Supplementary material: Anti-target mapping enhances selectivity of machine learning-predicted CK2 inhibitors

Huiyan Ying[1], Weikaixin Kong[1], Aron Schulman[1,2], Nikola Panajotovikj[3], Zia Tanoli[1], Jordi Mestres[3], Tero Aittokallio[1,2,4,5], Mitro Miihkinen[1,2,*]

[1]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki 00290, Finland

[2]iCAN Digital Precision Cancer Medicine Flagship, University of Helsinki and Helsinki University Hospital, Helsinki 00290, Finland

[3]Chemotargets SL, Parc Científic de Barcelona, Barcelona, Catalonia, Spain

[4]Institute for Cancer Research, Department of Cancer Genetics, Oslo University Hospital, Oslo 0310, Norway

[5]Oslo Centre for Biostatistics and Epidemiology (OCBE), Faculty of Medicine, University of Oslo, Oslo 0372, Norway

*corresponding author

1.1 Pan-cancer and Cancer Cell Line Analyses of CK2

Transcriptomic data (log2[TPM+1]) across 33 cancer types were obtained from The Cancer Genome Atlas (TCGA) via the UCSC Xena platform. Cancer type abbreviations and sample details are summarized in Supplementary Table S1. Differential expression analyses of CSNK2A1 and CSNK2A2 between tumor and adjacent normal tissues were performed using Wilcoxon rank-sum tests with false discovery rate (FDR) correction via the Benjamini–Hochberg method, and the results were visualized as boxplots.

For functional dependency assessment, CRISPR-Cas9 gene knockout dependency (CERES) scores and RNA-seq expression data for 1,103 cancer cell lines were retrieved from the DepMap portal (24Q4 Public release). Scatter plots visualizing baseline gene expression versus dependency scores were constructed using ggplot2, with the five most CK2-dependent cell lines (lowest CERES scores) highlighted and annotated via ggrepel.

All analyses were performed in R (version 4.4.2), employing packages including ggpubr, ggplot2, survival, forestplot, fmsb, and ggrepel.

1.2 The effect of bioactivity value types on regression model accuracy.

Various bioactivity types, including IC_{50}, K_d, K_i, and their combinations, could be used to train regression models. The different bioactivity types used as input will affect the accuracy of the regression models (Nature Communications, 12, 3307 (2021)). To find the most suitable bioactivity value types leading to high prediction accuracy, we used each bioactivity value type and the their combinations as input to train the regression models. For the combination, when a compound has multiple bioactivity types, the most confident bioactivity type was selected according to the confidence ranking of the various bioactivity categories

(Computational and Structural Biotechnology Journal, 18, 2020, 3819-3832). As the number of compounds with K_d value was relatively small, we did not use it as the sole bioactivity type to train the models, but rather combined it with other bioactivity types. In addition, apart from the simple combination of different bioactivity types, we also calculated the interaction score that summarizes IC_{50}, K_d, K_i into a single value (represented as IS_biochemical and IS_cell based) (Brief Bioinform. 2020 Jan 17;21(1):211-220). Compared to other activity types, the interaction score also considers the protein target family and assay format, which could provide more comprehensive information for the compound-target interactions.

After training the 5 regression models with different bioactivity types, we observed that IC_{50} alone or in combination with other bioactivity data types, resulted in highly accurate predictions; especially with K_d, even if the input dataset size was reduced (Supplentary Figure 2). To further investigate whether the IC_{50} alone or in combination contributed most to the high prediction accuracy, we eliminated dataset-size effects by randomly sub-sampling 1,000 compounds from each bioactivity-type dataset. Then, the datasets were split into training dataset and test set with the proportion of 3:1. The models were re-trained in the training dataset and the model prediction accuracy was scored in the test dataset. We observed that the IC_{50} bioactivities alone led to the highest prediction accuracy (Supplentary Figure 3). IC_{50} is commonly used as the training bioactivity value in many regression models (J. Chem. Inf. Model. 2014, 54, 10, 2751–2763. Current Bioinformatics, 17, 8, 2022, pp. 698-709(12)). Thus, we finally selected IC_{50} as the bioactivity value for the model training and prediction.

1.3 The effect of various chemical fingerprints on regression model accuracy.

Molecular fingerprints are widely used to describe the structural characteristics of molecules, particularly in relation to their biological activities (Journal of Cheminformatics, 12, 6, 2020). The chemical fingerprints provide an abstract representation of specific structural features. Generated directly from chemical structures, molecular fingerprints can be converted into 2D fragments, enabling an effective representation of chemical fragments in ML models(J. Chem. Inf. Model. 2022, 62, 6035−6045). In this study, five commonly used fingerprints with various lengths were compared to characterize the molecules and their impacts on the regression model accuracy. The regression models were trained with different chemical molecule fingerprints as the input feature, and the IC_{50} values as the bioactivity type. After that, the regression model accuracy were scored in the test dataset (Supplentary Figure 4). We observed that most of the chemical fingerprints performed similarly well, but there was a notable decrease on the model prediction accuracy when using the estate fingerprint. As the length of estate fingerprint has only 79 bits, it may be too short to effectively characterize molecules. It has been shown that limited number of features leads to a significant loss of information, resulting in poor predictions (J. Chem. Inf. Model. 2012, 52, 2840−2847). Through comparison, the standard fingerprint led to the highest prediction accuracy in most of the regression models, and therefore it was selected as the optimal chemical fingerprint to represent the molecules.

1.4 The performance of error prediction model in the new test set

To evaluate the application domain of bioactivity prediction model (BPM), we set up an error prediction model (EPM) (PMID: 24152204) using SVM method and the same parameters with BPM. The predicted labels of EPM were the errors(Equation 3 in Method part) from BPM, and the input features of EPM were the standard fingerprints. The previous test set

was divided into a calibration set and a new test set according to the ratio 1:1. Then the calibration set was used to train the EPM. When setting a specific confidence level (Alpha value), the samples with large predicted error values in the new test set would be removed and we only do bioactivity prediction of left samples. The result was shown in Supplentary Figure 5A-5D. To obtain robust results, the above process was repeated 60 times in every confidence level. When using the EPM method, the Spearman coefficient (Supplentary Figure 5A), Pearson coefficient (Supplentary Figure 5B) and RMSE (Supplentary Figure 5D) could achieve better results significantly. The NRMSE could also decrease when Alpha value=0.1, 0.2 or 0.3 (Supplentary Figure 5C). However, when Alpha value was less than 0.5, this metric increased. This was because when only a small part of compounds were left in the new test set, the limited sample size would make the NRMSE have less denominator values(Ya_{max}-Ya_{min}) in Equation 2, which is not related to model performance. To sum up, NRMSE is a kind of metric related to sample size and EPM could further improve BPM performance by only doing prediction on compounds located in the model application domain.

1.5 Off-target toxicity analysis and Sankey visualization

We selected three representative CK2 inhibitors (CX-4945, DMAT, and TBB) to investigate their off-target interactions with CDKs and related toxicity profiles. Main CDK off-targets of CK2 inhibitors were identified by combining results from ChEMBL, STITCH, and SwissTargetPrediction databases. For each CDK, toxicity associations were retrieved from the Comparative Toxicogenomics Database (CTD) and the top 20 toxicities ranked by Inference Score were selected.

The CK2i-CDK off-target–toxicity relationships were visualized as a Sankey diagram using the R packages: ggalluvial, ggplot2, and dplyr, which allowed us to map the progression from CK2 inhibitors to CDK off-targets and associated toxicity phenotypes.

Supplementary Table S1. The list of 33 cancer types and their abbreviations in the pan-cancer analysis from The Cancer Genome Atlas (TCGA) used in this study.

Cancer Type	Abbreviation
Adrenocortical carcinoma	ACC
Bladder Urothelial Carcinoma	BLCA
Breast invasive carcinoma	BRCA
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC
Cholangiocarcinoma	CHOL

Colon adenocarcinoma	COAD
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC
Esophageal carcinoma	ESCA
Glioblastoma multiforme	GBM
Head and Neck squamous cell carcinoma	HNSC
Kidney Chromophobe	KICH
Kidney renal clear cell carcinoma	KIRC
Kidney renal papillary cell carcinoma	KIRP
Acute Myeloid Leukemia	LAML
Brain Lower Grade Glioma	LGG
Liver hepatocellular carcinoma	LIHC
Lung adenocarcinoma	LUAD
Lung squamous cell carcinoma	LUSC
Mesothelioma	MESO
Ovarian serous cystadenocarcinoma	OV
Pancreatic adenocarcinoma	PAAD
Pheochromocytoma and Paraganglioma	PCPG
Prostate adenocarcinoma	PRAD
Rectum adenocarcinoma	READ
Sarcoma	SARC
Skin Cutaneous Melanoma	SKCM
Stomach adenocarcinoma	STAD
Testicular Germ Cell Tumors	TGCT
Thyroid carcinoma	THCA
Thymoma	THYM
Uterine Corpus Endometrial Carcinoma	UCEC
Uterine Carcinosarcoma	UCS
Uveal Melanoma	UVM

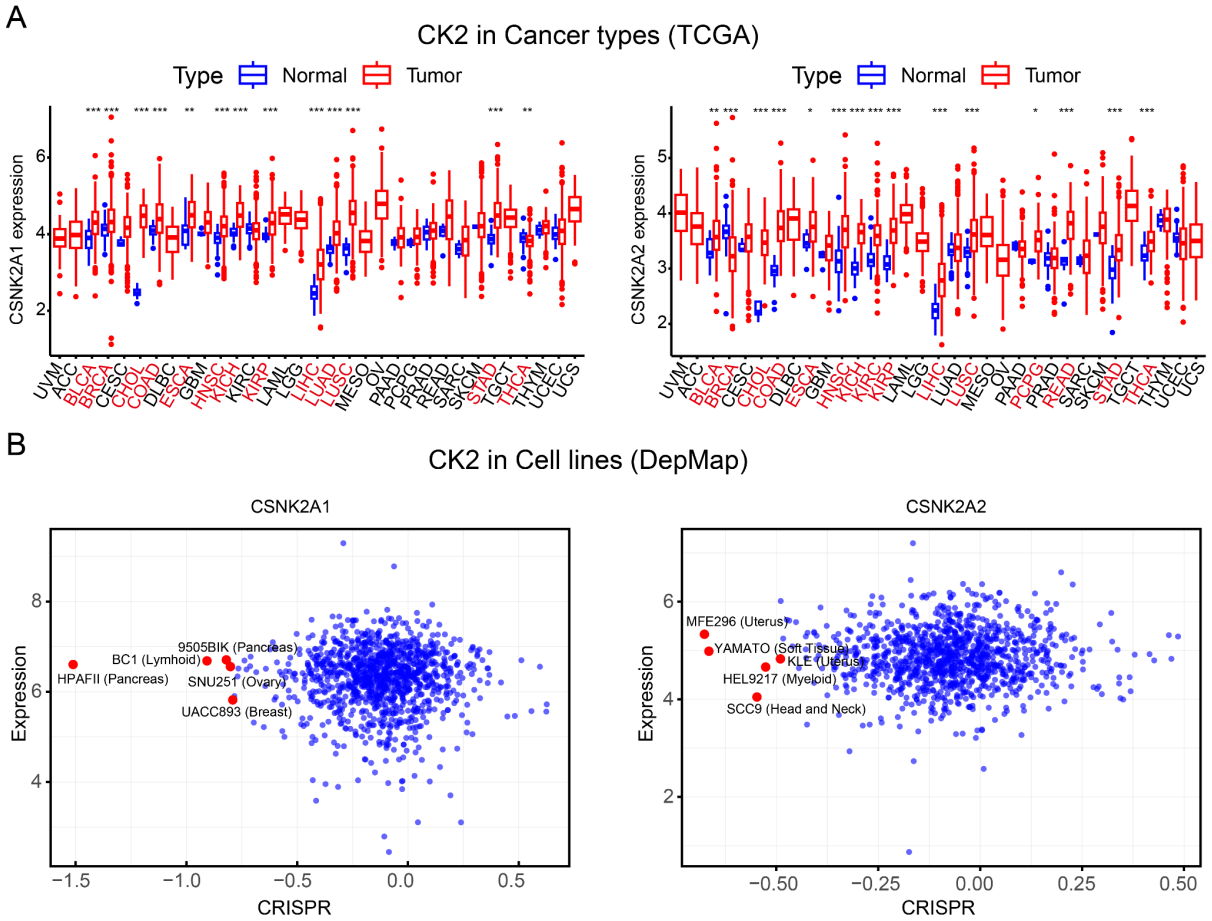
Supplementary Table S2. The studied CK2 inhibitors.

No	Compound Name	CAS No.	Core scaffold / Binding site	CK2 IC50 (nM)	Clinical Stage	PMID
1	CX-4945 (Silmitasertib)	1009820-21-6	Naphthyridine/ ATP-site	0.3-1	Phase II	PMID: 21159648
2	TBB	17374-26-4	Benzimidazole/ ATP-site	150-900	Preclinical	PMID: 11988074
3	DMAT	749234-11-5	Tetrabromo benzimidazole/ ATP-site	130-140	Preclinical	PMID: 17629615
4	TBI	-	Tetrabromo benzotriazole/ ATP-site	1300	Preclinical	PMID: 21755460
5	TDB	-	Tetrabromo-benzimidazole/ ATP-site	32	Preclinical	PMID: 28230762
6	TMCB	905105-89-7	Benzimidazole/ ATP-site	500	Preclinical	PMID: 26786382
7	SGC-CK2-1	2470424-39-4	Pyrazolo-pyrimidine/ ATP-site	4.2-36	Chemical probe	PMID: 35056914
8	CX-5011	1333382-30-1	Pyrimidine/ ATP-site	＜10	Preclinical	PMID: 23145120
9	CX-5279	-	Pyrimidine/ ATP-site	＜10	Preclinical	PMID: 21870818
10	CIGB-300	1072877-99-6	Peptide substrate	NA	Phase I–II	PMID: 30318085

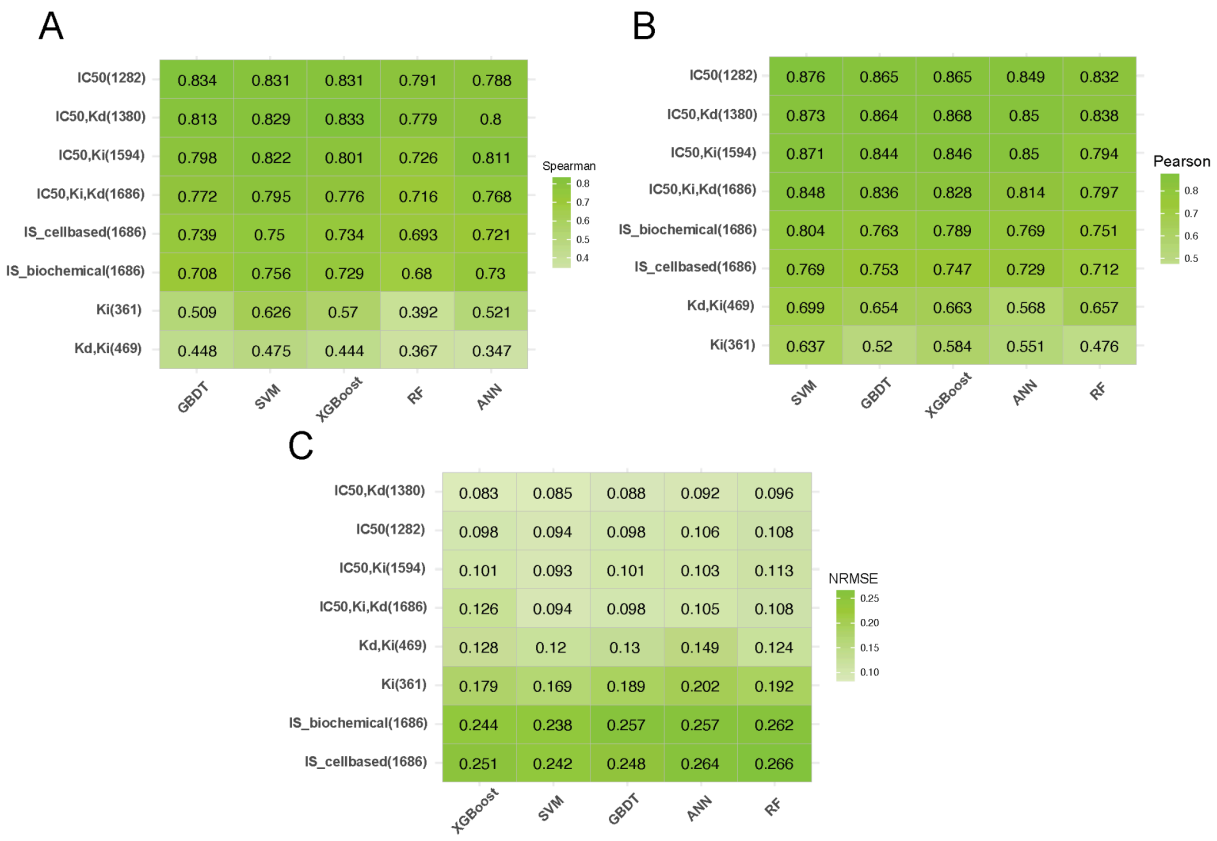
Supplementary Table S3. The Reported binding affinities (IC_{50}/K_d) of CX-4945 toward CDK family kinases retrieved from public kinase profiling databases (ChEMBL, BindingDB).

CDK family	CX4945
CDK1	IC50=56 nM Kd=3035 nM

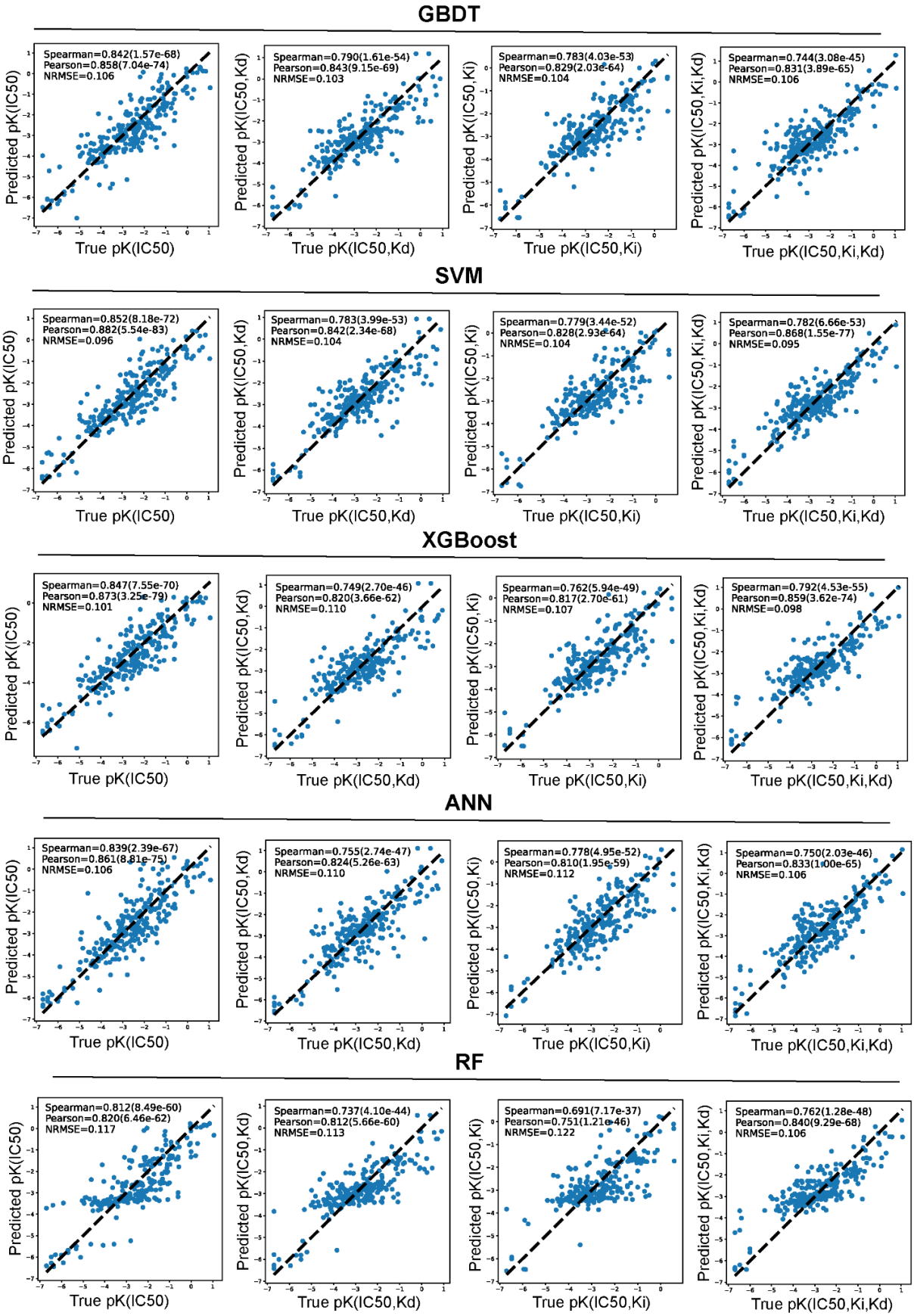
CDK2	IC50=1800 nM Kd=1693 nM
CDK3	Kd=30000 nM
CDK4	Kd=30000 nM
CDK5	Kd=514 nM
CDK6	Kd=30000 nM
CDK7	Kd=7102 nM
CDK8	NA
CDK9	Kd=30000 nM
CDK10	NA
CDK11	NA
CDK12	Kd=30000 nM
CDK13	Kd=30000 nM
CDK14	NA
CDK15	NA
CDK16	Kd=1208 nM
CDK17	Kd=30000 nM
CDK18	NA
CDK19	NA
CDK20	NA



Supplementary Figure 1. **Cancer genetics of CK2. (A) Pan-cancer expression of CK2 genes (CSNK2A1 and CSNK2A2)**. Differential expression analysis of CSNK2A1 and CSNK2A2 across 33 cancer types in TCGA. Box plots compare tumor (red) and adjacent normal tissues (blue). Cancer types with statistically significant differential expression between tumor and normal tissue (Wilcoxon test, FDR p < 0.05) are highlighted in red. **(B) Cancer cell line dependency on CK2 genes (CSNK2A1 and CSNK2A2).** Scatter plots of showing baseline gene expression levels (log2[TPM + 1]) versus CRISPR knockout dependency scores for CSNK2A1 and CSNK2A2 across 1,103 cancer cell lines from different tissue lineages (DepMap, https://depmap.org/portal/). Five cell lines with strong CK2 dependency (more negative CRISPR scores) are highlighted in red and annotated with their cell line name and lineage. This plot illustrates how baseline expression levels relate to cell growth dependency upon gene knockout.

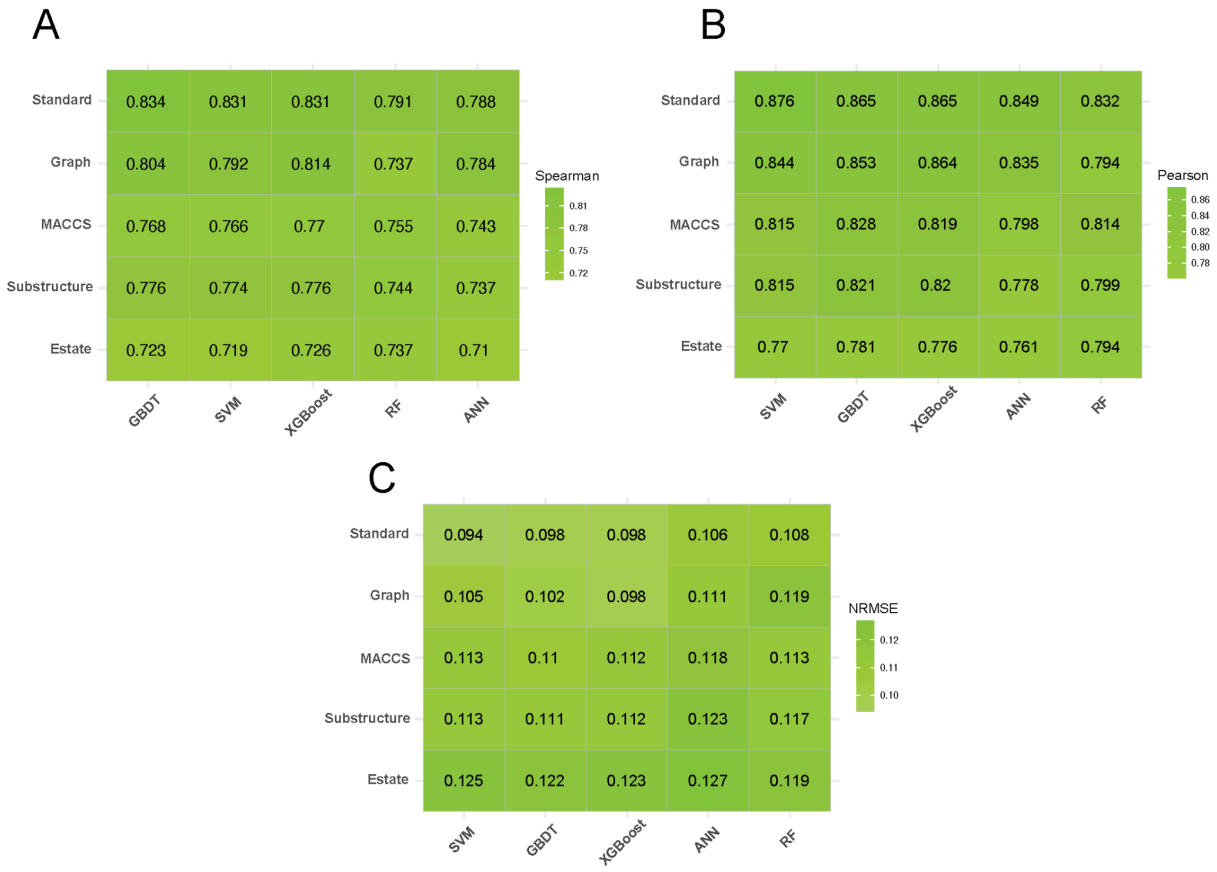


Supplementary Figure 2. **The XXX model prediction accuracy in the test dataset when trained with various bioactivity types.** The total number of compounds in the datasets are written in parentheses. A: Spearman correlation; B: Pearson correlation; C: Normalized Root Mean Square Error (NRMSE).

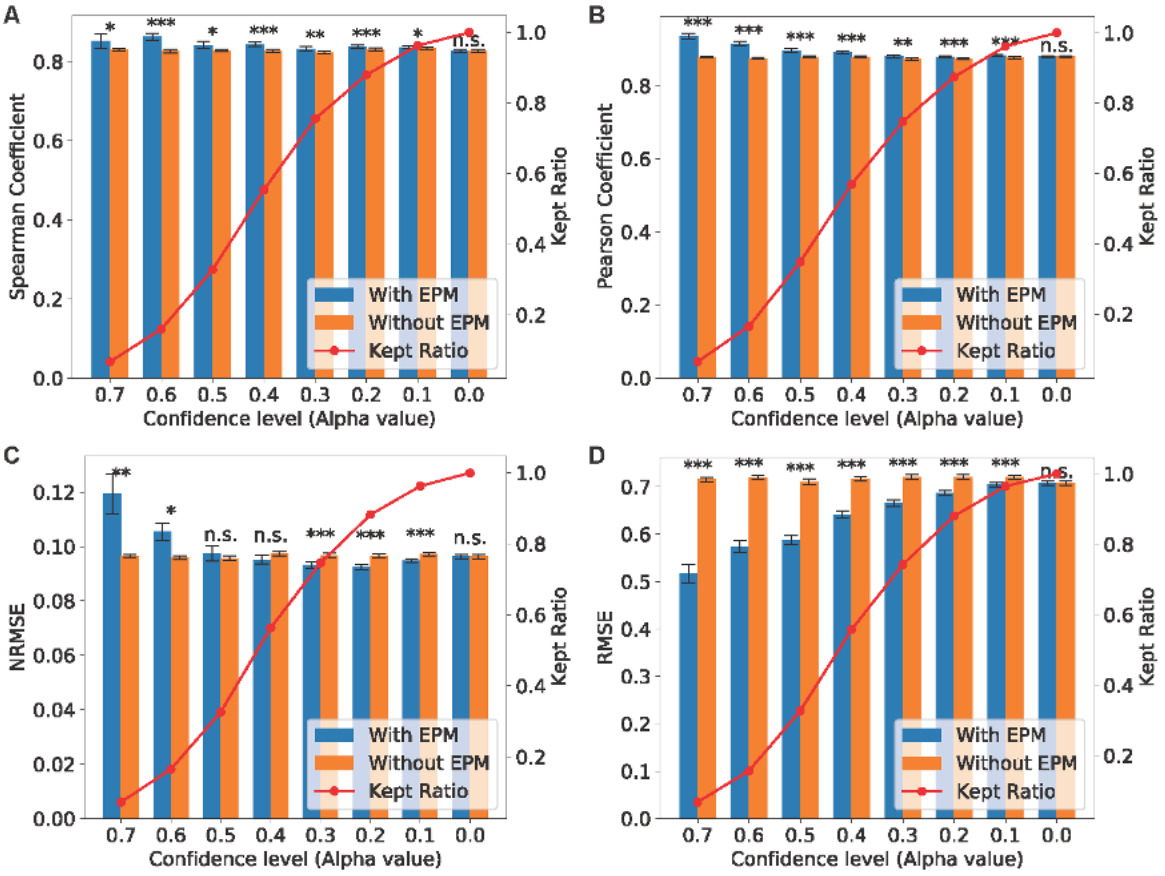


Supplementary Figure 3. Scatter plot between predicted and measured bioactivity values in test set (250 compounds) based on the regression models trained with selected bioactivity

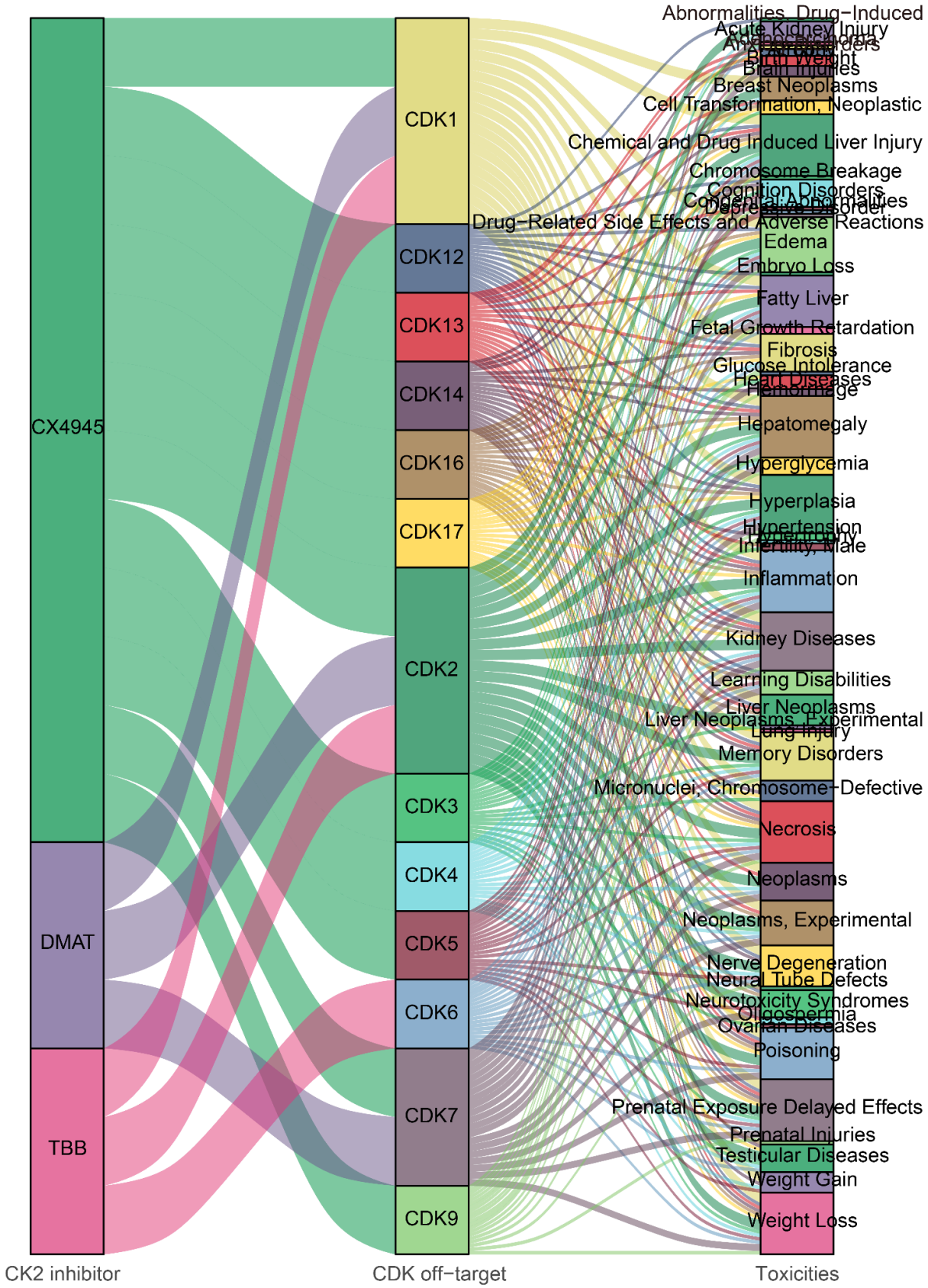
value types. 1000 compounds in each dataset were randomly selected from the full datasets with different bioactivity values (IC_{50} only; combined bioactivity values of IC_{50}, K_d; combined bioactivity values of IC_{50}, K_i; combined bioactivity values of IC_{50}, K_i, K_d).



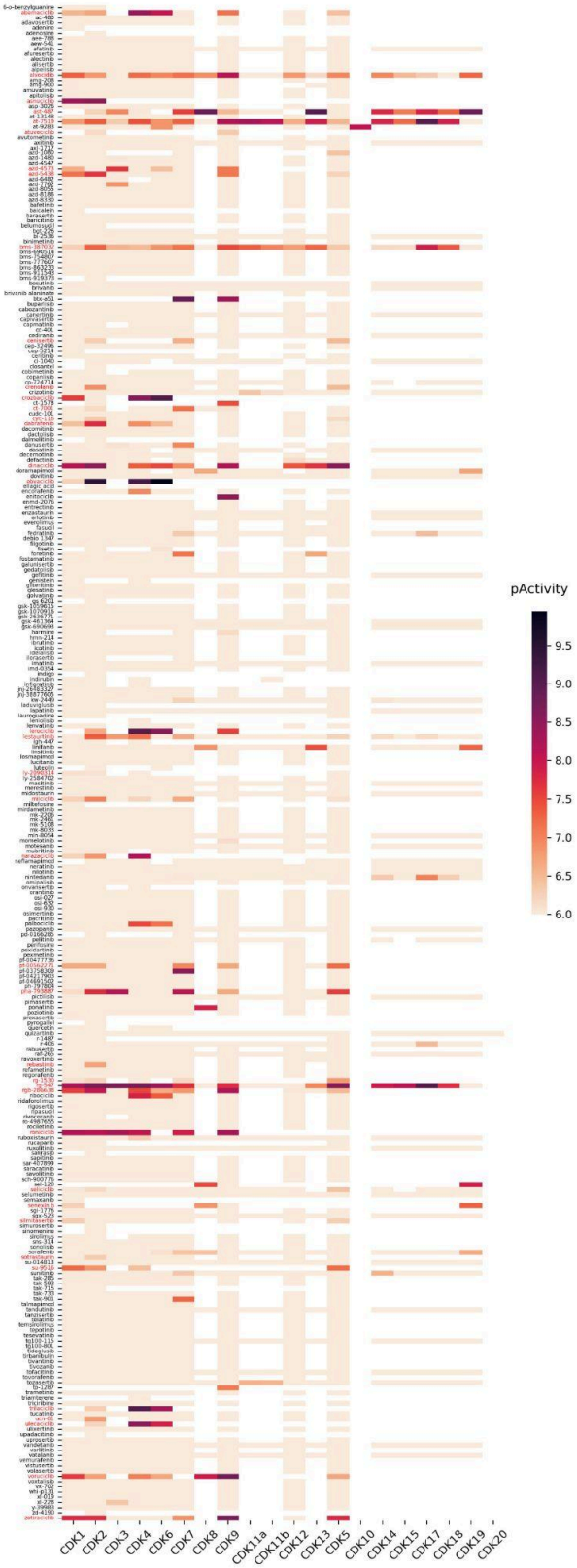
Supplementary Figure 4. **The XXX model prediction accuracy in the test dataset when using various molecule fingerprints as input features to train the regression models.** A: Spearman correlation; B: Pearson correlation; C: Normalized Root Mean Square Error (NRMSE).



Supplementary Figure 5. **The error prediction model performance in the new test set.** Dataset split and error model construction were repeated 60 times at every confidence level. The Wilcoxon test was used to calculate p-values.*p<0.05, **p<0.01, ***p<0.001. A,B,C,D…



Supplementary Figure S1. Sankey diagram for the putative mechanistic links between CK2 inhibitors (CX4945, DMAT, TBB), CDK off-targets, and associated adverse effects.


pActivity
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
CDK1
CDK2
CDK3
CDK4
CDK6
CDK7
CDK8
CDK9
CDK11a
CDK11b
CDK12
CDK13
CDK5
CDK10
CDK14
CDK15
CDK17
CDK18
CDK19
CDK20

Supplementary Figure 7. **Publicly available compound bioactivity data for the CDK protein family.** The data was collected from ChEMBL, BindingDB, and DrugTargetCommons. Compound names colored in red have pActivity > 6 against CDK1 or CDK2. pActivity is equivalent to $-\log_{10}$(molar IC_{50}, EC_{50}, K_i or K_d).