Supporting Information

Computational Bioactivity Fingerprint Similarities to Navigate the Discovery of Novel Scaffolds

Guo-Li Xiong[1, 2], Yue Zhao[1, 2], Lu Liu[1], Zhong-Ye Ma[1], Ai-Ping Lu[3], Yan Cheng[4], Ting-Jun Hou[5*], Dong-Sheng Cao[1,2,3*]

[1]Xiangya School of Pharmaceutical Sciences, Central South University, Changsha, Hunan 410003, China

[2]Hunan Key laboratory of Diagnostic and Therapeutic Drug Research for Chronic Diseases, Central South University, Changsha, Hunan 410013, China

[3]Institute for Advancing Translational Medicine in Bone and Joint Diseases, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong SAR, China

[4]Department of Pharmacy, The Second Xiangya Hospital, Central South University, Changsha, Hunan 410003, China

[5]Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China

*To whom correspondence should be addressed. Dongsheng Cao. Tel: +86 731 8982 4761; Email: oriental-cds@163.com

Correspondence may also be addressed to Tingjun Hou. E-mail: tingjunhou@zju.edu.cn

[†] the first two authors should be regarded as joint First Authors.

Table of Contents

Figure S1. Box plots of the AMRR values within Top100 (A) and Top 5000 (B) compounds for different screening datasets S3

Figure S2. The influence of threshold on SRR (A) and AMRR (B) of CBFP S4

Figure S3. Workflow for the construction of CBFP representation S5

Table S1. The global AMRR (%) results of ten molecular representations S6

Table S2. PARP-1 inhibitors approved by FDA S7

Table S3. Top 10 potential targets predicted by three target fishing methods (Compound 6) S8

Table S4. Top 10 potential targets predicted by three target fishing methods (Olaparib) S9

Table S5. Details of the 17 targets in the query dataset S10

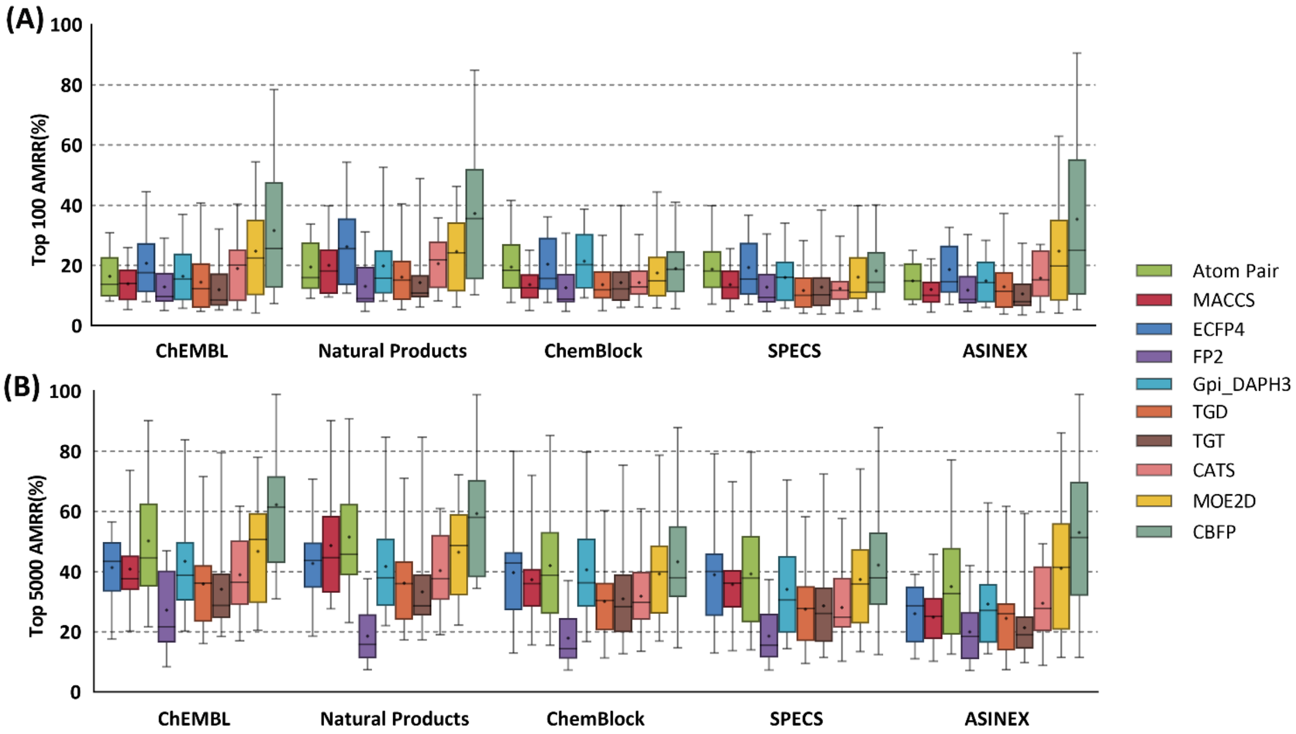


Figure S1. Box plots of the AMRR values within Top100 (A) and Top 5000 (B) compounds for different screening datasets. The horizontal lines indicate the median, and the plus signs represent the mean SRR values of the 1660 query molecules.

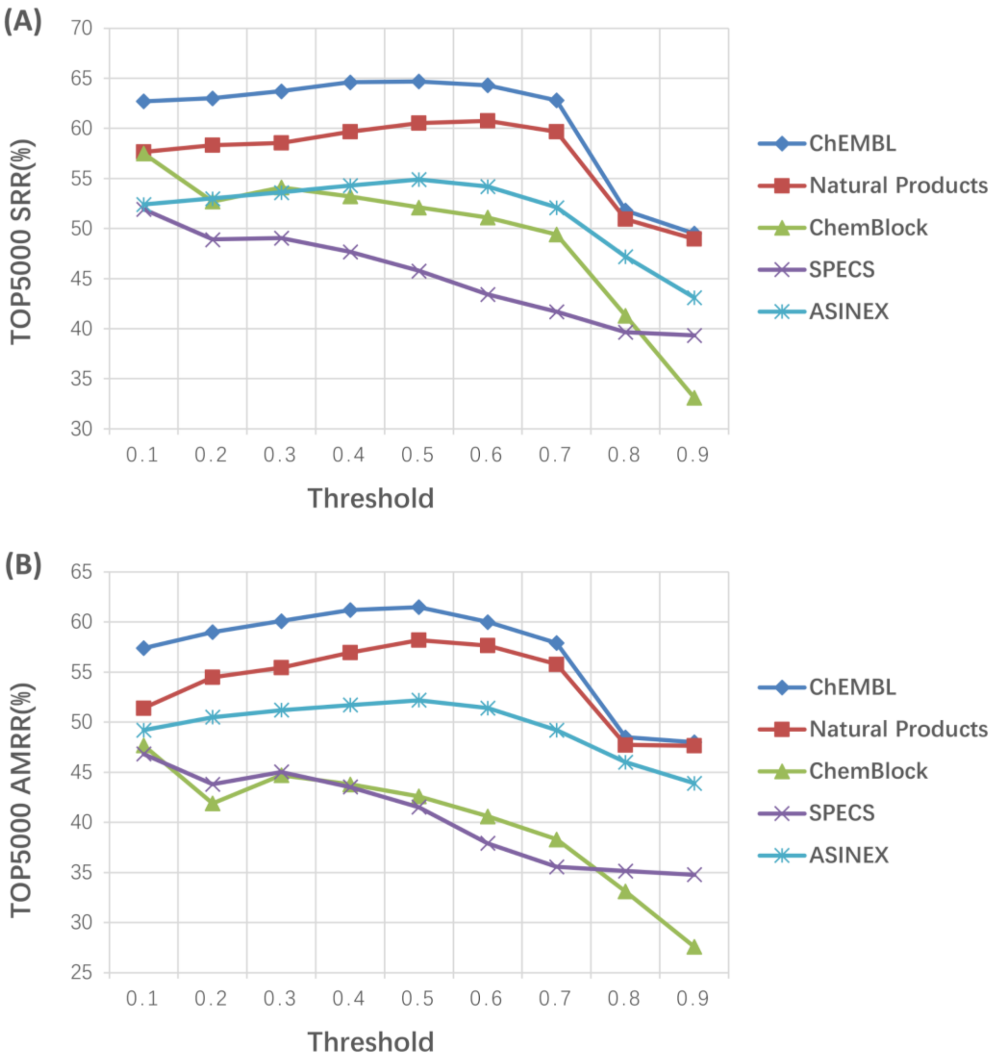


Figure S2. The influence of threshold on SRR (A) and AMRR (B) of CBFP representation. According to the comprehensive performance of the five screening datasets, 0.5 was selected as the cutoff to translate the 832-bit feature vector into standard binary fingerprint.

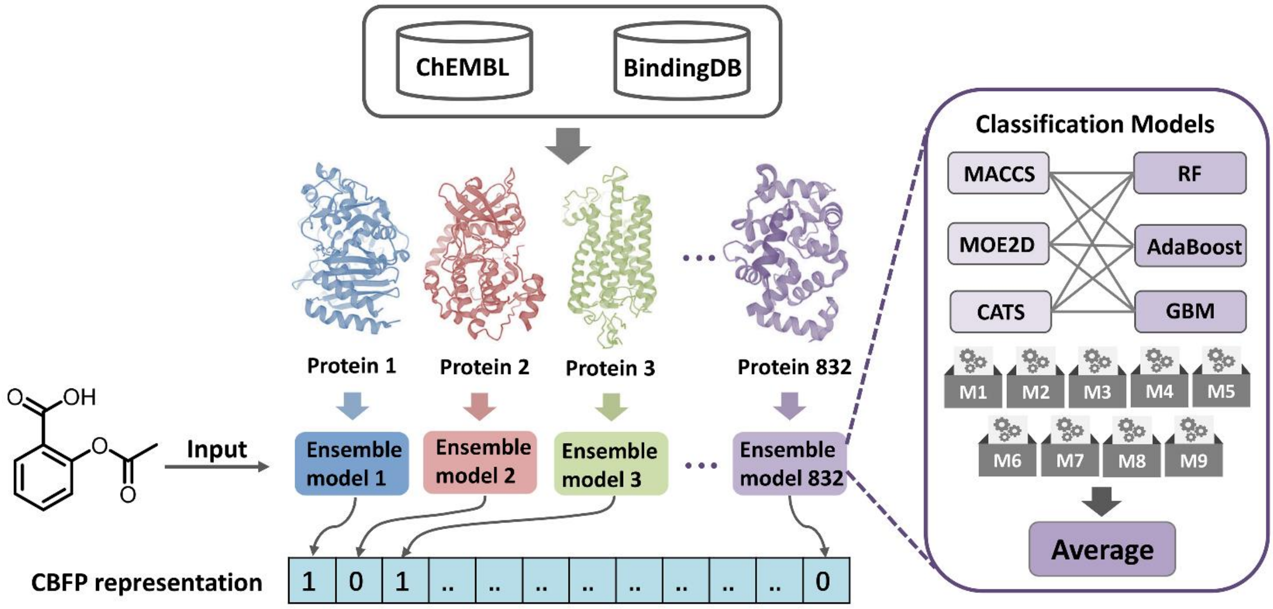


Figure S3. Workflow for the construction of CBFP representation.

Table S1. The global AMRR (%) results of ten molecular representations for different screening datasets

Molecular	ChEMBL		Natural Products		ChemBlock		SPECS		ASINEX	
representation	Top100	Top5000	Top100	Top5000	Top100	Top5000	Top100	Top5000	Top100	Top5000
AtomPair	16.4±7.2	41.3±10.2	19.4±8.2	42.7±12.2	19.5±9.4	39.6±15.7	18.8±9.1	38.9±16.4	14.8±6.4	26.0±9.7
MACCS	13.9±6.3	40.8±14.8	20.0±9.8	48.7±19.0	13.7±5.6	37.2±14.7	13.7±5.9	35.7±14.1	12.0±5.3	24.9±10.2
ECFP4	20.7±11.0	50.3±20.5	26.2±13.4	51.4±19.2	20.3±9.7	42.0±19.2	**19.3±9.2**	39.2±18.0	18.6±8.7	35.1±17.9
FP2	12.9±6.6	27.2±13.2	13.1±7.4	18.6±9.1	12.6±7.3	17.9±8.5	12.7±7.5	18.6±8.8	11.8±6.6	20.1±10.7
Gpi_DAPH3	16.4±9.0	43.4±18.2	19.8±11.7	41.8±19.0	**21.5±9.6**	40.7±17.1	16.0±8.2	34.1±17.0	14.8±7.1	29.1±14.4
TGD	14.5±9.7	35.8±14.6	16.1±9.4	36.1±14.5	13.6±6.9	30.1±12.9	11.6±6.9	27.6±13.2	12.9±8.7	24.4±13.4
TGT	11.9±7.2	34.2±15.0	14.3±9.8	33.3±15.6	14.2±8.6	30.9±15.3	12.8±8.7	28.7±15.4	10.5±6.2	21.4±11.9
CATS	18.9±10.1	38.9±13.4	20.5±8.3	40.3±12.1	14.3±6.3	31.9±12.5	12.4±6.2	28.0±12.6	15.8±7.6	29.4±13.0
MOE2D	24.7±15.3	46.8±17.4	25.8±12.8	46.5±15.7	17.5±10.3	39.3±15.9	16.2±10.0	37.4±16.4	24.7±16.7	41.0±23.2
CBFP	**31.6±21.0**	**62.2±21.3**	**37.2±22.6**	**59.3±22.1**	19.0±10.1	**43.3±17.9**	18.3±10.2	**42.2±18.3**	**35.4±26.2**	**53.0±27.5**

*The bold value represents the maximum value of this column.

Table S2. PARP-1 inhibitors approved by FDA

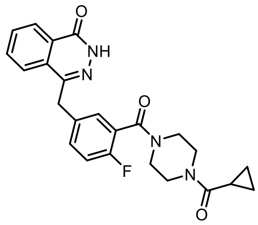
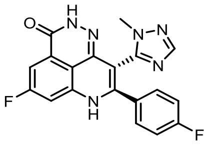
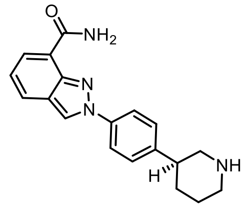
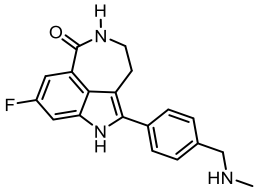
Chemotype	Name	Structure	Developer	Indications	Approved time
Phenazinone derivative	Olaparib		AstraZeneca	Advanced Ovarian Cancer, recurrent epithelial ovarian cancer, fallopian tube cancer or primary peritonea cancer HER2-negative metastatic breast cancer	2014, 2017 and 2018
	Talazoparib		Pfizer	Advanced or metastatic breast cancer	2018
benzimidazole formamide derivative	Niraparib		Merck	Ovarian cancer, fallopian tube cancer, peritoneal cancer and triple-negative breast cancer	2017
tricyclic lactam indoles derivative	Rucaparib		Clovis	Advanced Ovarian Cancer, recurrent epithelial ovarian cancer, fallopian tube cancer or primary peritoneal cancer	2016 and 2018

Table S3. Top 10 potential targets predicted by three target fishing methods (Compound 6).

Rank	SEA	SwissTargetPrediction	TargetNet	PharmMapper
1	Beta-glucuronidase	Matrix metalloproteinase 9	Estrogen receptor	Capsid protein
2	Dioxygenase	Matrix metalloproteinase 1	Platelet-derived growth factor receptor beta	NONE
3	Queuine tRNA-ribosyltransferase catalytic subunit 1	Toll-like receptor (TLR7/TLR9)	Aldose reductase	Phosphate acetyltransferase
4	Tankyrase-1	Tyrosine-protein kinase receptor FLT3	Protein kinase C epsilon type	Sodium/glucose cotransporter
5	Poly [ADP-ribose] polymerase 1 （Bovine）	Platelet-derived growth factor receptor	Hydroxycarboxylic acid receptor 2	Protein degV
6	Tankyrase-2	Serine/threonine-protein kinase Chk1	Transcription factor p65	Hemoglobin subunit alpha
7	Complement C1r subcomponent	Serine/threonine-protein kinase WEE1	Estrogen receptor beta	Phosphatidylethanolamine-binding protein 1
8	Protein Wnt-3a	Cyclin-dependent kinase 5/CDK5 activator 1	Amine oxidase [flavin-containing] B	Tyrosyl-tRNA synthetase
9	Calcium-dependent protein kinase 4	Leukotriene A4 hydrolase	Amine oxidase [flavin-containing] A	Preprotein translocase subunit secY
10	**Poly [ADP-ribose] polymerase 1 （Homo）**	Carbonic anhydrase I	Macrophage migration inhibitory factor	Bacteriorhodopsin

Table S4. Top 10 potential targets predicted by three target fishing methods (Olaparib)

Rank	SEA	SwissTargetPrediction	TargetNet	PharmMapper
1	**Poly [ADP-ribose] polymerase 1**	Poly [ADP-ribose] polymerase 2	**Poly [ADP-ribose] polymerase 1**	**NONE**
2	Poly [ADP-ribose] polymerase 6	Poly [ADP-ribose] polymerase 3	Cytochrome P450 2C19	Osmotically inducible protein C
3	Poly [ADP-ribose] polymerase 3	Poly [ADP-ribose] polymerase 6	Cytochrome P450 3A4	Uncharacterized protein YMR074C
4	Poly [ADP-ribose] polymerase 4	**Poly [ADP-ribose] polymerase-1**	Potassium voltage-gated channel subfamily H member 2	Type IV secretion system protein virB11
5	Poly [ADP-ribose] polymerase 2	Tankyrase-1	Cytochrome P450 2C9	Deoxyhypusine synthase
6	Poly [ADP-ribose] polymerase 10	Tankyrase-2	Non-receptor tyrosine-protein kinase TYK2	Phosphate acetyltransferase
7	Tankyrase-2	Acetyl-CoA carboxylase 1	COUP transcription factor 2	Uncharacterized protein MG296 homolog
8	Tankyrase-1	Acetyl-CoA carboxylase 2	Dipeptidyl peptidase 4	Troponin C, slow skeletal and cardiac muscles
9	High affinity cGMP-specific 3',5'-cyclic phosphodiesterase 9A	Acyl coenzyme A:cholesterol acyltransferase 1	Proto-oncogene tyrosine-protein kinase receptor Ret	Cytochrome b5
10	B1 bradykinin receptor	ADAM17	Alkaline phosphatase, tissue-nonspecific isozyme	Beta-elicitin cinnamomin

Table S5. Details of the 17 targets in the query dataset

Target ID	**Target Name**	**No. of molecules**	**No. of scaffolds**
EGFR	epidermal growth factor receptor erbB1	55	11
CA2	carbonic anhydrase II	70	14
DRD2	dopamine D2 receptor	95	19
CB1	cannabinoid CB1 receptor	80	16
GRHR	gonadotropin-releasing hormone receptor	65	13
SERT	serotonin transporter	75	15
KOR1	κ opioid receptor	60	12
ESR2	estrogen receptor β	70	14
HIVPR	human immunodeficiency virus type 1 protease	160	32
CFX	coagulation factor X	85	17
HIVRT	human immunodeficiency virus type 1 reverse transcriptase	100	20
NK1R	neurokinin 1 receptor	60	12
ADORA3	adenosine A3 receptor	130	26
MC4R	melanocortin receptor 4	100	20
DHFR	dihydrofolate reductase	65	13
VEGFR2	vascular endothelial growth factor receptor 2	160	32
MCHR	melanin-concentrating hormone receptor 1	230	46
Total		**1660**	**322**