuk GE, Saber H, et al. Trends in the approval of cancer therapies by the FDA in the twenty-first century. *Nat Rev Drug Discov* 2023;**22**:625–40.
- Vasan N, Baselga J, Hyman DM. A view on drug resistance in cancer. *Nature* 2019;**575**:299–309.
- Kumar R, Chaudhary K, Gupta S, Singh H, Kumar S, Gautam A, et al. CancerDR: cancer drug resistance database. *Sci Rep* 2013;**3**:1445.
- Yang MJ, Tao BZ, Chen CJ, Jia WQ, Sun SL, Zhang TT, et al. Machine learning models based on molecular fingerprints and an extreme gradient boosting method lead to the discovery of JAK2 inhibitors. *J Chem Inf Model* 2019;**59**:5002–12.
- Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742–54.
- Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. Coleman. ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* 2012;**52**:1757–68.
- Irwin JJ, Shoichet BK. ZINC—A free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005;**45**:177–82.
- Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 2019;**47**:930–40.

34. Huber F, van der Burg S, van der Hooft JJJ, Ridder L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J Cheminf* 2021;**13**:84.
35. Sahin E, DePinho RA. Axis of ageing: telomeres, p53 and mitochondria. *Nat Rev Mol Cell Biol* 2012;**13**:397–404.
36. Amaral JD, Xavier JM, Steer CJ, Rodrigues CM. The role of p53 in apoptosis. *Discov Med* 2010;**9**:145–52.
37. Wu K, Yee NA, Srinivasan S, Mahmoodi A, Zakharian M, Mejia Oneto JM, et al. Click activated protodrugs against cancer increase the therapeutic potential of chemotherapy through local capture and activation. *Chem Sci* 2021;**12**:1259–71.
38. Srinivasan S, Yee NA, Wu K, Zakharian M, Mahmoodi A, Royzen M, et al. SQ3370 activates cytotoxic drug *via* click chemistry at tumor and elicits sustained responses in injected and non-injected lesions. *Adv Ther* 2021;**4**:2000243.
39. Srinivasan S, Yee NA, Zakharian M, Alečković M, Mahmoodi A, Nguyen TH, et al. SQ3370, the first clinical click chemistry-activated cancer therapeutic, shows safety in humans and translatability across species. *bioRxiv* 2023:534654.
40. Chawla SP, Batty K, Alečković M, Bhadri VA, Bui N, Guminski A, et al. 1499P phase I clinical & immunologic data of SQ3370 in advanced solid tumors. *Ann Oncol* 2022;**33**:S1232.