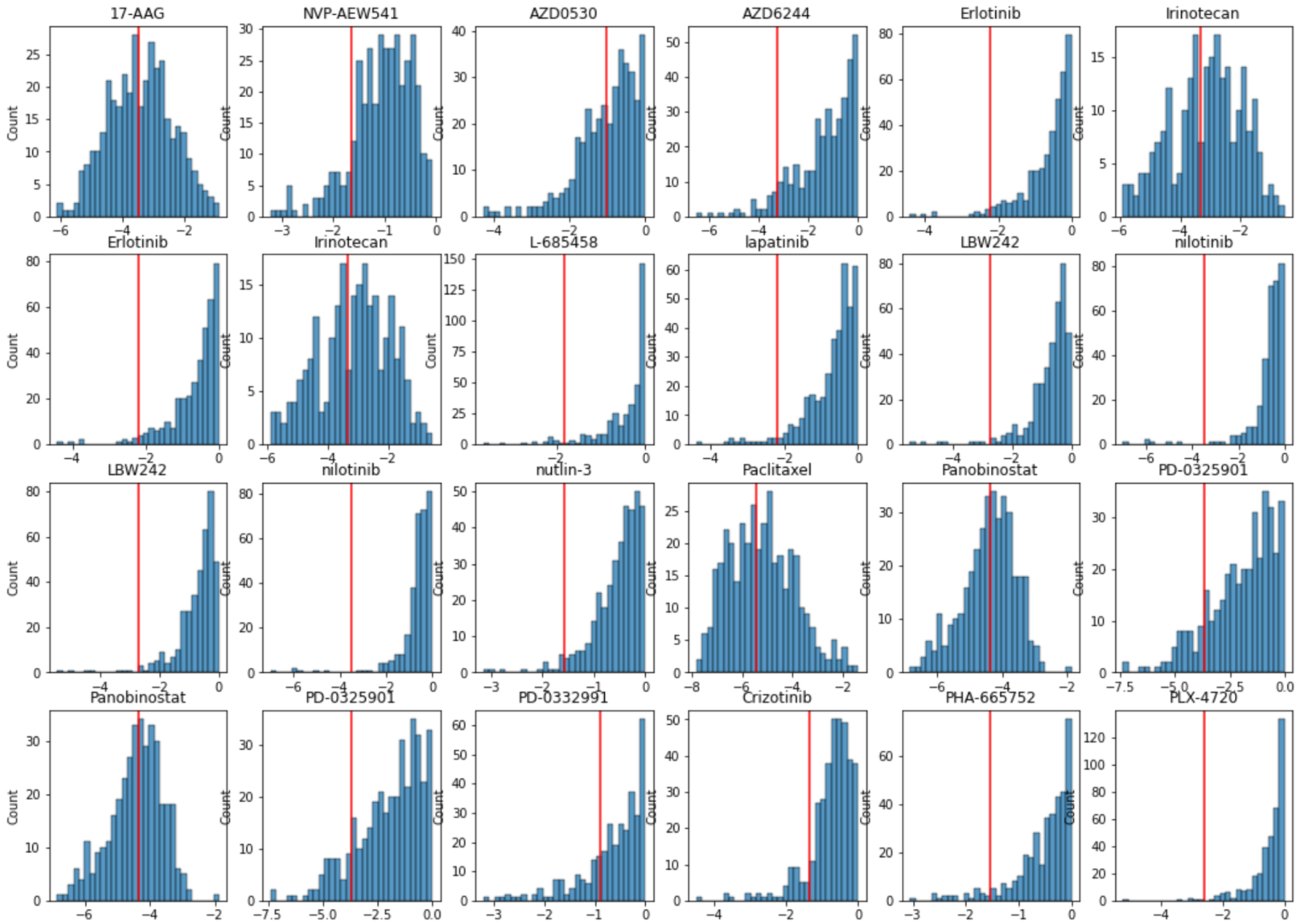
Supplementary

Supplementary Fig. 1: Histogram of Drug Response Values for the 24 Training Drugs Over 471 Cancer Cell Lines



Histogram of Drug Response Values for the 24 Training Drugs Over 471 Cancer Cell Lines. For each drug plot, the x-axis is the negative activity area values for drug response, and the y-axis is the total count. The red line on each plot is the cut-off computed by the m-cluster algorithm.

Supplementary Fig. 2: Drug Binding Sites Detection Across Various Protein Structures

a

	ABL1_3CS9_A	ABL1_3CS9_B	ABL1_3CS9_C	ABL1_3CS9_D	ABL1_5MO4_A
O8	Yes	Yes	Yes	Yes	Yes
N9	Yes	Yes	Yes	Yes	Yes
C24	No	No	No	No	No
N26	Yes	Yes	Yes	Yes	-
C25	-	No	No	No	No
C29	Yes	Yes	Yes	Yes	Yes
C30	-	Yes	Yes	-	Yes
C33	Yes	-	Yes	-	Yes
N35	No	No	No	No	No
C36	-	Yes	Yes	Yes	Yes
C38	Yes	-	Yes	Yes	Yes

b

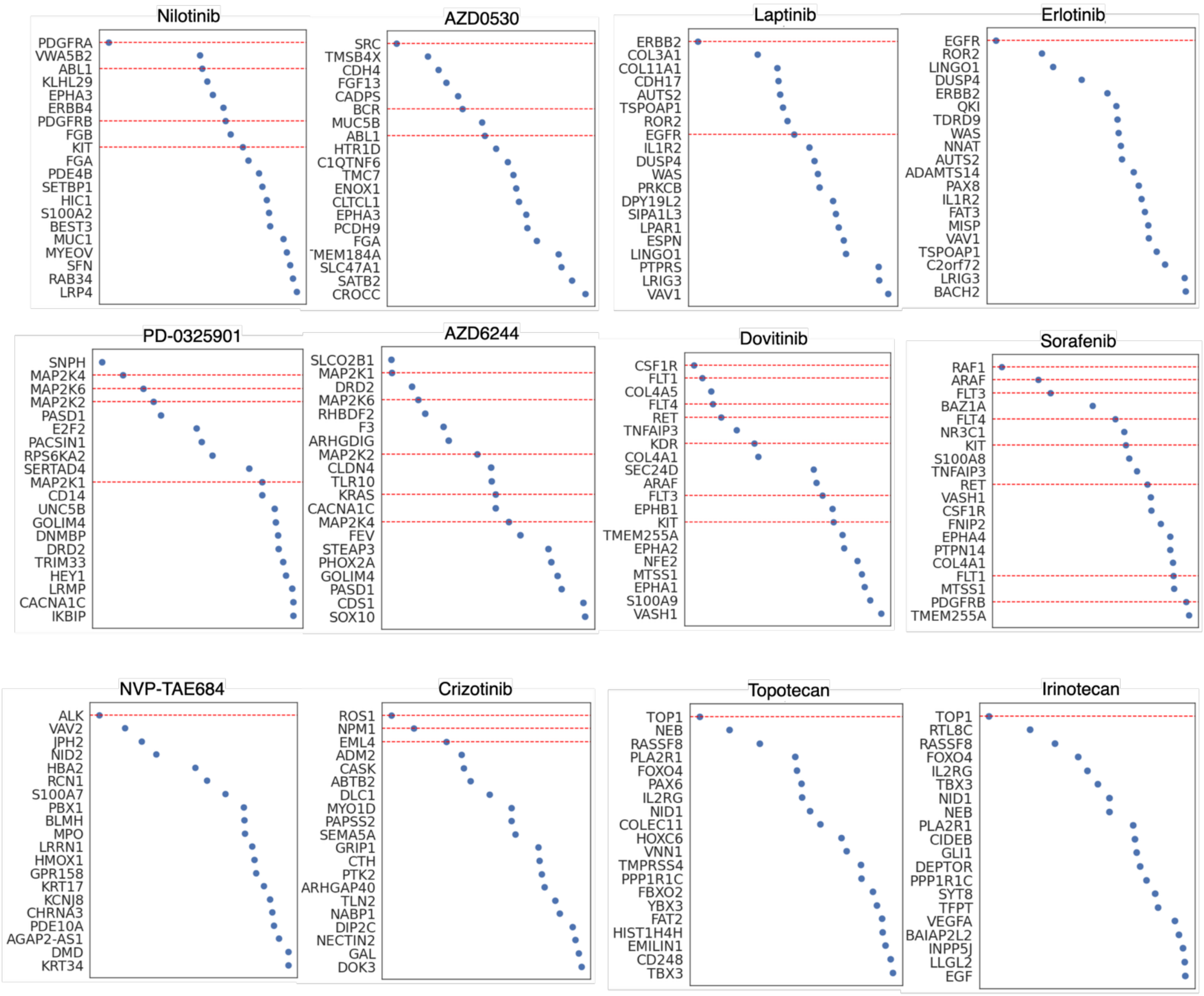
	ALK_2XB7_A
O5	Yes
N16	Yes
C17	Yes
C18	Yes
N20	Yes

c

	ALK_2XP2_A	ALK_2YFX_A	ALK_4ANQ_A	ALK_4ANS_A	ALK_5AAA_A	ALK_5AAB_A	ALK_5AAC_A	MET_2WGJ_A	ROS1_3ZBF_A
C5	Yes	Yes	-	-	-	-	Yes	-	Yes
C6	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-	Yes
C7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-	Yes
C12	-	-	Yes	Yes	Yes	Yes	-	Yes	Yes
C13	No	No	No	No	No	No	No	No	No
N14	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
C15	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-
C28	-	No	No	No	-	No	No	-	-
N29	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

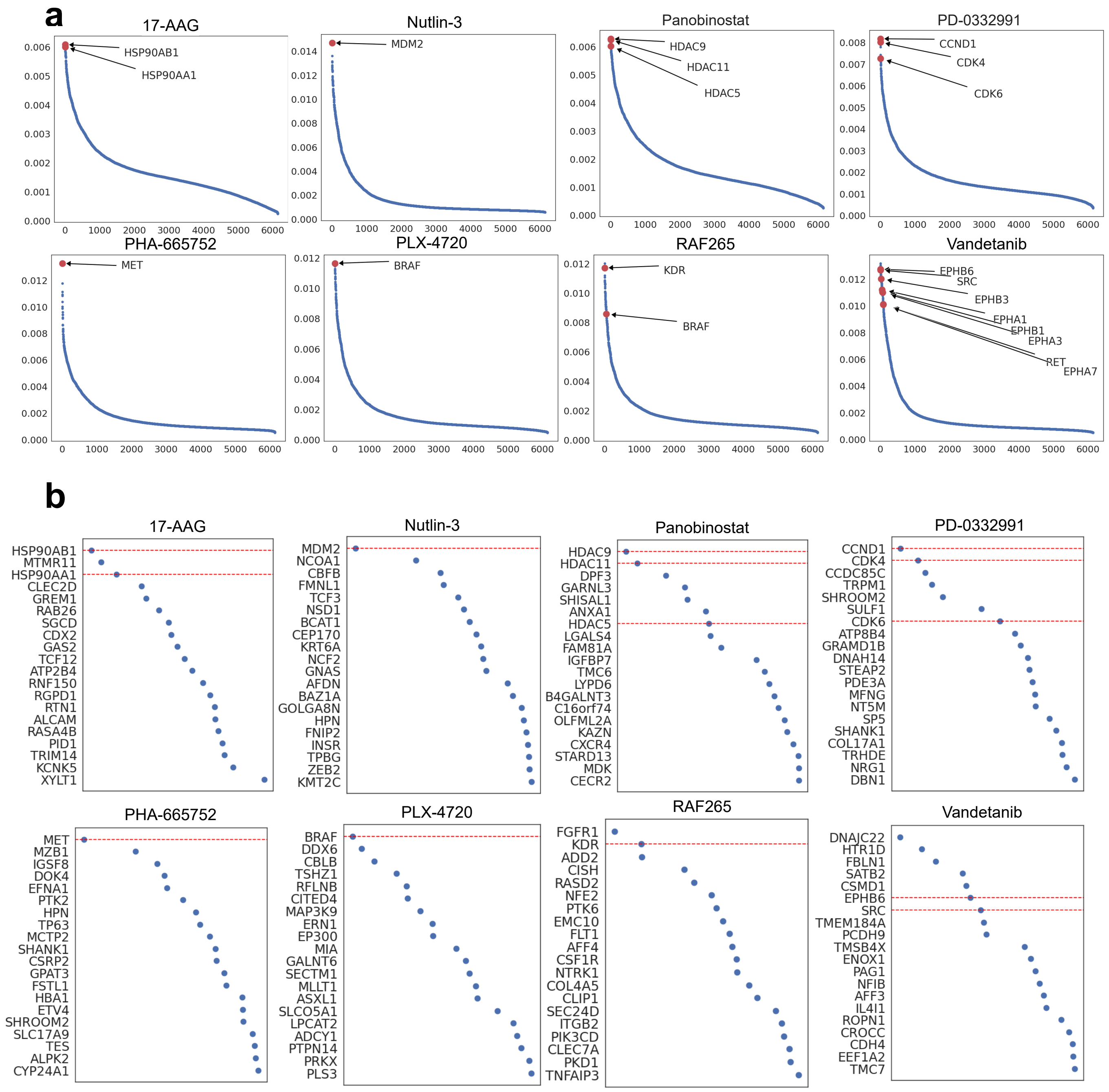
Tables show all the actual binding atoms (rows in the table) for drugs: NILOTINIB, NVP-TAE684, and CRIZOTINIB. Each column in the table represents one protein pose. For every entry in the table, “Yes” means the model detected that specific binding atom, and "No" means the model failed to detect that binding atom. **a.** The binding sites detection table for the Drug: NILOTINIB. **b.** The binding sites detection table for Drug: NVP-TAE684. **c.** The binding sites detection table for Drug" CRIZOTINIB.

Supplementary Fig. 3: Cross-Attention Gene Ranking with Top 20 Genes



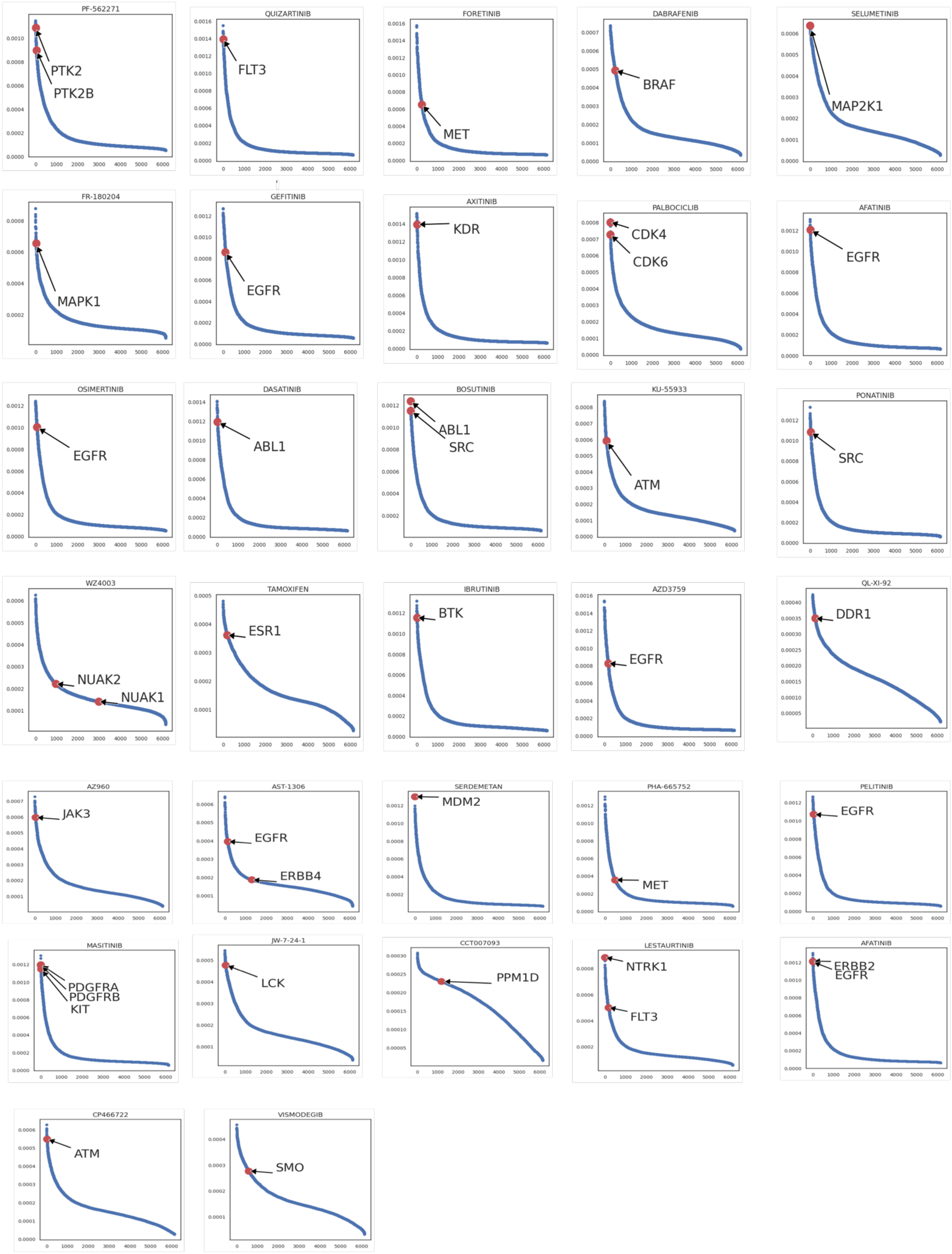
The zoomed-in top 20 cross-attention gene rankings corresponding to Fig. 3d. The dashed red lines correspond to the targeted genes for the drugs. Note that the total number of gene candidates is 6144. MIDI could still rank the target genes at the very top.

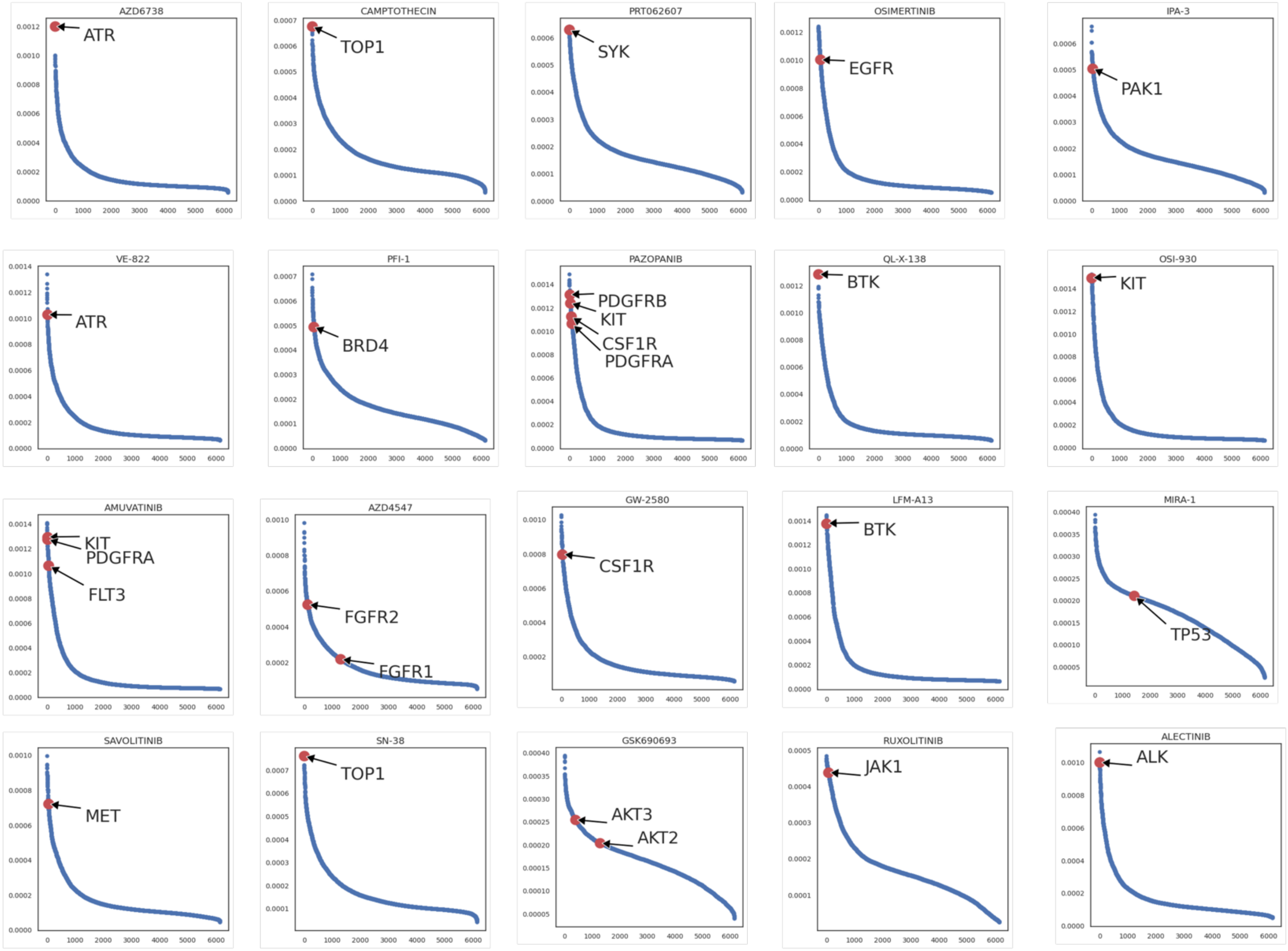
Supplementary Fig. 4: Cross-Attention Gene Ranking for Other Drugs in CCLE



a. Gene cross-attention score for each drug in ranking order. **b.** The top 20 genes with the highest cross-attention score for each drug. All of the targeted genes are ranked at the top.

Supplementary Fig. 5 Gene Ranking Visualization for the 52 GDSC validation Drugs

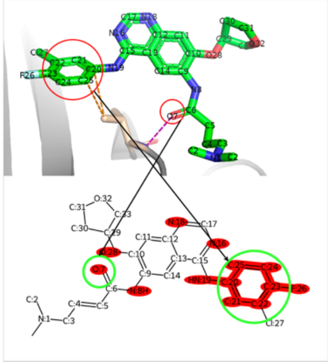

PF-562271
PTK2
PTK2B
QUIZARTINIB
FLT3
FORETINIB
MET
DABRAFENIB
BRAF
SELUMETINIB
MAP2K1
FR-180204
MAPK1
GEFITINIB
EGFR
AXITINIB
KDR
PALBOCICLIB
CDK4
CDK6
AFATINIB
EGFR
OSIMERTINIB
EGFR
DASATINIB
ABL1
BOSUTINIB
ABL1
SRC
KU-55933
ATM
PONATINIB
SRC
WZ4003
NUAK2
NUAK1
TAMOXIFEN
ESR1
IBRUTINIB
BTK
AZD3759
EGFR
QL-XI-92
DDR1
AZ960
JAK3
AST-1306
EGFR
ERBB4
SERDEMETAN
MDM2
PHA-665752
MET
PELITINIB
EGFR
MASITINIB
PDGFRA
PDGFRB
KIT
JW-7-24-1
LCK
CCT007093
PPM1D
LESTAURTINIB
NTRK1
FLT3
AFATINIB
ERBB2
EGFR
CP466722
ATM
VISMODEGIB
SMO



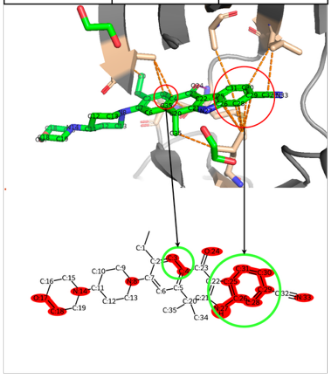
Target gene prediction on the 52 validation drugs from the GDSC dataset. The x-axis: the whole 6144 genes in ranking order. The y-axis: attention scores for the genes. The red dots: the target genes corresponding to the drugs. The drug names are recorded at the top of each plot. The prediction accuracy is shown in Extended Data Fig. 3f and Supplementary Fig. 7, where it shows that most of the predicted targeted genes are ranked within the top 200.

Supplementary Fig. 6: Drug Binding Site Detection for 24 of The Validation GDSC Drugs

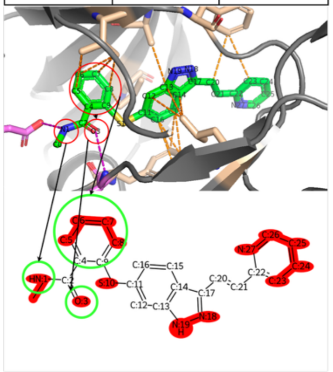
Drug Name	Target Gene	PDB_ID
AFATINIB	EGFR	6ZU8



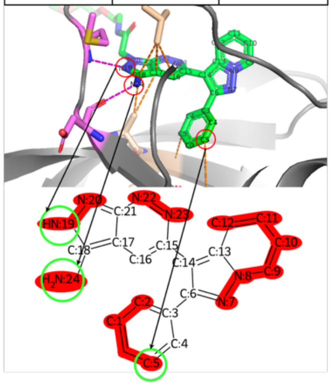
Drug Name	Target Gene	PDB_ID
ALECTINIB	ALK	3A0X



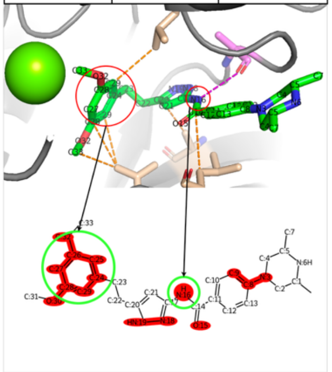
Drug Name	Target Gene	PDB_ID
AXITINIB	KIT	4AG8



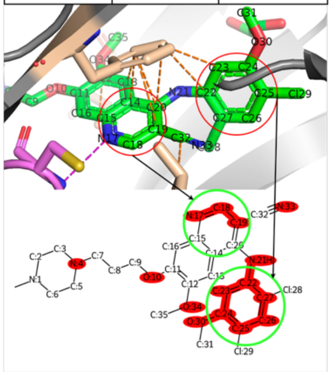
Drug Name	Target Gene	PDB_ID
FR-180204	ERK	5V61



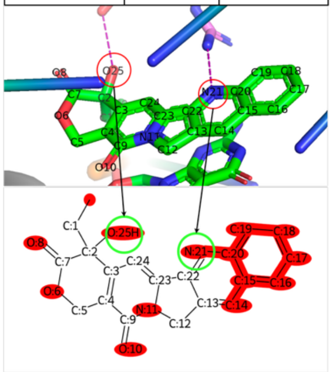
Drug Name	Target Gene	PDB_ID
AZD4547	FGFR	4RWK



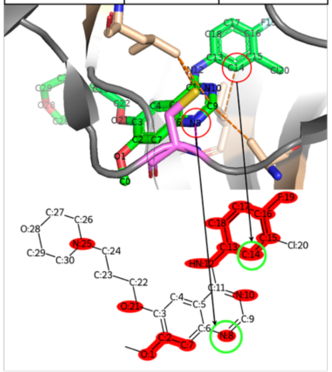
Drug Name	Target Gene	PDB_ID
BOSUTINIB	SRC	5VC3



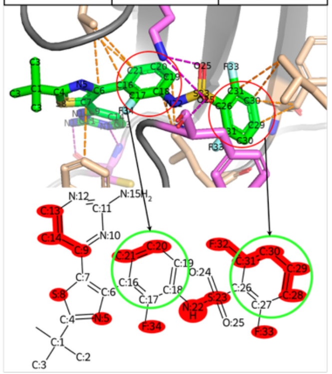
Drug Name	Target Gene	PDB_ID
CAMPTOTHECIN	TOP1	1T8I



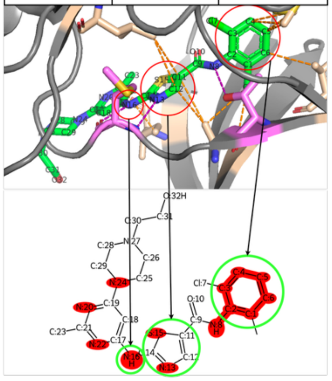
Drug Name	Target Gene	PDB_ID
GEFITINIB	EGFR	5Y80



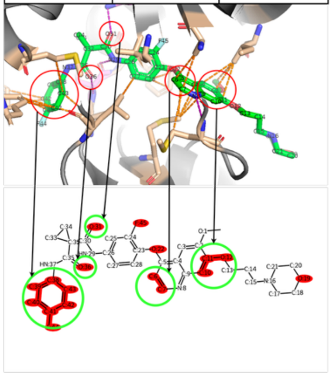
Drug Name	Target Gene	PDB_ID
DABRAFENIB	BRAF	6V2U



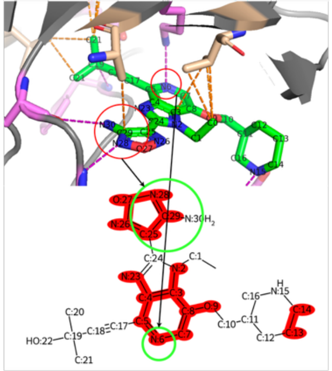
Drug Name	Target Gene	PDB_ID
DASATINIB	ABL,SRC	2Y6O

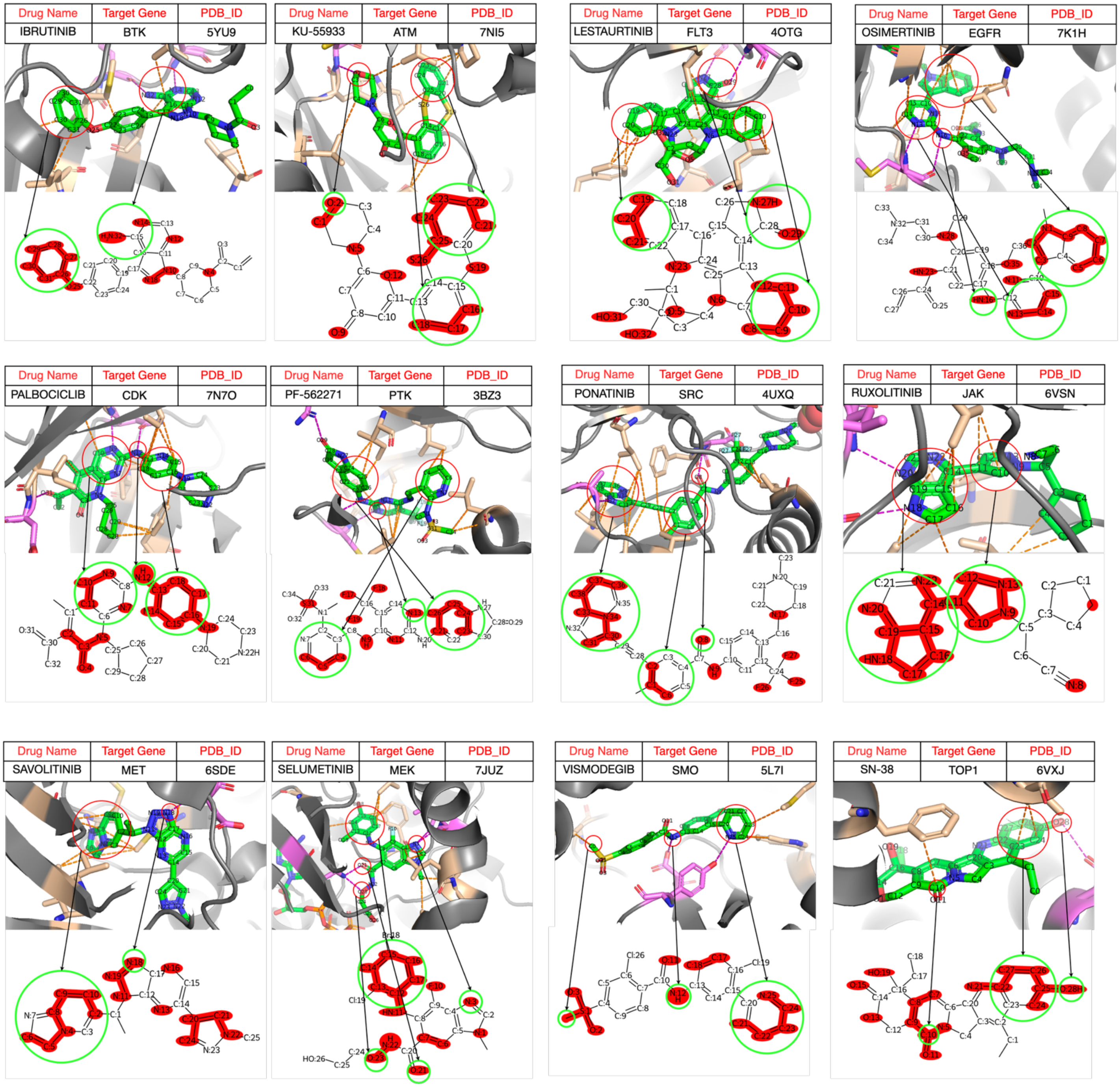


Drug Name	Target Gene	PDB_ID
FORETINIB	MET	6SDC



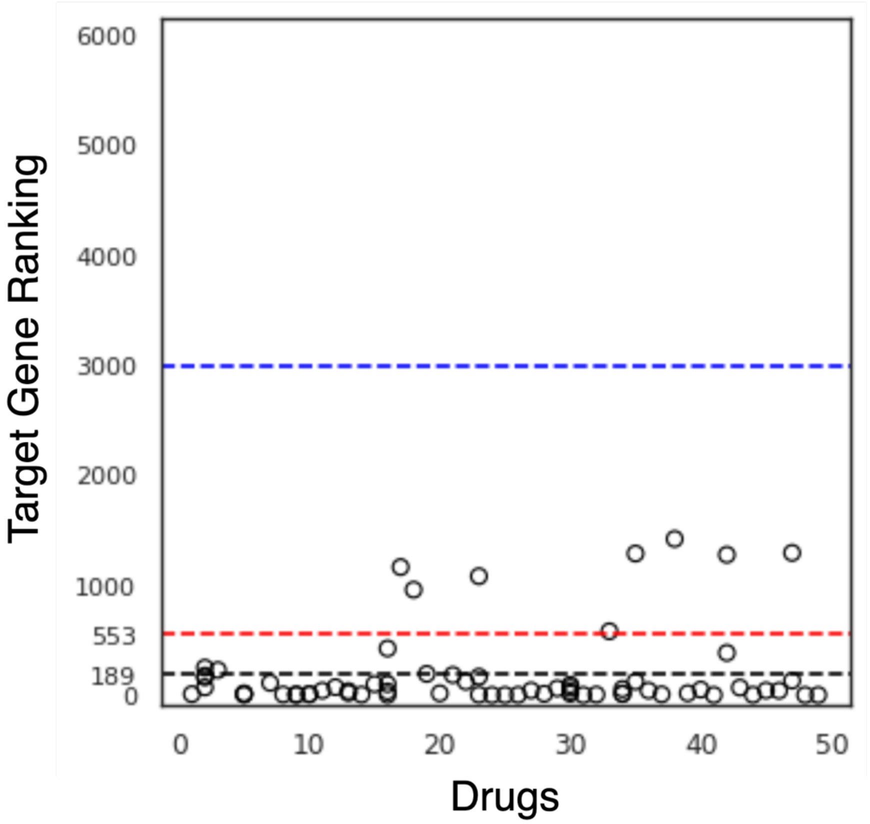
Drug Name	Target Gene	PDB_ID
GSK690693	AKT	3D0E





The drug-protein binding site detection of 24 GDSC validation drugs. We highlight the actual binding region in the PDB structure with red circles, and the corresponding detected binding region from MIDI in green circles. We report the accuracy of the atom detection result in Supplementary Table 6.

Supplementary Fig. 7: Scatter Plot of Target Gene Ranking for the 52 GDSC Validation Drugs



Target gene rankings prediction from MIDI for 52 validation drugs from the GDSC dataset. The x-axis represents randomly ordered validation drugs, and the y-axis represents the targeted genes for the drugs. The block dots are the target gene rankings for drugs, the black dashed line shows the mean value, and the red dashed line shows the first standard deviation. The plot shows that the mean value of targeted gene detection is 189 out of the whole 6144 gene candidates, demonstrating the strong target gene identification ability of MIDI.